Coherence in Machine Translation

Karin Sim Smith

Department of Computer Science

University of Sheffield

A thesis submitted for the degree of

Doctor of Philosophy

August 2017

Acknowledgements

Firstly, I would like to express my gratitude to my supervisor, Lucia, for having offered me the opportunity to do the PhD in the first place, and her support during my time at Sheffield University. I would like to thank my colleagues in the Natural Language Processing group at Sheffield for their friendship during these past few years. Special thanks goes to David, who has listened patiently and offered much practical advice. Also to Wilker, who taught me a lot in the short time I worked with him. I am also grateful to Mark Stevenson and Gordon Fraser on my panel, for their helpful suggestions. I would like to thank Bonnie Webber for her encouragement and insight, particularly on the matter of discourse connectives during that section of my research, and latterly for her very pertinent and helpful comments which improved the final version of this thesis. Finally, my deep gratitude goes to thank Nick for having encouraged me on this path and supporting me through it (financially, emotionally and practically). Thanks also go to my parents for proofreading the thesis, and my Mum for helping out at home during any trips away.

Abstract

Coherence ensures individual sentences work together to form a meaningful document. When properly translated, a coherent document in one language should result in a coherent document in another language. In Machine Translation, however, due to reasons of modeling and computational complexity, sentences are pieced together from words or phrases based on short context windows and with no access to extra-sentential context.

In this thesis I propose ways to automatically assess the coherence of machine translation output. The work is structured around three dimensions: entity-based coherence, coherence as evidenced via syntactic patterns, and coherence as evidenced via discourse relations.

For the first time, I evaluate existing monolingual coherence models on this new task, identifying issues and challenges that are specific to the machine translation setting. In order to address these issues, I adapted a state-of-the-art syntax model, which also resulted in improved performance for the monolingual task. The results clearly indicate how much more difficult the new task is than the task of detecting shuffled texts.

I proposed a new coherence model, exploring the crosslingual transfer of discourse relations in machine translation. This model is novel in that it measures the correctness of the discourse relation by comparison to the *source* text rather than to a reference translation. I identified patterns of incoherence common across different language pairs, and created a corpus of machine translated output annotated with coherence errors for evaluation purposes. I then examined lexical coherence in a multilingual context, as a preliminary study for crosslingual transfer. Finally, I determine how the new and adapted models correlate with human judgements of translation quality and suggest that improvements in gen-

eral evaluation within machine translation would benefit from having a coherence component that evaluated the translation output with respect to the source text.

"In its communicative function, language is a set of tools with which we attempt to guide another mind to create within itself a mental representation that approximates the one we have."

Scott Delancey

Contents

List of Figures

List of Tables

List of Acronyms

BLEU Bilingual Evaluation Understudy

EDU elementary discourse units

EM Expectation Maximisation

HMM Hidden Markov Model

HT Human Translation

HTER Human-targeted Translation Error Rate

LM Language Model

LSTM Long-Short Term Memory

ML Machine Learning

MT Machine Translation

ML Machine Learning

NLG Natural Language Generation

NMT Neural Machine Translation

PBMT Phrase Based Machine Translation

PE post-edited

PDTB Penn Discourse Tree Bank

RNNLM Recurrent Neural Network Language Model

RST Rhetorical Structure Theory

ST Source Text

SL Source Language

SMT Statistical Machine Translation

RST Rhetorical Structure Theory

TT Target Text

TM Translation Model

UD Universal Dependencies

WSD Word Sense Disambiguation

Chapter 1

Introduction

The amount of multilingual information available on the internet has fuelled a need for rapid online translation. While professional translators would struggle to meet the need, Machine Translation (MT) has grown as a faster, cheaper way of providing translations. This research comes at a time when progress in MT has been rapid to the extent that it is now becoming a more viable option, both as raw MT output for the general public, and as part of a pipeline for providing access to multilingual data.

However, the quality of Statistical Machine Translation (SMT), arguably the most widely used paradigm at the time that this work was done and the focus of this thesis, is still often far from perfect, and we hypothesise that one of the main problems with it is the failure of current SMT approaches to handle discourse, at various levels. This is largely due to the manner in which they work with a small context window, isolated from the surrounding text and the relevant discourse. Discourse has long been recognised as a crucial part of translation (Hatim and Mason, 1990), but when it comes to SMT, discourse information has been mostly neglected to date. For MT to progress to the next level, a strategy to address discourse is vital. This will render MT more acceptable, resolving ambiguous discourse relationships, retaining co-references, and overcoming other shortcomings which result from the fact that discourse context is largely ignored.

Recently increasing amounts of effort have been going into addressing discourse explicitly in SMT, covering lexical cohesion (Wong and Kit, 2012; Tiedemann, 2010; Carpuat, 2013; Carpuat and Simard, 2012; Xiong et al., 2013b; Gong

et al., 2015), discourse connectives (Cartoni et al., 2012; Meyer and Popescu-Belis, 2012; Steele, 2015; Steele and Specia, 2016), anaphora (Guillou, 2016; Hardmeier et al., 2013b) and negation (Fancellu and Webber, 2014; Wetzel and Bond, 2012).

Our aim is to focus on coherence, an issue that has not yet been exploited in the context of MT, and to research how coherence models can be used in evaluating MT. In particular, we are interested in the transfer of coherence from the source text to the target text, *without* a reference translation.

1.1 The Problem

To illustrate the problem of coherence in SMT, we take an example drawn from the 10th Workshop on Machine Translation (Bojar et al., 2015), and show machine translated output from one of the highest scoring systems[1]. We chose the French-English language pair as one of the best performing ones, where quality is of a level that makes focussing on discourse possible. (The source text (French) and the reference translation are included alongside the machine translation output in Appendix A.2).

> *We want cheap goods and was surprised that manufacturers produce in "industrial" it highlights super u but all the signs, and even the small shops are concerned. How to eat good, organic with 1,650 (average wage in france)? But to answer your post, if it is possible to eat properly with a small wage. After everyone is free to set its priorities where it sees fit. For my part, the food is one.*

From reading the MT output, it is clear that there is an issue with coherence, in terms of grammaticality, fluency and adequacy: The incorrect use of cohesive devices means that the reader has to work harder to make sense of the text: In this example, the first sentence has a lack of verb agreement ('We want cheap goods and *was* surprised'), incorrect noun form ('produce in "industrial"', meaning 'industrialised nature of production' or 'industrialised production'), an incorrect noun ('signs' instead of 'brands' or 'names'), wrong verb ('concerned'

[1] LIMSI-CNRS_mosesSoulMoreFeatures_primary

instead of a verb like 'affected'), adjective instead of adverb ('How to eat *good*'), a preposition instead of an adverbial phrase ('*After* everyone is free', which should presumably read 'After all'), lack of verb agreement again ('its' following 'everyone'), and a superfluous definite article ('the' in 'the food'). While some of these errors are grammatical, they arguably still affect the coherence of the text as they make it harder for the reader to understand.

1.2 Coherence

A useful description by Louwerse and Graesser (2005) is of 'cohesion as continuity in word and sentence structure, and coherence as continuity in meaning and context'. As Jurafsky and Martin (2009) similarly summarise, cohesion is the glue linking textual units, while coherence is the meaning relation between them. This may take many forms, of which the various types of cohesion form part, as do coherence relations, as possible connections between utterances (Jurafsky and Martin, 2009) and their discourse structure. Therein lies some of the difficulty of the task as while, for example, lexical cohesion can be more easily detected and addressed, the semantics, pragmatics and contextual indicators are much more difficult to determine.

While there has been recent work in the area of lexical cohesion in SMT, as a sub-category of coherence, looking at the linguistic elements which hold a text together, there seems to be little work in the wider area of coherence as a whole. Coherence is indeed a harder discourse element to define in the first place. While it does include cohesion, it is wider in terms of describing how a text becomes semantically meaningful overall, and additionally spans the entire document.

Halliday and Hasan (1976)'s classic book on cohesion identifies five types of cohesion which are present in coherent texts, namely reference, substitution, ellipsis, conjunction, and lexical cohesion. All of these contribute to the overall coherence of a text, and how easy it is for the reader to follow, but cohesion in itself is insufficient to ensure coherence. Within a text the argument structure has to be such that the reader can follow.

To illustrate, we borrow the example of Blakemore (2002):

As to your payment by direct debit, you do not need to take any action. British Gas will use existing wires, cables and meters for your electricity supply. In addition, from 1 October 2001 the variable base rate has also changed from 6.75% to 6.50% a year. If you want to cancel later, please call us on the same number. This will be collected on or just after October 2001, and in each subsequent month from your bank/building society.

Here the text is grammatically correct and there are elements of cohesion, such as connectives ('In addition') and lexical cohesion ('this'), but the text makes no sense, it lacks coherence.

Defining coherence is difficult, and there are many aspects to it. It encompasses a combination of all the five types of cohesion mentioned above, which jointly support the logic, which runs through the text. But, as Blakemore (2002, p.5) says, discourse markers need to be analysed not in isolation, but in terms of their influence on utterance interpretation.

Coherence is undeniably a cognitive process, and we limit our remit to the extent that this process is guided by linguistic elements discernible in the discourse.

1.3 Statistical Machine Translation

SMT can be seen as a Machine Learning (ML) problem (Aziz, 2014) which essentially takes a parallel text, divides the source text into words and phrases, and transforms them into the target text via certain rules guided by statistics. This forms the 'translational equivalence' or transfer model; the resulting pairs of phrases or grammatical productions define the search space. These are generally derived via alignments from parallel texts. Once learned, the system can then be used to translate unseen texts (Aziz, 2014) .

The model needs to discriminate between various possible translations, and so needs a mechanism for determining which is the best translation to use. Parameterization defines a function that gives a score to given input-output mappings (translation pairs) and thus enables the ranking of all possible outputs. This function includes various features which contribute to the translation: In a

phrase-based SMT system, the most popular type of SMT which we focus on in this thesis, these could be the Translation Model (TM) (matched phrase probabilities for source and target pairs), the reordering model (preventing excessive reordering) and the Language Model (LM) (a probability distribution over likely word sequences in the target language).

These are then combined via a log linear model (Equation 1.1), whereby the weightings of these individual features are adjusted to improve the translation quality (Koehn et al., 2003). The values of these features are determined via parameter estimation. First any feature functions derived from generative models are estimated, determining the word translation probabilities based on the training corpus. Once the model is parameterised, the parameters of the log linear model are estimated- including these generative features- to set feature weights for it. This is done via discriminative training, directly aiming to adjust the weights to improve translation quality according to a given evaluation metric. Decoding is then the process for finding the highest scoring translation under this model from an exponential number of possible translations for any given input (Lopez, 2008).

The format of the log linear model is:

$$P(e|f) = f(x) = exp \sum_{i=1}^{n} \lambda_i h_i(x) \tag{1.1}$$

where the probability of translating any given source sentence f into target sentence e is a function of the different components, such as the reordering model, the language model and the translation model, among others, and as represented by h. These are weighted by parameters from training, represented as λ.

The decoding algorithm incrementally computes scores from partial translations using the features mentioned above. It uses these scores to estimate the best path to completion (Koehn et al., 2003).

Finally, the MT output is automatically evaluated. This is traditionally done by comparing it to a reference translation, typically a human translation deemed 'correct', and measuring how close it is to that translation.

1.3.1 Lack of linguistic context in modelling

The translational equivalence model (described in Section 1.3) may take various forms. Until recently, the most widely-used SMT approach has been the phrase-based model, which has been constructed as a purely statistical model with no linguistic input, where it attempts to construct a target sentence by concatenating the translations of contiguous sequences of words or phrases. The phrase-based model can be generalised, and adapted to work with syntactic units instead of phrases. Although such a syntax-based model is syntactically informed, it is limited.

The phrase pairs or transfer rules are extracted from parallel data at training time. The probability estimates for the rules have to be learned, and sparsity must be reduced, to make it estimable. This involves making independence assumptions and dropping the surrounding dependencies. In extracting the separate phrases or transfer rules, much of the linguistic context is lost. As a result, anaphoric references are determined probabilistically rather than based on the referent, and many forms of cohesion can be lost. Phrases are devoid of their context, which helps disambiguation.

The phrase table in the TM may include $P(f|e)$ for a particular phrase, indicating the probability of translating any given French word or phrase as a particular English word or phrase, such as the probability that we equate the French word 'boucher' to the English word 'butcher' or to the word 'block'. So when constructing the building blocks for the model and deriving alignments, they are isolated chunks or words, taken out of context. This results in independently translated phrases which are concatenated– with no reference to coherence.The word and phrase alignments are derived from training data, so any frequently occurring phrases which form part of the cohesion in the source text will not necessarily have been aligned and scored as such.

The LM then later attempts to recover some of the linguistic context and does this via scoring rules in context. In its current form, the LM is still too weak to influence coherence, as it is limited by a small window of context, and being monolingual, there is no transfer of a contextual nature from the source text.

1.3.2 Lack of linguistic context at decoding time

Finally there is a lack of linguistic information in the decoder, where it attempts to search through a space of solutions to efficiently find a probable solution. In the decoder only one sentence at a time is processed, in isolation from surrounding sentences, due to the computational complexity inherent in SMT. This means that at decoding time all the inter-sentential links are lost.

The decoder takes the input sentence and generates lattices, representing possible translation excerpts for constructing the output sentence. It has to find a combination of rules that are compatible and cover the input sentence, then incrementally generate the output sentence (Koehn et al., 2003). It works on the basis of independently translated phrases which are joined up via the language model with a very limited context window. Coherence is not explicitly taken into account during this entire process: there is no logic or communicative intent being traced.

The decoding proceeds, building the target sentence left to right with limited reordering of phrases. (In principle any ordering is permitted, but this is not possibly in practice as it introduces excessive noise and significantly adds to computational complexity.) As a result, ordering may be wrong– either linguistically incorrect, or incoherent (certain ordering is deemed more coherent, as it is more easy for the brain to follow). Any lexical cohesion that is recreated by the language model will at most be influenced vaguely via previously mentioned features learned by the model from training data, but not directly identified and transferred.

The decoding problem itself is computationally costly (NP-complete, in fact Koehn (2010)) and so heuristics have been introduced to ensure that the decoder completes, and comes up with a potential output in a feasible amount of time. The decoder works out all potential translations for a given input sentence, and determines the one with the best score. This means that the number of options being considered in the search space grows exponentially in line with the length of the sentence. It could end up with a huge number of optional translations each time a new word in the sentence is considered, to the extent that the problem becomes intractable.

To address this, the lowest scoring options are pruned out, reducing the hypotheses to a more reasonable number. This may mean that some good translations are rejected at an early stage, simply because they seemed less probable early on. These may also be more coherent ones, but there is no intuition of coherence in this strategy. The computational complexity is already such that there is no way the decoder in its current form can consider a larger context window.

1.4 Coherence in SMT

1.4.1 Measuring coherence

While detecting coherence is intuitive for a human translator, it is hard to codify with a view to automatic learning. As mentioned in Section 1.2, coherence includes a cognitive element, and we limit ourselves to detecting and learning the linguistic elements which are discernible and which guide the cognitive process.

In their computational theory of discourse structure, Grosz and Sidner (1986) suggest that discourse structure includes three separate but interrelated components, which can be regarded as contributing to coherence:

1. the structure of the actual sequence of utterances in the discourse (called the **linguistic structure**);
2. the structure of purposes (called the **intentional structure**);
3. the state of focus (called the **attentional state**).

They state:

> 'This theory provides a framework for describing the processing of utterances in a discourse. Discourse processing requires recognizing how the utterances of the discourse aggregate into segments, recognizing the intentions expressed in the discourse and the relationships among intentions, and tracking the discourse through the operation of the mechanisms associated with attentional state.'

Indeed, previous computational mechanisms for assessing coherence have attempted to cover **discourse relations**, **intentional structure** and **entity-based coherence**, which cover (1), (2) and (3) respectively. As detailed in depth by Poesio et al. (2004), in doing so they have made simplifications which no longer entirely fit the original theoretical basis.

We will further investigate these components in an SMT context and adapt the models to advance the assessment of coherence in SMT. We are interested in capturing aspects of coherence as defined by Grosz and Sidner (1986) above, based on the attentional state, intentional structure and linguistic structure of discourse. As a result, we believe that a coherent discourse should have a context and a focus, be characterised by appropriate coherence relations, and be structured in a logical manner.

In terms of measuring coherence, previous experiments assessing coherence computationally have been in a monolingual setting, where the scenario has been to derive a correct summarization, or to determine a correct sentence ordering in a shuffled text. Moreover, they have often been on sentences which are themselves coherent in the first place. This differs from our SMT scenario firstly, in that it concerns only one language, and secondly, that the task is different. It is more clear-cut: if a text has been automatically summarized or shuffled, the overall logic has potentially been broken. The challenge then is to rediscover the logic pattern. In our scenario the situation is more nuanced, as the elements of coherence may be there to some degree, but they may have been distorted due to other changes which have occurred in the decoding process, resulting in an incorrect use of cohesive devices. Moreover generally the individual sentences in previous coherence experiments have themselves been coherent, whereas MT output may not be so.

The document excerpt below, extracted from **WMT** submissions on test data[1] illustrates this point (parallel versions of full document in Appendix A.3):

> *The matter NSA underlines the total absence of debates on the piece of information*

[1] http://www.statmt.org/wmt14, submission output from system onlineC.0 for French-English language pair

How the contradictory attitude to explain of the French government, that of a quotation offends itself in public while summoning the ambassador of the United States October 21, and other forbids the flying over of the territory by the bolivian presidential airplane, on the basis of the rumor of the presence to his edge of Edward Snowden? According to me, there are two levels of response from the French government. When François Holland telephones Barack Obama or when the minister of the foreign affairs Laurent Fabius summons the ambassador of the United States, they react to a true discovery, that is the one of the extent of the American supervision on the body of the communications in France. Not is it surprising to read in the columns of the World to some weeks of interval on one hand the reproduction of the American diplomatic correspondence and on the other hand a condemnation of the listen Quay of Orsay by the NSA? Not there would be as a vague hypocrisy on your part?

Here we can see cohesive elements, such as '' The fact that the French word *le renseignement* has been wrongly translated as *piece of information* in the title, means that the reader is left struggling to piece together the train of thought. Beyond the word order issue in next sentence (*How the contradictory attitude to explain...*), the wrong lexical choice of *a quotation* instead of the construct *d'un côté...de l'autre* loses the contrastive relation which structures that sentence. The phrase *the presence to his edge of Edward Snowden* makes no sense, due to the fact that *à son bord* in the French source text has been mistranslated as *to his edge*, instead of *on board*. The negation is mistranslated: Not is it... and *Not there would be...*, which undermines the coherence. The MT sticks too closely to the French construct (*N'est-il...*, see Appendix A.3) and copies the form of the latter, rather than the meaning.

The text is clearly not coherent. Comparing it to the text below, which is the human, reference translation, we can see the difference in terms of coherence. How we try to capture this is the challenge.

NSA Affair Emphasizes Complete Lack of Debate on Intelligence
Why the contradictory attitude of the French government? On the one

hand, it publicly takes offence and summons the Ambassador of the United States on October 21 and, on the other, it forbids the Bolivian president's plane to enter its air space on the basis of a rumor that Edward Snowden was on board? In my opinion, there are two levels of response from the French government. When François Hollande telephones Barack Obama, or when Foreign Minister Laurent Fabius summons the Ambassador of the United States, they are responding to a real discovery, that of the scale of America's surveillance of communications within France generally. And is it not surprising to read in the pages of Le Monde, on the one hand, a reproduction of diplomatic correspondence with the US and, on the other, condemnation of the NSA's spying on the Ministry of Foreign Affairs on the Quai d'Orsay, within a matter of weeks? Is there not an element of hypocrisy on your part?

1.4.2 Coherence transfer model

We presume that a coherent source text should be translated into a coherent target text, however there are issues to be considered.

Coherence patterns vary for different languages. Assessing coherence in its many facets in the source text is involved in itself, and transferring to the target text is complex. Coherence in one language may not be the same in another. For example, some languages use more items of lexical cohesion than others (Lapshinova-Koltunski, 2015b).

Reference resolution is directly impacted by the linguistic differences, where one language may have three genders and the other simply two. Moreover, the discourse units may be ordered differently in one language from another. Also, languages have different syntax structures.

Language-pair-specific coherence issues in an SMT context In addition, mismatches between particular language pairs in MT cause different manifestations of incoherence in the output. We examine the types of coherence-related

errors which occur in different language pairs (Chapter 5), and discover that some of these error patterns are more relevant for particular language pairs.

1.4.3 Integrating coherence in SMT

After we have detected coherence elements in the source text, there remains the issue of how to make sure the translation is also coherent. Options to integrate coherence into SMT include doing it via modelling, decoding or evaluation.

During Modelling Currently the only way to influence this process in SMT is by features, whose weight is computed at training time, and introduced in the log-linear model, as described in Section 1.3. This can only have a limited and generic influence. It can influence lexical choice, but not explicitly ensure intrasentential transfer of discourse elements. Features operate at a word/phrase level, on word(s) extracted and weighted at training time, without the source or (more particularly) the entire target sentence to evaluate. Features are hard to craft, since they are trained out of context. Moreover when they are used by the decoder, they are implemented with only a small target text window. Attempting to create a feature function that could influence coherence in a generic way, trained on separate data, and with a limited target text window is very challenging.

After decoding For most setups the other alternative to influence coherence would be via *reranking* of n-best lists which are derived during the *decoding* described in Section 1.3. N-best list ranking attempts to introduce more complexity into the model by incorporating global features, which were not possible at decoding time. These will be used to rank a limited number of options from the baseline model, those deemed the best. Ultimately, though, these only present the options derived under the current decoding conditions, so are potentially limited in terms of coherent options available.

Another option is the possibility of integrating features into a different framework which allows access to the entire target document. This means that we can score according to functions measuring coherence. Docent (Hardmeier et al.,

2013a) takes the MT output from a standard decoder as a baseline draft and enables changes to that. These changes can be computed on a sentence or document level, although the nature and scope of them is restricted without extensive adaptation.

During evaluation For purposes of automatic evaluation, metrics which do ngram matching or alignment (between the MT output and one or more reference translations) are used for assessing the quality of MT output. They often use a single reference translation. As such they are limited in assessing discourse level issues. Moreover, comparing with a single reference fails to account for the fact that there are numerous equally valid ways of correctly translating a text. Even if a human evaluation were to indicate an improvement induced by a discourse model, the fact that most existing metrics fail to value discourse phenomena means that they are limited in detecting discourse changes and that these could even degrade the score. Interestingly, as shown by (Smith et al., 2016), the BLEU score of a document can go up while actual quality goes down.

Another issue with many of the current metrics is that they depend on a reference translation, which should be a correct human translation. There may be many correct alternative translations of any one text, however, which is a major shortcoming with this way of assessing MT output. Besides relying on a reference translation, it only provides a superficial evaluation. This is a problem which is evident in practice, when mistranslations lead to business losses, as explained by (Levin et al., 2017): "Typically it is very difficult to detect such errors because doing so requires some understanding of the sentence meaning."

Aside from the benefits of automatic evaluation, without an automatic metric there is no way to optimise parameters in a system which does include discourse features. There are numerous difficulties with evaluation of discourse phenomena, as detailed by Hardmeier (2012), particularly if it is to work without annotation. Attempting to address the issue with a precision/recall based measure proved problematic, as Hardmeier and Federico (2010) found when applying it to assess pronoun translations in both MT and source, in that the MT output can vary widely. As mentioned by Guzmán et al. (2014), 'there is a consensus in the MT community that more discourse-aware metrics need to be proposed for this area

to move forward'. This is also reflected in the fact that researchers in the domain have latterly created test sets to capture and measure the translation of discourse phenomena (Sennrich, 2017; Bawden et al., 2017; Isabelle et al., 2017). We focus on coherence metrics which can be used for evaluation and reranking.

1.5 Scope and Aims

While MT output has been improving and becoming increasingly widely used, the quality is still lacking, and there is an increasing awareness of the need to integrate more linguistic information, including coherence, into SMT. Previously, coherence has been assessed in a monolingual context, using excerpts sourced from coherent text (either as shuffling or summarisation tasks). We believe that existing monolingual coherence models are not suitable for assessing coherence in MT output, and show how we can adapt these to capture coherence of SMT output in a more meaningful manner. We establish how coherence should be represented in a crosslingual context, and how it can be evaluated in a meaningful manner in MT.

This thesis explores the application of existing coherence models to MT output, and reports on our extensions to these models to render them better suited for the task. It proposes an entirely new model, based on crosslingual discourse relations. Finally, we introduce a metric, which can serve alongside existing metrics and can measure coherence independent of a reference translation.

The main aims of our work can be summarised as follows:

(A_1) Existing models are insufficient to measure coherence in MT output and monolingual methodologies for measuring coherence are possibly inadequate in a crosslingual context. We aim to benchmark these models, and subsequently adapt and extend them to our domain.

(A_2) We intend to show that discourse relations can be used in a crosslingual setting to capture coherence in SMT (previous work on discourse relations in SMT has been monolingual, and on MT output alone). We develop a model to capture discourse relations crosslingually, establishing crosslingual

mappings using embeddings, and use them to give an indication of the successful transfer of a discourse relation.

(A_3) Coherence patterns vary for different languages, which affects how we measure coherence in MT. We establish how it can be measured in a multilingual context. In addition, we aim to show how mismatches between particular language pairs in MT cause different manifestations of incoherence in the output.

(A_4) We plan to use models measuring coherence as a complementary metric to existing ones, illustrating that they are also useful for evaluation of MT *without* need for a reference translation.

1.6 Contributions

1. Implementation of Coherence Models: *Cohere: A toolkit for local coherence* (Sim Smith et al., 2016a), which incorporates a reimplementation of the entity grid and entity graph, and an extension of a syntax-based model which outperforms the existing one. (A_1)

2. Extension of Models: *The Trouble with Machine Translation Coherence* (Sim Smith et al., 2016b). Analysis of adapted coherence models in an MT setting. We show that assessing coherence in SMT is a far harder task for existing models than trying to reorder shuffled texts. (A_1)

3. Discourse Relations in a Crosslingual Setting: *Assessing Crosslingual Discourse Relations in Machine Translation*. We deploy crosslingual embeddings adapted for multiword discourse connectives and incorporate discourse relation mappings between source and target texts. We propose a novel approach that assesses the translated output based on the *source* text rather than the reference translation and focuses on measuring the extent to which discourse elements in the source are preserved in the MT output.(A_2)

4. *A Coherence Corpus in Machine Translation* (Sim Smith et al., 2015). This includes corpus analysis, examining the types of coherence errors that frequently occur in SMT. We also establish that different language pairs result in varying types of coherence errors. (A_3) (Sim Smith, 2017).

5. *Examining lexical coherence in a multilingual setting* (Sim Smith and Specia, 2017). This covers preliminary analysis indicating how lexical coherence is achieved on a crosslingual and multilingual basis. We detail modifications necessary for using the entity-based experiments in a cross-lingual scenario. (A_3)

6. We show that coherence models can serve as a reference-independent evaluation metric. (A_4)

7. *On Integrating Discourse in Machine Translation*: we make recommendations for the future, suggesting that progressing evaluation in MT beyond reference-based metrics and integrating and element of semantics would lead to greater integration of discourse phenomena

A full list of our publications is included in the Appendix.

1.7 Structure of thesis

Having established how we scope coherence for the purposes of this work, in addition to the difficulties faced in an SMT context, we detail the structure of this thesis. In the next Chapter we start by reviewing recent work in the area of discourse coherence in SMT, and evaluating existing, general coherence models. Subsequently, in Chapter 3 we illustrate how the task of assessing coherence in an MT setting is a much harder one than some previous monolingual settings (A_1). We explore crosslingual discourse relations in Chapter 4, and measure transfer of discourse relations from source to target text (A_2), illustrating that evaluation without a reference is possible (A_4). In Chapter 5 we investigate manifestations of incoherence in MT which arise in SMT due to linguistic variation between languages (A_3), and explain our work to create an artificial corpus for evaluating the models. In fact, this is a general framework which can be used to generate

a corpus for particular requirements, i.e. a particular language pair and genre. In Chapter 6, we turn to our preliminary work examining lexical coherence in a multilingual setting, exploring commonality and patterns of lexical coherence. Finally, in Chapter 7, we set out how our models can now be used to evaluate MT output (A_4). Our conclusions form Chapter 8.

Chapter 2

Discourse in SMT and existing Coherence Models

While the survey by Hardmeier (2012) provides a good overview of discourse in SMT at the time, his survey has been superseded by newer research and his case study is specifically on anaphora resolution. Our survey will not attempt to cover the same ground in terms of a general historical overview, but will cover more recent research in the general field of discourse, specifically as it relates to discourse phenomena in the SMT context (Section 2.1), and the subsequent focus in Section 2.2 will be on coherence, an issue largely ignored currently in SMT. To date these two sections of work remain separate, because while there is work on various discourse aspects in SMT, there is none on coherence- all coherence work has been in a monolingual context.

2.1 Discourse in SMT

Recent years have seen a flurry of work, much of it in association with the Workshop on Discourse in MT (Webber et al., 2013, 2015). In considering discourse phenomena in MT, we firstly touch on the issue of document-level discourse (Section 2.1.1), then structure our survey broadly around the groupings defined by Halliday and Hasan (1976), looking at work in grammatical cohesion (Section 2.1.2), lexical cohesion (Section 2.1.3) and discourse connectives and structure

(Sections 2.1.4 and 2.1.5). We finally touch on negation (Section 2.1.6), which also directly affects coherence in SMT, before turning to existing coherence models in Section 2.2.

2.1.1 Document level discourse

As discussed in detail in Section 1.3, most decoders work on a sentence by sentence basis, isolated from context, due to both modelling and computational complexity. This directly impacts the extent to which discourse can be integrated in SMT. An exception to this are approaches to multi-pass decoding, such as Docent (Hardmeier et al., 2013a). Docent is a document level decoder, which has a representation of a complete target text translation, to which changes can be made to improve the translation. It uses a multi-pass decoding approach, where the output of a baseline decoder is modified by a small set of extensible operations (e.g. replacement of phrases), which can take into account document-wide information, while making the decoding process computationally feasible.

To date attempts to influence document level discourse in SMT in this manner have been limited. Hardmeier (2012) initially investigated an element of lexical cohesion, rewarding the use of semantically related words (determined via LSA), integrated as an extra feature function. Subsequently, Sara Stymne and Nivre (2013) attempted to incorporate readability constraints into Docent, in effect jointly influencing the translation and simplification. This is directly relevant for our efforts to incorporate discourse elements at sentence and document level, and in theory opens up options for testing out features for improving coherence.

A similar document level framework was recently developed by Martìnez Garcia et al. (2017), who created a new operation to ensure that changes could be made to the entire document in one step. They use word embeddings to promote lexical consistency at document level, by implementing a new feature for their document-level decoder. In particular, they try to encourage consistency for the same word to be translated in a similar manner throughout the document. They deploy a cosine similarity metric between word embeddings for the current translation hypothesis and the context to check if they are semantically similar. Despite the fact that a bilingual annotator judging at document level found the

improved output to be better than the baseline 60% of the time, and equal 20% of the time (i.e. the improved output is better or the same for 80% of the documents), there was *no statistical significance* in the automatic evaluation scores (Martìnez Garcia et al., 2017).

2.1.2 Grammatical cohesion

Reference resolution Voigt and Jurafsky (2012) specifically examine referential cohesion of MT in the literary domain, carrying out a comparative study between newswire texts and literary ones. They find that literary texts have denser reference chains than news articles, and while human translations reflect this, machine translations do not. They also comment on the surprising amount of cohesion that is still achieved in MT, given that it works on a sentence-by-sentence basis, and note the fact that there is no means of ensuring referential consistency in MT.

Anaphora resolution Anaphora resolution, as reference resolution to something or someone previously mentioned, is a very challenging issue in MT which has been studied by several researchers over the past few years (Novák, 2011; Le Nagard and Koehn, 2010; Hardmeier and Federico, 2010; Guillou, 2012; Hardmeier et al., 2013b). It is something that SMT currently handles poorly, again due to the lack of intersentenial references. Anaphoric references are affected in several ways. The context of the preceding sentences is absent, meaning that the reference is undetermined. Even once it is correctly resolved in the source text (by additional pre-training or a second-pass), reference resolution is directly impacted by linguistic differences. For example, the target language may have multiple genders for nouns while the source only has one. The result is that references can be missing or wrong. Initial work by Guillou (2012) highlights the differences of coreference depending on the language pair. As detailed in greater depth by Hardmeier (2012), attempts have previously been made to address this, however results were not hugely successful. These include the earlier research by Hardmeier and Federico (2010), and also by Le Nagard and Koehn (2010), who try using coreference resolution to resolve pronouns, finding that the coreference

resolution systems were insufficient. Novák and Žabokrtský (2014) developed a crosslingual coreference resolution between Czech and English, with mixed results of improvements in addition to instances which are worse, indicating the complexity of the problem. Subsequently Hardmeier et al. (2013b) have attempted a new approach with anaphora resolution by using neural networks which independently achieves comparable results to a standard anaphora resolutions system, but without the annotated data.

Luong and Popescu-Belis (2016) focus on improving the translation of pronouns from English to French by developing a target language model which determines the pronoun based on the preceding nouns of correct number and gender in the surrounding context. They integrate by means of reranking the translation hypotheses and improving over the baseline of the DiscoMT 2015 shared task.

Luong and Popescu-Belis (2017) develop a probabilistic anaphora resolution model which they integrate in a Spanish-English MT system, to improve the translation of Spanish personal and possessive pronouns into English using morphological and semantic features. They evaluate the Accuracy of Pronoun Translation (APT) using the translated pronouns of the reference translation and report an additional 41 correctly translated pronouns from a base line of 1055.

Pronoun prediction More recently, pronoun prediction in general has been the focus of increased attention, resulting in the creation of a specific WMT Shared Task on 'Cross-lingual Pronoun Prediction' (Guillou et al., 2016), and to the development of resources such as test suites (Guillou and Hardmeier, 2016) for the automatic evaluation of pronoun translation. This has led to varied submissions on the subject, predicting third person subject pronouns translated from French into English; (Novák, 2016; Loáiciga, 2015; Wetzel et al., 2015). Most recently, we have seen an entire thesis on incorporating pronoun function into MT (Guillou, 2016), the main point being that pronouns should be handled according to their function- both in terms of handling within SMT and in terms of evaluation.

However progress has been hard, and Hardmeier (2014) suggests that besides evaluation problems, this is due to a failure to fully grasp the extent of the pronoun resolution problem in a crosslingual setting, and that anaphoric pronouns

in the source text cannot categorically be mapped onto target pronouns. If these issues can be successfully addressed, it will mark significant progress for MT output in general, and indirectly for coherence.

In her thesis Loaiciga Sanchez (2017) focuses on pronominal anaphora and verbal tenses in the context of machine translation, on the basis that a pronoun and its antecedent (the token which gives meaning to it), or a verbal tense and its referent, can be in different sentences and result in errors in MT output, directly impacting cohesion. She reports direct improvements in terms of BLEU scores for both elements. Again one cannot help wondering whether the improvement in terms of quality of the text as a whole is actually much higher than reflected in the improvements over BLEU score.

Verb tense In specific work on verbs, Loaiciga et al. (2014) researches improving alignment for non-contiguous components of verb phrases by POS tags and heuristics. They then annotated Europarl and trained a tense predictor which they integrate in an MT system using factored translation models, predicting which English tense is an appropriate translation for a particular French verb. This results in a better handling of tense, with the added benefit of an increased BLEU score.

Again on verbs, but this time with a focus on the problems that arise in MT from the verb-particle split constructions in English and German, Loáiciga and Gulordava (2016) construct test suites and compare how syntax and phrase-based SMT systems handle these constructs. They show that often there are alignment issues (with particles aligning to null) which lead to mistranslations, and that the syntax-based systems performed better in translating them.

2.1.3 Lexical Cohesion and WSD in SMT

Lexical Cohesion There has been work in the area of lexical cohesion in MT assessing the linguistic elements which hold a text together and how well these are rendered in MT.

Tiedemann (2010) attempts to improve lexical consistency and to adapt statistical models to be more linguistically sensitive, by integrating contextual de-

pendencies via a dynamic cache model. However, he found that his cache based adaptive models failed to improve the translation quality, suggesting that this may be due to the simplistic language model, the difficulties in optimizing, or to the propagation of errors. Gong et al. (2011) conducts a similar experiment but more finely tuned, with a dynamic cache, a static cache and a topic cache. This work is a document level approach to improve lexical choice and consistency, inspired by the practices of human translators in selecting appropriate lexical items. It results in an improved Bilingual Evaluation Understudy (BLEU) score. The previously mentioned *semantic document language model* (Hardmeier, 2012), integrates an element of lexical cohesion, rewarding the use of semantically related words (determined via LSA) and integrated as an extra feature function.

Carpuat and Simard (2012) find 'consistency does not correlate with translation quality', which is directly contrary to the findings by Wong and Kit (2012) (see below) and to our own findings, namely that consistency is crucial for lexical cohesion and general coherence. However this claim is made in reference to statistics obtained from using a very small training corpus versus a large one (for the Chinese-English language pair). It is no surprise that the small model has more repeated phrases, since it will have a small phrase table. This does not mean that it is consistently correct, and from the low BLEU score it clearly does not match the reference, which presumably has correct terminology. This also indicates that it would fail miserably on any words not seen in training. They do point out that consistency does not guarantee correctness either. Their study considers both in-domain data and general web data. A look at their results shows that, for example, while the number of repeated phrases in the reference translation was 25300, for the SMT this was 79248. This is a huge discrepancy, and surely indicates a considerable reduction in lexical choice in the case of the SMT, if this is indeed to be understood as meaning that phrases were repeated in the SMT where the reference translation used different forms. While there is a case for arguing that MT systems can be more consistent than human translators for using a set terminology (Carpuat and Simard, 2012), that would only be valid for a very narrow field, perhaps a highly technical domain, and an SMT system trained on in-domain data. Otherwise it would only account for a few set terms, unlike the dramatic difference illustrated here, and could well indicate a simplis-

tic style of text which potentially missed much of the semantic and pragmatic richness inherent in the source text.

Wong and Kit (2012) study lexical cohesion as a means of evaluating the quality of MT output at document level, but in their case the focus is on it as an evaluation metric. Their research supports the intuition we found, i.e. that human translators intuitively ensure cohesion, which in MT output often is represented as direct translations of source text items that may be inappropriate in the target context. They conclude that MT needs to learn to used lexical cohesion devices appropriately.

These findings are echoed by Beigman Klebanov and Flor (2013) in their research; although the latter consider pairs of words and define a metric calculating the *lexical tightness* of MT versus Human Translation (HT). The fact that they had to first improve on the raw MT output before the experiment, indicates that it was of insufficient quality in the first place, however this is perhaps due to the age of data (dating to 2008 evaluation campaign), as MT has progressed considerably since then.

Some research has been done on topic models such as by Eidelman et al. (2012), where they compute lexical probabilities which are conditioned on the topic. The model is constructed on training data using Latent Dirichlet Allocation, and using topic dependent lexical probabilities as features. This favours a more context dependent vocabulary choice, in effect more like lexical cohesion, which will however influence coherence indirectly.

Xiong and Zhang (2013) attempt to improve lexical coherence via a topic-based model, using a Hidden Topic Markov Model to determine the topic in the source sentence. They extract a coherence chain for the source sentence, and project it onto the target sentence to make lexical choices during decoding more coherent. They report very marginal improvement with respect to a baseline system in terms of automatic evaluation. This could indicate that current evaluation metrics are limited in their ability to account for improvements related to discourse. Xiong et al. (2013a) focus on ensuring lexical cohesion by reinforcing the choice of lexical items during decoding. They subsequently compute lexical chains in the source text, project these onto the target text, and integrate these into the decoding process with different strategies. This is to try and ensure that

the lexical cohesion, as represented through the choice of lexical items, is transferred from the source to target text. Gong et al. (2015) attempt to integrate their lexical chain and topic based metrics into traditional BLEU and METEOR scores, showing greater correlation with human judgements on MT output.

In their work on comparative crosslingual discourse phenomena, Lapshinova-Koltunski (2015a) find that the use of various lexical cohesive devices can vary from language to language, and may depend also on genre. In a different context, Mascarell et al. (2014) experiment with enforcing lexical consistency at document level for coreferencing compounds. They illustrate that for languages with heavy compounding such as German, translations of coreferencing constituents in subsequent sentences are sometimes incorrect, due to the lack of context in SMT systems. They experiment with two SMT phrase-based systems, applying a compound splitter in one of them, caching constituents in both systems, and find that besides improving translations the latter also results in fewer out-of-vocabulary nouns. Guillou (2013) investigates lexical cohesion across a variety of genres in HT, in an attempt to determine standard practice among professional translators, and compare it to output from SMT systems. She uses a metric (HerfindahlHirschman Index) to determine the terminological consistency of a single term in a single document, investigating consistency across words of different POS category. She finds that in SMT consistency occurs by chance, and that inconsistencies can be detrimental to the understanding of a document.

One of the problems with repetition is indeed automatically recognising where it results in consistency, and where it works to the detriment of lexical variation.

Word Sense Disambiguation The very nature of languages is such that there is no one-to-one mapping of a word in one language to a word in another. A particular word in the source could be semantically equivalent to several in the target, and there is a need to disambiguate. Word Sense Disambiguation (WSD) is one of the areas where SMT output must be improved in order to ensure the correct semantic framework for coherence.

Mascarell et al. (2015) use trigger words from the source text to try to disambiguate translations of ambiguous terms, where a word in the source language can have different meanings and should be rendered with a different lexical item

in the target text depending on the context it occurs in.

Carpuat (2013)'s work on WSD in MT provides a semantic evaluation of MT lexical choice, treating it like a cross-lingual WSD task, which means that the evaluation of MT output is focussed on this one issue. She reports that a SMT phrase based system does not perform as well as a cross-lingual WSD one, indicating that improvements can be leveraged here.

Xiong and Zhang (2014)'s sense-based SMT model tries to integrate and reformulate the WSD task in the translation context, predicting possible target translations in a similar manner to an earlier work by Vickrey et al. (2005). The latter discovered that correct translation of a word in context improves performance on a 'simplified machine translation task', where they determine the correct semantic meaning of the source word to more accurately select target word. This in itself represents significant progress in furthering coherence in SMT. Zhang and Ittycheriah (2015) experiment with three types of document level features, using context to try and improve WSD. They use context on both target and source side, and establish whether the particular alignments had already occurred in the document, to help in disambiguating the current hypothesis. Experimenting with the Arabic-English language pair, they show an increased BLEU score and a decreased error rate.

2.1.4 Discourse connectives

Discourse connectives, also known as discourse markers, are cues which signal the existence of a particular discourse relation, and are vital for the correct understanding of discourse. Yet current MT systems often fail to properly handle discourse connectives for various reasons, such as incorrect word alignments, the presence of multiword expressions as discourse markers, and the prevalence of ambiguous discourse markers. These can be incorrect or missing in the translation (Meyer and Popescu-Belis, 2012; Meyer and Poláková, 2013; Steele, 2015; Yung et al., 2015). In particular, where discourse connectives are ambiguous, e.g. some can be temporal or causal in nature, the MT system may choose the wrong connective translation, which distorts the meaning of the text. It is also possible that the discourse connective is implicit in the source, and thus needs to

be inferred for the target. While a human translator can detect this, current MT systems cannot.

The COMTIS[1] project (2010-2013) focused on various aspects of discourse, and produced a number of papers[2], one of which specifically mentioned coherence (Cartoni et al., 2012). In general, the project covered various aspects of coherence particularly concentrating on disambiguation of connectives, anaphora and verb tenses, attempting to automatically identify these as Inter Sentential Dependencies (ISDs) (Cartoni et al., 2012). They attempted both adding labelled connectives to the existing phrase table, and also training a system to learn from the labelled information. The successor of COMTIS, MODERN[3], specifically focusses on discourse relations and referring expressions and integrating these into MT. COMTIS lay the foundations for general multilingual discourse work, discovering for example the issues with comparability of the corpora which form the basis for much of the current SMT work (Cartoni et al., 2011).

Much of their work on discourse connectives involved *translation spotting*, identifying common translations of particular discourse connective pairs (Cartoni et al., 2013). They then annotated corpora with the discourse sense, in order to disambiguate for translation purposes and to improve the SMT models (Meyer and Popescu-Belis, 2012; Meyer, 2011; Meyer and Poláková, 2013). They also find that there are mismatches, where discourse connectives are implicit in one language yet explicit in another, which influences the quality of MT output.

They developed an ACT metric (Hajlaoui and Popescu-Belis, 2013) to measure the correctness of discourse connectives in MT output. Looking at a limited list of ambiguous connectives, they establish static mappings of corresponding discourse connectives in both languages. Their metric focuses on these seven ambiguous English connectives and the translations of these into French and Arabic. In their work they use a reference translation and alignments to determine the correctness of the discourse connective in the target text.

Li et al. (2014a) research ambiguity in discourse connectives for the Chinese-English language pair. In subsequent work, they report on a corpus study into

[1]Improving the Coherence of Machine Translation Output by Modeling Intersentential Relations

[2]http://www.idiap.ch/project/comtis/publications

[3]Modelling Discourse Entities and Relations for Coherent Machine Translation

discourse relations and an attempt to project these from one language to another Li et al. (2014b). They find that there are mismatches between implicit and explicit discourse connectives. For the same language pair, Yung et al. (2015) research how discourse connectives which are implicit in one language (Chinese), may need to be made explicit in another (English). This is similar to work by Steele (2015) who uses placeholder tokens for the implicit items in the source side of the training data, and trains a binary classifier to predict whether or not to insert a marker in the Target Text (TT). This notion of *explicitation*, and the opposite *implicitation*, is the subject of research by Hoek et al. (2015), which find that implicitation and explicitation of discourse relations occurs frequently in human translations. There seems to be a degree to which the implicitation and explicitation of discourse relations depends on the discourse relation they signal, and on the language pair in question. In addition, there is the fact that in their role as mediator, a human translator may often choose to make explicit a discourse relation that is implicit in the source (Hatim and Mason, 1990).

2.1.5 Discourse Relations

Discourse relations have long been recognised as crucial to the proper understanding of a text (Knott and Dale, 1994), as they provide the structure between units of discourse (Webber et al., 2012). Discourse relations can be implicit or explicit. If explicit, they are generally signalled by the discourse connectives (previous section).

While Marcu et al. (2000) and Mitkov (1993) previously investigated coherence relations as a means of improving translation output and ensuring it was closer to the target language this was taken no further at the time. One of the aims of the work by Marcu et al. (2000) was to "re-order the clauses and sentences of an input text to achieve the most natural rendering in a target language", due to the mismatch between source and target, and the limitations of MT systems. Taking inspiration from Rhetorical Structure Theory (RST), Tu et al. (2013) proposed an RST-based translation framework on basis of elementary discourse units (EDU)s, in an attempt to better segment the source text in a meaningful manner, and ensure a better ordering for the translation. This approach is more

sensitive to discourse structure, and introduces more semantics into the SMT process. Their research uses a Chinese RST parser and they aim to ensure a better ordering of EDUs, although the framework still has a limited sentence-based window.

There have been previous experiments specifically assessing discourse relations in an MT context. Guzmán et al. (2014) used discourse structures to evaluate MT output. They hypothesize that the discourse structure of good translations will have similar discourse relations to those of the reference. They parse both MT output and the reference translation for discourse relations and use tree kernels to compare HT and MT discourse tree structures (i.e. monolingually). They improve current evaluation metrics by incorporating discourse structure on the basis that 'good translations should tend to preserve the discourse relations' of a reference (Guzmán et al., 2014).

2.1.6 Negation

There has also been work on negation in MT, decomposing the semantics of negation and with an error analysis on what MT systems get wrong in translating negation (Fancellu and Webber, 2015a). For the language pair which they considered (Chinese-English) the conclusion was that determining the scope of negation was the biggest problem, with reordering the most frequent cause. Subsequently, Fancellu and Webber (2015b) show that the translation model scoring is the cause of the errors in translating negation. In general, MT systems often miss the focus of the negation, which results in incorrectly transferred negations that affect coherence.

2.2 Existing Coherence Models

There are various forms of monolingual coherence models, some modelling specific aspects of coherence and other, more general, neural network ones, which aim to capture coherence patterns automatically. None of these have been deployed in an MT context before our work. We briefly describe them in the following sections.

2.2.1 Entity-based coherence models

The entity-based approach derives from the theory that entities in a coherent text are distributed in a certain manner (Lapata, 2005; Barzilay and Lapata, 2008), as identified in various discourse theories, in particular in Centering Theory (Grosz et al., 1995). Centering theory builds on one of the three subcomponents mentioned earlier in Section 1.4.1, that of *attentional structure* (Grosz and Sidner, 1986). Clarke and Lapata (2010) used Centering to further their research for coherence in document compression, tracking the backward-looking centre of each sentence, i.e. the highest ranking element in the current sentence that is also in previous sentence. Kehler (1997) assesses several approaches that apply this to pronoun interpretation. As detailed by Poesio et al. (2004), these models apply a linguistic theory to a computational setting. As a result they simply *approximate* an implementation of Centering theory, making simplifications such as evaluating the entity patterns at sentence level, while the theory itself refers to *utterances* (Grosz et al., 1995).

Entity Grids Entity grids are constructed by identifying the discourse entities in the documents under consideration, and constructing a 2D grid whereby each column corresponds to the entity, i.e. noun, being tracked, and each row represents a particular sentence in the document (Lapata, 2005; Barzilay and Lapata, 2008). This theory holds that coherent texts are characterised by salient entities in strong grammatical roles, such as subject or object. The entity grid model has also been usefully extended to encompass both a local and global model, in addition to entity-specific enhancements (Elsner et al., 2007; Elsner and Charniak, 2011b). Elsner and Charniak (2011a) further adapt and apply it to the domain of chat disentanglement. This approach has been applied to assess readability in student essays (Burstein et al., 2010), and in combination with discourse relations (Pitler and Nenkova, 2008). Burstein et al. (2010) use the entity-grid for student essay evaluation, which is a scenario closer to ours. They used a range of additional features to take account of grammaticality, type token ratios, etc. These proved useful for discriminating good from bad quality essays, but it is unclear how much difference the array of additional features made.

Filippova and Strube (2007) apply the entity grid approach to German, and investigate whether grouping related entities and thus incorporating semantic relatedness was adequate, in the absence of syntactic information. The results with German were not as successful as English, however they judged the entity clustering promising. Research has indeed indicated that the syntactical weightings of the standard grid setup will not hold in German due to the German clausal structure, where word order of the subclause is affected (Cheung and Penn, 2010).

In a slightly different vein, although still entity-based, Somasundaran et al. (2014) see lexical chains as 'a sequence of related words that contribute to the continuity of meaning based on word repetition, synonymy and similarity', and consider how lexical chains affect discourse coherence quality. They use lexical chaining features such as their length, density, and link strength to detect textual continuity, elaboration, lexical variety and organisation, all vital aspects of coherent texts. Moreover, the interaction between chains and discourse cues can also show whether cohesive elements have been organised in a coherent fashion. Results again indicate that the best performance is achieved by combining these features with other discourse features.

Entity Graphs Guinaudeau and Strube (2013) converted a standard entity grid into a bipartite graph which tracks the occurrence of entities throughout the document, including between non-adjacent sentences, and achieving equal performance without training. They use it to capture the same entity transition information as the entity grid model, although they only track the occurrence of entities, and additionally can track cross-sentential references. They also claim that they can calculate the local coherence directly, without the need for feature vectors and a learning phase (Guinaudeau and Strube, 2013). Subsequent recent extensions to the entity graph include a method of normalization (Mesgar and Strube, 2014) and incorporating word embeddings (Mesgar and Strube, 2015).

2.2.2 Syntax-based models

Motivated by the strong impact syntax has in text coherence, Louis and Nenkova (2012) propose a coherence model which is based on syntactic patterns. It at-

tempts to measure one of the three subcomponents mentioned earlier in Section 1.4.1, that of *intentional structure* (Grosz and Sidner, 1986). The premise is that a document has an overall discourse purpose, and is composed of sentences which each have a communicative goal. This will vary according to genre. Syntax patterns are extracted from documents marked up with parse trees, in Penn Discourse Tree Bank (PDTB) format, and they establish coherence patterns typical to specific discourse types which identify the intentional discourse structure. The focus of their work was in using this knowledge via patterns, in terms of prominent syntactic constructions, to distinguish coherent from non-coherent texts. They use an Hidden Markov Model (HMM) to learn the document-wide patterns, and apply this on a local and global level to predict coherence. Their experiment again tested on ranking ordered versus shuffled texts, and achieves this to a certain extent, however there is no evidence that it actually measures the intentional structure of discourse.

2.2.3 Discourse relational models

Lin et al. (2011) use a discourse parser to determine discourse relations and their types across adjacent sentences. They construct a grid similar to the entity grid (see Section 2.2.1), but tracking all open class words (not just nouns) and recording the discourse relation in the cell for each (stemmed) lemma. They evaluate the coherence of the text from discourse role transitions and patterns on the basis that there is a *preferential, canonical, ordering* of discourse relations that leads to improved coherence. This represents the *linguistic structure* mentioned in Section 1.4.1. The results indicate similar or improved performance over a regular entity grid, with significantly improved performance when both are combined. They conclude that the combined model is linguistically richer as both models capture different aspects of coherence. This notion of *ordering* to produce a coherent text was previously researched by Lapata (2003), in experiments where they learn constraints on ordering and discover via various features the importance of syntactic and lexical information. Pitler and Nenkova (2008) experiment with discourse relations for assessing readability, and on the basis that discourse relations are considered 'a major factor in text coherence', show that discourse relations are

linked to text quality. They also find that both explicit and implicit relations are necessary, and conclude that 'using a combination of entity coherence and discourse relations produces the best performance'.

Some interesting research which predates the rise of SMT is that of Ghorbel et al. (2001) on using discourse structure of parallel texts to ensure coherence and cohesion. They adopt a semantic and pragmatic approach, inspired by RST, to map parallel spans of text represented as tree structures. They do this by extracting the *salient path* of a tree, navigating from root to terminal via nucleus nodes, which is seen as helping the readers follow the sense and content of the tree. While they develop this approach for the problem of text alignment, it seems not only insightful but, if reapplied in a different context, potentially an interesting way of mapping coherence transfer between source and target texts.

2.2.4 Neural network models

Recently, among the large amount of work involving neural network models there have been variations applied to the problem of coherence, establishing whether deep learning models have the ability to capture coherence (in a monolingual setting). Li and Hovy (2014) developed a coherence model based on distributed sentence representation. They used recurrent and recursive neural networks to perform sentence ordering and readability tasks. They leverage semantic representations to establish coherent orderings, using original texts as positive examples and shuffled versions as negative examples, for optimising the neural networks. Li et al. (2015) train a hierarchical Long-Short Term Memory (LSTM) to explore neural Natural Language Generation, and assess whether the local semantic and syntactic coherence can be represented at a higher level, namely paragraphs. In their model, one LSTM layer represents word embeddings, another represents sentences, and another paragraphs. They are then able to regenerate the text to a degree that indicates that neural networks are able to capture certain elements of coherence. Lin et al. (2015) use a hierarchical Recurrent Neural Network Language Model (RNNLM) to combine a word level model with a sentence level model for document modelling. They claim that their model captures both intra- and inter-sentential sequences.

Ji et al. (2016) develop various Document Context Language Models (DCLM), to combine both local and global information into the RNNLM architecture. They try three different models, with the stated aim of integrating contextual information from the RNNLMs of the previous sentence into the language model of the current sentence. They find that all variations outperform the standard RNNLM with their context-to-context DCLM performing best. All of them pass on contextual information via hidden states, this one directly impacts the generation of each word in current sentence specifically (the others impact the output or try an attentional mechanism).

In general, these models attempt to automatically learn the elements which contribute to coherence. As Manning (2015) states with reference to the deluge of deep learning models which have appeared recently, particularly since 2015, 'it would be good to return some emphasis in NLP to cognitive and scientific investigation of language rather than almost exclusively using an engineering model of research'. In our work we focus on trying to identify the linguistic elements involved in crosslingual coherence.

2.3 Summary

As we have seen in this review, the discourse contributions in MT recently (Section 2.1) have been various forms of referential cohesion, lexical cohesion, discourse connectives, and negation. While there are some existing coherence models (Section 2.2), such as the entity-based ones, syntax models and experiments with discourse connectives, there are none specifically for predicting or improving coherence in the area of SMT. While the task of automatically evaluating text coherence has been addressed previously, within applications such as multi-document text summarisation or in terms of optimal ordering within shuffled texts, our aim is to further investigate these components in an MT context without the use of a reference translation. We ultimately expect to be able to evaluate coherence in MT.

Previous research in coherence has found a multi-faceted approach works best (Poesio et al., 2004), and this confirms our belief that there are different aspects of coherence which need to be captured. In Chapter 3 we detail our experiments

extending entity-based models in a more focused manner to an SMT context with some of the other coherence models (described in Sections 2.2.2 and 2.2.3), adapting them to better suit the new task.

Chapter 3

Coherence Models in Machine Translation

As detailed in Section 2.2, previous coherence models have been developed for monolingual contexts, assessing the coherence of texts which are either extractive summarizations or are artificially formed from shuffled sentences of existing, coherent texts. In this chapter we examine how these models perform on the new task of assessing coherence of MT output, reimplementing the most popular coherence models in the literature, including two entity models (Sections 3.1.1 and 3.1.2) and a syntax-based model (Section 3.2.1) in our experiments. We then report our improvement over the syntax-based model (Section 3.2.2), which outperforms the state-of-the-art in the original (shuffling) task.

We illustrate the difference between assessing the output from MT systems and assessing the coherence of shuffled texts in a highly consistent, structured corpus. Here, our objective with these models is to assess whether the coherence models allow us to discriminate between HT and MT. Our hypothesis is that a good coherence model should be able to score human translations as having higher coherence than their counterpart machine translations in most cases. We also hypothesize that patterns of syntactic items between adjacent sentences can be better modelled through a latent alignment.

3.1 Entity-based models

We previously established (Section 6.1) that the entity-based approach derives from the idea that coherent texts are characterised by salient entities in strong grammatical roles, such as subject or object. The focus of the entity-based approach is on using this knowledge via patterns in terms of prominent syntactic constructions to distinguish coherent from non-coherent texts.

3.1.1 Entity-grid approach

The entity-based grid was first proposed by Lapata (2005) and Barzilay and Lapata (2005) with the aim of measuring local coherence in a monolingual setting. Generally the task consists of automatically assessing coherence either in an experiment ranking alternative automatic text summaries, or ranking alternative sentence orderings (Barzilay and Lapata, 2008). Here *incoherent* documents are created artificially, by randomly shuffling the sentences of the original document to create permutations of it. The task is then either to correctly reorder these automatically, or simply to discriminate the more coherent version.

Entity grids are constructed by identifying the discourse entities in the documents under consideration and representing them in 2D grids whereby each column corresponds to the entity (i.e. noun) being tracked, and each row represents a particular sentence in the document in order. An example can be seen in Figure 3.1, where each row represents consecutive sentences, and the columns (`e1`, etc.) represent different entities. In this example, `e7` represents *Kosovo*, which was repeated in sentences `s2`, `s3` and `s4`, in the roles of `subject` (S), `other` (X), and `subject` (S), respectively.

Once all occurrences of nouns and the syntactic roles they represent in each sentence are extracted, an *entity transition* is defined as a consecutive occurrence of an entity with given syntactic roles. These are computed by examining the grid vertically for each entity. For example, an `SS`, a `Subject-to-Subject` transition, indicates that an entity occurs in a subject position in two consecutive sentences. An `SO`, on the other hand, indicates that while the entity was in a subject role in one sentence, it became the object in the subsequent sentence. Probabilities for each entity transition can be easily derived by calculating the frequency of a

	e1	e2	e3	e4	e5	e6	e7
s1	-	-	-	-	-	-	-
s2	-	-	-	-	-	-	S
s3	-	-	-	-	-	-	X
s4	-	-	O	-	-	-	S
s5	S	-	-	-	-	-	-
s6	-	-	-	X	-	-	-

Figure 3.1: Example of an entity grid: sentences are rows, entities are columns, Entities are recorded in position of: Subject (S), Object (O), or other (X).

particular transition divided by the total number of transitions which occur in that document. The assumption is that incoherent texts have more breaks in the entity transitions, and thus lower scores.

The entity grid has been implemented as a generative model (Lapata, 2005), and a discriminative one (Barzilay and Lapata, 2005). Initially we experimented with the discriminative model, however we now reimplement the generative model, to examine whether it is more suitable for our task. Equation 3.1 shows this formulation, where m is the number of entities, n is the number of sentences in a document D and $r_{s,e}$ is the role taken by entity e in sentence s. This model makes a Markov assumption, under which an entity's role is independent of all but its h preceding roles, where h is the length of the transitions.

$$p(D) = \frac{1}{m \cdot n} \prod_{e=1}^{m} \prod_{s=1}^{n} p(r_{s,e} | r_{(s-h),e} \dots r_{(s-1),e}) \tag{3.1}$$

Probabilities for these transitions can be easily derived by calculating the frequency of a particular transition, normalised by the total number of transitions which occur in that document (the implementation in the previous chapter), or else generatively (as here) by estimating the probabilities of individual transition events.

The Brown Coherence Toolkit[1] represents Elsner and Charniak (2011b)'s work on entity grid models and implements variations of a generative model in English. However, it has been specially trained for tasks of:

- *Discrimination* (testing the model's ability to distinguish between a human-authored document in its original order, and a random permutation of that

[1] http://cs.brown.edu/~melsner/manual.html

document).

- *Insertion* (finding the optimal place to insert each sentence into the document, given the correct ordering of the other sentences).
- *Ordering* (finding the ordering of sentences which is maximally coherent according to the model) as per Elsner et al. (2007).

Our task is different from all three of these modes, as those tasks are performed on texts which are coherent at the outset, so none of these modes are appropriate. We implement a generative entity-grid model where the coherence of a text is calculated as per Lapata (2005), see Equation 3.1. The original model presumes that grids of coherent texts have a few dense columns and many sparse ones, and that entities occurring in the dense columns will more often be subjects or objects. It assumes that these characteristics are less common in texts exhibiting lower coherence (Lapata, 2005).

Quantitative results for the experiments with the entity-grid model are given in Table 3.4 and discussed in Section 3.4.

3.1.2 Entity graph approach

As mentioned in Section 2.2.1, Guinaudeau and Strube (2013) framed the entity grid into a graph format, using a bipartite graph which they claim had the advantage both of avoiding the data sparsity issues encountered by Barzilay and Lapata (2008) and of achieving equal performance on measuring overall document coherence without the need for training (the grid required training to compute the entity transition probabilities). They use it to capture the same entity transition information as the entity grid experiment, although they only track the occurrence of entities, avoiding the nulls or absences of the other (tracked as '-' in the entity grid framework). Additionally, the graph representation can track cross-sentential references, not just those in adjacent sentences. Here too we track the presence of all entities, taking all nouns in the document as discourse entities, as recommended by Elsner and Charniak (2011b).

The coherence of a text in this model is measured by calculating the average outdegree of a projection, progressively summing the shared edges between the

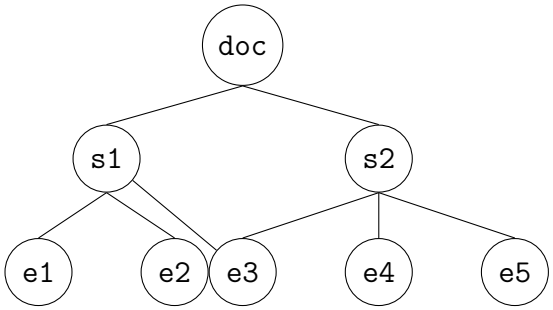

Figure 3.2: Entity Graph: `s1` and `s2` are sentences, `e1-e5` are entities occurring in the document.

sentences in the document. Edges are created where there are shared entities between sentence nodes. For illustration, Figure 3.2 shows a document extract, displaying two sentences with a total of five entities between them, of which one (`e3`) is shared, i.e. occurs in both sentence one (`s1`) and sentence two (`s2`).

This metric takes the format of a directed graph, encompassing the edges between one sentence and any subsequent ones in the text. Formally, the coherence of a document in Guinaudeau and Strube (2013) is shown in Equation 3.2. This is a centrality measure based on the average outdegree across the `N` sentences represented in the document graph. The outdegree of a sentence s_i, denoted $o(s_i)$, is the total weight leaving that sentence, a notion of how connected (or how central) it is. This weight is the sum of the contributions of all edges connecting s_i to any $s_j \in D$. The total contribution $W_{i,j}$ of a pair of sentences (s_i, s_j) is a simple weighted average, namely, $W_{i,j} = \sum_{e \in E_{i,j}} w(e, s_i) \cdot w(e, s_j)$, where $E_{i,j}$ is the set of entities common to the pair, and $w(e, s_i)$ quantifies the importance of the role of `e` in s_i.

$$\begin{aligned} s(D) &= \frac{1}{N} \sum_{i=1}^{N} o(s_i) \\ &= \frac{1}{N} \sum_{i=1}^{N} \sum_{j=i+1}^{N} W_{i,j} \end{aligned} \tag{3.2}$$

They define three types of graph projections: *binary*, *weighted* and *syntactic*,

which affect the weights given to each edge. Binary projections simply record whether two sentences have any entities in common. Weighted projections take the number of shared entities into account, rating the projections higher for more shared entities. A syntactic projection includes syntax, where syntactic information is used to weight the importance of the link by calculating an entity in role of subject (S) as a 3, an entity in role of object (O) as a 2, and other (X) as a 1. These are projected between two sentences, following the sequential order of the text, as sets of shared entities.

We projected the entity relationships onto a graph-based representation, experimenting in various settings. Our objective was to assess whether the graph metric gives us a better appreciation of differences in entity-based coherence across languages. This representation can encode more information than the entity-grid as it spans connections not just between adjacent sentences, but among all sentences in the document.

We reimplemented the algorithm in Guinaudeau and Strube (2013) using the *syntactic projection* in this instance, and ran experiments with the same objective and datasets as for the grid model.

Quantitative results for the experiments with the entity graph model are also given in Table 3.4 and discussed in Section 3.4.

3.2 Syntax-based models

3.2.1 Syntax-based model

Motivated by the strong impact syntax has in text coherence, Louis and Nenkova (2012) propose both a local and a global coherence model based on syntactic patterns. Our implementation focuses on their local coherence model. It follows the hypothesis that, in a coherent text, consecutive sentences will exhibit syntactic regularities, and that these regularities can be captured in terms of co-occurrence of syntactic items. In their approach, which they describe as addressing the *intentional structure* of Grosz and Sidner (1986)'s theory, their assumption is that a document has an overall discourse purpose and is composed of sentences which each have a communicative goal. There will be particular patterns in

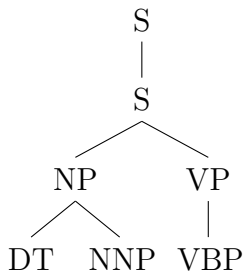

Figure 3.3: Extract d-sequences or grammar productions, annotated with the leftmost child node. So at a sequence of depth 3 : $\mathrm{NP}_{\mathrm{DT}}$ $\mathrm{VP}_{\mathrm{VBP}}$ or a sequence of depth 2 : S_{NP}. Alternatively we can extract context-free grammar productions : S $\rightarrow$ NP VP)

adjacent sentences which are in line with the communicative goal. This syntactic coherence reflects the fact that a sentence of a particular type (e.g. speculation) is likely followed by another type (e.g. endorsement), and that these patterns can be automatically detected.

The syntax patterns extracted can either be context-free grammar productions (e.g. S $\rightarrow$ NP VP) or *d*-sequences (a sequence of sibling constituents at depth *d* starting from the root, possibly annotated with the left-most child node they dominate, e.g. $\mathrm{NP}_{\mathrm{NN}}$ $\mathrm{VP}_{\mathrm{VB}}$). By way of illustration we include an extract from a parse tree (see Figure 3.3), from which either d-sequences or grammar productions can be extracted at varying depths.

The model conditions each sentence on the immediately preceding sentence, where both are seen as *pairs* of syntactic patterns from adjacent sentences. Each sentence is assumed to be generated one pattern at a time and patterns are assumed to be independent of each other. The parameters of the model are *unigram* and *bigram* patterns over a vocabulary of syntactic items (i.e. productions or *d*-sequences) which are directly observed from training data by relative frequency counting. $|V|$ is the size of the vocabulary of syntactic items.

$$p(D) = \prod_{(u_1^m, v_1^n) \in D} \prod_{j=1}^{n} \frac{1}{m} \sum_{i=1}^{m} \frac{c(u_i, v_j) + \alpha}{c(u_i) + \alpha|V|} \tag{3.3}$$

The coherence of a document under the model is given by Equation 3.3, where

(u_i^m, v_j^n) represents adjacent sentences, and $c(\cdot)$ is a function that counts how often a pattern (or a pair of patterns) was observed in the training data. To account for unseen syntactic patterns at test time, their model is smoothed by a constant α.

In our experiments, we derived the syntactic items in the form of the d-sequence, defined as the leaves of the parse tree at a given depth (we experiment at depths 2, 3, 4), and annotated with the left-most leaf. The choice of d-sequences results in what we believe to be an informative representation. Further experiments could use grammatical productions as an alternative.

3.2.2 Syntax-based model with IBM 1

The model introduced by Louis and Nenkova (2012) assumes that every pattern in a preceding sentence may equally be responsible for a pattern in the following sentence. We experiment with a truly generative model with a similar parametrisation to that of Louis and Nenkova's model. We introduce alignments between syntactic patterns in adjacent sentences as a latent variable. Our model is similar to the IBM model 1 (Brown et al., 1993), where the current sentence is generated by the preceding one, one pattern at a time, with a uniform prior over alignment configurations. The latent alignment variable allows us to model the fact that some patterns are more likely to trigger particular subsequent patterns.

In IBM model 1, a latent alignment function a maps patterns in v_1^n (current sentence) to patterns in u_0^m (preceding sentence), where u_0 is a special NULL symbol which models instances where there is no direct alignment. Here n is the current sentence and m the preceding sentence. The score of a document is given by Equation 3.4.

$$P(D) = \prod_{(u_1^m, v_1^n) \in D} p(v_1 \dots v_n, a_1 \dots a_n | u_0 \dots u_m) \tag{3.4}$$

As the alignment is hidden, we marginalise over all possible configurations, which is tractable due to an independence assumption (namely that items align independently of each other). Equation 3.5 shows this tractable marginalisation.

$$p(D) = \prod_{(u_1^m, v_1^n) \in D} \prod_{j=1}^{n} \sum_{i=0}^{m} p(v_j|u_i) \tag{3.5}$$

We use Expectation Maximisation (EM) to estimate the parameters in Equation 3.5 (Brown et al., 1993). As we observe more data this model converges to better parameters. A similar solution was proposed in a different context by Soricut and Marcu (2006) in their work on word co-occurrences.

To avoid assigning zero probability to documents containing unseen patterns, we modify the training procedure to treat all the singletons as pertaining to an unknown category (UNK), thus reserving probability mass for future unseen items.[1] In addition to this special UNK item, we also include NULL alignments, which together with UNK will smooth the bigram counts.

We report results for both syntax experiments in Sections 3.3.4 and 3.3.5.

3.3 Experiments and Results

3.3.1 Datasets

To estimate the parameters of the entity-grid and syntax-based models (i.e. distribution over entity role transitions and syntactic patterns), we use the most recent portion of English LDC Gigaword corpus, randomly checking the quality and excluding two sections deemed to be of inadequate quality.[2] Table 3.1 displays information about the size of these datasets.

To test our models on the translation task, we use news WMT14 test data as corpus (Bojar et al., 2014), considering submissions from all participating MT systems (including statistical, rule-based, and hybrid) in the translation shared task for three language pairs, namely, 13 German-English (de-en) systems, 9 French-English (fr-en) systems and 13 Russian-English (ru-en) systems.

We make the assumption that the HT (reference) is a coherent text, and that the MT output may or may not be coherent. While the former is a fair assump-

[1] The hypothesis, backed by Zipf's law, is that unseen items are singletons that we have not yet observed, and that singletons we have observed would remain so if we observed some more data.

[2] https://catalog.ldc.upenn.edu/LDC2003T05

Table 3.1: Number of documents and sentences in the training (Gigaword) and test (WMT14) sets.

Corpus	Portion	Documents	Sentences
Gigaword	12/2010	41,564	774,965
WMT14	de-en	164	3,003
WMT14	fr-en	176	3,003
WMT14	ru-en	175	3,003
Accidents & Earthquakes	Earthquakes	99	1,254
Accidents & Earthquakes	Accidents	100	1,228

tion, we acknowledge that many outputs from MT systems may be somewhat coherent. However, we are not aware of any datasets with translated data which have been annotated for coherence. This is a challenging task in itself, since judging coherence is a complex and subjective task which requires, at the very least, well trained annotators.

For the shuffling task we also use the MT data, taking the HT as the coherent texts and shuffled versions of them to create incoherent ones.By way of comparison on this shuffling task, we use the corpus widely used for coherence prediction, the Earthquakes and Accidents corpus[1], which consists of short articles (averaging 10.4 and 11.5 sentences in length, respectively (Barzilay and Lapata, 2008)), many short sentences and a well-defined structure. The corpus contains 100 documents relating to accidents, and 99 relating to earthquakes, in addition to the shuffled permutations.

We used the Stanford CoreNLP toolkit (Manning et al., 2014) to parse the English MT and HT output.

3.3.2 Metrics

We evaluated the results according to a number of metrics, defined as follows. Let m be a model, $d \in D$ a document, r the reference or original (non-shuffled) version and s the shuffled or MT output. Then let $\text{win}_m(d_r, d_s)$ return 1 if model m scores reference document d_r higher than a shuffled or MT document d_s, and 0 otherwise. We define tie as where they score the same. Finally, $\text{first}_m(d_r)$ returns

[1] http://people.csail.mit.edu/regina/coherence/CLsubmission/

1 if the reference ranks first, and $\text{solo}_m(d_r)$ returns 1 if the reference occupies a position alone in the ranking. Our various model evaluation methods are defined as follows:

ref$_>$ how often a model ranks reference documents strictly higher than any of their shuffled or MT counterparts: $\frac{1}{|D||S|}\sum_d\sum_s \text{win}_m(d_r, d_s)$

ref$_\geq$ how often a model ranks the reference no worse than any of their shuffled or MT counterparts: $\frac{1}{|D||S|}\sum_d\sum_s \text{win}_m(d_r, d_s) + \text{tie}_m(d_r, d_s)$

ref$_{1^*}$ how often the reference is ranked strictly higher than every other system: $\frac{1}{|D|}\sum_d \text{first}_m(d_r) \times \text{solo}_m(d_r)$

3.3.3 Model descriptions

Grid represents our generative implementation of the entity grid

Graph represents our implementation of the entity graph

LN-dx represents our implementation of the original syntax model, where x signifies the depth of the parse trees

IBM1-dx represents our IBM1 syntax model, where x signifies the depth of the parse trees

3.3.4 Results on shuffling task

To test our hypothesis that patterns of syntactic items between adjacent sentences can be better modelled through a latent alignment, we conducted the traditional shuffling experiment with our reference text and a randomly shuffled version of it[1]. The aim was to check whether our IBM1 formulation for the syntax model outperforms the original syntax model. Thus we are comparing grammatically correct and coherent sentences instead of MT output.

From our results in Table 3.2, it is clear that our adaptation is an improvement over the original syntax model by a large margin. In fact, in most cases

[1]The shuffled texts were created using python library random.shuffle.

Table 3.2: Model comparisons for shuffling experiment on WMT data, ref_{1^*} is "accuracy" used in previous work, $\text{ref}_{\geq}$ is how often a model ranks the reference no worse than any of the MT submissions.

fr-en	ref_{1^*}	$\text{ref}_{\geq}$	de-en	ref_{1^*}	$\text{ref}_{\geq}$	ru-en	ref_{1^*}	$\text{ref}_{\geq}$
IBM1-d3	82.95	85.23	GRID	79.27	80.49	IBM1-d3	79.43	80.00
GRID	75.00	77.84	IBM1-d3	76.83	76.83	GRID	74.86	76.00
IBM1-d4	71.59	73.86	IBM1-d2	71.34	71.34	IBM1-d2	74.86	75.43
IBM1-d2	64.77	67.05	IBM1-d4	63.41	63.41	IBM1-d4	64.57	65.14
GRAPH	50.00	53.98	GRAPH	62.80	65.24	GRAPH	50.29	54.29
LN-d3	46.59	59.66	LN-d4	53.66	62.20	LN-d4	46.29	57.71
LN-d4	41.48	54.55	LN-d2	47.56	59.15	LN-d3	45.14	57.14
LN-d2	38.64	55.11	LN-d3	46.95	54.88	LN-d2	40.57	56.00

it also outperforms the entity grid. Noteworthy is the fact that the ref_{1^*} metric discriminates how often a model ranks the unshuffled documents strictly higher than any other version, not just equal to them, as the $\text{ref}_{\geq}$ does.

The difference between our experiment and those reported elsewhere (Louis and Nenkova, 2012; Barzilay and Lapata, 2008) is that those other experiments have been performed on the aforementioned Earthquakes and Accidents corpus, which is quite specific in nature, as described in 3.3.1. By way of comparison, we also include results on the aforementioned corpus under our models (Table 3.3). Here the $\text{ref}_{\geq}$ metric results for our reimplementation of the syntax model are close those of the original local model with d-sequences (Louis and Nenkova, 2012). Moreover, results for previous grid experiments were obtained using supervised training where the parameters are trained on this same Earthquakes and Accidents corpus, then tested on a heldout section of the same dataset. We adopted a more automated approach, training on more general data, with a view to being applied more widely. This does, however, affect the results, particularly given the nature of the Earthquakes and Accidents corpus.

3.3.5 Results on translation task

We performed our experiments with data drawn from the WMT14 evaluation campaign. This campaign (as well as others in MT) is designed and implemented at the sentence level, and there is no human-annotated, gold-standard data that

Table 3.3: Model comparisons for shuffling experiment on Earthquakes and Accidents corpus, ref_{1*} corresponds to "accuracy" as reported in previous work, $\text{ref}_{\geq}$ is how often a model ranks the reference no worse than any of the shuffled counterparts.

Earthquakes	ref_{1*}	$\text{ref}_{\geq}$
IBM1-d2	80.88	80.88
IBM1-d3	77.10	77.10
Grid	66.21	66.21
Graph	60.53	60.58
LN-d2	57.62	71.73
LN-d3	57.00	67.69

Accidents	ref_{1*}	$\text{ref}_{\geq}$
Graph	86.51	86.51
IBM1-d3	72.61	72.61
IBM1-d2	67.32	67.37
Grid	50.25	50.25
LN-d4	46.58	55.89
LN-d2	38.82	57.15

suits the task of assessing MT coherence (or even general quality) at the document level. Because of that, we can only evaluate our models in terms of how well they distinguish human translated documents (references) from machine translated ones, assessing the submissions by participants of the shared task. This evaluation is conducted under the assumption that the reference documents are coherent. An obvious benefit of such a strategy is that we can assess models automatically and objectively without the need for any particular type of annotation (e.g. reference translations). On the other hand, it is not realistic to assume that every MT is incoherent, thus a limitation of such an experimental setting is that little can be concluded from the particular ranking of MT systems produced by any given coherence model. For this reason, we refrain in this instance from making comparisons between MT systems in our analysis. To provide a concise summary of our findings, we aggregate the results for all MT systems in this section. A breakdown of results per MT system is provided in Chapter 7. Table 3.4 shows the performance of our models according to different evaluation methods (scores are percentages of instances where the HT is correctly identified as higher than the MT), ranked by the first method.

Our results show that all the models tested are more limited in their ability to assess coherence in an MT context, since the task is more difficult than that of distinguishing shuffled from original texts. The models can score machine translated texts as equal to reference translations, and in some cases, even higher than the reference translations. The latter is particularly true for rule-based MT systems, since these systems seem to be more consistent in their use of entities

Table 3.4: Model comparisons for translation task according to different evaluation methods (scores are percentages of instances where the HT is correctly identified as higher than the MT), ranked by the first method.

de-en	$\text{ref}_{>}$	$\text{ref}_{\geq}$	ref_{1^*}
Graph	67.03	68.62	28.66
IBM1-d2	53.52	53.56	12.20
IBM1-d3	53.05	53.05	17.68
IBM1-d4	45.12	45.12	8.54
LN-d3	43.67	60.55	8.54
LN-d4	43.34	53.38	10.37
Grid	37.71	37.71	6.10
LN-d2	30.35	67.78	6.71

ru-en	$\text{ref}_{>}$	$\text{ref}_{\geq}$	ref_{1^*}
Graph	60.84	63.21	20.57
IBM1-d3	58.02	58.02	10.86
IBM1-d2	57.41	57.54	13.14
IBM1-d4	52.57	52.57	10.29
LN-d3	48.62	63.47	9.14
LN-d4	47.21	58.42	8.57
LN-d2	34.59	65.58	4.00
Grid	31.38	31.38	5.14

fr-en	$\text{ref}_{>}$	$\text{ref}_{\geq}$	ref_{1^*}
IBM1-d4	58.24	58.66	20.45
Grid	55.54	56.68	22.16
IBM1-d3	54.19	54.62	17.61
LN-d4	45.17	55.82	14.77
Graph	41.62	45.60	11.93
IBM1-d2	41.41	42.19	13.64
LN-d3	41.26	59.23	15.34
LN-d2	26.63	66.26	10.23

and syntactic patterns.

In our experiments with Grid, the MT displays no more sparse columns than the reference counterpart. It would seem that given how pre-eminent the focused nouns are, these are captured in the MT output. There are however differences in transition patterns, in that some patterns are more common in the MT than the HT, such as 'OO', or other patterns with strong object positions. This seems to indicate a more simplistic style by MT systems. In general, however, it is more difficult to distinguish an MT from HT text, unlike detecting the sudden breaks in transitions or shifts of focus which occur in artificially shuffled texts. Judged in our scenario as discriminating coherent from incoherent text, it essentially appears that MT handles entity-based coherence adequately, insofar as there is a similar pattern of entities. Whether these are the *correct* entities is undetermined.

As with the grid experiment, it became apparent again with Graph that the coherence judgement between an MT and an HT text is much more subtle

than between an ordered and shuffled text. Or between a potentially disjointed, automatically-generated summary, and a human summary. The coherence scores were not always automatically higher for the HT documents and lower for the MT ones, although the reference generally scored higher than the MT system. This would seem to indicate either that the model is inadequate or that MT does in fact correctly record the occurrence of particular entities consistently. Again, whether these are the correct entities is another matter. If a text has been automatically summarized or shuffled, the overall logic has clearly been broken, and the challenge then is to rediscover the logic pattern. In MT the situation is more nuanced, as the elements of coherence may be there to some degree, yet it may still be lacking in coherence due to other changes which have occurred in the decoding process.

Our extension of the syntax-based model – IBM1 – consistently outperforms LN according to all metrics. That is because IBM1 learns a distribution over hidden alignments between syntactic items. These alignments give more grounding for certain syntactic patterns. However neither syntax-based model does particularly well, and it is unclear if the syntactic structure is correlated with intentional structure of a text.

Overall, we found that the best coherence model was able to score the human translations higher than any particular MT system for 67%. Some models are clearly more heavily affected by the use of methods that disregard ties. The LN model typically clusters the reference together with MT systems. The other models, especially IBM1 and Graph are less affected by differences in evaluation methods. While the figures change across methods, the trend in the ranking of models is maintained.

In general, IBM1 and Graph come out the strongest in terms of scores, with Grid performing poorly (except for the `fr-en` language pair). Overall Graph performs better than Grid, perhaps because it offers a broader view of entity-based coherence, in that it captures links between all entities in all sentences in the text, including links over non-adjacent sentences. Also, as such it is not as dependent on consecutive transitions. If we disregard ties, Graph features as the best model for two out of the three language pairs, i.e., except for `fr-en`, with IBM1 performing similarly well. Interestingly, there is a difference between

language pairs, which deserves further investigation.

It is worth emphasising that among our three language pairs, `fr-en` is arguably the one which is generally of the highest MT quality. Low translation quality may have affected the performance of the models differently as they rely on linguistic information to different extents. GRID, which performed the best for `fr-en`, relies heavily on the correct identification of nouns and their syntactic roles in sentences. Therefore, for the other languages, an excessive number of ungrammatical or unnatural translations – and unreliable syntactic roles as a consequence – may have affected the model more significantly. Moreover, the `fr-en` language pair is closer than the other two, and therefore more likely to be similar syntactically in the output, which could improve performance of the GRID model: If the MT output remained similar syntactically to the source language, then GRID would not perform as well for other language pairs (it is known that the syntactic assumptions which hold for English do not do so for German). Although this potentially affects GRAPH too, it does not depend on entity transitions but models connections among all sentences in a document. Moreover, a closer inspection of the data showed that the quality of the `fr-en` reference translation was not as good as the `de-en` reference translation. Coupled with better MT output for the `fr-en` language pair, this would make it a more difficult task for the models to differentiate between HT and MT.

While the GRID model does well in the shuffling experiment, it does not do so well with the MT output, coming near the bottom. Clearly shuffling and reordering are entirely different tasks, as illustrated by the differences in the scores between Table 3.2 and Table 3.4. By comparison, the ability of GRAPH (as the other entity-based method) to distinguish between HT and MT output is presumably due to it being more robust in terms of tracking entities and the fact that it does not rely on the syntactic transitions (between sentences), unlike GRID. Moreover, while the transitions modelled in GRID are in fact over sentences, the original theory intended this to be over utterances (Poesio et al., 2004). Overall, despite our extensions to create the IBM1 model, which lead to a direct improvement, the syntax model seems inadequate for properly measuring the intentional structure.

3.4 Conclusions

Work on measuring text coherence has thus far been commonly limited to somewhat artificial scenarios such as sentence shuffling or insertion tasks. These operations naturally tend to break the overall logic of the text. In this Chapter we have investigated local coherence models for a very different scenario, on texts which are automatically translated from a given language by systems of various overall levels of quality, to see whether we can discriminate between HT and MT texts. We have shown that this is a different, and more difficult task. Coherence in this scenario is much more nuanced, as elements of coherence are often present in the translations to some degree, and their absence may be connected to various types of translation errors at different linguistic levels. We also proposed a new model which explores syntax following a more principled method to learn the syntactic patterns. This extension outperforms existing ones in the monolingual shuffling task on news data, and performs more credibly than the original one in our new task. The question arises, whether these syntax models are adequate proxies for measuring aspects of coherence pertaining to intentional structure, even when adapted. Or whether they are limited to scenarios such as detecting or reordering shuffled text. By way of a supplementary test to determine if our models are measuring coherence, and not simply the differences between the MT and HT, we intend to test them on an artificial corpus containing injected coherence errors (Chapter 5).

In the next chapter (Chapter 4), we take a crosslingual approach, investigating whether we can measure the extent to which discourse relations in the Source Text (ST) are transferred to the TT. This is part of our investigation into whether these models (discourse relations, syntax, entity) can serve as proxies for the *linguistic structure*, the *intentional structure*, and the *attentional state* of a text.

Chapter 4

Crosslingual Discourse Relations in Machine Translation

In this Chapter we turn our attention to discourse relations which along with the entity and syntax models of the previous chapter to cover the three tenets of the Computational Theory of Discourse Structure (Grosz and Sidner, 1986) described in Section 1.4.1; the *linguistic structure*, the *intentional structure*, and the *attentional state*. As such, they should form a good basis for assessing coherence in MT computationally. In this experiment we adopt a crosslingual approach, researching whether we can evaluate the transfer of the semantics of discourse relations from ST to TT. Also known as *coherence relations* these are vital for the coherence of a text, capturing the inner logic of a text, often via the connectives that act as a signal to the reader (Stede, 2011).

We propose a novel approach that assesses the translated output based on the *source* text rather than the reference translation and measures the extent to which coherence elements (discourse relations, specifically) in the source are preserved in the MT output.

4.1 Crosslingual Discourse Relations

Despite the fact that discourse relations have long been recognised as crucial to the proper understanding of a text (Grimes, 1975; Longacre, 1996) current MT

systems often fail to properly handle discourse relations for various reasons, such as incorrect word alignments, the presence of multiword expressions as discourse markers, and the prevalence of ambiguous or implicit discourse markers. Most MT systems, certainly SMT ones, do not take account of discourse relations explicitly, and discourse connectives are simply treated as any other words to be translated.

As discussed in Section 2.1.5, previous research on assessing discourse relations in MT has covered work incorporating discourse structure in the evaluation of MT output by comparing the discourse tree structure of the MT to that of the gold standard (reference translation) (Guzmán et al., 2014), or on a (also reference-based) discourse-connective specific metric (Hajlaoui and Popescu-Belis, 2013). The latter is closest to our work, but focuses on a narrow selection of ambiguous connectives, and uses a reference translation. Taking seven ambiguous connectives, it scores the MT output based on whether it has the same or equivalent connectives to the ones used in the reference. It uses a static list of equivalents for each of these (seven) connectives, and judges whether the sense of the target connective is compatible with that of the source based on how it has been translated in the reference.

While using a reference for evaluating a candidate translation is the norm, it is an inflexible and potentially restrictive way of evaluating translations. There can be a number of ways a text can be correctly translated, but usually only one reference is available. Moreover the prerequisite of a gold standard reference for evaluation is limiting, as evaluations can only be performed on small, pre-defined test sets. Aside from the benefits of automatic evaluation, without an automatic metric there is no way to optimise parameters in an SMT system which does include discourse features.

In this Chapter we attempt to establish how a particular discourse relation in the source text can be rendered in the target text, and then to evaluate how well this is translated in the MT output. This is not in comparison with a gold standard, but based on whether the MT output carries the intended meaning of the source text. In that respect, our task is more challenging: instead of performing string matching to detect the presence of certain connectives in the MT output and reference translation, we need to detect the discourse relations in

the *source* text and determine whether these relations are correctly transferred crosslingually to the target language – *without* a reference translation

We assume that the semantics of a discourse relation should transfer from source to target language, as this has been broadly established (da Cunha and Iruskieta, 2010; Laali and Kosseim, 2014). In other words, while the actual segmentation into discourse units may vary from language to language (Mitkov, 1993), the meaning of the actual discourse relation is constant across languages. Therefore, on the basis that discourse relations are semantically similar across languages, we assess how well explicit discourse relations are transferred from the source to the target language. These often take the form of lexical cues, or discourse connectives, which signal the existence of a particular discourse relation. This is the case particularly for the lexically-based PDTB (Rashmi Prasad and Nikhil Dinesh and Alan Lee and Eleni Miltsakaki and Livio Robaldo and Aravind Joshi and Bonnie Webber, 2008), and to a lesser extent the hierarchical RST (Knott and Dale, 1994), as illustrated by work on lexical cues in RST (Khazaei et al., 2015). Compiling static lists of equivalent discourse markers in two languages is a cumbersome approach, given the variety of discourse markers in both languages, the range of relationship each can encode, and the fact that there are many alternative lexicalisations for discourse connectives (Prasad et al., 2010). Moreover, as established in previous research crosslingual discourse relations (Meyer and Popescu-Belis, 2012; Meyer et al., 2011; Li et al., 2014b), there can be mismatches due to ambiguous discourse markers. We therefore train crosslingual discourse embeddings which we hope will capture equivalences in discourse connectives across languages in a more flexible manner due to the context which they encapsulate.

Essentially a metric, this experiment incorporates a likelihood score from the crosslingual embeddings, in addition to a weighted score for the correctness of the particular discourse relation. We evaluate it by assessing how well different outputs render the discourse relation as captured from the source text. We first compare the scores of MT output over that of a post-edited version of the same, and find our metric scores the post-edited (PE) greater or equal to the MT for 78% of the documents. We also evaluate system submissions from the WMT14 campaign, finding that comparing our rankings to the human rankings results in

some notable exceptions regarding rules-based systems. Finally, we experiment with integrating our features into a Quality Evaluation framework (Bojar et al., 2016), resulting in decreased prediction error over a strong baseline. Our metric is novel in that it evaluates the MT directly against the source text. It uses crosslingual embeddings trained to handle multiword discourse connectives, and incorporates discourse relation mappings between source and target texts.

In Section 4.4 we describe our initial alternative attempts to measure the transfer of crosslingual discourse relations, before describing the actual methodology used; in Section 4.2 we describe the pipeline to build this metric, including how we created the crosslingual embeddings to track specifically the discourse connectives, the methodology employed to evaluate the target relations based on the source text and our overall scoring metric. In Section 4.3 we present the results from our experiments.

4.2 Methodology

As can be seen from the diagram in Figure 4.1, our discourse score is composed of two components: Discourse Relation (DR) and Discourse Connective (DC). For measuring the correctness of the Discourse Connective, we take the score given from pretrained word embeddings for translation of the cue. In Discourse Relation, the semantics of the discourse relations themselves are estimated by comparing (a) the discourse relation of the source text, as assessed from usages defined in LexConn (Roze et al., 2010), with (b) the relation of the target text, as assessed via a discourse tagger for English (Pitler and Nenkova, 2009). We establish whether the source text discourse relation (identified via syntactic lexical cues combined with LexConn) corresponds to the discourse relation present in the MT output (based on the English discourse tagger). These two components together give an indication of how well the explicit discourse relation is transferred from source to target text. We used the Stanford CoreNLP toolkit (Manning et al., 2014) to parse both the French (in order to ascertain the discourse usage of a potential cue) and the English MT output (to provide as input to the English discourse tagger of Pitler and Nenkova (2009)). We detail these two components in Sections 4.2.2 and 4.2.3.

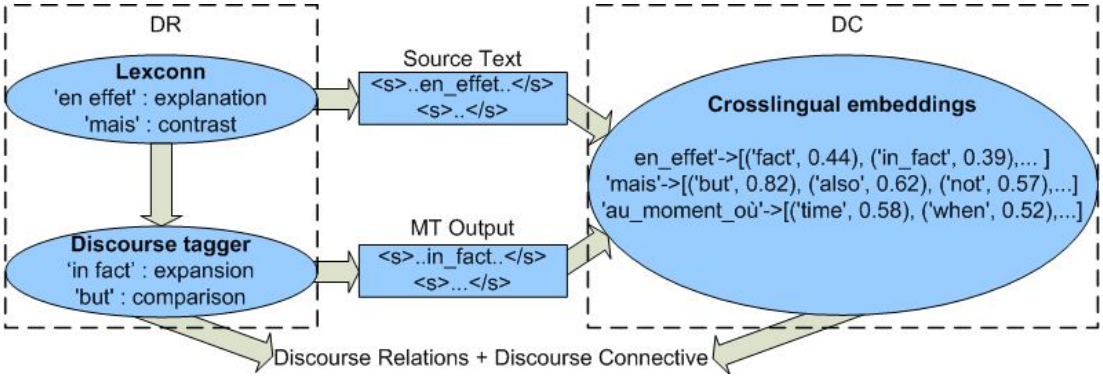

Figure 4.1: Dis-Score incorporates a discourse relation component and a discourse connective component. The DR component (left) uses the LexConn definitions combined with syntax rules to determine the French connective, comparing to the relation derived by the tagger for the English. The DC component (right) incorporates the probability a particular French connective is translated as the one found in the MT output.

4.2.1 Datasets

We used French-English as our language pair, as the quality of the MT output needs to be of a certain level before attempting to discern transfer of discourse relations. In addition, we needed a dataset with post-edited translations for evaluation purposes. As we discuss next, two types of data are needed: a large amount of human-translated parallel data to build word embeddings models, and a lesser amount of MT output parallel data for testing of the models.

Training Data For training our discourse-specific bilingual embeddings we require a parallel corpus. In order to ensure that the correct use of discourse connectives is captured in the training of our embeddings, we use a filtered version of the Europarl corpus (Koehn, 2005), as provided by IDIAP[1] (Cartoni et al., 2011). This consists of a filtered source text to ensure that only the excerpts which were originally written in French (in our case) are taken as source. The parallel text is then formed by the original French sentences and their English reference translation. Given that many discourse markers may be composed of several words (see Section 4.2.2 for a full explanation), we hyphenate the discourse cues

[1]https://www.idiap.ch/dataset/europarl-direct

in the training data – as per previous work on training phrases (Mikolov et al., 2013). We also train the embeddings with a non-hyphenated version, and with the full Europarl dataset (Koehn, 2005). This results in three embedding models (hyphenated, non-hyphenated, and full Europarl) (Section 4.2.2).

LIG Test Data As our main test data we again used the LIG corpus (Potet et al., 2012) of French-English translations. This dataset was chosen since it includes a PE version, suitable for our task. In all it comprises 361 parallel documents, a total of 10,755 tuples: <FR, MT, PE, HT>, including the source text (FR), the machine translated output (MT), the post-edited output (PE) and the reference translation (HT), drawn from various WMT editions. The translations were produced by a phrase-based SMT system. The instructions to those performing the post-edition were to make the minimum amount of corrections necessary for a publishable translation (Potet et al., 2012).

We show results using the MT-PE documents, taking the PE version for comparison instead of the reference HT. Using the PE for comparison will avoid mismatches that are due to variances (e.g. style) in freely created reference translations. HT will of course include much greater variation, which our metric cannot easily capture, as will be explained later. Our hypothesis is that our metric should score the PE more highly than the MT, with many instances where both score equally, due to the fact that the MT will correctly render the discourse connectives in those texts, while other sentences in the source will not have explicit connectives at all.

WMT Test Data For comparison, we also score all the French-English submissions from the 2014 WMT shared translation task (Bojar et al., 2014) with our model.[1] We then show the correlations between the ranking of MT systems that participated in the shared task and our metric, in comparison with the best scoring reference-based metrics from the 2014 WMT metrics task (Macháček and Bojar, 2014). The top metric in this shared task (DiscoTK (Guzmán et al., 2014)) also includes a discourse relation component. We take the French-English translations, comprising 175 documents, a total of 3,003 sentences per system

[1]Most recent editions of WMT do not include French.

submission, with eight system submissions in total.

4.2.2 Discourse Connectives

Training In French, many discourse cues are composed of several words (Roze et al., 2010), such as *au moment où*. In fact, Laali and Kosseim (2014) found that in French larger order ngram connectives are more prevalent than in English and that, for example, LexConn contains 69 4-gram connectives. To a lesser extent this is true also in English (e.g. *of course*, *as well as*, *in addition*). In order to have a phrase representing the full cue returned in our embeddings, we hyphenate the discourse cues in the training data. To capture these discourse cues, we train bilingual word embeddings as per Luong et al. (2015), using the MULTIVEC toolkit (Bérard et al., 2016) with the following modifications. We used LexConn (Roze et al., 2010), a French lexicon of 328 discourse connectives, to identify the French discourse connectives in our training corpus. We then hyphenated all the identified discourse connectives before training bilingual embeddings. We did the same for the English corpus, hyphenating all the connectives which appeared in the list of 226 compiled by Knott and Dale (1994). This ensures that for multi-word connectives, our embeddings return the full discourse connective. For example, for the French discourse connective *parce_que*, the model correctly returns *because*, among others. We also train embeddings with a non-hyphenated version of the corpus, and with the full Europarl dataset (Koehn, 2005). We then back off to each of the above in turn (hyphenated, non-hyphenated, full Europarl), where the first model has no equivalent for our searched connective. The parallel data we use for our hyphenated embeddings consists of 214,972 sentence pairs, while the embeddings we trained with the full Europarl consists of 154,915,709 sentence pairs.

Testing Once we have identified the existence of a French discourse connective in the source text, we then use syntactic cues based on findings by Laali and Kosseim (2014) to verify if the cue is being used in a discourse context. These involve identifying whether the potential connective in the source text has a syntactic tag of correct category (e.g. ADV, C, MWADV, MWC, CS etc.), for

example that the word *alors* is being used in a discourse sense, as part of *alors que*, not simply as a comment word. The same happens in English: the word *and* can be used in a discourse sense (to join two clauses), or a non-discourse sense (as part of a listing). This helps us to determine if the lexical items from LexConn identified in the French source text are being used as a discourse connective. Laali and Kosseim (2014) determined that syntax could be used to filter out constituents that were not discourse connectives. This has previously been done successfully in English (Pitler and Nenkova, 2009).

4.2.3 Discourse Relations

Discourse relation theories include the hierarchical RST (Mann and Thompson, 1988) and lexically-based PDTB (Rashmi Prasad and Nikhil Dinesh and Alan Lee and Eleni Miltsakaki and Livio Robaldo and Aravind Joshi and Bonnie Webber, 2008). RST is a theory whereby the text is decomposed into a tree structure, recursively creating subtrees of subsegments, from the most basic EDU (Elementary Discourse Unit) upwards. So each sentence is composed of several EDUs, which are connected by a discourse relation. In PDTB the intention is to identify a discourse relation between two arguments, without presupposing any hierarchical structure. Discourse relations can be implicit or explicit. If explicit, they are generally signalled by discourse connectives. Implicit relations can be derived from the context, but have no given markers. We restrict this experiment to sentence-level relations, as did Guzmán et al. (2014), on the basis that MT systems and MT evaluation are at sentence-level.

In the absence of a French discourse parser to identify the occurrence of a connective and give an indication of the discourse relation being used in the source (French) text we use LexConn. As mentioned already, it includes total list of 328 connectives. Any connectives occurring in our training data will be represented in the crosslingual discourse embeddings we create. In addition to determining the discourse versus non-discourse usage of a particular word or phrase as described above, we also use syntax to disambiguate connectives where there are multiple senses for the given connective. For example, in English the word *since* can have either a causal or a temporal sense. This applies in French too. To do this,

we identify all the connectives in LexConn which occur more than once with different discourse relations assigned to the various instances of the connective. This resulted in an initial list of 80, out of a total list of 328 connectives usages, for which we checked for occurrences in the ANNODIS corpus (Afantenos et al., 2012) which would let us ascertain their correct discourse relation in a given context. ANNODIS is a French resource, consisting of a collection of documents from Wikipedia, French news articles and reports with discourse annotations (Afantenos et al., 2012). Those potentially ambiguous connectives which did not occur we discarded, as we had no context to evaluate, moreover they were more likely to be infrequent. We then filtered further, discarding connectives where there was no instance of both discourse relations in ANNODIS, or else where the two ambiguous relations came under the same higher level categorisation (such as for *mais*, which can be indicative of the relations *contrast* or *violation*, and when mapped under our four broad mappings (see below) results in same category, *comparison*). Our final list of ambiguous connectives (requiring disambiguation for our purposes) included: *après*, *aussi*, *alors que*, *depuis que*, *en*, *tandis que*, *même*, *si*, *tout d'abord*.

We then devised disambiguation rules based on heuristics or syntax for these remaining ambiguous connectives. For example, for the connective *aussi*; if it was in sentence initial position and therefore upper case, could be regarded as pertaining to *result*, otherwise to discourse relation *parallel*. This particular rule was directly based on information from LexConn. For most of the many connectives captured in our embeddings there is no ambiguity, and we take the discourse relation as defined by LexConn. These rules are simply to cover the few ambiguous ones, to ensure that we compute our score against the correct discourse relation.

For assessing the discourse relation on the target side (English), we used the discourse tagger developed by Pitler and Nenkova (2009), which identifies the top level of PDTB discourse relations, namely Temporal, Comparison, Contingency and Expansion. The relations of LexConn (30 in total) can be manually mapped roughly to the second level of the PDTB, so we manually establish the corresponding mappings, assigning relevant ones to the four PDTB ones which the tagger identifies. Ideally we would like to have a more fine-grained approach, but in the absence of a discourse parser for the French side, this was the most

robust approach we could devise. Given the variation of discourse relations in LexConn, a more detailed mapping would be more difficult without a great deal of additional analysis. Moreover, discourse parsers still fail to reach high levels of accuracy for relationship identification, and so we opted for a more flexible approach in this initial experiment. We make the assumption that if the tagger cannot identify a relation in the MT, then it is probably not properly rendered, and so will not be scored.

4.2.4 Dis-Score Metric

Our metric, Dis-Score, is composed of the probability given to a potential discourse connective in English for any particular French connective from the specifically pretrained bilingual embeddings (DC), combined with a score reflecting the correctness of the discourse relation match (DR), weighted by γ. We calculate the value for γ by doing grid search cross validation on the LIG corpus using the scikit-learn toolkit (Pedregosa et al., 2011), which results in γ being set to a value of 0.045. DR is therefore weighted more heavily than DC, even if this seems like an extremely low weight, in fact DC can be quite small, so it actually represents an upweighting. This is summed over each sentence of the document, and normalised by the number of sentences in the document. Formally, the score is as follows, where Dis-Score(D) is our overall score for document D, N is the number of sentences in the document, and M is the number of discourse connectives in a sentence. ST is the source text, and so the scoring function tracks the number of discourse connectives and relations in the French text.

$$\text{Dis-Score(D)} = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M}\frac{DC}{DC_{ST}} * \gamma\frac{DR}{DR_{ST}} \tag{4.1}$$

4.3 Results and Discussion

We evaluate our Dis-Score metric in several ways. Firstly, on the LIG dataset, we compute scores for MT and PE and hypothesise that the PE should score higher than the MT, although there will be many ties, where the MT correctly renders the discourse relation, or where there is no explicit discourse relation in

Output	Number of Wins	Percentage of Wins
PE	113/361	31%
MT	80/361	22%

Table 4.1: Number and proportion of times PE wins over the MT version in the LIG corpus according to the Dis-Score metric at the document-level. In 168 out of 361 cases, the metric score was the same for MT and PE.

the source text. There may be instances where the PE renders the discourse relation in an implicit manner, while this is unlikely to be the case for the MT, and which therefore results in a situation where the MT may sometimes score higher. Secondly, we evaluate the outputs from the WMT14 system submissions, and establish how our scores compare with the official system and segment rankings. Finally, we integrate our scores as features in a machine translation Quality Estimation model (Biçici and Specia, 2015), and show that the addition of our features reduces prediction error.

4.3.1 LIG Test Set

In Table 4.1 we report our results on the LIG corpus comparing the metric scores for the PE and the MT. Under our metric, the scores for the PE are greater or equal to the MT for 281 of the 361 documents (78%) in the LIG corpus. Half of the documents are tied according to the metric, where both documents score equally. This is to be expected, as the MT successfully renders a significant amount of the connectives. Moreover, there may be sentences on the source side where there are no discourse connectives present. For the LIG corpus as a whole, there are 2998 sentences where one or more explicit discourse relation is detected, out of a total 10756 sentences. On closer analysis, for the documents where the MT scored more highly than the PE, sometimes the tagger failed to identify the connective in the PE, despite recognising the same connective in the MT.

Interestingly, the MT scores better than the HT under our model. We found that there are numerous instances where the relation is rendered in a more subtle manner in the HT, and can be inferred from implicit discourse relations. For example, we found: *mais cela n'aura servi à rien* translated as *to no avail* in the HT, which our model did not score. Here the MT scored for having an explicit

connective *but*, which is the most probable connective for the French *mais*. This supports the findings by Meyer and Webber (2013), that up to 18% of discourse connectives are not rendered explicitly in human translations. As such, they are missed by our current configuration, which does not take account of implicit relations.

However, we consider our metric effective for measuring the extent to which the MT output captures explicit discourse relations from the source text, and renders them in the target text. Due to the nature of MT, it closely follows the source text. This automatically increases the embeddings score, where MT selects most probable equivalent, for cases where it has correctly translated it. Whereas even when the HT does use an explicit discourse relation, such is the natural variability in human translation that the connective may not have been the most probable translation. For example, the French *mais* would have highest probability score for English connective *but*, while *however* or *nevertheless* are equally good choices. This is also the reason why we chose to compare the MT with the PE. Finally, using LexConn to identify the discourse connectives in the source text is occasionally problematic, as these cues are open class words and, as already noted (Laali and Kosseim, 2014), some are not captured in LexConn.

Hoek et al. (2015) found that both implicitation and explicitation of discourse relations occurs frequently in human translations, i.e. explicit relations made implicit and vice versa. In their role as mediator, human translators often make explicit a discourse relation that is implicit in the source (Hatim and Mason, 1990). In addition, more research needs to be done to ascertain how often discourse connectives in French are not directly rendered in English. Ideally, a well-tuned model should take account of the amount of implicitation and explicitation typical of the language pair in question. It would be good to move the focus in our metric even further from discourse connectives to discourse relations, although this requires a discourse parser on both sides. For the evaluation of MT output in its current shape a one-to-one relationship is the best we can aim for.

As an additional way of evaluating our discourse metric, we use our scores (final and two components) as features in a **Quality Estimation (QE)** framework. QE is a more general field of reference-less machine translation evaluation (Blatz et al., 2004; Specia et al., 2009) which is based on training a model to pre-

dict a quality label for a text from human-labelled instances described through various features. As our initial model, we reproduce the strong official sentence-level "baseline" model from the WMT shared tasks (see Bojar et al. (2016) for the latest edition). It uses 17 well performing features and Support Vector Regression for training. The features are more superficial but complementary to our discourse scores, including, for example, language model probabilities for the translations and counts of words in source and translations. As quality labels, in our experiments we use the HTER scores (Snover et al., 2010), i.e., the edit distance between the MT output and its PE version. This score ranges from 0 – MT and PE are identical – to 1 – all words in MT are different from PE). This is the only annotation we have for the LIG dataset. The additional, sentence-level, features from our metric are (see Section 4.2.4): DC, DR, $Dis - Score$ $(\frac{DC}{DC_{ST}} * \gamma \frac{DR}{DR_{ST}})$.

We randomly split the dataset into 70% for training and the remaining for test. This resulted in 7,528 sentence instances to train the QE model and 3,227 sentence instances to test it. In order to extract features and build prediction models, we used the freely available QuEst++ tool (Specia et al., 2015). The performance of the system is measured in terms of Pearson correlation between predicted and true HTER scores. The correlation for the original (baseline) model was 0.117, while the correlation for the model including the discourse scores is 0.145.

4.3.2 WMT Test Set

For further evaluation, we follow Hajlaoui and Popescu-Belis (2013) and also show how the system submissions from 2014 WMT shared translation task score under our metric. It should be noted however that we measure discourse relations in isolation, focussing on explicit connectives, whereas MT output has other problems which will affect the WMT rankings. These more general problems are the target of most reference-based, n-gram matching metrics, such as BLEU (Papineni et al., 2002). Therefore, directly comparing our results to standard metrics would not lead to a fair analysis. On the other hand, by taking the overall ranking of the MT systems (generated from human evaluation) and checking how our

system	Dis-Score	Human ranking	DiscoTK-party	DiscoTK-light
UEDIN	0.437 (3)	1	0.829	1
STANFORD	0.414 (7)	2	0.768	0.957
KIT	0.414 (7)	2	0.756	0.939
ONLINE-B	0.417 (6)	2	0.738	0.855
ONLINE-A	0.448 (2)	3	0.651	0.814
RBMT1	0.430 (4)	4	0.200	0.227
RBMT4	0.459 (1)	5	0.013	0.047
ONLINE-C	0.421 (5)	6	-0.063	0.004

Table 4.2: Human ranking of 2014 WMT MT system submissions compared to Dis-Score and top WMT14 metric rankings.

metric and other metrics would rank the same systems, we can gain insights on where metrics fail or are complementary. Therefore, we also show two versions of DiscoTK, the top scoring metric in WMT14.

Both DiscoTK-light and DiscoTK-party include discourse structure, which is not covered by other metrics. The latter was also the best performing at the WMT14 metrics task (Bojar et al., 2014). DiscoTK-light combines variations of discourse structure from comparing RST discourse trees of MT and HT using a convolution tree kernel, while the DiscoTK-party metric combines the latter with other metrics operating at different levels (lexical, etc.) (Guzmán et al., 2014). As previously mentioned, DiscoTK parses the MT output into RST discourse trees and compares it to the HT tree, whereas we compare the MT with the ST, and check whether the cues and semantics for a particular discourse relation are comparable (instead of comparing the structure).

As can be seen from the results in Table 4.2, our metric would lead to a different ranking, where a rule-based system (RBMT4) would rank the highest. This is perhaps not surprising, given that rule-based MT systems tend to model linguistic structures (including discourse) more explicitly through rules. DiscoTK approximates the human ranking very well, but it should be noted that DiscoTK-party is combined with a number of other metrics, and uses machine learning models trained on human rankings.

We display the system-level correlations with human judgements, as per the Pearson correlation for WMT14, and display results for our system in Table 4.3, as well as some of the others for comparison. Given that our metric only looks at

Metric	fr-en
Dis-Score	-0.213*
DiscoTK-light	0.965*
DiscoTK-party	0.970*
DiscoTK-light+DisScore	0.969*
DiscoTK-party+DisScore	0.975*

Table 4.3: Results on WMT14 at system level: Pearson correlation with human judgements. Our Dis-Score alone, and in linear combination with various DiscoTK 2014 WMT submissions.

Metric	Average	wmt12	wmt13	xties
Dis-Score	0.012*	-0.941*	0.263*	0.250*

Table 4.4: Results on WMT14 Fr-En at segment level: different variations of Kendall's τ rank correlation with human judgements.

one isolated discourse component, we do not expect the correlation to be high on its own. This was the case for Guzmán et al. (2014) with DiscoTK-light, although to a lesser extent. We combine our scores linearly with DiscoTK variants as per Guzmán et al. (2014), discovering that when combined with DiscoTK-light, the resultant score is close to that of DiscoTK-party. In addition, when combined with DiscoTK-party, it ranks second overall for that language pair.

We also display the segment-level correlations with human judgements in Table 4.4. There are many segments with score of 0 since they have no detected connectives, and the per-segment variation of Dis-Score is high, since some sentences have several discourse relations, while others have none. Humans judgements consider other aspects and quality as a whole, whereas our metric just measures transfer of explicit discourse connectives. As such the correlation is low, at 0.012, but is positive. The highest segment-level correlation was 0.433. Under the Kendall's τ variant used for wmt13, which handles ties differently (Bojar et al., 2014), the correlation is higher at 0.263. The number of non-zero segments varies from system to system, but ranges from 504 to 566 segments out of a total of 3003.

Comparing the rankings (Table 4.2), the main difference is that the rule-based systems do better under our model, and move from the lower half of the table to

MT system	Translation
source	Une “petite avarie”, circonscrite à l’espace de confinement du réacteur, s’était *alors* produite sur le navire amiral de la flotte française.
RBMT1	“Small damages, circumscribed with the space of containment of the engine, had *then* occurred on the flagship of the French fleet.
STANFORD	A “small damage,” confined to the area of the reactor’s containment, had occurred on the flagship of the French fleet.
UEDIN	A“small damage,” confined to the reactor’s containment area, was produced in the flagship of the French fleet.
ref	A “small amount of damage”, confined to the area of the reactor chamber, *then* occurred on the French fleet’s flagship.

Table 4.6: Examples of translations from different MT systems, where RBMT1 correctly preserves *then*.

the top. To give an intuition of why this is the case, we include some examples in Table **??**. For this sentence, the RBMT1 system scored for having *so that* for a translation of the French *pour que*, which was recognised as discourse relation *goal* which correctly mapped to Contingency. The other two systems (which are among the highest ranking) displayed in the table do not have this discourse cue, which is important for the understanding of the text. As can be seen from the reference, which has been included for illustration purposes, the discourse relation is implicit in the human translation, and rendered by *to enable*.

Looking at another example in Table 4.6, the RBMT1 system scored better under our metric than other higher ranked systems (at WMT14) because it included the discourse cue *then* as a translation of the French *alors* for a temporal or causal discourse relation. The other systems displayed in Table 4.6 have lost the cue.

4.4 RST and lexical cues

We initially experimented with – and abandoned– an alternative strategy to capture discourse relations, using lexical cues from an annotated corpus, inspired by Khazaei et al. (2015). In our case the cues were culled from ANNODIS (Afantenos et al., 2012) for the French, and from the RST corpus for the English (Mann and Thompson, 1988). Unlike Khazaei et al. (2015), we then annotated the established cues with RST relations and trained embeddings on the annotated data, as has been done for POS tags previously (Levy and Goldberg, 2014; Paetzold and Specia, 2016). Due to the lack of parallel data annotated with discourse relations, we tried training our embeddings with BILBOWA (Gouws et al., 2015), which is trained on monolingual data and extracts a bilingual signal from a smaller amount of parallel data. However, like Upadhyay et al. (2016), we found that the performance was very poor (even on the unannotated results). We subsequently trained bilingual embeddings with parallel data using Luong et al. (2015)'s method via the MULTIVEC toolkit (Bérard et al., 2016), which gave good results. It turned out, though, that for the annotated data the cues were clearly too sparse. ANNODIS (Afantenos et al., 2012) in particular is a small dataset, and while we had hoped we could train bilingual embeddings without a large amount of parallel data, it is insufficient for training embeddings in this manner. As a result, this experiment was not successful, resulting in OOVs or poor matches, due to the sparsity involved; the lexical cues did not occur frequently enough in the context of particular relations for training the bilingual embeddings with specific annotations.

After some research into options for French, which was our Source Language (SL) in this experiment, we had determined that there was no suitable French discourse parser available. However, there were several for English. Our idea was then to project the relation and establish whether it correlated with that of the SL (where the discourse parser on English MT output gives an indication of the discourse relation, and we use this relation as potential guide for projection). This would then be incorporated in a scoring function, reflecting whether an equivalence existed between the discourse relation identified by the English discourse parser, and projected to French- and verified by the apparent French

discourse relation, derived according to lexical cues and usages in LexConn (Roze et al., 2010). The main task would be to identify an appropriate discourse connective, i.e. reflecting the correct relation. Whether or not we can actually label the particular discourse relation is less important, given that discourse parsers are not very accurate on that part of the task.

Instead of using an RST parser and RST-based cues for our embeddings, however, we then moved onto investigate the lexically-based PDTB, using a tagger based on PDTB which identifies the top level of PDTB relations, as explained in detail in the this chapter. A simpler but more robust approach which still has crosslingual embeddings at its core, and which we hope will capture the semantics of the discourse relations in a more flexible yet accurate manner.

4.5 Conclusions

We have shown how the crosslingual transfer of discourse relations can be measured in MT, in terms of handling connectives as cues signalling the discourse relations, as well as assessing the subsequent semantic transfer of the discourse relation. Our model is measured against the source text and does not use a gold reference or alignments. As such it represents a way we can measure transfer of one element which contributes to the coherence of a text. For more flexible and realistic evaluation of a translation, we need to move away from the current approach towards assessing the translated output conditioned on the source text. This will need to be a multifaceted semantic approach, of which assessing the transfer of discourse relations from source to target is but one element which requires evaluation.

Our work introduces a way in which this can be done, successfully scoring the PE greater or equal to the MT 78% of the time. We believe our work is novel, in that we do this using crosslingual word embeddings pretrained for multiword discourse connectives and incorporate discourse relation mappings between source and target text. While we recognise that this only covers explicit, not implicit relations, and rewards translations that are closer to the source than some better human translations, we believe it is suited to evaluating MT and is a novel and constructive effort to address an evaluation gap. By necessity, it is dependent

on a parser and tagger, which sometimes do not correctly assess the constituents or discourse relations. Ultimately using a discourse parser on the French source text would lead to greater accuracy, particularly if we could then map the discourse relations in more detail. Ideally we could then also move beyond tracking the transfer of intrasentential relations to track the transfer of intersentential relations, at document level.

In the next Chapter (5), we examine the type of coherence errors which actually occur in MT output, and use these as a basis for creating a corpus engineered to contain genuine coherence errors, as an alternative evaluation strategy.

Chapter 5

A Corpus to Measure Coherence in Machine Translation

As illustrated already (Chapter 2), the issue of coherence in MT has received little attention to date, and an initial major issue we face in this area is the lack of labelled data. While coherent (human authored) texts are abundant, and incoherent texts could be taken from MT output, the latter generally also contains other errors which may not be specifically related to coherence. This makes it difficult to identify and quantify issues of coherence in those texts. In this chapter we introduce our initiative to create a corpus consisting of data artificially manipulated to contain errors of coherence common in MT output. We detail the systematic way we have analysed MT errors and extracted particular coherence-related errors that are specific to MT output, and have subsequently injected them into an otherwise coherent and grammatically correct text. Our goal is to create a corpus which consists of data artificially manipulated to contain errors of coherence common in MT output. Such a corpus could potentially be used as training data for coherence models in supervised settings.

We will then use it as one method for evaluating our coherence models. We explain the motivation behind this initiative, considering similar previous work in Section 5.1. We then examine the issues of incoherence in MT systems, illustrating with real errors found via manual analysis (Section 5.2), and our proposed methodology for this experiment (Section 5.3). In Section 5.4 we detail

the pipeline for extracting genuine errors of specific types (detailed in Sections 5.5, 5.6 and 5.7) which have occurred in MT output, and injecting them into our corpus. We subsequently test our models on the newly created artificial corpus and report on the results (Section 5.8), then we identify the limitations of the approach (Section 5.9).

5.1 Motivation

Previous computational models for assessing coherence in a monolingual context (see Section 2.2) have generally used automatically summarised texts, or texts with sentences artificially shuffled as their 'incoherent' data for purposes of evaluation[1]. These are instances of artificially created labelled data, where the logical order of the text has been distorted, affecting particular aspects of coherence. For our task, however, it is inadequate as MT preserves the sentence ordering, but suffers from other aspects of incoherence. Moreover, while the MT output can potentially be considered 'incoherent', it contains a multitude of problems, which are not all due to lack of coherence. Our aim is to create a corpus which exhibits errors related to coherence, but does not have the grammatical or stylistic errors which would otherwise be present in regular MT output. This will mean that we can assess our coherence models by isolating other issues that are unrelated to coherence, thus ensuring that they do not simply differentiate between MT output and HT output but are specifically targeting coherence issues.

As far as we are aware, no attempts have been made to create a corpus exhibiting incoherence, other than by shuffling ordered sentences. There has been work in other areas to introduce errors in correct texts. For example, Felice and Yuan (2014) and Brockett et al. (2006) inject grammatical errors common to non-native speakers of English in good quality texts. The work by Sennrich (2017) automatically created test sets of contrastive pairs (MT and reference) specifically for evaluating the quality of Neural Machine Translation (NMT) output, by introducing syntactic and semantic errors via rules. Felice and Yuan (2014) use existing corrected corpora to derive the error distribution, while Brockett et al. (2006) adopt a deterministic approach based on hand-crafted rules. Logacheva

[1] http://people.csail.mit.edu/regina/coherence/CLsubmission/

and Specia (2015) inject various types of errors in human translations to generate negative data for MT quality estimation purposes, but these are at the word level and the process was guided by post-editing data. They derived an error distribution by inspecting post-edited data. We can try inducing a distribution of errors for coherence in a similar way, but will need a large amount of post-editings of entire documents. We also have the added difficulty of trying to isolate which of the edits relate to coherence errors.

5.2 Issues of incoherence in MT systems

Current MT approaches suffer from a lack of linguistic information at various stages (modelling, decoding, pruning (as described in Section 1.3)) which results in a lack of coherence in the output. Below we describe and illustrate a number of issues that impact coherence and which are not handled well in MT. They have been identified in our own work (during work for Chapter 6) and by others (Section 2.1). While there are certainly other possible issues such as word order and WSD, the broad classification used in our analysis helps in the task of understanding the coherence issues we need to target within the context of MT.

Datasets The examples of incoherence given in this Section have been identified in our own error analysis done in either of the following corpora. The second corpus is additionally used in our subsequent experiment (from Section 5.3), and so is described in more detail below):

- the **newstest** data (source and output) from the WMT corpora[1], of which we select examples focusing on French and German as source, and English as output. The output consists of submissions made by those MT systems participating in the Workshop on Machine Translation.

- the **LIG** corpus (Potet et al., 2012) of French-English translations:

  In all it comprises 361 parallel documents, a total of 10,755[2] quadruples: <FR, MT, PE, HT>, comprising the source text (FR), the machine trans-

[1] http://www.statmt.org/wmt12/ and http://www.statmt.org/wmt13/
[2] After removing some null and duplicate lines from the original 10,881.

version	text
ST	*'Cette anne, c'était **au tour de** l'Afrique de nommer le président et elle a nommé la Libye.'*
MT	*'This year, it was **at the tour of** Africa to appoint the president and has appointed Libya.'*
REF	*'This year it was Africa**'s turn** to nominate the chairman, and they nominated Libya.'*

Table 5.1: Example of lexical cohesion error: *au tour de* is mistranslated in the MT, instead of *turn*, it has *tour* (Potet et al., 2012).

lated output (MT), the post-edited output (PE) and the reference translation (HT), drawn from various WMT editions. This dataset was also chosen as our corpus for injecting errors into (see Section 5.3) since it includes a PE version, suitable for our task. The translations were produced by a phrase-based SMT system. The instructions to those performing the post-editing were to make the minimum amount of corrections necessary for a publishable translation (Potet et al., 2012), the same corpus as described in Section 4.2.1.

Lexical coherence MT has been shown to be somewhat consistent in its use of terminology in some research (Carpuat and Simard, 2012), and this can be an advantage for texts drawn from narrow domains with significant training data (see Section 2.1), but MT systems may output direct translations of ST items that may be inappropriate in the target context. Moreover, while a specific TT word may correctly translate a ST word in one context, it may require a totally different word in another context. In our training data the most common occurrence of the French noun *boucher* may correspond to the English word *butcher*. This increases the probability of the translation equivalence *butcher*, yet in the translated text it could on occasion be used as a verb indicating *to block* (for example, *road block*).

As illustration, we include an example in Table 5.1. Here the wrong translation of the French word *tour* was used, and renders the sentence incoherent. While the usage in the ST is obvious from context, this is not the most common translation of the word, and this is why the system got it wrong. As Wong and Kit (2012) note, the lexical cohesion devices have not only to be recognised, but also must

version	text
ST	*'L'extrême droite européenne est caractérisée par **son** racisme...'*
MT	*'The extreme right is characterised by **his** racism...'*
REF	*'A common feature of Europe's extreme right is its racism...'*

Table 5.2: Example of reference resolution error: the pronoun *son* is wrongly translated as *his* (Potet et al., 2012).

be used appropriately. And this may differ from the ST to the TT.

Reference resolution As stated by Elsner and Charniak (2008), "Pronoun coreference is another important aspect of coherence- if a pronoun is used too far away from any natural referent, it becomes hard to interpret, creating confusion. Too many referents, however, create ambiguity." Similarly, incorrect referents are also misleading and affect coherence. Reference resolution, pronoun prediction and anaphora resolution (referring to someone or something previously mentioned), are very challenging issues in current MT approaches and have given rise to considerable research (Novák, 2011; Le Nagard and Koehn, 2010; Hardmeier and Federico, 2010; Hardmeier et al., 2013b; Guillou, 2012). This is again due to the fact that inter-sentential references are lost in most decoders, which they translate one sentence at a time. In fact, often the window under consideration is smaller than an entire sentence, and references within it may still be wrong. Reference resolution is affected in several ways. The context of the preceding sentences is absent, and so the reference is undetermined. Even once it is correctly resolved (by additional pre-training or a second-pass), reference resolution is directly impacted by linguistic differences, for example, the target language may have multiple genders for nouns while the source only has one. The result is that references can be missing or wrong. In the example we show in Table 5.2, the error occurs within the same (short) sentence, so is not beyond the sentence boundary. Here the pronoun *son*, referring to the racism of the extreme right, is wrongly rendered as *his*.

Discourse connectives Discourse connectives are vital for the correct understanding of discourse. Yet in MT systems these can be handled incorrectly or

version	text
ST	*Je me rappelle qu'il disait que **si** Mincy avait permis Bayamon de remporter un championnat, Gausse allait nous aider essayer d'en remporter un autre.*
MT	*I remember that he said that **if** Mincy had enabled Bayamon to win a championship, respect would help us try to win another.*
REF	*I remember him saying that **if** Mincy had given Bayamon one championship, Gausse would help get another.*

Table 5.3: Example of erroneous discourse connective: *si* should have been better translated as *while*(http://www.statmt.org/wmt13).

missing altogether (Meyer and Poláková, 2013; Meyer and Popescu-Belis, 2012; Steele, 2015). In particular, where discourse connectives are ambiguous, e.g. those which can be temporal or causal in nature, the MT system may choose the wrong connective translation, which distorts the meaning of the text. For example, the English word *since* can be both temporal and causal in nature. Depending on its meaning in context, it could equate to a different word in another language. So if it was being used in a temporal sense, the translation *depuis que* would be appropriate in French, whereas if it were being used in a causal sense, that translation would be wrong and the translation *parce que* would be more correct. To illustrate, we show an example in Table 5.3 where the French connective *si* can signal *condition* (most of the time) or it can signal *concession* (less often). In this case it is being used in the second sense, but this has not been correctly translated– interestingly not in the reference either, which was not of particularly good quality here (*while* would have been a more appropriate translation).

It is also possible that the discourse connective is implicit in the source, and thus need to be inferred. It may also be legitimately implicit in the target. While a human translator can detect this, an MT system cannot. Or sometimes a connective can consist of several words, for example *on the other hand* in English, and is only partly rendered. While they are small, cue words guide the reader and help create the logic in the text.

Clausal ordering It is well-established that the ordering of textual units di-

rectly affects coherence (Lapata, 2003). This is the case for clauses, as well as sentences, where there are 'canonical orderings' (Mann and Thompson, 1988). Different languages have different structures, and so the target language may require reordering to be more coherent. Extensive reordering is penalized in Phrase Based Machine Translation (PBMT) systems, which means that ordering in the TT may be distorted, ending up too close to the canonical order of the ST and leading to an incoherent sentence formation. Consider the example below, ignoring lexical errors and poor word ordering, and focusing on the boldfaced fragment where the natural logic of the English clauses is distorted in the MT: the clauses are reversed, undermining the coherence of the text as a whole:

> *The pool collector discovered remarkable specimens Meder during their research in Hungary. In addition to sumptuous spas such as the 100-year-old Széchenyi bath in Budapest, she found a nitrate-containing waters in a cosy cave (cave bath in Miskolc-Tapolca), as well as a thermal spa, whose Becken are filled with alkali-containing water and are in a bottle-shaped building (Városi Termálfürdö in Jászberény).* **Offer spectacular views, however, many heated outdoor pools in the Switzerland and Austria***: while you have the entire city in the eye of the Zurich roof swimming pool, you can look in the outdoor swimming pool in a bath in St. Anton at the Arlberg snow-covered - and: in the steam room, there is a window from which you can watch the bustle on the ski slopes.*

This is apparent from comparing the MT output to the REF: the reference translation has a clausal pattern which is more coherent to the English reader.

> Pool collector Meder discovered some notable examples during research in Hungary. She fond nitrate-rich water in a karst cave (the Cave Bath in Miskolctapolca) and a thermal bath filled ith alkaline water in a bottle-shaped building (Városi Termálfürdö in Jászberény) in addition to magnificent therapeutic baths such as the 100-year-old Széchenyi baths in Budapest. **Many of the heated outdoor baths in Switzerland and Austria, on the other hand, offer spectacular views.** You can view all of Zurich from a rooftop bath or look

> out on the snow-capped Arlberg from an outdoor pool in St. Anton. And there is a window in the steam bath giving a view of the action on the ski piste.

Particularly in hierarchical or tree-based MT systems, the order of clauses within sentences may have become reversed, or may be unnatural for the TT. This can affect the understanding of the sentence, the overall logic of it in the context of the surrounding sentences, or simply require a reread which itself is indicative of impaired coherence.

5.3 Methodology

The proposed framework will take as input well-formed documents that are determined 'coherent' and then artificially distort them in ways (detailed below) that directly affect coherence in the manner that an MT system would. The resulting texts will make a corpus of 'incoherent' texts for assessing the ability of our coherence models to discriminate between coherent and incoherent texts.

This will be done in a flexible manner, such that the incoherent documents can be created for a variety of (coherent) input texts. Moreover they can be created for specific types of errors. The quality of MT output varies greatly from one language pair and MT system to another. For example, the output from a French-English MT system trained in very large collections is superior to that of, an English-Finnish system trained on smaller quantities of data (Bojar et al., 2015). The error distributions will therefore vary depending on the language pairs and MT systems. The errors themselves will also vary, depending on the language pair, in particular for aspects such as discourse markers and discourse structure, and on the MT system. Some of these errors are more relevant for particular language pairs, e.g. negation for French-English, which is otherwise a well-performing language pair. We propose to inject errors programmatically in a systematic manner, as detailed below. The aim is to create a flexible framework so that we can generate an artificial corpus for a particular language pair using output from a particular MT system, wherever a parallel PE text exists.

Ideally we would like to establish the distribution of errors from their occur-

rences in MT output, although it is very problematic to determine an appropriate error distribution based on observations. The distributions will be specific to given language pairs and MT systems. Moreover, one of the main issues is identifying errors of coherence in the first place: manual inspection and annotation for coherence is very hard to formalise as a task, time consuming and costly. If we could do this, then our original problem would be solved– that of ultimately detecting and improving the coherence of MT output. This is specifically why we need this corpus. Therefore, the errors themselves and consequently the distribution of errors in our corpus will be based on post-edits, on the basis that these are made to ensure the comprehensibility of the text.

We initially considered recreating the particular errors artificially: We proposed replacing entities with alternatives, using phrase tables from an MT system to generate likely entity variations. We planned to replace certain discourse connectives, reflecting connectives which have perhaps both temporal and causal meaning in one language and are commonly mistranslated. We considered modifying the order of sibling nodes in the syntax tree (e.g. reversed) at the appropriate level in order to alter the order of clauses.

However, in this experiment we have devised a method which will allow a more systematic and realistic approach. In particular, we identify these errors systematically by comparing MT output with the parallel PE version. In general, the post-edited version of MT output should not include stylistic changes, but should be limited to corrections which ensure comprehension (a *corrected* version of the MT, in effect). Given that our aim is to create a corpus which exhibits errors related to coherence, but does not have the grammatical errors[1] or stylistic differences which would otherwise be present in regular MT output, we use this PE version as our corpus into which we inject errors. We use it as the 'coherent' text to which we can compare our MT version and identify specific coherence errors. These will be extracted from the MT and will then be injected into our otherwise correct PE version, resulting in a parallel corpus where the 'incoherent' version should only contain coherence errors. As Popović et al. (2016) points out, post-edits are closer to the MT output than human-authored reference translations,

[1]Some do not think that grammatical errors come under coherence, although we would argue that they may well do so, depending on the error.

and therefore more suitable for automatic evaluation or tuning.

By necessity, this method relies on good quality post-editing; the post-edits should ideally be done by a professional translator, or at least a bilingual who can assess whether the coherence of the source text is being transmitted to the target text. And the post-edits should only be those deemed necessary to ensure coherence, not simply subjective vocabulary choices. Given that our method is automatic, it will also include some unavoidable errors, which are particularly related to the automatic word alignments that we rely on. However, we believe this is still a worthwhile exercise, and is the closest we can hope get to a corpus with specific MT coherence-related errors without resorting to annotation which is very difficult to define. The errors injected are authentic, as is the error distribution of the annotated corpus. It is comparable to existing artificial attempts to induce coherence errors, such as the well-known shuffling experiments (whereby sentences in coherent texts are shuffled to create incoherent ones).

As detailed already in Section 5.2, there are various coherence errors which occur in SMT in particular, due to the way the SMT decoder works, with translation being done sentence by sentence, with little or no access to context and often no particular modelling of crosslingual variations. We hope that we can test the effectiveness of our various models in the task of measuring coherence. We therefore focus our error injection on three types of errors, i.e. errors relating to entity-based lexical coherence, discourse connectives and clausal ordering. While the original syntax model aimed to trace intentional structure, the results were inconclusive (Chapter 3), even in our extended implementation. However our our IBM1 syntax model can potentially track clausal ordering when taken at a certain level of parse tree depth, as it considers syntactic patterns consisting of bigram sequences of sibling constituents. So we will test whether the IBM syntax model is capable of measuring the coherence of clausal ordering. Our ultimate aim is to create a corpus with these types of coherence errors, as systematically and correctly as is possible automatically.

5.4 Pipeline

Having previously described the types of errors that we hoped to capture (Section 5.2), we now describe the pipeline we derived for automatically detecting and injecting them, and what errors actually were injected by this process where we automatically corrupt a PE version to produce authentic errors.

The LIG corpus (see Section 5.2) is suitable for our experiment, as it consists of parallel MT-PE documents, with a target language of English. The errors are extracted by comparison and analysis of these parallel documents. We first derive Human-targeted Translation Error Rate (HTER) alignments (Snover et al., 2006) between the MT and the PE versions, and pinpoint the elements which have changed, in order to help us determine the nature and degree of the post-editing performed.

The following sections describe the manner in which we derive the different classes of error. In each case errors are derived by comparing the MT and PE in various ways, and then tracking the corrected error to inject into the corpus, which is a copy of PE version– and so in theory should have no grammatical errors left. An alternative would have been to inject the errors into the reference, but it was deemed that this would be unrealistic, as the style of the reference would be considerably different from the MT from which the errors derived, and the continuity in lexical choice would differ too much. Two versions of the corpus are created, one tagged, in which the errors are identified with xml tags, and one version plain text. For the tagged version, the tags identify the type of injected error.

We used the Stanford Parser (Manning et al., 2014) to parse both versions, which on the whole seems to handle the poorer quality of MT output adequately. This allows us to identify the nouns (for the entity errors) and create parse trees as input to the discourse tagger by Pitler and Nenkova (2009), which identifies the discourse connectives and their usage.

5.5 Entity errors

For all the documents in the corpus, we identify all the nouns in each parallel sentence, and determine whether there are any that have been *deleted* from the MT and are absent in the PE, or *inserted* in the PE, but absent in the MT, or both of these possibilities (i.e. *substitution*). The ones that have been deleted from the MT by the post-edition are deemed therefore erroneous, and thus we will reserve them for injecting back in later. The nouns that have been inserted will be tracked for removal (to revert the post-edit, in this case either addition or substitution). Having tracked the document and line for each error, we subsequently inject these errors into the seed corpus.

When injecting the entity error– a deleted/replaced/inserted noun– we have to account for the fact that other post-edits may have taken place, changing the sentence structure and the location of that noun. We use the HTER alignments to identify the position of the word, and check that the word to be replaced/removed is actually a noun. If this is not the case, we attempt to find the nearest candidate (nearest noun). Ultimately we could create many more heuristics in this situation, but hand crafting them is time-consuming. In general, we are able to identify the position for injection. In our tagged version of the corpus we keep track of the word which has been deleted/replaced. For instances where the word order within the sentence has been altered substantially (determined by a threshold parameter passed when deriving the clausal ordering errors, see Section 5.7), then no attempt is made to insert that instance of entity error, as it will have already been injected via the clausal replacement (see Section 5.7). An example of an injected entity error in our corpus is:

> *Governments must offer more* `<error type=entity edit=0 item=possibility>` *chance* `</error>` *for success to young people by improving the access to and of care and education .*

Here the word *chance* had been replaced with *possibility* in the post-edited version. They are both possible translations for *la chance* in French. A further example extracted from the error-injected corpus, in a document on sustainable forests:

Nobody can finish his day without having used a `<error type=entity edit=0 item=product>`*sideshow*`</error>` *from the forest .*

Here the word *sideshow* in the original MT has been inserted into the post-edited version. It undermines the coherence of the text, in that it renders it incomprehensible. We inject the erroneous word from the MT back in, replacing *product*, which had been deleted in the post-editing.

5.6 Connective errors

To identify errors related to discourse connectives we use a discourse tagger (Pitler and Nenkova, 2009). It identifies the potential discourse connectives in a sentence, marking those which are not serving a discourse purpose (e.g. whether 'and' has a discourse connective usage in that context or not), in addition to identifying the discourse sense of the connective in question (e.g. whether 'since' is used as a temporal or causal connective). We extract and compare the connectives for the MT and PE, again determining where a connective has been deleted/inserted/substituted. For each document we compare the list of connectives identified by the tagger in the MT with those identified in the PE. This is the same tagger we used in Chapter 4 to identify the discourse relation used in English. It uses syntactic features, "dividing sentences into elementary discourse units among which discourse relations hold". (Pitler and Nenkova, 2009) We follow the same pattern, taking the post-edited version of the sentence, using the list of potential connective errors, identifying the correct location, and injecting the connective error into it. Again, a check is made to ascertain if the post-edits have been so substantial in the sentence under consideration that it will be hard to inject the error in the correct position, in which case we skip that instance.

For example:

The southern and eastern Europe will suffer inflation, it is unavoidable, `<error type=connective edit=del item=because>` `</error>` *as the regions develop and industrialize, the terms of trade improve, and under the auspices of a monetary union, regional inflation can become a barrier to overcome.*

In the above instance the connective *because* had been inserted in the post-edition to ensure that the causal sense is clear. So in order to recreate that coherence error, we delete it out of the post-edited seed corpus.

5.7 Clausal ordering errors

The category of clausal ordering errors identifies instances where the coherence has been impacted through an erroneous ordering of clauses within the sentence. We again use the HTER alignments for this exercise, determining where the reordering has been substantial– this is controlled via a threshold parameter. In general, it seemed that where the HTER alignments differ by more than four positions, then the error was of a more substantial nature. Our assumption is that this had an effect on the coherence of the sentence, given that the post-editor had deemed necessary to alter the structure to that extent. This tended to be in places where the natural language structure of English had been distorted, compromising coherence. As has been widely recorded elsewhere (Lin et al., 2011; Louis and Nenkova, 2012; Guzmán et al., 2014), the way in which sentences are structured and arguments laid out is very important for coherence. By keeping the threshold as high as this (i.e. four positions), we are overlooking simple reversals of adjective-noun positions (which often occur when translating between Romance languages and English, for example). For this category of clausal ordering errors, the entire sentence was replaced in the artificial corpus, i.e. the entire MT sentence used in place of the PE sentence. While this is somewhat simplistic, we needed an automatic solution in this exercise. Moreover, the order had generally changed to such an extent that trying to identify which are clausal and which are lexical replacements becomes very difficult to do automatically. We experimented with various threshold parameters. A threshold of four, whereby differences in alignments exceed four positions (i.e. five or more) were tracked, resulting in 920 clausal ordering errors for 10759 sentences of text (for 361 documents). We also tested with a threshold three (i.e. alignments which differ by more than four positions), which led to 1671 out of 10759 sentences. With a threshold of five there are 510 errors out of 10759 sentences. These are not a substantial amount to distinguish the PE from the version with inserted

version	text
MT	*<error type=clausal>Throughout the 1990s, they are virtually out of Kosovo managed by the Serbs in creating parallel institutions.</error>*
PE	Throughout the 1990s, they virtually left Serb-managed Kosovo by creating parallel institutions.
MT	*<error type=clausal>Today, it seems that I have finally right, if not to panic, at least concern to me.</error>*
PE	Today, it seems, I am finally right, if it's not time to panic, at least it's time for me to worry.
MT	*<error type=clausal>Globally, the danger is that the lawlessness shown by Putin exports.</error>*
PE	Globally, the danger is that the lack of respect for the law shown by Putin could spread.

Table 5.4: Injected clausal errors, with threshold set to four.

errors. The results can be seen from Table 5.7. With the threshold set at four the sentences displayed in Table 5.4 were deemed incoherent, and re-introduced into the seed corpus.

In each instance the clausal structure is distorted and the meaning is unclear as a result. The errors detected are not strictly ones of clausal ordering, however, but general word order problems. The type of errors we anticipated capturing and injecting were not always matched in reality for this type of error, as illustrated by the examples above.

Increasing the threshold ensured that we only capture larger changes in word position, and therefore more dramatic errors. It also, however, means that there are fewer errors overall. While this level of error (more dramatic) could be easier for the coherence models to pick up, the problem is that there are then so few errors to differentiate overall. The results from various thresholds are displayed in Table 5.7.

Type of error	Number of errors injected
lexical errors	9716
connective errors	1117
clausal errors	920

Table 5.5: Corpus description of error types, detailing the number of errors deemed to be of particular types injected into the artificial corpus.

5.8 Results

In Table 5.5 we detail the number of errors deemed to be of particular types which are injected into the artificial corpus. We investigate how GRID, GRAPH, LN, IBM1 and Dis-Score perform with this corpus. To determine whether those coherence models (representing entity, syntax and discourse relations) are indeed measuring coherence, not simply the differences between the MT and HT, we test them on this artificial corpus, containing the injected coherence errors.

We try different versions of the corpus, injecting just entity errors, just structural errors, or injecting all three types of error together. This allows us see whether injecting the specific type of errors is a productive exercise and whether our models are doing what we expect. Our results are displayed in Tables 5.6-5.8 in the following subsections. Table 5.6 displays results where only entity errors were injected. Table 5.7 displays results for varying thresholds of clausal errors. Finally, Table 5.8 displays errors for corpus with all 3 types of error injected. For the syntax-based models, we experimented at various depths of the parse tree and display the best result only. In each case, we display the upper bound, which indicates the proportion of documents with an error injected.

5.8.1 Entity errors

We would hope that the entity-based coherence models identify this type of error. As can be seen from the results in Table 5.6, the GRAPH metric does very well, identifying the error-injected version 75.62% of the time. The GRID, however, does not perform so well, and is unable to discriminate between the injected and clean corpus. This is puzzling, since the entity transitions will have been broken in places, but that seems to have been insufficient for the model to discriminate. 358

model	$\text{ref}_{>}$	$\text{ref}_{\geq}$
upper bound	99.00	
GRAPH	75.62	76.45
GRID	34.35	34.90

Table 5.6: Ability of entity models to detect corpus with entity errors only. Upper bound is given by 358 out of 361 documents with no lexical errors. Comparing injected PE vs PE.

model	$\text{ref}_{>}$	$\text{ref}_{\geq}$
upper bound	85.04	
IBM1-D4	50.69	50.69
LN-D4	38.23	52.08
model	$\text{ref}_{>}$	$\text{ref}_{\geq}$
upper bound	70.91	
IBM1-D4	45.15	51.80
LN-D5	4.43	97.51

Table 5.7: Ability of syntax models to detect injected corpus with clausal ordering errors, at threshold 4 (above) and 5 (below). The upper bound is 85.04% and 70.91% respectively. Comparing injected PE vs PE.

documents out of 361 contained an inserted lexical error. Whereas the GRAPH metric benefits from the fact that it is directed, so the links to all successive entities are affected also.

5.8.2 Clausal ordering errors

We had presumed that due to the nature of the IBM1 model, modelling syntax patterns at various levels, it would respond to the injection of clausal errors. From the results in Table 5.7, this is clearly not the case. However perhaps there are insufficient errors injected. Moreover, we are only injecting random sentences, never adjacent sentences. As a result this model possibly does not have enough to work with, given that it models *bigrams* of patterns in adjacent sentences. This was also a cruder method of error injection, compared to the other two. As such, it is possibly not effective enough. Certainly, from the results for the $\text{ref}_{>}$ metric in Table 5.7 the syntax models are unable to detect the clausal errors artificially injected in this way. With the threshold set at four, 307 documents (out of 361)

contained an error (85%). At a threshold of five, there were errors injected in 256 documents out of 361, so in 70.91%. The upper bound is therefore higher than for entity errors. Interestingly, comparing the scores under both the $\text{ref}_{>}$ and the $\text{ref}_{\geq}$ metric for the LN model and IBM1 model and looking at the second set of results (for the threshold of 5 in the lower half of Table 5.7), the LN model scores the PE higher or equal to the injected corpus much more frequently, as this metric rewards ties, whereas the $\text{ref}_{>}$ metric penalises them. As such, the LN score increases significantly when ties are explicitly rewarded, and it performs better when the ranking is between just two outputs which vary very little.

5.8.3 Connective errors

Due to the fact that the MT does manage to correctly render a large proportion of the connectives, we expect there to be many ties (discussed already in Section 4.3). In fact 236 out of 361 of the documents result in a tie. Which is the reason that the $\text{ref}_{>}$ metric seems so low, at 34.63% (in Table 5.8). However, out of the 146 documents where there was no tie in the scores, the PE scored higher than the injected version on 125 of the cases, or 86%. Hence in this case the $\text{ref}_{\geq}$ (where Dis-Score scores 94.18%) actually does not simply indicate a lack of ability to discriminate. The upper bound for Dis-Score with metric $\text{ref}_{>}$ is actually 86%, since 51 documents contain no connective errors, and Dis-Score solely detects these.

The really interesting issue here is that Dis-Score is scoring against the *source* text, and is evaluating whether the PE and injected version have an equivalent connective for the estimated discourse relation in the French source text. The other models were measuring the output alone.

5.8.4 All errors combined

We assess the ability of the models to detect the errors in the artificial corpus when comparing it against the PE, the idea being that they are detecting coherence errors alone. The results for both are displayed in Table 5.8.

model	$ref_{>}$	$ref_{\geq}$	**upper bound**
GRAPH	77.29%	78.12%	99.00%
IBM1-D3	49.58%	50.97%	85.04%
GRID	40.44%	40.72%	99.00%
DIS-SCORE	34.63%	94.18%	86.00%
LN-D4	14.40%	83.93%	85.04%

Table 5.8: Ability of models to detect injected corpus with all 3 error types (entity, clausal, connective errors). Comparing injected PE (our artificial corpus) vs PE. We report the upper bound as the number of documents injected with that type of error.

version	text
MT	*The* `<error type=entity edit=0 item=goal>` *aim* `</error>` *of scientific advisory committees is to provide impartial advice and reflect on political* `<error type=entity edit=0 item=processes>`*process*`</error>` *.*
MT	*In any* `<error type=connective edit=0 item=case>` *However* `</error>` *, it was too late to protest since the ten eastern countries had actually become full European Union members.*

Table 5.9: Example of limitations of the approach.

5.9 Limitations of the approach

As mentioned previously, there are shortcomings due to the automatic nature of the process of injecting errors. In particular, the errors extracted (and subsequently injected) depend on the post-edits. This may lead to unnecessary or subjective changes.

For example, see Table 5.9 where the first injected error is actually subjective– there is no real need to change the noun in this instance (from *aim* to *goal*). This word should NOT have been post-edited according to the post-editing guidelines used in LIG. However this has occurred due to the nature of the post-edits. The second error in this sentence is also arguably legitimate– a case of the French noun being plural where English would more likely be singular, although this does not actually impair coherence.

In addition, there is occasionally an issue with the alignments. For example,

version	text
MT	*In fact, different generation were the decisive factor in the race for the presidential election, replacing the feelings regional dominrent every presidential race before it.*
PE	*In fact, generation differences were the decisive factor in the presidential election race, replacing the regional feelings that dominated every presidential race before it.*
INJECTED PE	*In fact , generation* `<error type=entity edit=del item=differences> </error>` *were the decisive factor in the presidential election race , replacing the regional feelings that dominated every presidential race before it .*

Table 5.10: Injected error of lexical type resulting from reversal of post-edit.

the second row in Table 5.9, where the word *However* appeared in the MT, and was therefore due to replace *In any case*, but has actually just replaced the latter part of the phrase.

Sometimes the post-edit has used a turn of phrase that is better for the target language - see the example in Table 5.10, changing *different generation* to *generation differences*, but the MT construct (which is overly affected by the source) is then corrupted by reintroducing what has been picked up as an entity error. So, looking at the MT in top row, which was modified via post-editing to the middle entry of the table. This has been modified, removing the word *differences* which was inserted in the post-edit, to now read as on the bottom row.

Here the word *differences* had been added in the post-edit, and therefore removed from MT, but in this instance leaves that sentence worse. Unfortunately, without a high degree of additional heuristic insertion rules, this is hard to avoid.

5.10 Conclusions

In this chapter, we began by describing how problems which are related to lack of coherence are manifested in MT output. We then detailed how we automatically distorted a PE version of a corpus, manipulating the data in systematic ways to create a corpus of artificially generated incoherent data. While not error-free, this represents an automatic way of creating a corpus with some of the errors

we believe are symptomatic of MT output, and which impair coherence. The experiment was not successful at capturing clausal ordering errors which occur in MT output. However, injecting just the lexical and discourse errors, it could serve as a first pass, suitable for refinement via human annotation. Sentences could also be used in isolation, as part of a test suite to measure the effectiveness of MT systems at handling this type of discourse phenomenon. The process could be improved with use of better monolingual alignments and additional linguistic analysis for identifying the exact positions of entities and discourse connectives to be replaced. The pipeline code can be used with any input corpus which meets the prerequisites described in Section 5.4. This means it can be used to recreate a coherence corpus for specific purposes, such as a corpus for German-English focused on lexical coherence. The types of errors in the latter will differ from a French-English corpus focusing on discourse connective errors. In addition, we believe that evaluation within MT has to move beyond benchmarking against a single reference translation, as linguistic variation is so great. One way of doing this is by assessing MT output in alternative ways.

In terms of evaluating our models, the experiment was productive insofar as it indicates that the Graph and Dis-Score metrics were able to detect the lexical coherence and discourse connective errors injected. The syntax models were unable to detect clausal ordering errors injected in this fashion, which we attribute to the fact that the models themselves are limited and the errors injected were not in fact only *clausal ordering errors*. Given that our extended IBM1 model performed better in the previous experiment (Chapter 3), we can conclude that it was able to discriminate between syntax patterns in MT and HT output. This merits further investigation, with potential to influence syntax in MT output.

In the next Chapter (Chapter 6), we apply a simplified version of the entity-based models (described in Section 2.2.1) to a multilingual context for the first time, examine differences in the general patterns across languages, and establish how we can integrate that insight in order to measure whether the lexical coherence in the ST is transferred to the TT.

Chapter 6

Examining Lexical Coherence in a Multilingual Setting

This chapter presents an exploratory study which represents our early research on how lexical coherence is realised in a multilingual context, with a view to identifying patterns that could be later used to improve overall translation quality in MT models. It lays the groundwork for a crosslingual lexical coherence metric. Ideally a coherent source document when translated properly should result in a coherent target document. However, coherence does vary in how it is achieved in different languages. Unlike a human translator, who translates the document as a whole in context ensuring that the translated document is as coherent as the source document, most MT systems, and particularly SMT systems, translate each sentence in isolation and have no notion of discourse principles such as coherence and cohesion. We explore the two entity-based frameworks (an entity-grid model and an entity graph metric, described previously in Chapter 3) in a **multilingual** setting to understand how lexical coherence is realised across different languages. These frameworks have previously been used for assessing coherence in a monolingual setting. We apply them to a multilingual setting for the first time, assessing whether entity based coherence frameworks could help measure and ensure lexical coherence in an MT context. We examine linguistic differences in the general patterns across three languages (French, German and English), to determine which aspects of coherence are preserved crosslingually and

which ones are language dependent. We then establish how we could integrate that insight with a view to measuring whether the lexical coherence in the ST is transferred to the TT, in a similar way as we did for discourse relations in Chapter 4.

In the following section (Section 6.1) we describe adapting entity based coherence models to a multilingual context. Then we detail our experimental settings (Section 6.2) for the two main parts of this research. Firstly (Section 6.3), we present a multilingual comparative entity-based **grid** for a corpus comprising various documents covering three different languages, using data and settings as described in Section 6.2. We examine whether similar patterns of entity transitions are exhibited, or whether they varied markedly across languages. Secondly (Section 6.4), we apply an entity **graph** in a multilingual context, using the same corpus. We assess whether this different perspective offers more insight into crosslingual coherence patterns. Our goals are to understand differences in the implementation of lexical coherence entity models across languages so that in the future we can establish whether this can be used as a means of ensuring that the equivalent lexical coherence is transferred from source to machine translated documents. Our conclusions are set out in Section 6.5.

6.1 Exploring entity-based coherence

Entity-based coherence aims to measure the *attentional state*, formalised via Centering Theory (Grosz et al., 1995), as discussed in Section 2.2.1. Previous computational models for assessing entity-based coherence have been deployed in a monolingual setting, (Lapata, 2005; Barzilay and Lapata, 2008; Elsner et al., 2007; Elsner and Charniak, 2011b; Burstein et al., 2010; Guinaudeau and Strube, 2013) as detailed in Chapter 2. The focus of previous work was in using this knowledge (of the attentional state), via patterns of prominent syntactic constructions, to distinguish coherent from non-coherent texts. In our research detailed here, we investigate differences in the general patterns, particularly across languages. Our final goal – which remains as future work – is to track the attentional focus in the source text, and attempt to measure the extent to which this is correctly rendered in the target text.

While previous work on entity grids (Lapata, 2005; Barzilay and Lapata, 2008) has found factors such as the grammatical roles associated with the entities affect local coherence, this research was on English texts, a language with a relatively fixed word order. However languages vary, and as described in Poesio et al. (2004) the parameters of the experiment may need to be adapted. Cheung and Penn (2010) suggest that topological fields (identifying clausal structure in terms of the positions of different constituents) are an alternative to grammatical roles in local coherence modelling, for languages such as German, and showing that they are more effective than grammatical roles in an ordering experiment. The syntactic patterns used in the aforementioned entity grid research do not apparently hold for Japanese, Italian or Turkish either (Poesio et al., 2004). Indeed, as Filippova and Strube (2007) reported when applying the entity grid approach to group related entities and incorporate semantic relatedness, "syntactic information turned out to have a negative impact on the results". Our initial experiments will take all nouns in the document as discourse entities, as recommended by Elsner and Charniak (2011b), and investigate how they are realised crosslingually. This will work for the languages we have chosen: French, German and English.

For this set of experiments we therefore apply a slightly simplified version of the grid, recording the presence or absence of particular (salient) entities over a sequence of sentences. In addition to being the first cross-lingual study of the grid approach, this experiment also aims at examining the robustness of this approach without a syntactic parser. While the grammatical function may have been useful as an indicator in the aforementioned monolingual research, this does not necessarily hold in a multilingual context. Simply tracking the existence or absence of entities – and how they move in and out of focus – allows for direct comparison across languages.

Entity distribution patterns vary according to text domain, style and genre, which are all valuable characteristics to capture and attempt to transfer from source to target text languages where appropriate. The distribution of entities over sentences may vary from language to language too. The challenge from an MT point of view would be to ensure that an entity chain is carried over from source to target text, despite differences in syntax and sentence structure, and taking account of linguistic variations. We experiment with both an entity grid

and an entity graph. Entity grids are constructed by identifying the discourse entities in the documents under consideration and representing them in 2D grids whereby each column corresponds to the entity (i.e. noun) being tracked, and each row represents a particular sentence in the document. Entity graphs represent the same information in a graph format, where nodes represent the sentences and entities. Edges are created where there are shared entities between sentence nodes. Both are explained in detail in Chapter 3.

While some research has indicated that MT frameworks are good at lexical cohesion (Carpuat and Simard, 2012), in that they are consistent, others have reported different results (Wong and Kit, 2012), since MT systems can persist in using a particular translation which is incorrect (Guillou, 2013). We believe that investigating entity-based frameworks in a multilingual setting may shed some light on the issue. In particular, we had initially hoped to ascertain whether they help in the disambiguation of lexical entities, where in an MT setting the translation of a particular source word, e.g. *bank* in English, could be translated as either *la rive*[1] or *la banque*[2] in French, depending on the context. Currently most SMT systems determine which word to use based on the probabilities established at training time (i.e. how frequently *bank* equated to *la rive* and how frequently it equated to *la banque*), and a short surrounding context window. While this choice should be determined by the whole context, the problem is that most systems translate one sentence at a time, disregarding the wider context. Furthermore, while some lexical ambiguities can be resolved at sentence level, translations may be one phrase at a time, influenced by the options present in the LM.

6.2 Experimental settings

For our multilingual experiments, we used parallel texts from the WMT10 corpus[3] (Callison-Burch et al., 2010) with three languages: English, French, and German. In particular, we used the WMT10 test data, comprising 90 news ex-

[1] bank of a river
[2] bank as a financial institution
[3] http://www.statmt.org/wmt10/

cerpts extracted over various years. The direction of translation varies for different documents, as discussed in Section 6.3.

For comparison, we also take the French and English documents from the **LIG** corpus (Potet et al., 2012) of French into English translations. It contains a total of $10,755$[1] quadruples of the type: *<source sentence, reference translation, automatic translation, post-edited automatic translation>*. These form a concatenated group of 361 documents, of which we use 119 in our preliminary study, using the source (French) and reference translation (English). These are news excerpts drawn from various WMT years. In both these corpora, reference translations are provided by professional translators.

We used Stanford CoreNLP (Manning et al., 2014) to identify the noun phrases in each language. For the grid experiment, we set the salience at 2, i.e. recording only entities which occurred more than twice, and derived models with transitions of length 3 (i.e. over 3 adjacent sentences). We computed the mean of the transition probabilities, i.e. the probability of a particular transition occurring, over all the documents.

6.3 Multilingual grids

We report the entity transition distributions computed on the WMT10 data set for German, French and English in Table 6.1. Here $XX-$ indicates that an entity occurred in two consecutive sentences, but was absent in the following one. Similarly, $X-X$ indicates that an entity occurred in one sentence, was absent in the second, but occurred again in the next one. The dataset is small (90 documents) and therefore just indicative, but shows variation between the languages. Transitions are extracted across sentences, throughout the document (see Section 3.1.1 for more details). Of particular interest here are the compound words prevalent in German, and how these affect the entity grid. By counting the number of columns over all 90 grids, we can establish how many entities were tracked in each grid. As we see from Table 6.2, French logs the highest number of entities over all the grids, and German the least. In German we find the prevalence of compound nouns reduces the entity count.

[1] After removing some null and duplicate lines from the original $10,881$.

In order to illustrate the differences between the distributions of these entity transitions over the different languages, we then computed Jensen-Shannon divergence scores for French-English and for German-English, both displayed in Figure 6.1. This is defined as:

$$JSD(P||Q) = \frac{1}{2}D(P||M) + \frac{1}{2}D(Q||M) \tag{6.1}$$

$$where$$

$$M = \frac{1}{2}(P + Q)$$

Paying attention to the scale, it is clear from Figure 6.1 that the German and English divergence is greater overall than the divergence for French and English. For example the entity transitions which showed the highest variation were $XX-$, which was 0.045 for the difference between French and English and over 0.1 for German and English. Similarly, there is a difference between the entity transition XXX where the variation over the same pairs was 0.02 and 0.08. This indicates that for the German-English pair the pattern of entities occurring in three consecutive sentences was different from the French-English pair, and is informative for translations from these different languages into English.

Transition	**German**	**French**	**English**
XXX	0.001445	0.002382	0.000441
$X-X$	0.006240	0.006917	0.003184
$XX-$	0.005905	0.008853	0.003130
$-XX$	0.004142	0.006155	0.001672

Table 6.1: Multilingual entity transitions (mean of 90 WMT newstest2008 documents)

	German	**French**	**English**
Total no. entities overall	7435	9194	8481
Sentences overall	2030	1964	2013

Table 6.2: Statistics on extracted entity grids for WMT newstest2008 documents

There is a clear pattern across the entity transitions over the three languages

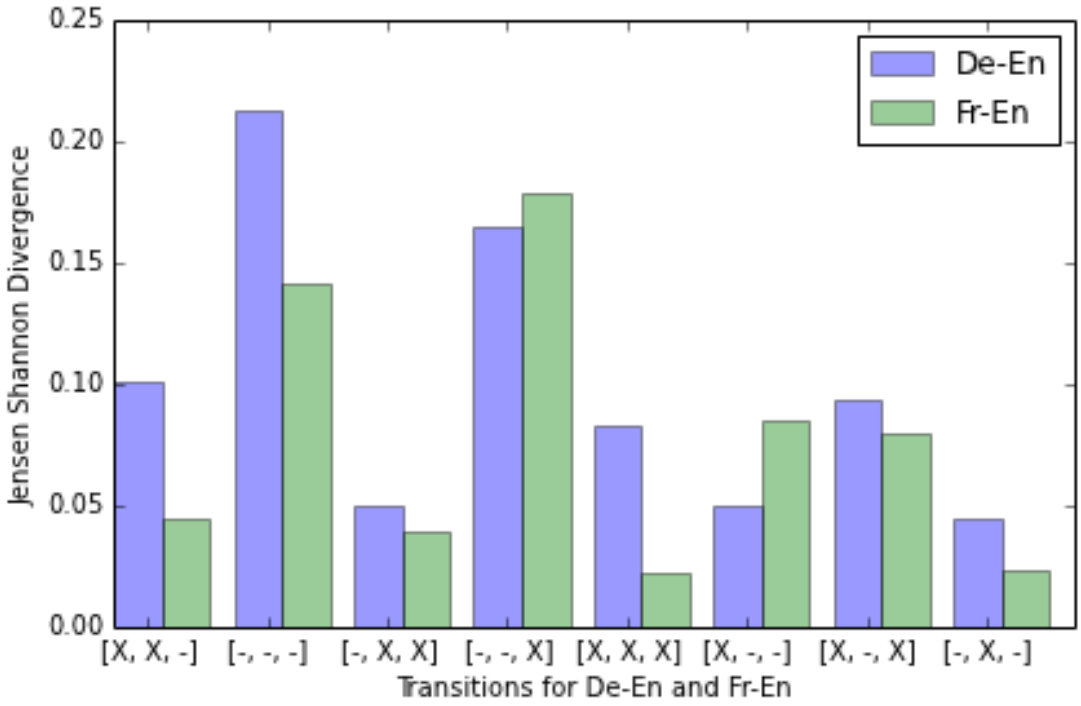

Figure 6.1: Jensen-Shannon divergence over distribution of entity transitions (length 3) for German-English and French-English (WMT newstest2008)

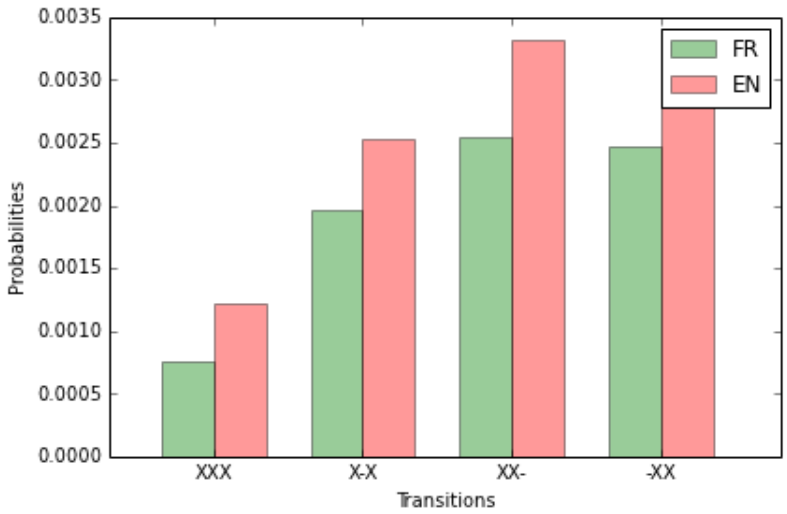

Figure 6.2: Comparative probabilities of transitions for the initial 119 documents of the LIG corpus

studied, although this is not a large dataset and as such it is just indicative. While the transition pattern $[-,-,-]$ alone is not very informative, it illustrates how for German-English the divergence is greater than for French-English. There are more compounds in the German text which result in fewer columns in the entity grid (Table 6.2), and comparatively fewer transitions of $[-,-,-]$. For French-English the patterns are less divergent.

By way of comparison, examining the LIG corpus, English had a higher probability over the various entity transitions in general, as illustrated in Figure 6.2. On closer analysis, it would appear that there are various issues at play. Firstly,

	French	**English**
Total no. entities overall	20886	17922
Sentences overall	3460	3420

Table 6.3: Statistics on extracted entity grids for initial 119 LIG documents

there is the matter of sentence boundaries, which affects the transition probabilities. Unlike the original WMT segmentation, which is enforced to ensure strictly parallel structures, we have used the natural sentence segmentation for these ex-

periments[1]. This results in a different number of sentences (Tables 6.2 and 6.3) and has a direct impact on the transition numbers. Our aim here is not to enforce strictly parallel sentences but to establish issues which need to be taken into account when trying to compare the lexical coherence in this manner. Across many of the documents in the **newstest2008**, the French version had fewer sentences within segments than the corresponding segments in German or English. This increases the number of transitions from sentence to sentence. French also exhibited more entities per document (Table 6.2 and 6.3). So the transitions are more concentrated. Both of these factors account for some of the higher levels of entity transitions in French over English and German in the WMT **newstest2008** documents. As can be seen from Table 6.2, the WMT **newstest2008** documents is for English and German to have more, shorter sentences. So elements of discourse which were in one sentence in French were occasionally split over two sentences in German or English, and thus an entity transition was over two consecutive sentences in French, but had a sentence between them in the other two languages. As a result, the XXX transition count was typically higher for French.

Of course, we can enforce the constraint of strictly parallel sentences, as in the WMT markup, but it is interesting to see the natural linguistic variation (albeit expressed as an individual translator's choice here). In this instance we are comparing the same texts, on a document by document basis, so comparing the same genre and style, yet there is a consistent difference in the probabilities. This would appear to indicate, amongst other things, that the manner in which lexical coherence is achieved varies from language to language. While just a preliminary study with a small dataset, it is supported by other research findings (Lapshinova-Koltunski, 2015a), which indicate that the amount of lexical coherence can vary from language to language.

6.3.1 Linguistic trends

The datasets used in these exploratory experiments are not large enough to constitute the basis for any significant statistical analysis, which warrants a large

[1]We use the punctuation of the texts themselves, not the $\langle seg \rangle$ breaks added for WMT when constructing the grid, as that captures the transfer from sentence to sentence.

corpus study. We therefore simply highlight some linguistic trends which would should be taken into account in future work. Interestingly, another reason for the variation across languages is the fact that in French there are instances of a noun in the plural as well as singular. For example, in document 37 of the **LIG** corpus the French used two separate entities where the English had one: 'inequality', which occurred at sentences: 0, 1, 2, 3, 4, 12, 13, 14, 17, 18, 19, 21, 31, was rendered in French by 2 separate entities: 'inégalités' at 0, 1, 2, 4, 12, 14, 17, 18, 19, 31 'l'inégalité' at 2, 3, 13.

This phenomenon occurred elsewhere too: 'effort' in English occurred in the following sentences of document 24: 8, 9, 10, 11. In French we actually find 3 separate entities used, due to the way the parser dealt with the definite article: 'l'effort' at 8, 'effort' at 9, 11 and 'efforts' at 9, 10. While we can adapt our models (via lemmatisation) to account for the linguistic variation, it is important that we appreciate the linguistic variation in the first place, if we want to measure **appropriate** lexical coherence.

Another comparative linguistic trend we found was that sometimes an entity in English is actually rendered as an adjective in French, and therefore not tracked in the entity grid, such as document 5, where the source text, i.e. French, has 'crises cambiaires' rendered in the English as 'currency crises', and while 'currency' is identified as an entity in English, it is an adjective in French, thus not identified as an entity. Apart from affecting the transition probabilities, it would seem that some form of lexical chains is necessary to fully capture all the necessary lexical information in this multilingual setting. In the same document, 'currency' occurs 8 times as an entity in the English, yet in the French besides being rendered as an adjective twice, is rendered 4 times as 'caisse d'émission' and only once as 'monnaie'. This is reflected in the fact that for this document the English had 127 entities where the French had 152.

As already mentioned, in general German exhibited a lower entity count (Table 6.2). This count is affected by the amount of compound words in German, and how we decide to model them. Thus, for example, from a particular document on cars[1], the word 'car' features as a main entity, but whereas it appears 6 times in French ['voiture' at sentences 6, 8, 23, 31, 32, 33] and English ['car' at

[1]newstest2008, docid nytimes/2007/11/29/53302

sentences 5, 7, 22, 31, 32, 33] respectively, in German it only appears twice ['Auto' at sentences 7, 22]. However, 'car' is part of a collection of compound words in German, such as 'High-end-auto' at sentence 31 in the document, [31=X] and 'Luxusauto' at sentence [32=X]. As it occurs in a different form, it is, in this instance, tracked as a different entity altogether.

Similarly, German exhibited a high ratio of $X - X$ transitions, where an entity skips a sentence, then reoccurs. This is explained by the occurrence of more, shorter sentences, as described above, and also by the compounding factor. With shorter sentences there is a greater chance that entities are split between two sentences, where the French may have had one. This also leads to lower likelihood of a transition to the next sentence; the transition would instead skip one sentence (appear as $X - X$ transition instead of $XX-$ or XXX). Plus a particular entity may not appear in three consecutive sentences, as it may have done in the French or English versions, because in the middle sentence it is part of a compound word.

This illustrates the linguistic differences that need to be taken into account when examining comparative coherence in a multilingual context. This could lead to a decision to lemmatise before extracting grids or graphs, but in that case they are no longer strictly **entity** grids. We can apply linguistic processing to make the different grids comparable, but that should be sensitive to the linguistic variation, as overly processing to make them comparable will lose the natural expression in a particular language.

In some cases the quality of the text was also an issue. WMT data (from which the **LIG** corpus was also derived) is generated both from texts originally in a given language, e.g. English, and texts manually translated from other languages (e.g. Czech) into that language (say English). And in some cases the human translation of the documents was not particularly good. This was the case for some of the English documents translated from Czech in the **newstest2008** corpus. This has a direct influence on the coherence of the text, yet as noted by Cartoni et al. (2011), often those using this WMT corpus fail to realise the significance of whether a "source" text is an original or a translation. What also has to be taken into account is the language of the source text, and the tendency for it to affect the target text in style, depending on how literal the translation

	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**
DE	x	-	-	x	x	x	-	-	-	-	-	x	-	-	-	x	-	x
FR	x	-	-	x	x	x	-	-	-	-	-	x	-	-	-	x	-	x
EN	x	-	x	-	x	-	-	-	-	-	x	-	-	-	x	-	x	-
	18	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**31**	**32**	**33**	**34**	**35**
DE	-	x	-	x	-	-	-	-	-	x	-	-	-	-	x	x	-	-
FR	-	x	-	-	x	-	-	-	-	-	x	-	-	-	-	x	x	-
EN	x	-	-	x	-	-	-	-	-	x	-	-	-	-		x	-	-

Table 6.4: Occurrences of 'Brown' in various sentences of parallel document (dropping last sentences of document due to spacing)

is.

It is interesting to trace how the main entities in a given text are realised across the languages. In Table 6.4 each numbered column represents a numbered sentence in a parallel document – not the original WMT segmentation. We have cut the last few sentences from the table, in order to fit it in. We can clearly see how the main entity is represented through the document, albeit not at identical positions due to the sentence breaks. In this case the French and German entities were closely matched in position at the start of the document, and then the English and German by the end. However, the main point is that in general, there are the same number of occurrences, as the thread of discourse is traced through each document with exact positions dependent on sentence breaks.

6.4 Multilingual graphs

We also analyse the graph framework in a multilingual setting to try and garner additional insight into variations in coherence patterns in different languages. The intuition is that this framework could be more informative than the grid as it spans connections between not just adjacent sentences, but any subsequent ones. We used the weighted projection, which considers the frequencies of the various entities in the documents, which we determined was more appropriate than syntax in a comparative multilingual context, for reasons explained already. Our intuition is that the weighted projection gives the best appreciation of the cohesive links between sentences, as it gives a higher weighting where they are more frequent, unlike the unweighted one which simply logs the sentences in

which an entity occurs. We used the same WMT newstest2008 dataset as for the grid experiments. The graph coherence scores were computed for all parallel multilingual documents and the summed scores are displayed in Table 6.5. As standalone scores, they are meaningless, but serve to establish the parameters for our longterm goal of conditioning the TT on the ST.

	coherence score	**coherence score (no compound splitter)**
French	26	30
English	47	56
German	17	4

Table 6.5: Number of documents (out of the 90 in the WMT newstest2008 dataset) for a given language which scored the highest among the 3 languages

On closer analysis we encountered the same issue with German compounds as for the grid, whereby the entities in the German grid were more sparse, due to the fact that compound words accounted for several entities. To establish just how much difference this was making, we also try applying a compound splitter for German[1]. For a given entity, we check if it decomposes into several entities, and if so each is entered separately in the graph. This results in a more uniform coherence score over the 3 languages. Whereas German had the highest coherence score for only 4 out of the 90 documents when no compound splitter was applied (as seen in Table 6.5), this figure rose to 17 with a compound splitter. One clear point to be made from these scores is that in a crosslingual study of this kind, using a compound splitter for German allows for a more direct comparison.

Interestingly, looking at the coherence scores for all three languages under the entity graph, they exhibit remarkably similar graph profiles (Figure 6.3). The documents which result in a low score for English are similarly low for French and German. So it would seem that it is possible to assess lexical coherence as judged by this metric in a crosslingual manner, albeit as one aspect of coherence, not as sufficient to alone judge the overall coherence of the document. As Tanskanen (2006) points out, "cohesion may not work in absolutely identical ways in all languages, but the strategies of forming cohesive relations seem to display considerable similarity across languages".

[1] http://www.danielnaber.de/jwordsplitter

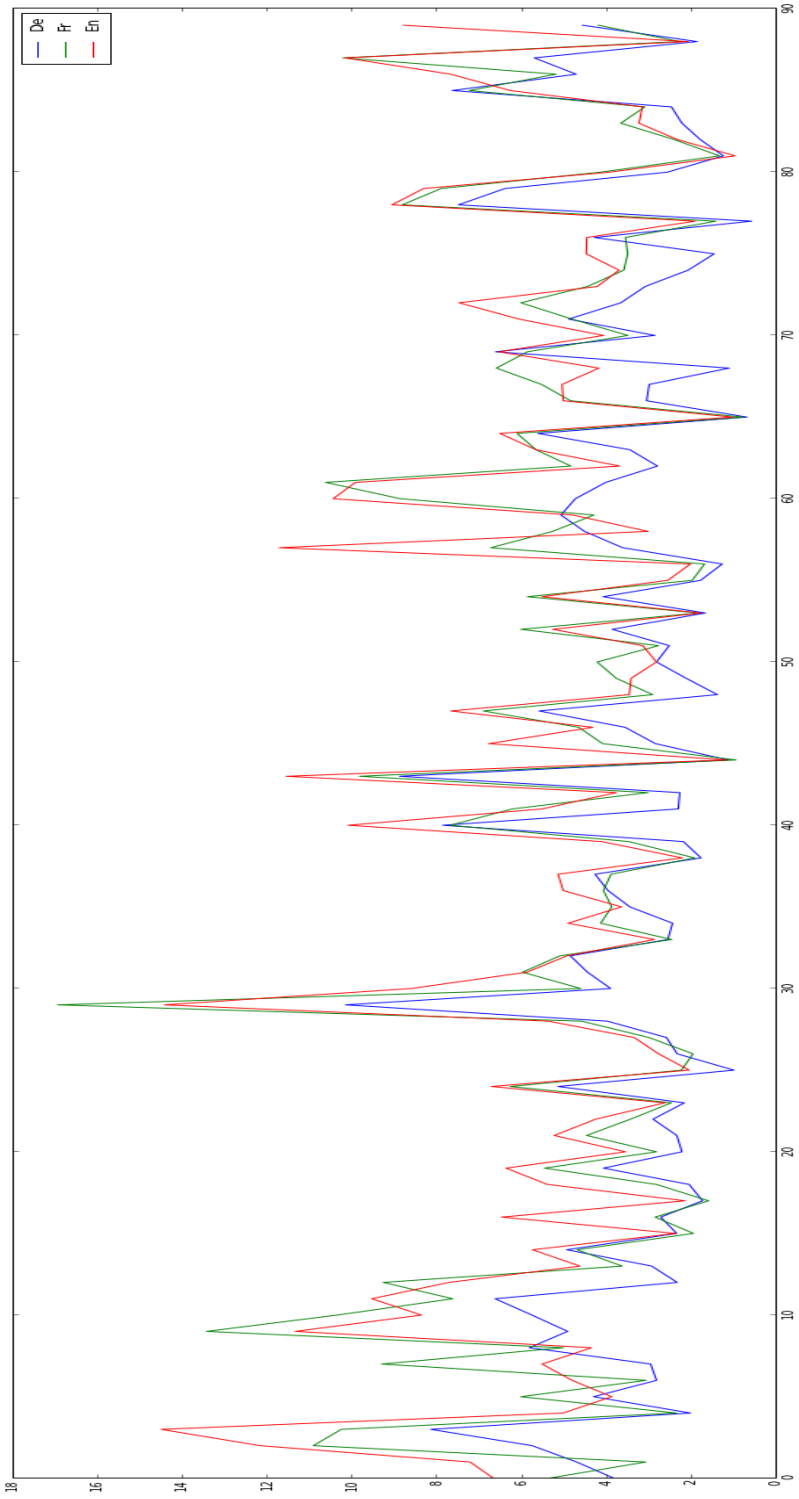

Figure 6.3: Multilingual graph coherence scores for WMT newstest2008 dataset, displaying the score (y-axis) for each document (x-axis)

While the graph profile in Figure 6.3 does follow the same pattern for the three languages, indicative of the nominal pattern across the documents, there is a variation in the scores. The English documents had the largest proportion of high coherence scores, scoring highest more often than French or German. This could be a general characteristic that English involves more coherence as expressed via simple entity-based coherence and that in German coherence is possibly achieved through other means. Lapshinova-Koltunski (2015b) illustrate that languages tend to vary in the way they use lexical coherence and other discourse features.

6.4.1 Source language implications

As mentioned already, it is important for this dataset that we understand what the source language is, and this is marked up on the documents within the WMT data set. This is relevant because it indicates which languages are original texts and which are translations. The first 30 documents are originally Hungarian and Czech (documents 0-29). The subsequent 15 ones are originally French (docs 30-44), the next 15 are Spanish (45-59), the next 15 are English (60-74) then German (75-90). This is interesting, as we can then see patterns emerging of naturally coherent texts. It also means that for a number of documents our French, German and English versions are all translations. One point to note is that ideally this should be extended over an additional corpus of parallel documents, to gain more data, as otherwise we just have 15 texts of each original language. In part of her comparative study, Lapshinova-Koltunski (2015b) used comparable documents instead of parallel ones, which has the benefit of naturally authored texts– as parallel by nature means that one side is a translation. But in not having strictly parallel documents, they are not totally comparable. In the meantime, we can see from Table 6.6 how these affect the scores assigned under this metric. While it is tempting to consider whether having an original German text means that the coherence is higher for German and more evenly scoring in general, or whether an English source text results in less coherence for the German, the number of documents in this preliminary work are very small and therefore not representative. This could be worthwhile pursuing as a corpus study, however.

The problem is in finding a dataset which consists of a large amount of parallel documents, not just parallel data.

	French highest	**English highest**	**German highest**
French original (docs 30-44)	3	8	4
English original (docs 60-74)	6	8	1
German original (docs 75-90)	4	6	5

Table 6.6: Breakdown of highest scoring documents according to the graph metric

Although the projection score is normalised in that the sum of projections is multiplied by 1/N where N is the number of sentences, there is an inevitable bias in favour of longer documents, for example, document 65 in our experiment using the WMT data has only 3 sentences and reads as a coherent one, yet due to the shortness has a low score. Document 29, by comparison, achieves a high score yet reads incoherently - it is originally Czech, and the translation is clumsy in parts. The high score is due to repetition of words like 'millions', 'krona' or 'year' or their equivalent in French and German. The extension by Elsner and Charniak (2011b) for 'unlinkable entities' addresses this particular issue, but of note is also the fact that the range of vocabulary will also influence the entity distribution patterns, as clearly a small range of lexical items which are constantly repeated will lead to a high number of transitions. Lexical choice may vary according to genre or stylistic guidelines, but this should ideally be captured and transferred across languages.

6.5 Conclusions

We observed distinct patterns in a comparative multilingual approach: the probabilities for different types of entity grid transitions varied, and were generally highest in French, lower in German, with English behind the two, indicating a different coherence structure in the three languages. It is clear that entity-based coherence varies from language to language. French may have multiple representations for what would potentially be one entity in English: the use of singular and plural forms of the noun as noticed in French, or adjectival forms

representing the entity. We have also detected differences in implementation due to the compound structure of German; in German while compound nouns affect the coherence score considerably, even with a compound splitter (as for the graph experiment) the coherence score from the graph is still generally lower. The standard format of the grid can therefore usefully be modified for a multilingual context, both to factor out syntactic differences, and to take account of compound words in German. Given that as pointed out already (Poesio et al., 2004), Centering Theory is *per utterance* not *per sentence*, as is the case for the standard entity grid implementation, we would in future choose to implement it *per clause*. However, as we have seen, the entity grid transition information itself is more prone to distortion over sentence breaks, unlike the more flexible graph implementation.

We have seen that the entity graph metric leads to a clear picture of entity-based coherence scores. This is perhaps more useful than the grid for comparative studies. Intuitively, it would seem that this different perspective, i.e. the graph model, offers more insight into crosslingual coherence patterns, in that it captures all the connections between entities throughout the entire document. We can also see better how entity-based coherence is achieved in different languages. Here the exact sentence breaks do not matter so much, and the score is based on how cohesive the document is as a whole.

A pertinent extension to this research includes expanding the graph to include lexical chains, in place of simple entities, or incorporating embeddings, which would allow for crosslingual variance in the semantic coverage of an individual lexical item. As pointed out by Elsner (2011, p.31) the standard monolingual entity grid penalizes coherent transitions between different entities which may be semantically related, and he addresses this with a vector representation and a measurement of lexical similarity. The same applies to the graph. This would potentially better account for the compound structure of German, and the use of singular and plural forms of the noun as noticed in French, or adjectival forms representing the entity. It is valuable to register and identify the differences and bear them in mind for future development, particularly for crosslingual transfer.

We have established how we can track the attentional focus in the source and target text, and what parameters need consideration in future work to measure

the extent to which this is correctly rendered. However MT systems can persist in using a particular translation which is incorrect for a given entity, and a monolingual entity graph or grid cannot capture this. To address this issue, we need to integrate some semantics and to condition on the ST. The notion of semantics has already been by raised by Poesio et al. (2004), who mentions that "an analysis in terms of underlying semantic connections between events or propositions is more perspicuous than one in terms of entity coherence" (p.354). In fact, the extension proposed by Elsner (2011) integrates the probability of a word given a particular topic. This could be repurposed to also condition on the ST. The challenge from an MT point of view would be to ensure that the equivalences are maintained, so an entity chain is carried over from source to target text, despite differences in syntax and sentence structure. However, even this is insufficient to ensure that the document is fully coherent – more linguistically based elements are necessary to do that.

In the next chapter (Chapter 7), we describe integration and evaluation of our models, and present the challenges we see for MT evaluation: we consider the advantages of integrating communicative intent and semantics, while conditioning on the source text instead of the reference.

Chapter 7

Coherence Models to Improve MT

As previously mentioned, we believe that a coherent discourse should have a context and a focus, be characterised by appropriate coherence relations, and be structured in a logical manner. This is what we have aimed to evaluate with our coherence models in Chapters 3-4. As described in Section 1.4.3, integration and evaluation are other ways we envisage to evaluate our models in an MT context. We detail our efforts in integrating and evaluating our work (in Sections 7.1 and 7.2), and the obstacles we have found. We then revisit the goal of translation (Section 7.3), in addition to examining how this can be achieved in MT (Section 7.4) and setting out what we see as the way forward (Section 7.5).

7.1 Integration

Document level decoder In order to integrate discourse features into an SMT framework, we need a bigger context window than would be the case for current feature functions in a standard PBMT system (see modelling in Section 1.4.3), particularly for document level features. We planned to integrate our coherence discourse features via a document-level decoder. We explored this option with Docent (Hardmeier et al., 2013a), which is a document-level decoder that has a representation of a complete TT translation, to which changes can be made to

improve it. It uses a multi-pass decoding approach, where the output of a baseline decoder is modified by a small set of extensible operations (e.g. replacement of phrases), which can take into account document-wide information, while making the decoding process computationally feasible. To date, attempts to influence document-level discourse in SMT in this manner have been limited.

Docent includes a *proposal component*, which proposes improvements, and a *scoring component*, which determines which ones are accepted. The initial state is improved by a hill-climbing decoding algorithm, whereby a new state is generated by non-deterministically applying one of a small set of operations that randomly replace/delete/add phrases. If it meets the necessary criteria, determined by the *scoring component*, it is accepted as the new state. The operations make changes to a single sentence at a time. The operations in Docent are stochastic, and by themselves are unlikely to influence document-level coherence for our purposes. Existing scoring components in Docent use standard SMT optimisation metrics, such as BLEU, which are unlikely to capture changes at discourse level. We tried integrating features from our entity graph model to explore this option but initial results were that it seemed ineffective. Our intuition was that in order to integrate document-level coherence features, the changes would need to be systematic and would require writing a new operation. In terms of our models, while there could be a lexical change proposed and accepted for one sentence in the document, it would not uniformly impose a consistent lexical term, for instance, unless by writing a new operation. Also the performance implications of the proposed operations are such that it is prohibitive to integrate linguistic operations such as parsing without adapting them further (by means, say, of annotating items in the phrase table via supertags, such as those in Birch et al. (2007)). This was not judged a promising path to explore.

A similar document level framework was very recently developed by Martìnez Garcia et al. (2017), who also introduced a new operation to ensure that changes could be made to the entire document in one step (see Section 2.1.1). They found in their recent substantial and innovative research that automatic metrics *"are mostly insensitive to the changes introduced by our document-based MT system"*, despite human annotators preferring the translations from their new operation 60% of the time, with an additional 20% where their preference was

the same (Martìnez Garcia et al., 2017). This seems a clear illustration that the evaluation process is flawed. We further discuss this question in Section 7.5.

Given that automatic metrics are deployed to determine the weights of the features during the tuning process, discourse features such as those of Martìnez Garcia et al. (2017) that do not impact the score will not be weighted highly. In general, features integrated in an SMT system to attempt to directly impact coherence are therefore unlikely to have much effect.

Constraints in SMT The popularity of SMT in the past couple of decades has largely been to the exclusion of deeper linguistic elements. Performance of SMT systems surpassed previous rules-based systems, and progress was described by the famous quote by Frederick Jelinek :"Every time I fire a linguist, the performance of the speech recognizer goes up". This dominance of SMT was detrimental to the exploration of many linguistic elements. As reported by Hardmeier (2015), "the development of new methods in SMT is usually driven by considerations of technical feasibility rather than linguistic theory". As described already in Section 1.3, most decoders work on a sentence by sentence basis, isolated from context, due to both modelling and computational complexity. This directly impacts the extent to which discourse can be integrated. Considerable progress has been made in the field of SMT, culminating in models which yield surprisingly good output given the limited amount of crosslingual information they have. While SMT comprises a complex and finely tuned system, it is linguistically impoverished, superficially concatenating phrases which have previously been found to align with those of another language when training, with no reference to the intended meaning in context. The more recent NMT approach has been proven to capture elements of context (syntactic and semantic), which are now helping to make NMT output more fluent than that of SMT (Bojar et al., 2016). However, these elements are not modelled explicitly via any linguistic theory. As a result, it can give rise to considerable semantic errors.

In the past, all of these constraints in SMT have restricted integration of linguistic elements and hindered progress to another level. With the success of NMT and the significant paradigm change it brings, much more context can potentially be integrated– but the risk is that we do not embrace this opportunity

to advance to a deeper linguistic level of translation. As illustrated by recent comparative research into output from PBMT and NMT systems (Popović, 2017; Burchardt et al., 2017), the latter is capable of producing output which is far more fluent. At the same time, MT is increasingly being used in pipelines for other tasks, such as speech translation (Waibel and Fugen, 2008). In order to fulfil its role, MT needs to capture and transfer the communicative intent of the ST into the TT. We believe it is worth revisiting the basics of translation theory (see Section 7.3) to establish the purpose of MT, with a view to taking MT to a level where it can better fit requirement.

7.2 Evaluation

The three types of models we experimented with in Chapters 3-4 capture different elements of coherence. While we expect their scores to be complementary, some MT systems may well do better at some aspects and not so well at others, so will score differently under the separate models. We illustrate how the different systems from the WMT14 submissions (Bojar et al., 2014) score under the different types of models, including scores for the Entity-graph (as the best performing entity one), our IBM1-Syntax model (as the best performing syntax one, albeit not actually measuring *intent*, as originally foreseen), and Dis-Score (our crosslingual discourse metric). This is clearly illustrated in Figure 7.1, where we visualize the scores per model of each system, summing over the 175 documents in that WMT14 submissions test set, and with all scores scaled to fall between 0 and 1 (inclusive).

As can be seen from the raw scores from the coherence models (Table 7.1), they clearly measure different aspects and result in different rankings for the submitted system outputs. Some systems will handle lexical coherence better than others, while some may better capture and transfer discourse connectives.

From the evaluation metric submissions for WMT14 (Macháček and Bojar, 2014), we also display the results of DiscoTK-party and REDSys system level scores by way of comparison (Table 7.1), as the two metrics with the best correlation to human rankings for that language pair (fr-en), and therefore judged the best performing. There is variation over the score rankings, particularly for the

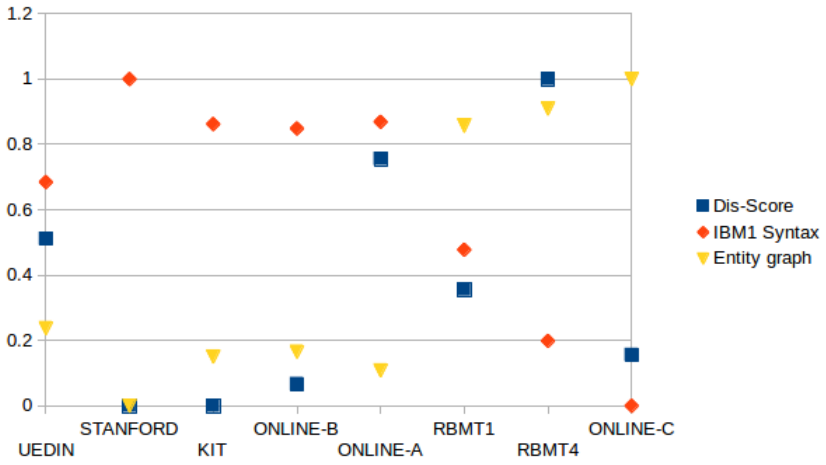

Figure 7.1: Comparative scores for the WMT system submissions under our different coherence models.

Entity-graph and Dis-Score, but less so for our IBM1-Syntax model. The most obvious observation to be made from these results, is that it would appear that the rules-based systems (rbmt1 and rbmt4) are scored higher under our entity and crosslingual discourse relational models (moving from the lower half to upper half of the scoreboard), which may well be due to the fact that these systems work at a higher level than the other systems. We see this as indicative of the fact that possibly some of the strengths of these more linguistic models have perhaps been overlooked by current methods of evaluation. We further illustrate the differences by visualizing the scores under the different models - shown in the Kernel Density Plots in Figures 7.2-7.4. In these plots we see how the distribution over scores varies for the eight Fr-En WMT system submissions from model to model. While there is variation, and the leading systems can be identified, the profiles are remarkably similar.

Correlation with human judgements and current metrics Our models are only measuring *aspects of coherence*, and are insufficient as a standalone metric given that there are other issues which need evaluated to judge the accuracy, fluency and grammatical correctness of a translated document. Moreover, in fact the human judgements on WMT are not themselves at document level, and so are not therefore directly comparable. Human evaluation at WMT is on a window of

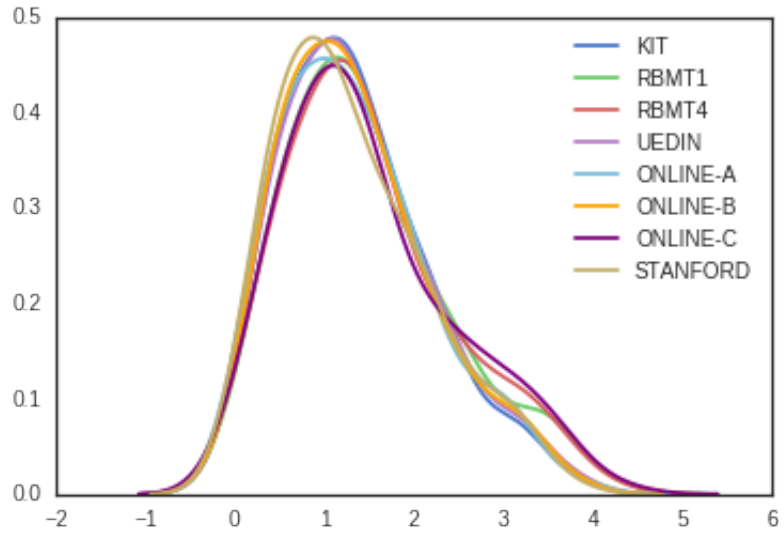

Figure 7.2: Distribution of scores for the WMT system submissions under our Entity Graph metric.

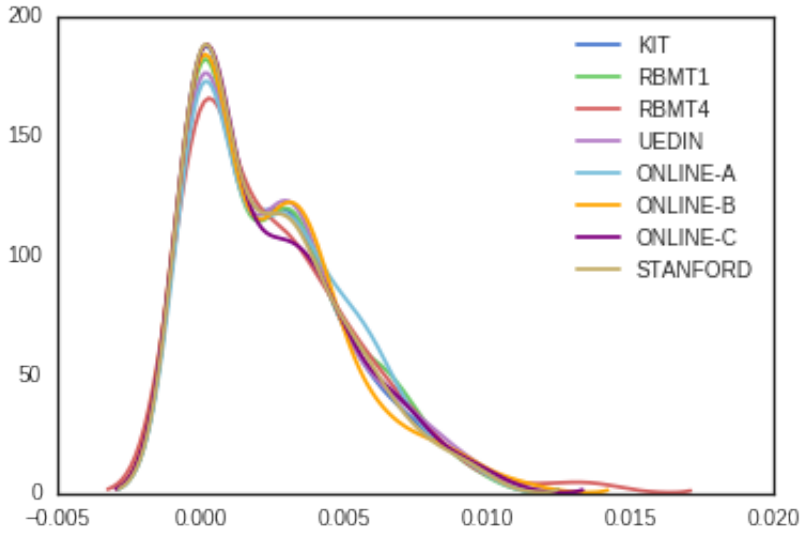

Figure 7.3: Distribution of scores for the WMT system submissions under our Dis-Score metric.

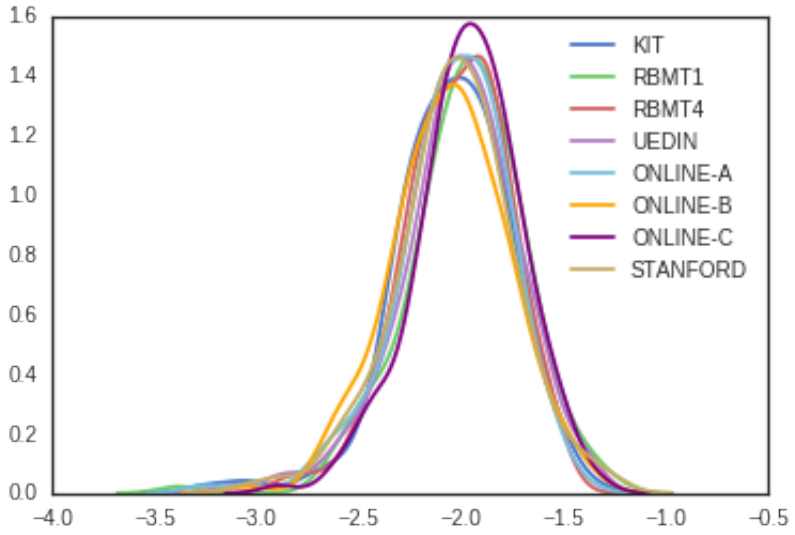

Figure 7.4: Distribution of scores for the WMT system submissions under our IBM1-Syntax model.

a couple of source sentences, with no target translation context, and therefore do not give credit to models which overall may have a more consistent or coherent output at document level (we continue this point in Section 7.5).

We report the correlations with human judgements in Table 7.2. As mentioned previously, we sum over the 175 documents in that WMT14 submissions test set, and scale all scores to fall between 0 and 1 (inclusive). We cannot expect a high correlation between our models alone and human rankings, because we are only capturing certain aspects of coherence, and not other measures of adequacy and correctness. Moreover the human assessors have not been asked to directly account for coherence in their sentence-level rankings. The results from the IBM1-Syntax model correlate very well with human rankings (0.941) whereas those of Entity-graph do so very poorly (-0.933). This may well be related to the fact that the human judgements are sentence-level, whereas the Entity-graph considers the pattern of entities in the document as a whole. Dis-Score correlations were discussed in Chapter 4.

To see whether our metrics are productive, as judged in terms of whether they are complementary to other metrics, we combine them linearly with the DiscoTK-party and tBleu metrics. The DiscoTK variations metrics are

System	Dis-Score	Syntax	Entity	Human	Disco	REDSys
UEDIN	0.437 (3)	-1623.66 (5)	238.97 (4)	1	0.829	0.0174
STANFORD	0.414 (7)	-1598.66 (1)	231.05 (8)	2	0.768	0.0171
KIT	0.414 (7)	-1609.59 (3)	236.07 (6)	2	0.756	0.0171
ONLINE-B	0.417 (6)	-1610.66 (4)	236.57 (5)	2	0.738	0.0172
ONLINE-A	0.448 (2)	-1609.05 (2)	234.66 (7)	3	0.651	0.0169
RBMT1	0.430 (4)	-1640.10 (6)	259.65 (3)	4	0.200	0.0153
RBMT4	0.459 (1)	-1662.24 (7)	261.40 (2)	5	0.013	0.0147
ONLINE-C	0.421 (5)	-1677.99 (8)	264.41 (1)	6	-0.063	0.0144

Table 7.1: Human ranking of 2014 WMT MT system submissions compared to raw scores from coherence models and top WMT14 metric rankings. Disco here is DiscoTK-party-tuned.

based on discourse structure, where the DiscoTK-party includes other phenomenon (for extended description refer to Chapter 4). tBleu, on the other hand, is an advanced BLEU metric and is therefore based on ngram matches, with the *t* signifying that it is more tolerant and results in higher correlation with human judgement (Libovický and Pecina, 2014). As such it is interesting to see whether our metric is complementary to an ngram matching one.

As already discovered in Chapter 4, combining our Dis-Score to the DiscoTK-party metric increases the correlation directly (see DiscoTK-party+DisScore). Looking at the correlation of IBM1-Syntax with DiscoTK-party, it increases the correlation from 0.970 to 0.973, even if it does not directly measure *intentional structure* to any significant degree.. Clearly our models are of benefit in that they are capturing useful information which can complement even the metrics which already include some discourse information. Particularly interesting is the increased correlation when combining DisScore with tBleu, an increase from 0.952 to 0.963.

7.3 Translation as communication

Human evaluation is judged by WMT as the best way of evaluating the performance of MT systems. However for them to be able to properly perform their task, the humans acting as judges need to be aware of the remit of that task.

Translation theory has evolved over the years, from the functional and dy-

Metric	fr-en
Dis-Score	-0.213 (0.012 or 0.263)
Entity graph	-0.933
IBM1 syntax	0.941
DiscoTK-party	0.970
DiscoTK-party+DisScore	**0.975**
DiscoTK-party+Entity graph (scaled)	0.228
DiscoTK-party+IBM1 syntax (scaled)	**0.973**
tBLEU	0.952
tBLEU+Dis-Score	**0.963**
tBLEU+Entity (both scaled)	0.360
tBLEU+IBM1 syntax (both scaled)	0.566

Table 7.2: Results on WMT14 at system level: Pearson correlation with human judgements. Our metrics alone, and in linear combination with DiscoTK-party 2014 WMT submissions. For Dis-Score we report the segment level correlation in brackets– Kendall's τ variants for WMT14 and WMT13. The reported empirical confidence intervals of system level correlations were obtained through bootstrap resampling of 1000 samples (confidence level of 95 %) (Macháček and Bojar, 2014).

namic equivalence of Nida and Taber (1969), to Baker (1992)'s view of equivalence (word, grammatical, textual, pragmatic equivalence), Hatim and Mason (1990)'s view of the translator as a communicator and mediator, and Relevance theory applied to translation (Gutt, 1989).[1] Nowadays there is a broad essential agreement on the importance of discourse analysis: on the need to extract the communicative intent and transfer it to the target language- in an appropriate manner, taking account of the cultural context and the genre.

While there is now a great need for translation, which cannot be met by humans (in terms of the cost or number of human translators), MT can be usefully deployed for gisting, and for some language pairs even as a good quality first draft. However if it is to be more, for example if it is to be used as part of a pipeline for a series of tasks, then it needs to embrace its role in terms of *communicative intent*. Used in pipelines such as voice translators, where Speech Acts are relevant, or as vital components of a multimodal framework, we cannot ignore the fact that the communicative intent is currently not a core building block in MT.

[1]Cognitive Linguistics is a further development which is beyond this work

Translation inherently involves communication. As has been said by others previously (Becher, 2011), MT could benefit from mimicking the way a human translator works. Translators makes several passes on a text. They begin by reading the ST and extracting the communicative intent– establishing what the author of the text is trying to say. They identify any cultural references and any acronyms or terminology relevant to the domain. For the former, they need to be aware of the significance of the references and their connotations. They then attempt to transfer these in an appropriate manner to the TT, taking account of their TT audience. While MT is far from this and of necessity some of these tasks are done at training time, it has to at least begin to grapple with semantics, if it is to perform a meaningful role.

7.4 Semantics

In terms of proposing how this might look for evaluation purposes, we would suggest that semantic parsing may offer one way forward. While this is not available in many languages, and may start off as a limited evaluation method, there are ways in which this can be done.

Progress in the field of semantics has been considerable recently, and in particular work based on Universal Dependencies (UD)[1] would seem to offer new opportunities which MT evaluation could benefit from: UD are annotations of multilingual treebanks which have been built to ensure crosslingual compatibility. The latest version (2.0) covers 50 languages. Recent work by (Reddy et al., 2016) to build on this and transform dependency parses into logical forms (for English) opens up opportunities for crosslingual semantic parsing. While still a field in development, it is one option to be explored if we want to evaluate the semantic transfer in MT. We could foresee that initially at least it could be achieved by developing text cases (see Section 7.5) on the back of annotations, ensuring that the basic semantics of a sentence in one language (the ST) matches that of another (the TT). While ultimately this requires the MT to be of a good standard for parsing, in the case of NMT with a good language pair, this is now the case, and indeed has to be for any meaningful attempt to integrate discourse. The

[1] http://universaldependencies.org/

existence of semantic parsers has now opened the way for metrics which provide an automatic semantic evaluation of MT output (Lo, 2017), albeit based on the reference in this case.

In the short term, test cases can be devised that do not involve a parser, merely test the ability of a system to effect semantic transfer. Reddy et al. (2017) give a concrete example using their semantic interface based on UD for a multilingual question-answering experiment, where they generate ungrounded logical forms for several languages in parallel and map these to Freebase parses which they use for answering a set of standard questions (translated for German and Spanish). They simplify to ensure crosslingual compatibility, but essentially illustrate how semantic parsing can work crosslingually. For an indepth explanation of the process, see Reddy et al. (2017).

Using these as a test bed and running against WMT systems as additional evaluation could be very useful, perhaps indicating which systems are more capable of capturing and translating the *meaning* of the source. In the long run, ideally the aim of MT is to capture the meaning of the ST, and then based on that generate the TT (a kind of concept-to-text-generation). This would of course involve a shift in paradigm for MT.

In practice, however, this lack of semantics is a problematic issue: as the MT researchers from Booking.com describe in their work on NMT in the real world, mistranslations which may seem insignificant, can be hugely problematic in a business scenario, such as the difference between 'free parking' and 'parking is available'. In their experiment, they used professional translations to judge the translations, based on adequacy and fluency, and introduce additional rules to address the lack of 'sentence *meaning*' Levin et al. (2017).

7.5 Beyond reference-based evaluation of MT output

Hardmeier (2012) already touches on the problem of current automatic evaluation methods. In particular, he mentions the shortcomings of ngram-based metrics and the issue of sentence level evaluation, when in fact much of discourse is document

level: *"However, it could be argued that the metric evaluation in the shared task itself was biased since the document-level human scores evaluated against were approximated by averaging human judgments of sentences seen out of context, so it is unclear to what extent the evaluation of a document-level score can be trusted."* The problems with BLEU are well known already, and are also interestingly illustrated in research by Smith et al. (2016), proving that optimizing by BLEU scores can actually lead to a drop in quality. Another major problem is the fact that the evaluation of MT output is still largely based on comparison to a single reference or gold standard translation. A reference, or gold standard translation, is *one* version. A text can be translated in *many* ways, all of which will reflect the translator's interpretation of what the ST is saying. To constrain the measure of correctness to a single reference is only consulting *one* interpretation of the ST. There could be equally good (or better) examples of MT output which are not being scored as highly as they should, simply because they employ a different lexical choice. While in some scenarios evaluation is based on multiple references, this is rare, and costly.

Moreover, reference-free evaluation is valuable for other reasons: MT is also being used extensively online, where no direct assessment is feasible due to the lack of a reference translation. This poses a real problem, as illustrated recently when Facebook had to issue an apology over a mistranslation which had led to someone's arrest [1].

The field of Natural Language Generation (NLG) has a similar problem, and researchers are reaching a similar conclusion, as is clear from Novikova et al. (2017) where they describe their work as 'a first step towards reference-less evaluation for NLG by introducing grammar-based metrics'. And another paper where they 'investigate a reference-less quality estimation approach (..) which predicts a quality score for a NLG system output by comparing it to the source meaning representation only' (Dusek et al., 2017). The context is slightly different but the words *source meaning* are relevant to both MT and NLG– and the logical groundtruth in evaluation.

Recently, there has even been a trend towards totally ignoring the ST during

[1] https://www.theguardian.com/technology/2017/oct/24/facebook-palestine-israel-translates-good-morning-attack-them-arrest

evaluation of WMT submissions, where *'human assessors are asked to rate a given translation by how adequately it expresses the meaning of the corresponding reference translation'* (Bojar et al., 2016). So human assessors are asked to rate a given translation by how close it is to the *reference* translation, with no regard to the *source* text. The process is treated as a monolingual direct assessment of translation fluency and adequacy. We would argue that self-evidently adequacy should be based on how well the meaning of the ST has been transferred to the TT, and that to ignore the ST (simply relying on the one rendering of it) is to lose that direct dependency, whereas a proper measure of adequacy is whether the translation captures and transfers the communicative intent from ST to TT.

Moreover, the human assessment of the output has recently become *'researcher based judgments only'*– which is also problematic, in that the researchers in question are not generally trained in translation, and some are monolingual. This means that they will not necessarily capture nuances, ambiguity in the source, or discourse information, such as the implicit discourse relations of the reference translation, for example, and know to look for them in the MT output. Not knowing the source language means that they cannot assess the correctness of the output if it alters from the reference.

Moving forward As mentioned by Guzmán et al. (2014), 'there is a consensus in the MT community that more discourse-aware metrics need to be proposed for this area to move forward'. Both Popović (2017) and Burchardt et al. (2017) directly or indirectly touch on the issue of evaluation. As part of her analysis Popović (2017) attempts to classify the type of errors made by each system. Burchardt et al. (2017) introduces a test suite which, while it is common and invaluable in software engineering, is not widespread for this domain and is a most constructive development. With the suite of tests they aim to cover different phenomena, and how the systems handle them, asserting they aim to focus on new insights rather than how well the systems match the reference (Burchardt et al., 2017).

In the past there have been examples of unit testing for evaluation of MT quality, in particular King and Falkedal (1990) who developed theirs for evaluation of different MT systems themselves (before financial commitment to a specific

one). Nevertheless a substantial amount of the logic (behind using test suites for evaluation) is still valid: evaluating the strengths and weaknesses of output from various MT systems, with tests focussing on specific aspects (syntactic, lexical ambiguity etc) for particular language pairs.

In a more general vein, Lehmann et al. (1996) develop test suites for NLP in their Test Suites For Natural Language Processing work, which are intended for the general evaluation of NLP systems. Their test suites aimed to be reusable, focused on particular phenomena and consisting of a database which could identify test items covering specific phenomena. Similarly, the MT community could potentially develop relevant tests, with agreement on format and peer reviews.

This type of method could easily be adopted as a means of evaluation in the context of WMT tasks, and besides being much more informative, would help to pinpoint strengths and weaknesses, leading to more focussed progress. Existing test suites, such as the ones developed by Guillou and Hardmeier (2016) and Loáiciga and Gulordava (2016), could be integrated and added to, giving a more comprehensive and linguistically-based evaluation of system submissions. Unit tests could be added to by interested parties, with peer reviewing if appropriate. The resulting suite could eventually cover a whole range of discourse aspects, and an indication therefore of how different systems perform, and places where there is work to be done. The concept is not new and could build on previous initiatives and experience, such as Hovy et al. (2002) to ensure it is adaptable yet robust, providing a baseline for progress in particular aspects of discourse.

In terms of evaluation in training, one novel idea is the use of post-edits in evaluation (Popović et al., 2016)– this can be seen as more informative and reliable feedback, if done by a human translator, and can be directly used to improve the system.

7.6 Conclusions

In this chapter we considered the scores of our coherence models for the WMT submissions, illustrating the different phenomena they capture. We also showed the correlation between our scores and human rankings, as is standard in this domain. While we did not expect a high correlation as stand-alone metrics, com-

bining them with other metrics led to an increased correlation overall (markedly in some cases), indicating that they are nevertheless capturing valuable information. However we cannot ignore the fact that the SMT architecture does not make integration easy, and current automatic metrics do not value discourse information.

Considerable progress was made in the field of MT over the past two decades, culminating in models which give surprisingly good output given the limited amount of crosslingual information they have. While those were the models which were best-performing at the start of this research, NMT models are now the most performant, to the extent that in the past year they have been the best performing at WMT (Bojar et al., 2016), and although deeper than the linguistically superficial SMT, to evaluate progress we need to be able to measure the extent to which these models successfully integrate discourse.

There are numerous difficulties with evaluation of discourse phenomena, particularly if it is automatic. But the potential advantages of progressing beyond single reference-based evaluation are considerable– not least the ability to evaluate without first commissioning a reference translation each time. At a time when MT is being used in a pipeline in which dialogue acts play an important role, it is vital that evaluation of MT be based on something more substantial than string matching to a single reference, or judgements made without regard for ST. Once MT begins to integrate an element of semantics, which would tackle the issue of reference-free evaluation, it no longer makes sense to evaluate on a single reference. While the translator's role as mediator will not easily be replaced by machines– as yet it cannot capture the pragmatics or recreate the contextual richness for the target audience– nevertheless we must ensure we assess MT output based on a measure of *adequacy* compared to the *source*, if it is to fulfil its communicative intent. This applies to human and automatic evaluation.

Chapter 8

Conclusions

In this chapter we summarize the work we have covered in the thesis in Section 8.1. We then evaluate the extent to which we have realised our aims (Section 8.2), before mentioning some possible expansions for the future (Section 8.3).

8.1 Coherence in MT

Recently increasing amounts of effort have been going into addressing discourse explicitly in MT, as detailed in Chapter 2, resulting in a wealth of research on various aspects of discourse. Our contribution to this has been in the domain of coherence. Given that coherence is multifaceted, we elected to start our research on three dimensions of attentional structure, intentional structure and linguistic structure (Grosz and Sidner, 1986). We investigated whether by combining these components in an MT context we can adapt the models and advance the assessment of coherence in MT.

Work on measuring text coherence has thus far been monolingual and commonly limited to somewhat artificial scenarios, such as sentence shuffling tasks. These operations naturally tend to break the overall logic of the text. In Chapter 3, we investigated existing local coherence models for a very different scenario, evaluating machine translated texts. The quality of the output varies between the different MT systems. Coherence in this scenario is much more nuanced, as elements of coherence are often present in the translations to some degree,

and their absence may be connected to various types of translation errors, at different linguistic levels. We illustrated in this chapter that it is a much more difficult task than the traditional monolingual one, particularly in the case where the latter comprises coherent sentences which are shuffled (the case of automatic summarization is harder, but the original text is coherent in the first place). Our experiments in this chapter highlight how current artificial tasks are overly simplistic for measuring coherence. We further improve on the state-of-the-art syntax model, and apply current models for the evaluation of MT output for the first time.

In Chapter 4 we developed a model to evaluate the crosslingual transfer of discourse relations in MT, in terms of connectives as cues signalling the discourse relations, and assessing the subsequent semantic transfer of the discourse relation. We used crosslingual word embeddings pretrained for multiword discourse connectives, and incorporated discourse relation mappings between source and target text. This work is innovative in that it is based on the source text and the extent to which this has been appropriately transferred to the target. This has not been researched before.

We examined the different types of incoherence errors which occur in MT output for different language pairs in Chapter 5. Faced with the lack of labelled data (where negative examples are labelled for instances of incoherence) we attempted to create a corpus of artificially generated incoherent data. While not error-free, this provides an automatic and systematic way of creating a corpus with some of the errors encountered and resulted in a corpus suitable for final human annotation. This represents an innovation which has not been used previously, and shows areas for further work.

In Chapter 6, we observed clear trends in our comparative experiments where we applied the entity grid and entity graph to a multilingual setting, which has not been attempted before (as far as we are aware) . We concluded that entity-based coherence varies from language to language (including factors such as the compound structure of German, and the use of singular and plural forms of the noun, or adjectival forms representing the entity, as noticed in French). We noted that with adaptations we can see comparative patterns throughout the entire document where an entity chain is carried over from source to target text

(in our study of parallel, human-translated texts). The graph leads to a clear picture of entity-based coherence which is perhaps more useful than the grid for comparative studies and offers more insight into crosslingual coherence patterns, important to consider for crosslingual transfer.

In Chapter 7 we highlight the fact that as with other research on discourse phenomena in MT, there are issues preventing integration and evaluation of new insights. While those of integration have traditionally been the harder to solve, they may well be improved by the move to a new paradigm (NMT), whereas those of evaluation will need concerted effort to do so, although they are not computationally difficult. There is little incentive to integrate much of the research on discourse phenomena into an MT system while evaluation remains reference-based, and does not consider document level issues, nor adequacy of translation in terms of communicative intent.

At a time when MT is increasingly being used in a pipeline for other tasks, the communicative intent of the translation process needs to be properly integrated into the task. Moreover, in order to take MT to another level, it will need to judge output not based on a single reference translation, but based on notions of fluency and of adequacy – ideally with reference to the source text.

8.2 Evaluation of aims

We have proven that measuring coherence in MT is harder than the traditional task of reordering shuffled texts (A_1) and have improved on state-of-the-art syntax model (A_1). We have created the first crosslingual discourse relation model for assessing the translation of discourse relations in an MT context (A_2). Our new Dis-Score metric evaluates against source, not against a single reference translation (A_4). We have also analysed manifestations of incoherence across different language pairs and created a corpus as further method of evaluation (A_3). We have applied an entity-based grid and graph in a multilingual scenario for the first time, examining lexical coherence in a multilingual context (A_3).

8.3 Future work

We present the ways in which this work could be extended in future:

- We could extend the Dis-Score metric to cover inter-sentential relations, beyond the intra-sentential ones it currently covers. This would give a greater degree of structure, particular for document level MT. This could be achieved either by means of a discourse parser on the ST as well as the TT, or by building on top of work being done in the domain of crosslingual discourse parsing (Braud et al., 2017). Depending on the robustness of the approach, we could envisage moving beyond using a tagger which simply identifies the four top-level PDTB relations to using a more detailed level of discourse relations. An obvious extension is to include implicit discourse relations in the analysis, identifying discourse relations which are not explicitly signalled via a lexical cue, as our work in Chapter 4 covered only explicit discourse relations.

- The next step for the entity graph model would be adapting it to be more robust, taking account of related (and coherent) but non-identical entities, via embeddings or vectors, and potentially conditioned on topic relatedness, as per Elsner (2011). We could extend his implementation to also condition it on the source, building on work from Chapter 6. This would mean that it could measure the extent to which lexical coherence is transferred from source to target. This requires more research, including exploring crosslingual semantic relatedness via lexical chains or embeddings, perhaps supported with an extensive corpus study. It would also need to address the question of whether the context of the embeddings and the topic relatedness are sufficient to disambiguate between lexical items which have multiple meanings in a particular language, to address the problem whereby MT consistently uses the same, but the *wrong* translation.

- Ultimately, the lexical coherence of the graph model could be productively integrated with the crosslingual discourse relations of Dis-Score, to give a more comprehensive evaluation of coherence. This would be in line with the interrelated components of the discourse structure described by Grosz

and Sidner (1986), and is also supported in the point made by Poesio et al. (2004, p.354) regarding supplementing 'entity' coherence with 'relational' coherence.

- We could package the Dis-Score metric as open source resource, in order to allow others to use it. It could also be expanded beyond the French-English language pair it currently covers- which would become easier as discourse parsers become available for a larger number of languages. One idea also worth exploring would be whether ngram2vec (Zhao et al., 2017) embeddings would be able to capture multiword embeddings, saving the effort involved in identifying, pre-hyphenating and training the multiword discourse embeddings as we did.

- While perhaps not measuring intentional structure to any significant degree, our extended IBM1 model (Chapter 3) was able to discriminate between syntax patterns in MT and HT output. It also had an interestingly close correlation with human judgements (Chapter 7). This merits further investigation, as it could at least help move MT output towards more natural syntactic structures.

Appendix A

A.1 Publications

- Sim Smith, K. and Specia, L. (2017). Assessing Crosslingual Discourse Relations in Machine Translation. *To appear*

- Sim Smith, K. (2017). On integrating discourse in machine translation. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, Copenhagen, Denmark. Association for Computational Linguistics.

- Sim Smith, K., Aziz, W., and Specia, L. (2016a). Cohere: A toolkit for local coherence. In Chair), N. C. C., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

- Sim Smith, K., Aziz, W., and Specia, L. (2016b). The trouble with machine translation coherence. In *Proceedings of the 19th annual conference of the European Association for Machine Translation (EAMT)*, pages 178–189, Riga, Latvia.

- Sim Smith, K. and Specia, L. (2016). *New perspectives on cohesion and coherence: Implications for Translation*. Translation and Multilingual Natural Language Processing. Language Science Press, Berlin.

- Sim Smith, K., Aziz, W., and Specia, L. (2015). A proposal for a coherence

corpus in machine translation. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 52–58, Lisbon, Portugal. Association for Computational Linguistics.

- Steele, D., Sim Smith, K., and Specia, L. (2015), Sheffield Systems for the Finnish-English WMT Translation Task. In *Tenth Workshop on Statistical Machine Translation*, pages 172–176, Lisbon, Portugal. Association for Computational Linguistics.

A.2 Extracts for 1.1

MT output:

We want cheap goods and was surprised that manufacturers produce in "industrial" it highlights super u but all the signs, and even the small shops are concerned. How to eat good, organic with 1,650 (average wage in france)? But to answer your post, if it is possible to eat properly with a small wage. After everyone is free to set its priorities where it sees fit. For my part, the food is one.

French source:

On veut des produits peu chers et on stonne que les industriels produisent en industriels On met en avant super u mais toutes les enseignes et mme le petit commerce sont concerns Comment bouffer bon, bio avac 1650 (salaire moyen en france)? Mais pour rpondre votre post, si cest possible de manger correctement avec un petit salaire. Aprs chacun est libre de mettre ses priorits l o il lentend. Pour ma part lalimentation en est une.

Reference:

We want cheap products and we are surprised that the factory-produced ones are produced in factories. We put super u forward, but all brands and even small-scale trade are concerned. How to eat well, bio with 1,650 (average salary in France)? But to reply to your post, it is possible to eat well with a small salary. Afterwards, everyone is free to set their priorities where they like. For me food is one of them.

A.3 Extracts

A.3.1 Translation output from onlineC.0

The matter NSA underlines the total absence of debates on the piece of information

How the contradictory attitude to explain of the French government, that of a quotation offends itself in public while summoning the ambassador of the United States October 21, and other forbids the flying over of the territory by the bolivian presidential airplane, on the basis of the rumor of the presence to his edge of Edward Snowden? According to me, there are two levels of response from the French government. When François Holland telephones Barack Obama or when the minister of the foreign affairs Laurent Fabius summons the ambassador of the United States, they react to a true discovery, that is the one of the extent of the American supervision on the body of the communications in France. Not is it surprising to read in the columns of the World to some weeks of interval on one hand the reproduction of the American diplomatic correspondence and on the other hand a condemnation of the listen Quay of Orsay by the NSA? Not there would be as a vague hypocrisy on your part? The journalistic gait is not a moral positionnement, but the research of the interest and relevance of information that allow every citizen to forge itself an opinion. When raised WikiLeaks the sail on the analysis by the American diplomacy of political or other issues the entire world, we consider in fact that, to the look of the American power, that constitutes an important lighting. When we describe the American systems of interception in opposition to the French diplomacy to the United States, this is not in any case to outrage us of this practice, is to describe the world such as it is. Did France benefit from information furnished by the concerning NSA of the terrorist operations aiming our interests? Can one to deprive oneself American collaboration? Set it up since wholesale ten years of technological tools of interception very powerful by the United States, but also by France, officially was justified by the fight against the terrorism. Besides, in this domain, France and the United States notably set up of the procedures of cooperation and of exchanges of almost daily information and that are described party and of other

as essential. By way of example, the presence of Mohammed Merah in the tribal zones to Miranshah was signaled to the French thanks to the means of the NSA. France can be driven, for example, to transmit entire pads of given ones on the region of the Sahel to the American services, and, in compensation - one the already quickly said -, the Americans can give the news to the French on of other regions of the world. Therefore the bottom question behind this matter NSA is not so the capacity or the right of the countries to endow itself with interception tools, that the question of the total absence of debates previous, notably within the Parliaments, on the justification of such systems, the perimeter that must be it them, and, in the last analysis, the question of the attained to liberties. What do risk actually the United States? a degradation of their picture? One has beautiful to denounce them, I do not see any which manner they will be able to be punished. The risk run by the Americans can be doubles. The first one, it is when their allies - and that was the case recently - learn that their leaders, sometimes to the highest summit of their State, were overseen. This is the case of Brazil and Germany, two countries where diplomatic relations with the United States stretched themselves. Another effect can be him more economical: more and more of European or South American businesses grumble, to the light of the revelations, to entrust their confidential data to American contractors subjected to the American laws, and therefore to the mastery of the NSA. Last element: the vast movement of engaged revelations by media of the entire world, that contributes to set in motion a debate on the practices of supervision of the services of piece of information even then practically nonexistent, could push the legislators, including Americans, to reconsider the strengths that they gave to their piece of information services.

A.3.2 French source text

L'affaire NSA souligne l'absence totale de débat sur le renseignement

Comment expliquer l'attitude contradictoire du gouvernement français, qui d'un coté s'offusque en public en convoquant l'ambassadeur des Etats-Unis le 21 octobre, et de l'autre interdit le survol du territoire par l'avion présidentiel bolivien, sur la base de la rumeur de la présence à son bord d'Edward Snowden ?

Selon moi, il y a deux niveaux de réponse de la part du gouvernement français. Lorsque François Hollande téléphone à Barack Obama ou quand le ministre des affaires étrangères Laurent Fabius convoque l'ambassadeur des Etats-Unis, ils réagissent à une vraie découverte, qui est celle de l'ampleur de la surveillance américaine sur l'ensemble des communications en France. N'est-il pas surprenant de lire dans les colonnes du Monde à quelques semaines d'intervalle d'une part la reproduction de la correspondance diplomatique américaine et d'autre part une condamnation des écoutes du Quai d'Orsay par la NSA ? N'y aurait-il pas comme une vague hypocrisie de votre part ? La démarche journalistique n'est pas un positionnement moral, mais la recherche de l'intérêt et de la pertinence d'informations qui permettent à chaque citoyen de se forger une opinion. Lorsque WikiLeaks lève le voile sur l'analyse par la diplomatie américaine d'enjeux politiques ou autres dans le monde entier, nous considérons en effet que, au regard de la puissance américaine, cela constitue un éclairage important. Lorsque nous décrivons les systèmes d'interception américains à l'encontre de la diplomatie française aux Etats-Unis, ce n'est en aucun cas pour nous indigner de cette pratique, c'est pour décrire le monde tel qu'il est. La France a-t-elle bénéficié d'informations fournies par la NSA concernant des opérations terroristes visant nos intérêts ? Peut-on se priver de la collaboration américaine ? La mise en place depuis en gros dix ans d'outils technologiques d'interception très puissants par les Etats-Unis, mais aussi par la France, a officiellement été justifiée par la lutte contre le terrorisme. D'ailleurs, dans ce domaine, la France et les Etats-Unis notamment ont mis en place des procédures de coopération et d'échanges d'informations quasi quotidiens et qui sont décrits de part et d'autre comme essentiels. A titre d'exemple, la présence de Mohammed Merah dans les zones tribales à Miranshah a été signalée aux Français grce aux moyens de la NSA. La France peut être conduite, par exemple, à transmettre des blocs entiers de données sur la région du Sahel aux services américains, et, en contrepartie - on l'a déjà rapidement dit -, les Américains peuvent donner des informations aux Français sur d'autres régions du monde. Donc la question de fond derrière cette affaire NSA n'est pas tant la capacité ou le droit des pays de se doter d'outils d'interception, que la question de l'absence totale de débat préalable, notamment au sein des Parlements, sur la justification de tels systèmes, le périmètre qui doit être le leur, et, en fin de

compte, la question des atteintes aux libertés. Que risquent réellement les Etats-Unis ? une dégradation de leur image? On a beau les dénoncer, je ne vois pas de quelle manière ils pourront être punis. Le risque couru par les Américains peut être double. Le premier, c'est lorsque leurs alliés - et ça a été le cas récemment - apprennent que leurs dirigeants, parfois au plus haut sommet de leur Etat, ont été surveillés. C'est le cas du Brésil et de l'Allemagne, deux pays o les relations diplomatiques avec les Etats-Unis se sont tendues. Un autre effet peut être lui plus économique: de plus en plus d'entreprises européennes ou sud-américaines rechignent, à la lumière des révélations, à confier leurs données confidentielles à des prestataires américains soumis aux lois américaines, et donc à l'emprise de la NSA. Dernier élément: le vaste mouvement de révélations engagé par des médias du monde entier, qui contribue à enclencher un débat sur les pratiques de surveillance des services de renseignement jusqu'alors quasiment inexistant, pourrait pousser les législateurs, y compris américains, à reconsidérer les pouvoirs qu'ils ont donnés à leurs services de renseignement.

A.3.3 Reference translation

NSA Affair Emphasizes Complete Lack of Debate on Intelligence

Why the contradictory attitude of the French government? On the one hand, it publicly takes offence and summons the Ambassador of the United States on October 21 and, on the other, it forbids the Bolivian president's plane to enter its air space on the basis of a rumor that Edward Snowden was on board? In my opinion, there are two levels of response from the French government. When François Hollande telephones Barack Obama, or when Foreign Minister Laurent Fabius summons the Ambassador of the United States, they are responding to a real discovery, that of the scale of America's surveillance of communications within France generally. And is it not surprising to read in the pages of Le Monde, on the one hand, a reproduction of diplomatic correspondence with the US and, on the other, condemnation of the NSA's spying on the Ministry of Foreign Affairs on the Quai d'Orsay, within a matter of weeks? Is there not an element of hypocrisy on your part? The journalistic method is not to adopt a moral position, but to investigate the significance and relevance of information and enable every citi-

zen to form an opinion. When WikiLeaks reveals the American administration's monitoring of political and other matters somewhere in the world, we consider this to be significant enlightenment with regard to the American government. In describing the American methods of data interception in relation to the French diplomatic representation in the United States, we do not aim at expressing indignation about this practice, but rather at describing the world as it is. Has France benefited from the intelligence supplied by the NSA concerning terrorist operations against our interests? Can we do without collaboration with the Americans? The setting up of high-performance interception technology over practically the past ten years by the United States - and by France - has been officially justified by the fight against terrorism. Furthermore, in this regard, France and the United States in particular have implemented procedures, sometimes described as essential, for cooperating and exchanging information on an almost daily basis. For example, France was informed of the presence of Mohammed Merah in the tribal areas of Miranshah through the NSA's resources. Also France may, for example, have to transmit entire blocks of data on the Sahel region to the Americans and, in return - as already briefly mentioned - the Americans may provide information to the French about other parts of the world. Hence the question at the heart of the NSA affair is not so much the capacity or the right of a country to use interception tools, as the issue of the complete lack of prior debate - especially within parliaments - on the justification of such systems, the extent to which they should be used and, ultimately, the issue of the infringement of freedoms. What risk does the United States actually run? Ruining its image? However much we denounce the US, I see no way in which it can be punished. The risk run by the Americans could be twofold. The first is when their allies - as has been the case recently - learn that their governments have been spied on, sometimes at the highest level. This is the case in Brazil and Germany, two countries where diplomatic relations with the United States are strained. Another effect could be more commercial: in the light of the revelations, more and more European and South American countries are balking at the idea of entrusting their confidential data to American providers that are subject to American law and hence to the grips of the NSA. Finally, the widespread exercise in revelations conducted by the media across the world, which is contributing to the establishment of a debate

on surveillance practices by intelligence services that have been almost invisible until now, could force legislators - including those of America - to reconsider the powers they have granted their intelligence agencies.

References

Afantenos, S., Asher, N., Benamara, F., Bras, M., Fabre, C., Ho-Dac, M., Draoulec, A. L., Muller, P., Péry-Woodley, M.-P., Prévot, L., Rebeyrolles, J., Tanguy, L., Vergez-Couret, M., and Vieu, L. (2012). An empirical resource for discovering cognitive principles of discourse organisation: the ANNODIS corpus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Aziz, W. F. (2014). *Exact Sampling and Optimisation in Statistical Machine Translation*. PhD thesis, University of Wolverhampton.

Baker, M. (1992). *In Other Words: A Coursebook on Translation*. Routledge.

Barzilay, R. and Lapata, M. (2005). Modeling Local Coherence: An Entity-Based Approach. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 141–148, Ann Arbor, Michigan.

Barzilay, R. and Lapata, M. (2008). Modeling Local Coherence: An Entity-based Approach. *Comput. Linguist.*, 34(1):1–34.

Bawden, R., Sennrich, R., Birch, A., and Haddow, B. (2017). Evaluating discourse phenomena in neural machine translation. *CoRR*, abs/1711.00513.

Becher, V. (2011). *When and why do Translators add connectives? A corpus-based study*, volume 23.

Beigman Klebanov, B. and Flor, M. (2013). Associative Texture Is Lost In Translation. In *Proceedings of the Workshop on Discourse in Machine Translation*, pages 27–32, Sofia, Bulgaria. Association for Computational Linguistics.

Bérard, A., Servan, C., Pietquin, O., and Besacier, L. (2016). MultiVec: a Multilingual and Multilevel Representation Learning Toolkit for NLP. In *The 10th edition of the Language Resources and Evaluation Conference (LREC 2016)*.

Biçici, E. and Specia, L. (2015). QuEst for high quality machine translation. *The Prague Bulletin of Mathematical Linguistics*, 103(1):43–64.

Birch, A., Osborne, M., and Koehn, P. (2007). CCG Supertags in Factored Statistical Machine Translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, pages 9–16, Stroudsburg, PA, USA. Association for Computational Linguistics.

Blakemore, D. (2002). *Relevance and Linguistic Meaning: The Semantics and Pragmatics of Discourse Markers*. Cambridge Studies in Linguistics. Cambridge University Press.

Blatz, J., Fitzgerald, E., Foster, G., Gandrabur, S., Goutte, C., Kulesza, A., Sanchis, A., and Ueffing, N. (2004). Confidence estimation for machine translation. In *In Proceedings of the 20th COLING*, pages 315–321, Geneva.

Bojar, O., Buck, C., Federmann, C., Haddow, B., Koehn, P., Leveling, J., Monz, C., Pecina, P., Post, M., Saint-Amand, H., Soricut, R., Specia, L., and Tamchyna, A. (2014). Findings of the 2014 Workshop on Statistical Machine Translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA. Association for Computational Linguistics.

Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huck, M., Jimeno Yepes, A., Koehn, P., Logacheva, V., Monz, C., Negri, M., Neveol, A., Neves, M., Popel, M., Post, M., Rubino, R., Scarton, C., Specia, L., Turchi, M., Verspoor, K., and Zampieri, M. (2016). Findings of the 2016 Conference on Machine Translation. In *Proceedings of the First Conference on Machine Translation*, pages 131–198, Berlin, Germany. Association for Computational Linguistics.

Bojar, O., Chatterjee, R., Federmann, C., Haddow, B., Huck, M., Hokamp, C., Koehn, P., Logacheva, V., Monz, C., Negri, M., Post, M., Scarton, C., Specia,

L., and Turchi, M. (2015). Findings of the 2015 Workshop on Statistical Machine Translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 1–46, Lisbon, Portugal. Association for Computational Linguistics.

Braud, C., Coavoux, M., and Søgaard, A. (2017). Cross-lingual rst discourse parsing. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 292–304. Association for Computational Linguistics.

Brockett, C., Dolan, W. B., and Gamon, M. (2006). Correcting ESL errors using phrasal SMT techniques. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 249–256.

Brown, P. F., Pietra, V. J. D., Pietra, S. A. D., and Mercer, R. L. (1993). The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, 19(2):263–311.

Burchardt, A., Macketanz, V., Dehdari, J., Heigold, G., Peter, J.-T., and Williams, P. (2017). A Linguistic Evaluation of Rule-Based, Phrase-Based, and Neural MT Engines. In *Proceedings of the 20th Annual Conference of the European Association for Machine Translation. Annual Conference of the European Association for Machine Translation (EAMT-2017), May 28-31, Prague, Czech Republic*, volume 108. De Gruyter Open.

Burstein, J., Tetreault, J. R., and Andreyev, S. (2010). Using Entity-Based Features to Model Coherence in Student Essays. In *HLT-NAACL*, pages 681–684.

Callison-Burch, C., Koehn, P., Monz, C., Peterson, K., and Zaidan, O., editors (2010). *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*. Association for Computational Linguistics, Uppsala, Sweden.

Carpuat, M. (2013). *Proceedings of the Seventh Workshop on Syntax, Semantics and Structure in Statistical Translation*, chapter A Semantic Evaluation of Machine Translation Lexical Choice, pages 1–10. Association for Computational Linguistics.

Carpuat, M. and Simard, M. (2012). The Trouble with SMT Consistency. In *Proceedings of WMT*, pages 442–449, Montreal, Canada.

Cartoni, B., Gesmundo, A., Henderson, J., Grisot, C., Merlo, P., Meyer, T., Moeschler, J., Popescu-Belis, A., and Zufferey, S. (2012). Improving MT coherence through text-level processing of input texts: the COMTIS project.

Cartoni, B., Zufferey, S., and Meyer, T. (2013). Annotating the meaning of discourse connectives by looking at their translation: The translation-spotting technique. *D&D*, 4(2):65–86.

Cartoni, B., Zufferey, S., Meyer, T., and Popescu-Belis, A. (2011). How Comparable Are Parallel Corpora? Measuring the Distribution of General Vocabulary and Connectives. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*, BUCC '11, pages 78–86, Stroudsburg, PA, USA. Association for Computational Linguistics.

Cheung, J. C. K. and Penn, G. (2010). Entity-based local coherence modelling using topological fields. In *ACL*, pages 186–195.

Clarke, J. and Lapata, M. (2010). Discourse Constraints for Document Compression. *Computational Linguistics*, 36(3):411–441.

da Cunha, I. and Iruskieta, M. (2010). Comparing rhetorical structures in different languages: The influence of translation strategies. *Discourse Studies*, 12(5):563–598.

Dusek, O., Novikova, J., and Rieser, V. (2017). Referenceless quality estimation for natural language generation. *CoRR*, abs/1708.01759.

Eidelman, V., Boyd-Graber, J., and Resnik, P. (2012). Topic Models for Dynamic Translation Model Adaptation. In *Association for Computational Linguistics*.

Elsner, M. (2011). Generalizing local coherence modeling.

Elsner, M., Austerweil, J., and Charniak, E. (2007). A unified local and global model for discourse coherence. In *Proceedings of HLT-NAACL*, pages 436–443.

Elsner, M. and Charniak, E. (2008). Coreference-inspired coherence modeling. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, HLT-Short '08, pages 41–44, Stroudsburg, PA, USA. Association for Computational Linguistics.

Elsner, M. and Charniak, E. (2011a). Disentangling chat with local coherence models. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 1179–1189, Stroudsburg, PA, USA. Association for Computational Linguistics.

Elsner, M. and Charniak, E. (2011b). Extending the Entity Grid with Entity-Specific Features. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, HLT '11, pages 125–129, Stroudsburg, PA, USA. Association for Computational Linguistics.

Fancellu, F. and Webber, B. (2015a). Translating Negation: A Manual Error Analysis. In *Proceedings of the Second Workshop on Extra-Propositional Aspects of Meaning in Computational Semantics (ExProM 2015)*, pages 2–11, Denver, Colorado. Association for Computational Linguistics.

Fancellu, F. and Webber, B. (2015b). *Translating Negation: Induction, Search And Model Errors*, pages 21–29. Association for Computational Linguistics. Date of Acceptance: 24/03/2015.

Fancellu, F. and Webber, B. L. (2014). Applying the semantics of negation to SMT through n-best list re-ranking. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, EACL, pages 598–606, Gothenburg, Sweden.

Felice, M. and Yuan, Z. (2014). Generating artificial errors for grammatical error correction. In *Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, Gothenburg, Sweden.

Filippova, K. and Strube, M. (2007). Extending the Entity-grid Coherence Model to Semantically Related Entities. In *Proceedings of the Eleventh European Workshop on Natural Language Generation*, ENLG '07, pages 139–142, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ghorbel, H., Ballim, A., and Coray, G. (2001). Rosetta: Rhetorical and semantic environment for text alignment. In *Proceedings of Corpus Linguistics*, pages 224–233.

Gong, Z., Zhang, M., and Zhou, G. (2011). Cache-based Document-level Statistical Machine Translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 909–919, Stroudsburg, PA, USA. Association for Computational Linguistics.

Gong, Z., Zhang, M., and Zhou, G. (2015). Document-Level Machine Translation Evaluation with Gist Consistency and Text Cohesion. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 52–58, Lisbon, Portugal. Association for Computational Linguistics.

Gouws, S., Bengio, Y., and Corrado, G. (2015). BilBOWA: Fast Bilingual Distributed Representations Without Word Alignments. In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pages 748–756. JMLR.org.

Grimes, J. E. (1975). *The Thread of Discourse*. Mouton The Hague.

Grosz, B. J. and Sidner, C. L. (1986). Attention, Intentions, and the Structure of Discourse. *Comput. Linguist.*, 12(3):175–204.

Grosz, B. J., Weinstein, S., and Joshi, A. K. (1995). Centering: A Framework for Modeling the Local Coherence Of Discourse. *Computational Linguistics*, 21:203–225.

Guillou, L. (2012). Improving Pronoun Translation for Statistical Machine Translation. In *Proceedings of the Student Research Workshop at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pages 1–10, Stroudsburg, PA, USA. Association for Computational Linguistics.

Guillou, L. (2013). Analysing Lexical Consistency in Translation. In *Proceedings of the Workshop on Discourse in Machine Translation*, pages 10–18, Sofia, Bulgaria. Association for Computational Linguistics.

Guillou, L. (2016). *Incorporating Pronoun Function into Statistical Machine Translation*. PhD thesis, University of Edinburgh.

Guillou, L. and Hardmeier, C. (2016). PROTEST: A Test Suite for Evaluating Pronouns in Machine Translation. In Chair), N. C. C., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

Guillou, L., Hardmeier, C., Nakov, P., Stymne, S., Tiedemann, J., Versley, Y., Cettolo, M., Webber, B., and Popescu-Belis, A. (2016). Findings of the 2016 WMT shared task on cross-lingual pronoun prediction. In *Proceedings of the First Conference on Machine Translation (WMT16), Berlin, Germany. Association for Computational Linguistics*, Berlin, Germany.

Guinaudeau, C. and Strube, M. (2013). Graph-based Local Coherence Modeling. In *Proceedings of ACL*, pages 93–103.

Gutt, E.-A. (1989). *Relevance Theory: Guide to Successful Communication in Translation*. PhD thesis, University of London.

Guzmán, F., Joty, S., Màrquez, L., and Nakov, P. (2014). Using Discourse Structure Improves Machine Translation Evaluation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 1: Long Papers*, pages 687–698. The Association for Computer Linguistics.

Hajlaoui, N. and Popescu-Belis, A. (2013). Assessing the Accuracy of Discourse Connective Translations: Validation of an Automatic Metric. In *14th International Conference on Intelligent Text Processing and Computational Linguistics*, page 12. University of the Aegean, Springer.

Halliday, M. A. and Hasan, R. (1976). *Cohesion in English*. Longman, London.

Hardmeier, C. (2012). Discourse in Statistical Machine Translation. *Discours 11-2012*, (11).

Hardmeier, C. (2014). *Discourse in Statistical Machine Translation*. PhD thesis, Uppsala University, Department of Linguistics and Philology.

Hardmeier, C. (2015). On Statistical Machine Translation and Translation Theory. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 168–172, Lisbon, Portugal. Association for Computational Linguistics.

Hardmeier, C. and Federico, M. (2010). Modelling pronominal anaphora in statistical machine translation. In *IWSLT*, pages 283–289.

Hardmeier, C., Stymne, S., Tiedemann, J., and Nivre, J. (2013a). Docent: A Document-Level Decoder for Phrase-Based Statistical Machine Translation. In *51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, Proceedings of the Conference System Demonstrations, 4-9 August 2013, Sofia, Bulgaria*, pages 193–198.

Hardmeier, C., Tiedemann, J., and Nivre, J. (2013b). Latent Anaphora Resolution for Cross-Lingual Pronoun Prediction. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 380–391, Seattle, Washington, USA. Association for Computational Linguistics.

Hatim, B. and Mason, I. (1990). *Discourse and the translator*. Longman.

Hoek, J., Evers-Vermeul, J., and Sanders, T. J. (2015). The Role of Expectedness in the Implicitation and Explicitation of Discourse Relations. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 41–46, Lisbon, Portugal. Association for Computational Linguistics.

Hovy, E., King, M., and Popescu-Belis, A. (2002). Principles of Context-Based Machine Translation Evaluation. *Machine Translation*, 17(1):43–75.

Isabelle, P., Cherry, C., and Foster, G. F. (2017). A challenge set approach to evaluating machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 2486–2496.

Ji, Y., Cohn, T., Kong, L., Dyer, C., and Eisenstein, J. (2016). Document Context Language Models.

Jurafsky, D. and Martin, J. H. (2009). *Speech and Language Processing*. Prentice Hall, 2 edition.

Kehler, A. (1997). Current theories of centering for pronoun interpretation: A critical evaluation. *Computational Linguistics*, pages 23–3.

Khazaei, T., Xiao, L., and Mercer, R. E. (2015). Identification and Disambiguation of Lexical Cues of Rhetorical Relations across Different Text Genres. In *Proceedings of the First Workshop on Linking Computational Models of Lexical, Sentential and Discourse-level Semantics*, pages 54–63, Lisbon, Portugal. Association for Computational Linguistics.

King, M. and Falkedal, K. (1990). Using Test Suites in Evaluation of Machine Translation Systems. In *Proceedings of the 13th Conference on Computational Linguistics - Volume 2*, COLING '90, pages 211–216, Stroudsburg, PA, USA. Association for Computational Linguistics.

Knott, A. and Dale, R. (1994). Using Linguistic Phenomena to Motivate a Set of Rhetorical Relations. *Discourse processes*, 18(1):35–62.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of MT Summit X*, pages 79–86, Phuket, Thailand.

Koehn, P. (2010). *Statistical Machine Translation*. Cambridge University Press, New York, NY, USA, 1st edition.

Koehn, P., Och, F. J., and Marcu, D. (2003). Statistical phrase-based translation. In *Proceedings of NAACL/HLT*, pages 48–54.

Laali, M. and Kosseim, L. (2014). Inducing Discourse Connectives from Parallel Texts. In *COLING*, pages 610–619.

Lapata, M. (2003). Probabilistic text structuring: Experiments with sentence ordering. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 545–552, Stroudsburg, PA, USA. Association for Computational Linguistics.

Lapata, M. (2005). Automatic evaluation of text coherence: models and representations. In *Proceedings of IJCAI*, pages 1085–1090.

Lapshinova-Koltunski, E. (2015a). Exploration of Inter- and Intralingual Variation of Discourse Phenomena. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 158–167, Lisbon, Portugal. Association for Computational Linguistics.

Lapshinova-Koltunski, E. (2015b). Exploration of Inter-and Intralingual Variation of Discourse Phenomena. *DISCOURSE IN MACHINE TRANSLATION*, page 158.

Le Nagard, R. and Koehn, P. (2010). Aiding Pronoun Translation with Co-Reference Resolution. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 252–261, Uppsala, Sweden. Association for Computational Linguistics.

Lehmann, S., Oepen, S., Regnier-Prost, S., Netter, K., Lux, V., Klein, J., Falkedal, K., Fouvry, F., Estival, D., Dauphin, E., Compagnion, H., Baur, J., Balkan, L., and Arnold, D. (1996). TSNLP: Test Suites for Natural Language Processing. In *Proceedings of the 16th Conference on Computational Linguistics - Volume 2*, COLING '96, pages 711–716, Stroudsburg, PA, USA. Association for Computational Linguistics.

Levin, P., Dhanuka, N., Khalil, T., Kovalev, F., and Khalilov, M. (2017). Toward a full-scale neural machine translation in production: the booking.com use case.

Levy, O. and Goldberg, Y. (2014). Dependency-Based Word Embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume 2.

Li, J. and Hovy, E. H. (2014). A Model of Coherence Based on Distributed Sentence Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2039–2048. Association for Computational Linguistics.

Li, J., Luong, T., and Jurafsky, D. (2015). A Hierarchical Neural Autoencoder for Paragraphs and Documents. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1106–1115, Beijing, China. Association for Computational Linguistics.

Li, J. J., Carpuat, M., and Nenkova, A. (2014a). Assessing the Discourse Factors that Influence the Quality of Machine Translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 283–288. Association for Computational Linguistics.

Li, J. J., Carpuat, M., and Nenkova, A. (2014b). Cross-lingual Discourse Relation Analysis: A corpus study and a semi-supervised classification system. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 577–587. Dublin City University and Association for Computational Linguistics.

Libovický, J. and Pecina, P. (2014). Tolerant BLEU: a Submission to the WMT14 Metrics Task. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 409–413, Baltimore, Maryland, USA. Association for Computational Linguistics.

Lin, R., Liu, S., Yang, M., Li, M., Zhou, M., and Li, S. (2015). Hierarchical Recurrent Neural Network for Document Modeling. In *Proceedings of EMNLP*, Lisbon, Portugal. Association for Computational Linguistics.

Lin, Z., Ng, H. T., and Kan, M.-Y. (2011). Automatically Evaluating Text Coherence Using Discourse Relations. In *Proceedings of ACL*, pages 997–1006.

Lo, C. (2017). MEANT 2.0: Accurate semantic MT evaluation for any output language. In *Proceedings of the Second Conference on Machine Translation, WMT 2017, Copenhagen, Denmark, September 7-8, 2017*, pages 589–597.

Loáiciga, S. (2015). Predicting Pronoun Translation Using Syntactic, Morphological and Contextual Features from Parallel Data. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 78–85, Lisbon, Portugal. Association for Computational Linguistics.

Loáiciga, S. and Gulordava, K. (2016). Discontinuous Verb Phrases in Parsing and Machine Translation of English and German.

Loaiciga, S., Meyer, T., and Popescu-Belis, A. (2014). English-French verb phrase alignment in europarl for tense translation modeling. In *The Ninth Language Resources and Evaluation Conference*, number EPFL-CONF-198442.

Loaiciga Sanchez, S. (2017). *Pronominal anaphora and verbal tenses in machine translation*. PhD thesis, University of Geneva.

Logacheva, V. and Specia, L. (2015). The role of artificially generated negative data for quality estimation of machine translation. In *Proceedings of the 18th Annual Conference of the European Association for Machine Translation*, pages 51–58, Antalya, Turkey.

Longacre, R. E. (1996). *The grammar of discourse*. Topics in language and linguistics. Plenum Press, New York, array edition.

Lopez, A. (2008). Statistical Machine Translation. *ACM Comput. Surv.*, 40(3):8:1–8:49.

Louis, A. and Nenkova, A. (2012). A Coherence Model Based on Syntactic Patterns. In *Proceedings of EMNLP-CoNLL*, pages 1157–1168, Jeju Island, Korea.

Louwerse, M. M. and Graesser, A. C. (2005). *Coherence in Discourse*, pages 216–218. Encyclopedia of linguistics.

Luong, N.-Q. and Popescu-Belis, A. (2016). A contextual language model to improve machine translation of pronouns by re-ranking translation hypotheses. *Baltic Journal of Modern Computing*, 4(2):292.

Luong, N.-Q. and Popescu-Belis, A. (2017). Machine translation of spanish personal and possessive pronouns using anaphora probabilities. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, number EPFL-CONF-225949. Association for Computational Linguistics.

Luong, T., Pham, H., and Manning, C. D. (2015). Bilingual word representations with monolingual quality in mind. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 151–159.

Macháček, M. and Bojar, O. (2014). Results of the WMT14 Metrics Shared Task. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, Baltimore, Maryland, USA. Association for Computational Linguistics.

Mann, W. C. and Thompson, S. A. (1988). Rhetorcal structure theory: Toward a functional theory of text organization. *Text-Interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

Manning, C. D. (2015). Computational Linguistics and Deep Learning. *Computational Linguistics*, 41(4):701–707.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60.

Marcu, D., Carlson, L., and Watanabe, M. (2000). The Automatic Translation of Discourse Structures. In *Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference*, NAACL 2000, pages 9–17, Stroudsburg, PA, USA.

Martìnez Garcia, E., Creus, C., España Bonet, C., and Màrquez, L. (2017). Using Word Embeddings to Enforce Document-Level Lexical Consistency in Machine

Translation. In *Proceedings of the 20th Annual Conference of the European Association for Machine Translation. Annual Conference of the European Association for Machine Translation (EAMT-2017), May 28-31, Prague, Czech Republic*, volume 108. De Gruyter Open.

Mascarell, L., Fishel, M., Korchagina, N., and Volk, M. (2014). Enforcing consistent translation of German compound coreferences. In *KONVENS*, pages 58–65.

Mascarell, L., Fishel, M., and Volk, M. (2015). Detecting Document-level Context Triggers to Resolve Translation Ambiguity. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 47–51, Lisbon, Portugal. Association for Computational Linguistics.

Mesgar, M. and Strube, M. (2014). Normalized Entity Graph for Computing Local Coherence. In *Proceedings of TextGraphs@EMNLP 2014: the 9th Workshop on Graph-based Methods for Natural Language Processing, October 29, 2014, Doha, Qatar*, pages 1–5.

Mesgar, M. and Strube, M. (2015). Graph-based Coherence Modeling For Assessing Readability. In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics, *SEM 2015, June 4-5, 2015, Denver, Colorado, USA.*, pages 309–318.

Meyer, T. (2011). Disambiguating Temporal-contrastive Discourse Connectives for Machine Translation. In *Proceedings of the ACL 2011 Student Session*, HLT-SS '11, pages 46–51, Stroudsburg, PA, USA. Association for Computational Linguistics.

Meyer, T. and Poláková, L. (2013). Machine Translation with Many Manually Labeled Discourse Connectives. In *Proceedings of the 1st DiscoMT Workshop at ACL 2013 (51st Annual Meeting of the Association for Computational Linguistics)*, page 8, Sofia, Bulgaria.

Meyer, T. and Popescu-Belis, A. (2012). Using Sense-labeled Discourse Connectives for Statistical Machine Translation. In *Proceedings of the Joint Workshop*

on Exploiting Synergies Between Information Retrieval and Machine Translation (ESIRMT) and Hybrid Approaches to Machine Translation (HyTra), EACL 2012, pages 129–138, Stroudsburg, PA, USA. Association for Computational Linguistics.

Meyer, T., Popescu-Belis, A., Zufferey, S., and Cartoni, B. (2011). Multilingual Annotation and Disambiguation of Discourse Connectives for Machine Translation. In *SIGDIAL Conference*, pages 194–203. The Association for Computer Linguistics.

Meyer, T. and Webber, B. (2013). *Implicitation of Discourse Connectives in (Machine) Translation*, pages 19–26. Association for Computational Linguistics.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Mitkov, R. (1993). How could rhetorical relations be used in machine translation?(and at least two open questions). In *Proceedings of the ACL Workshop on Intentionality and Structure in Discourse Relations*.

Nida, E. A. and Taber, C. (1969). *The Theory and Practice of Translation*. E. J. Brill, Leiden.

Novák, M. (2011). Utilization of Anaphora in Machine Translation. In *WDS'11 Proceedings of Contributed Papers, Part I*, pages 155–160, Praha, Czechia. Matfyzpress.

Novák, M. (2016). Pronoun Prediction with Linguistic Features and Example Weighing. In *Proceedings of the First Conference on Machine Translation (WMT). Volume 2: Shared Task Papers*, pages 602–608, Stroudsburg, PA, USA. Association for Computational Linguistics.

Novák, M. and Žabokrtský, Z. (2014). Cross-lingual Coreference Resolution of Pronouns. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 14–24, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Novikova, J., Dusek, O., Cercas Curry, A., and Rieser, V. (2017). *Why We Need New Evaluation Metrics for NLG*, pages 2231–2242. Association for Computational Linguistics.

Paetzold, G. H. and Specia, L. (2016). Unsupervised Lexical Simplification for Non-Native Speakers. In Schuurmans, D. and Wellman, M. P., editors, *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*. AAAI Press.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Pitler, E. and Nenkova, A. (2008). Revisiting Readability: A Unified Framework for Predicting Text Quality. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 186–195, Stroudsburg, PA, USA. Association for Computational Linguistics.

Pitler, E. and Nenkova, A. (2009). Using Syntax to Disambiguate Explicit Discourse Connectives in Text. In *ACL 2009, Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the AFNLP, 2-7 August 2009, Singapore, Short Papers*, pages 13–16.

Poesio, M., Stevenson, R., Eugenio, B. D., and Hitzeman, J. (2004). Centering: A parametric theory and its instantiations. *Computational Linguistics*, 30(3):309–363.

Popović, M. (2017). Comparing Language Related Issues for NMT and PBMT between German and English. In *Proceedings of the 20th Annual Conference*

of the European Association for Machine Translation. Annual Conference of the European Association for Machine Translation (EAMT-2017), May 28-31, Prague, Czech Republic, volume 108. De Gruyter Open.

Popović, M., Arčan, M., and Lommel, A. (2016). Potential and Limits of Using Post-edits as Reference Translations for MT Evaluation. In *Proceedings of the 19th annual conference of the European Association for Machine Translation (EAMT)*, pages 218–229, Riga, Latvia.

Potet, M., Esperança-rodier, E., Besacier, L., and Blanchon, H. (2012). Collection of a Large Database of French-English SMT Output Corrections.

Prasad, R., Joshi, A., and Webber, B. (2010). Realization of Discourse Relations by Other Means: Alternative Lexicalizations. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, pages 1023–1031, Stroudsburg, PA, USA. Association for Computational Linguistics.

Rashmi Prasad and Nikhil Dinesh and Alan Lee and Eleni Miltsakaki and Livio Robaldo and Aravind Joshi and Bonnie Webber (2008). The Penn Discourse TreeBank 2.0. In *In Proceedings of LREC*.

Reddy, S., Täckström, O., Collins, M., Kwiatkowski, T., Das, D., Steedman, M., and Lapata, M. (2016). Transforming Dependency Structures to Logical Forms for Semantic Parsing. *Transactions of the Association for Computational Linguistics*, 4.

Reddy, S., Täckström, O., Petrov, S., Steedman, M., and Lapata, M. (2017). Universal semantic parsing. *arXiv preprint arXiv:1702.03196*.

Roze, C., Laurence, D., and Muller, P. (2010). LEXCONN: a French Lexicon of Discourse Connectives. In *Multidisciplinary Approaches to Discourse - MAD 2010*, Moissac, France.

Sara Stymne, Jörg Tiedemann, C. H. and Nivre, J. (2013). Statistical Machine Translation with Readability Constraints. In *Proceedings of NODALIDA*, pages 375–386.

Sennrich, R. (2017). How Grammatical is Character-level Neural Machine Translation? Assessing MT Quality with Contrastive Translation Pairs. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 376–382, Valencia, Spain.

Sim Smith, K. (2017). On Integrating Discourse in Machine Translation. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, Copenhagen, Denmark. Association for Computational Linguistics.

Sim Smith, K., Aziz, W., and Specia, L. (2015). A proposal for a coherence corpus in machine translation. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 52–58, Lisbon, Portugal. Association for Computational Linguistics.

Sim Smith, K., Aziz, W., and Specia, L. (2016a). Cohere: A Toolkit for Local Coherence. In Chair), N. C. C., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

Sim Smith, K., Aziz, W., and Specia, L. (2016b). The Trouble with Machine Translation Coherence. In *Proceedings of the 19th annual conference of the European Association for Machine Translation (EAMT)*, pages 178–189, Riga, Latvia.

Sim Smith, K. and Specia, L. (2017). *Examining lexical coherence in a multilingual setting.* Translation and Multilingual Natural Language Processing. Language Science Press, Berlin.

Smith, A., Hardmeier, C., and Tiedemann, J. (2016). Climbing Mount BLEU: The Strange World of Reachable High-BLEU Translations. In *Proceedings of the 19th annual conference of the European Association for Machine Translation (EAMT).*

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). A Study of Translation Edit Rate with Targeted Human Annotation. Proceedings of the 7th Conference of the Association for Machine Translation in the Americas.

Snover, M., Madnani, N., Dorr, B., and Schwartz, R. (2010). TER-Plus: Paraphrase, Semantic, and Alignment Enhancements to Translation Edit Rate. *Machine Translation*.

Somasundaran, S., Burstein, J., and Chodorow, M. (2014). Lexical Chaining for Measuring Discourse Coherence Quality in Test-taker Essays. In *Proceedings of COLING*.

Soricut, R. and Marcu, D. (2006). Discourse Generation Using Utility-Trained Coherence Models. In *Proceedings of the COLING/ACL*, pages 803–810, Sydney, Australia.

Specia, L., Paetzold, G., and Scarton, C. (2015). Multi-level translation quality prediction with QuEst++. In *ACL-IJCNLP 2015 System Demonstrations*, pages 115–120, Beijing, China.

Specia, L., Turchi, M., Cancedda, N., Dymetman, M., and Cristianini, N. (2009). Estimating the Sentence-Level Quality of Machine Translation Systems. In *Proceedings of the 13th Annual Conference of the European Association for Machine Translation*, EAMT, pages 28–37, Barcelona, Spain.

Stede, M. (2011). *Discourse Processing*. Synthesis Lectures on Human Language Technologies. Morgan and Claypool Publishers.

Steele, D. (2015). Improving the Translation of Discourse Markers for Chinese into English. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL, pages 110–117, Denver, Colorado.

Steele, D. and Specia, L. (2016). Predicting and Using Implicit Discourse Elements in Chinese-English Translation. In *Proceedings of the 19th annual con-*

ference of the European Association for Machine Translation (EAMT), pages 305–317, Riga, Latvia.

Tanskanen, S. (2006). *Collaborating towards Coherence: Lexical cohesion in English discourse*. Pragmatics & Beyond New Series. John Benjamins Publishing Company.

Tiedemann, J. (2010). Context Adaptation in Statistical Machine Translation Using Models with Exponentially Decaying Cache. In *Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing*, pages 8–15, Uppsala, Sweden.

Tu, M., Zhou, Y., and Zong, C. (2013). A Novel Translation Framework Based on Rhetorical Structure Theory. In *ACL (2)*, pages 370–374. The Association for Computer Linguistics.

Upadhyay, S., Faruqui, M., Dyer, C., and Roth, D. (2016). Cross-lingual Models of Word Embeddings: An Empirical Comparison. *CoRR*, abs/1604.00425.

Vickrey, D., Biewald, L., Teyssier, M., and Koller, D. (2005). Word-Sense Disambiguation for Machine Translation. In *In EMNLP*, pages 771–778.

Voigt, R. and Jurafsky, D. (2012). Towards a Literary Machine Translation: The Role of Referential Cohesion. In *Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature*, pages 18–25, Montréal, Canada. Association for Computational Linguistics.

Waibel, A. and Fugen, C. (2008). Spoken language translation. *IEEE Signal Processing Magazine*, 25(3).

Webber, B., Carpuat, M., Popescu-Belis, A., and Hardmeier, C., editors (2015). *Proceedings of the Second Workshop on Discourse in Machine Translation*. Association for Computational Linguistics, Lisbon, Portugal.

Webber, B., Egg, M., and Kordoni, V. (2012). Discourse structure and language technology. *Natural Language Engineering*, 18(4):437–490.

Webber, B., Popescu-Belis, A., Markert, K., and Tiedemann, J. (2013). *Proceedings of the ACL Workshop on Discourse in Machine Translation (DiscoMT 2013)*. Association for Computational Linguistics.

Wetzel, D. and Bond, F. (2012). Enriching Parallel Corpora for Statistical Machine Translation with Semantic Negation Rephrasing. In *Proceedings of the Sixth Workshop on Syntax, Semantics and Structure in Statistical Translation*, SSST-6, pages 20–29, Jeju, Republic of Korea.

Wetzel, D., Lopez, A., and Webber, B. (2015). *A Maximum Entropy Classifier for Cross-Lingual Pronoun Prediction*, pages 115–121. Association for Computational Linguistics.

Wong, B. T.-M. and Kit, C. (2012). Extending Machine Translation Evaluation Metrics with Lexical Cohesion to Document Level. In *Proceedings of EMNLP-CoNLL*, pages 1060–1068.

Xiong, D., Ben, G., Zhang, M., Lv, Y., and Liu, Q. (2013a). Modeling Lexical Cohesion for Document-Level Machine Translation. In *Proceedings of IJCAI*.

Xiong, D., Ding, Y., Zhang, M., and Tan, C. L. (2013b). Lexical Chain Based Cohesion Models for Document-Level Statistical Machine Translation. In *Proceedings of EMNLP*, pages 1563–1573.

Xiong, D. and Zhang, M. (2013). A Topic-Based Coherence Model for Statistical Machine Translation. In *Proceedings of AAAI*, pages 977–983.

Xiong, D. and Zhang, M. (2014). A Sense-Based Translation Model for Statistical Machine Translation. In *ACL (1)*, pages 1459–1469.

Yung, F., Duh, K., and Matsumoto, Y. (2015). Crosslingual Annotation and Analysis of Implicit Discourse Connectives for Machine Translation. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 142–152, Lisbon, Portugal. Association for Computational Linguistics.

Zhang, R. and Ittycheriah, A. (2015). Novel Document Level Features for Statistical Machine Translation. In *Proceedings of the Second Workshop on Dis-*

course in Machine Translation, pages 153–157, Lisbon, Portugal. Association for Computational Linguistics.

Zhao, Z., Liu, T., Li, S., Li, B., and Du, X. (2017). Ngram2vec: Learning improved word representations from ngram co-occurrence statistics. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*.