Domain and Genre Dependency in Statistical Machine Translation

Marco Brunello

Submitted in accordance with the requirements for the degree of

Doctor of Philosophy

University of Leeds

School of Languages, Cultures and Societies

September 2014

The candidate confirms that the work submitted is his own and that appropriate credit has been given where reference has been made to the work of others.

This copy has been supplied on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.

The contribution reported in Chapter 3 (excluding sections 3.2.1 and 3.4) is based on Brunello (2012).

Acknowledgements

First of all I would like to thank Serge Sharoff, Bogdan Babych and Martin Thomas for being my supervisors during these past few years, and also for giving me the opportunity to be involved in projects and teaching at the Centre for Translation Studies. Thank you also to all the former and current members of staff at the CTS I had the chance to work with, in particular Richard Forsyth for the support both at the academic and the human level. I would like to thank as well colleagues from other universities I had the chance to know during conferences and European projects meetings. A special thanks goes to TAUS for funding my research, giving me the opportunity to explore (and hopefully contribute to) a field I have been always interested in.

Thanks to my family - my mother Maria Luisa, my father Graziano and my sister Paola - and my girlfriend, Kathryn, for believing in me all the time, and for loving me unconditionally.

Now it would be very difficult to list all of them, but I would like to thank the countless number of people I managed to get to know during these four past years, which in a way or another left a mark in my life. In particular thanks to those who honoured me with their true friendship and have shared with me some nice moments.

I am now a different person than I was four years ago, and in case I became a better one it is definitely thanks to all of you.

Abstract

Statistical Machine Translation (SMT) is currently the most promising and widely studied paradigm in the broader field of Machine Translation, continuously explored in order to improve its performance and to find solutions to its current shortcomings, in particular the sparsity of big bilingual corpora in a variety of domains or genres to be used as training data. However, while one the main trends is still to rely as much as possible on already available large collections of data, even when they do not fit quite well specific translation tasks in terms of relatedness of content, the possibility of using less but appropriately selected training sets - depending on the textual variety of the documents that need to be translated case by case - has not been extensively explored as much so far.

The goal of this research is to investigate whether this latter possibility, i.e. the lack of availability of large quantities of assorted data, can have a possible solution in the application of strategies commonly used in genre and domain classification (including unsupervised topic modeling and document dissimilarity techniques), in particular performing subsampling experiments on bilingual corpora in order to obtain a good fit between training data and the texts that need to be translated with SMT.

For the purposes of this study, already existing freely available large corpora were found to be unsuitable for the selection of domain/document-specific subsamples, so two new parallel corpora - English-Italian and English-German - were compiled employing the "web as corpus" approach on websites containing translated content. Then some tests were made on documents belonging to different varieties, translated

with SMT systems built using subsamples of training data selected using document dissimilarity measures in order to pick up the most suitable documents as training data.

Such method has shown how the choice of subsampling strategy heavily depends on the text variety of each considered document, but it has also proven that better translation results can be obtained from small samples of training sets rather than using all the available data, which brings benefits also in terms of quicker training times and use of fewer computational resources.

Contents

Chapter 1

Introduction

The idea of building Machine Translation (MT) systems first emerged around 60 years ago (Weaver, 1955) and it has seen a remarkable growth during the last decades, when MT systems started being developed both in the academic field and in the private sector, becoming a widely employed technology by users as well. The emergence of MT even led many people to seriously consider that MT may soon take over and substitute human translation - even claiming that human translators would be left unemployed because of that. But MT is far from being a 100% reliable fully-functioning multi-purpose technology, and it still requires a certain degree of human interpretation of its output. As said in Koehn (2010, 20), the possibility of having fully-automatic high quality machine translation can be considered at the moment nothing more than a holy grail of MT, since so far it has been possible to develop fully-automatic MT systems only for a limited amount of specific (and of very codified) communicative situations, e.g. weather forecast, summaries of sports events, multinational companies documentation. This means that translation could be difficult, sometimes impossible, to be performed completely automatically in most cases.

So, rather than aiming at the quite unfeasible target of building a fully reliable all-purpose MT system, it may be possible to improve the performances of MT approaching the problem from alternative points of view, like the possibility to carry out topic-specific MT tasks. In Statistical Machine Translation (SMT) it is possible to create translation systems providing a certain quantity of bilingual

(and monolingual, in the target language) texts as training data to an SMT engine, so in order to obtain good performances for a specific SMT task it is crucial to employ (and where possible select) those training data which are most suitable for the text(s) one wants to translate. The main trend is to employ large quantities of parallel data in order to maximise the coverage of translation possibilities (Bloodgood & Callison-Burch, 2010). In many cases most of the data employed may be out-of-domain, and the translation performance is then adjusted tuning SMT systems towards specific translation tasks (see section 2.3.3). However recent trends have shown that it is possible to rely on reduced amounts of relevant training data, and the research here presented follows this direction: considering the textual variety of single texts that need to be translated, it may be useful to train MT systems case-by-case, using small amounts of accurately selected training data. In order to understand how to select the most suitable data for each specific translation situation, and whether using much smaller training sets than what usually happens in SMT makes sense, the research here presented contains an exploratory study implementing strategies borrowed from the "web as corpus" paradigm and genre/domain/document dissimilarity studies.

The advantage of this approach is twofold: on the one hand tailored MT systems may yield better translations, on the other hand using less but focused data means fewer time to train the SMT systems themselves. Such operational benefits would be very valuable when thinking of possible implementations of the strategy here described in actual scenarios like the translation industry, where companies may have limited amounts of time to carry out specific translation tasks.

1.1 Objectives

The problems this research seeks to try to address in specific are:

- **Data sparsity**

  The availability of openly accessible parallel corpora to be used as sources of training data for SMT is limited, particularly in terms of assortment,

most of them being mainly documentation coming from international organisations (regulations, statutes, transcription of parliamentary proceedings etc.). These parallel collections are usually employed as out-of-domain data, but due to their high topic specificity it would not be correct to consider them "general domain" data: they are used as such because at the current moment there is no such a thing like "general-purpose"[1] large multilingual corpora as BNC, UKWAC etc. are in monolingual corpus linguistics. This problem emerged since the very beginning of the research project here presented, since some of these collections have been considered but ended up being inappropriate for the purpose of studying language variability (See 3.1.2).

- **Is more data better data?**

  In theory the very limited domain specificity of the above mentioned parallel corpora would make them quite unsuitable for the translation of the majority of non-related texts. Nevertheless because of their free availability they are actually widely employed as material on which to build SMT baseline systems, then tuned towards specific user cases - mainly implementing in-domain monolingual and/or multilingual material through a series of strategies (domain adaptation). This means that the "more data is better data" rule has been usually followed as a default approach to SMT until very recent times, since in theory the more are data provided when training a new SMT system, thus having a better coverage for multiword expressions (minimising the chances of meeting unseen words etc.), the greater are the chances to obtain a better, more appropriate and fluent, translation. But even though "the de facto standard consists in training SMT systems with all the available data" (Gascó *et al.*, 2012), it seems that in certain cases adding more out-of-domain data can be not only not useful but even harmful (Haddow & Koehn, 2012), and several recent studies such as Elizalde Cecilia, Pouliquen Bruno, Mazenc Christophe (2012) started pointing out

[1] "General purpose" here is intended in the sense of a corpus containing documents whose content is diverse in terms of language variability.

that a "less (but ad hoc) data" approach may be beneficial. So the possibility of using smaller but accurately selected training sets is here explored, in order to see whether it is possible to obtain benefits from this strategy such as use of less computational resources, quicker training, more appropriate domain-specific translations.

- **Multilingual webcorpora**

  As just said, freely accessible parallel parallel corpora available on the web have some limitations. However on the Internet it is possible to find a large amount of bilingual/multilingual websites/pages in a variety of language pairs, published for disparate purposes. They can be collected and processed, extracting their plain text and aligning translated content at the sentence level, in order to build new parallel corpora. But there is a surplus of difficulties compared to the traditional monolingual corpus collection from the web, mainly concerning how to find and pair multilingual webpages, added to the usual "web as corpus" issues such as text quality, copyright matters etc. Several strategies to collect parallel corpora from the web have been developed during the last 15 years (see section 2.2.3), but most of them are not available to the community for various reasons: some of them were based on now deprecated technology, contractual constraints, the authors' choice not to publicly release them etc. Moreover the majority of these contributions do not provide wide information about the genres/domains of the retrieved parallel data, whereas it may be important to know the nature of possible training data with regards to their composition in terms of text types.

  Based on them a system able to collect parallel corpora from the web has been set up for this project, providing two new corpora in a variety of genres and domains, and their composition has been analysed - and so doing provided an overview about the most common typologies of multilingual websites on the web for the considered languages.

- **How to sample?**

Having a specific document that need to be translated and a collection of bilingual texts containing possible training data, it is necessary to decide which ones are the most suitable to perform as training data to develop an SMT system for this specific translation. The strategy here presented involves the comparison of that document with every single text from the whole collection of training candidates via document dissimilarity measures, followed by a selection of a subsample containing only those documents appearing most similar to the considered document.

- **Industrial scenarios**

  When working in the translation industry there can be the necessity of developing an MT system with the purpose of translating a certain kind of documents for a client, in a limited amount of time: in such situation the ability of intelligently selecting the most relevant training data may be beneficial to quickly train an SMT system good enough to provide a first set of translations which would be post-edited anyway. So this research may be relevant with regards to practical industrial settings, and this is going to be verified through the experiments here presented.

1.2 Thesis outline

This work is divided mainly in two parts: the first involves the exploration of existing resources, both SMT and text classification tools and material; the second describes the collection and analysis of two new parallel corpora (English-Italian and English-German) and the subsampling for document-specific SMT experiments. More specifically:

Chapter 2 provides an overview of the three main research fields of interest: SMT (the basics of the paradigm), parallel corpora from the web (in particular previous attempts and methods), and a summary of genre, domain and document dissimilarity studies - with a focus on those employed in this project.

Chapter 3 presents the reasons for new resources, including a classification study on Europarl, and the complete pipeline employed to download parallel corpora from the web. Topic modeling and document dissimilarity analyses on

the English-Italian and English-German corpora collected from the web are also provided.

Chapter 4 reports the final set up for the sampling experiments, including the description of preliminary tests, the selection of test documents and the final experiments themselves. Evaluation through automatic metrics and manual analysis of the results is then presented.

Chapter 5 contains a brief summary with the conclusions, achieved results, still open concerns and suggesting possible directions for future works.

Chapter 2

Background

The aim of this chapter is to provide the context of the research here reported, presenting an overview of the various fields of study that have been covered in order to carry out this research project. In particular the basics of SMT are reviewed in section 2.1, the matter about collecting parallel corpora from the web in 2.2 and the background of studies on text variability in 2.3.

2.1 Statistical Machine Translation

2.1.1 The SMT paradigm

With **Statistical Machine Translation (SMT)** we define a specific Machine Translation paradigm which makes use of statistical models in order to carry out automated translation tasks. It can be considered in juxtaposition to the **Rule-Based Machine Translation (RBMT)** paradigm, which can be referred as "the classical approach to MT", as its main alternative.

The two paradigms differ fundamentally in their core approach: RBMT systems are based on linguistic knowledge extracted from grammars and dictionaries in the source and target languages of interest, and automated translations are performed based on models containing this structured information. SMT systems instead learn how to perform previously unseen translations from already existing bilingual texts themselves, and in the form of unstructured information (e.g. tree-based models). From this point of view SMT is similar to another corpus-based

approach, **Example-Based Machine Translation (EBMT)**, but the peculiarity of SMT consists in building translation systems based on the analysis of parallel corpora with statistical models. The advantage of SMT consists then in the fact that it is not necessary to perform long and time consuming hand-crafted analyses which ideally have to take into account as many as possible translation rules between two languages (at least for an intended purpose) as it happens for example with RBMT, but - at least in its basic form - building an SMT systems mainly consists in training a model with a bilingual corpus in the two languages of interest to allow it to learn how to perform new translations based on (not necessarily linguistic) segment alignments. SMT has become one of the most actively studied and promising MT paradigms, already having in 2010 "about one thousand academic papers [...] published on the subject, about half of them in the past three years alone"(Koehn, 2010, xi).

Even though SMT has seen a remarkable growth and advance during the last few decades, the core concept of using information theory strategies to perform automated translations actually goes back to the 50's, and was well exemplified in a sentence by Warren Weaver when he said "When I look at a article in Russian, I say 'This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode' " (Weaver, 1955). However some decades passed before SMT emerged and established itself as one of the main paradigms of MT: in the early 90's the IBM T.J. Watson Research Centre developed **Candide** (Berger *et al.*, 1994), an SMT system able to perform French to English translations based on statistics calculated on a training set of bilingual sentences extracted from transcriptions of parliament proceedings collected in the Canadian Hansard corpus[1]. By the end of the 90's the interest on SMT grew considerably, with the Defense Advanced Research Projects Agency (DARPA) funding programs such as TIDES (Translingual Information Detection, Extraction and Summarization) and GALE (Global Autonomous Language Exploitation) which had a strong focus on the development of SMT. Another important landmark for SMT was the birth of private companies commercialising SMT systems, like

[1] http://www.isi.edu/natural-language/download/hansard.

Language Weaver (founded in 2002 and acquired in 2010 by SDL[2]).

So the majority of the most important developments of the SMT paradigm can be located mainly in the last few decades, which have seen continuous and growing progress in the field of SMT and led this paradigm to become an object of great interest in the MT community and beyond, reviving the whole discipline and also contributing to further development of related research interests, like MT evaluation. Let us now consider more technically the specifics of SMT and in general the starting points on which the thesis here presented is based.

2.1.2 Core method

The baseline approach of SMT is based on the analysis of probability distributions of segments contained in collections of bilingual texts in the two languages of interest. In its basic form an SMT system requires two essential elements:

1. A **translation model**, based on the sentence and word alignments of the content of a bilingual corpus;
2. A **language model**, which is a collection of sentences in the target language (either the target language side of the parallel corpus, or other monolingual data, or both), provided in order to ensure a fluent output.

As reported in Brown *et al.* (1993), the basic formula of SMT is

$$\tilde{e} = arg\max_{e \in e^*} p(e|f) = arg\max_{e \in e^*} p(f|e)p(e)$$

where `p(e|f)` is the probability distribution that a segment `e` in the target language is the translation of the segment `f` in the source language; the Bayes theorem is used in order to reformulate this in order to have a translation model `p(f|e)` and a language model `p(e)`.

[2] http://www.sdl.com/aboutus/news/pressreleases/2010/sdl_acquires_language_weaver.html.

This way an SMT system tries to estimate the most likely translations of each segment contained in the aligned sentences, based on previously made translations. These segments were initially words in the early SMT models (IBM Model 1 to 5), but then the research in the field started considering chunks of words (phrases) as more suitable for this purpose. The introduction of phrase-based models is justified by the fact that words may be not the best basic unit for SMT: the same concept could be expressed with different amounts of words depending on the language. However the word *phrase* here is not necessarily meant to define the concept of *multiword expression* in a grammatical sense, since phrase-based SMT does not make use of linguistic notions when creating translation models based on segment alignments. This way phrase-based systems has proven very useful to address certain issues such as translation of ambiguous expressions.

The description provided so far has been about the basics of the SMT approach. It is possible to apply several kind of improvements to the standard phrase-based SMT baseline, i.e. integrating linguistic information such as part-of-speech tags (factored models), employing syntactic enhancements (tree-based models) or combining it with strategies borrowed from other approaches (like RBMT) to build hybrid MT systems. Several other strategies have been (and are being) developed in order to solve several issues still affecting SMT (see section 2.3.3). These problems are going to be considered in the next paragraph.

2.1.3 Open issues

As shown in the previous section, SMT can provide several advantages compared to traditional RBMT. However there are several open issues in SMT that still need to find a proper solution.

One of these problems is the availability of suitable multilingual resources: whilst a state of the art SMT system like Moses (Koehn *et al.*, 2007) is publicly available under the GNU Lesser General Public License[1], to find training data to be fed into it is a more difficult task. Two requirements are important in order to get a good translation from SMT: having a good fit between the training data provided to the system and a specific text that needs to be translated, in

[1] http://www.gnu.org/licenses/lgpl.html.

order to ensure appropriate lexicon coverage, grammar constructions etc., and to provide a suitable amount of training data to produce understandable automated translations. But in many cases it is difficult to find a set of training data both in the intended text variety and size (this problem is explored in deeper detail in the next paragraph).

Together with this there is the fact that the chances of obtaining good quality SMT do not depend only on the availability of parallel data, which is even further lowered when dealing with medium and low density languages (i.e. languages whose digital resources are not available in large quantities as much as, for example, English, Spanish, French etc.), but also the intrinsic difficulty of certain language pair combinations: as shown in Koehn (2010, 18-20), Arabic-English or Chinese-English statistical translation outputs appear to be overall qualitatively not as good as the French-English translation, considering about 200 million words for the first two language pairs and 40 million words for the latter (See Figure 2.1). The typological distance between two languages, even more when they belong to different writing systems, makes certain language pairs more difficult to translate. However difficulties may emerge even when translating between related languages, for example performing SMT between two Germanic languages having different word order (like English and German) can affect the quality of the output.

Other problems are related to the text processing when preparing data for SMT training, for example sentence alignment: for various reasons a sentence in one language may correspond to two or more sentences in the other language, and vice versa. Currently available sentence aligners such as Hunalign (Varga *et al.*, 2005) and Gargantua (Braune & Fraser, 2010) are able to manage with such cases, but still this remains a non-trivial task: any mistakes during the sentence splitting process, to be done prior to sentence alignment, can further lead to the creation of wrongly aligned segments. Errors occurring during this task may impair following steps of text processing for SMT, above all word alignment, and so the ability of an SMT system to properly translate new sentences.

Figure 2.1: Examples of French-English, Chinese-English and Arabic-English machine translations (from Koehn *et al.* (2007)).

2.2 Parallel corpora from the web

Some matters about the use of bilingual texts in SMT have already been mentioned in the previous section, but they are going to be analysed more in detail here. In particular the concept of parallel text will be explained in 2.2.1; the current status of availability of ready-made and publicly available parallel corpora is described in 2.2.2; an overview of previous attempts and strategies to collect parallel corpora from the web is given in 2.2.3; known problems related to the collection of large amounts of texts from the Internet (in particular when dealing with bilingual texts) are listed in 2.2.4.

2.2.1 Definitions: bilingual data, bitext, multilingual corpus etc.

Parallel texts in previous sections have been defined as texts in a source language provided along with their translation in a target language, while the expression **multilingual text** is used when a text translated in more than one language, and they are usually aligned at the sentence level. In the field of translation studies they are also usually referred as **bitexts**, while in the translation industry the concept of **translation memory (TM)** is more specific and defines bitexts where segments (usually sentences, but they can be also paragraphs or other type of language units) are stored in databases, to be employed in **computer-assisted translation (CAT)** tools. TMs do not necessarily keep the order of sentences as in the original bitext they have been extracted from, and usually only a single record of repeated segments is kept. A typical format in which TMs are formatted is **Translation Memory eXchange (.tmx)**, which is a specific kind of XML standard where different attributes can be specified for each segment, such as language, author, notes etc.[1]

Parallel corpus is instead used to define a large collection of bilingual texts, which can be provided in different formats depending on their size, purpose etc. In order to build models able to translate from L1 to L2 Moses accepts as input

[1]Specifications for the TMX format can be found at http://www.ttt.org/oscarstandards/tmx.

two separate plain text format documents where each line/sentence (in L1) in the first file corresponds to its translation (in L2) in the second file. The main publicly available corpora are listed in the next section.

2.2.2 Existing resources

There are no strict standards about how big should precisely be a training set for SMT in order to provide statistics apt to perform previously unseen translations, but a broad generalization about this was made by Philipp Koehn saying that "machine translation models are typically estimated from parallel corpora with tens to hundreds of millions of words", and "language models may use even more data: millions and even trillions words have been used in recent research systems" (Koehn, 2010, 264). However, while it is nowadays possible to obtain monolingual corpora of millions/trillions words and in a variety of topics (especially from the web, see Baroni & Bernardini (2006)), the access to parallel corpora in the order of hundreds millions words can be quite limited. For example the Linguistic Data Consortium offers several parallel corpora[1] as a paid service, with fees in the order of thousand of dollars per corpus for non-members, but it is well resourced only for certain language pairs (mainly English, Arabic and Chinese) and text types (e.g. news and law). Some multilingual corpora are instead publicly available on the Internet, and they found wide use in the community since they are more accessible resources in terms of costs and in some cases they provide a reasonably wide choice of language pairs.

Europarl (Koehn, 2005) is probably the best-known resource when dealing with SMT between European languages. It provides textual material extracted from the proceedings of the European Parliament from 1996 to 2011[2], including texts in 21 European languages. Sizes of the several L1-English (or vice versa) parallel corpora of this collection are variable, being around 50 million words per language for French, Spanish, German, Italian, Portuguese and Duch and several smaller amounts for the remaining languages[3]. The texts contained in this multilingual corpus are provided both as monolingual with detailed XML markup

[1] http://www.ldc.upenn.edu.

[2] Version 7, released on 15th May 2012.

[3] For a complete overview of these data see http://www.statmt.org/europarl.

and paired with English in plain text format, with a series of tools (aligner, tokeniser, truecaser etc.) to process and use them into an SMT system like Moses. Although it is very topic specific this corpus represents one of the main resources for SMT users and developers and it is employed as a "general-purpose" baseline on which build possible improvements (these strategies and the topic-specificity of Europarl will be better examined later in this thesis).

The availability of official reports from national and international organizations as publicly accessible documents make them an easy-to-obtain source of parallel data. In addition to Europarl there are other notable examples of similar kind of texts: **JRC-Acquis**[1] (Steinberger *et al.*, 2006) is a corpus collecting the complete body of European Union law applicable to the member states, available in 22 European languages. Or, talking about other proceedings of parliamentary debates, another well known parallel corpus is the **Canadian Hansard**, an English-French corpus containing debates from the Canandian Parliament[2]. Another corpus made by institutional documents, this time coming from the United Nations, is **MultiUN**[3] (Eisele & Chen, 2010).

Most of these resources are part of the **OPUS Project**, an initiative aimed at collecting in a single website parallel corpora coming from "open sources" around the web, consistently provided in a variety of formats (XML, TMX, sentence-aligned plain text for Moses) and with detailed documentation[4].

2.2.3 Crawling the web for bitexts

Some of the above mentioned parallel corpora are not only made available on the Internet but also collected from the Internet itself, i.e. by downloading and aligning the parallel content of institutional websites. Similarly it is possible to crawl the web in order to find other websites with multilingual content, taking advantage of the "Web as Corpus" approach: the web can be considered a very large corpus, providing the widest possible variety of contents in terms of formats, topics, languages etc. (Kilgarriff & Grefenstette, 2003), and this definitely

[1] http://ipsc.jrc.ec.europa.eu/index.php?id=198.
[2] http://www.isi.edu/natural-language/download/hansard/.
[3] http://www.euromatrixplus.net/multi-un/.
[4] http://opus.lingfil.uu.se/.

includes a number of multilingual websites built for a variety of purposes. So it makes sense that the web would be exploited not just to build monolingual text corpora but also multilingual ones. This possibility started being taken into account around the last couple of decades, i.e. when the usefulness of the Internet for the development of tools and parallel corpora for translation studies scholars - or even for professional translators, and researchers in the field of MT - has been pointed out and encouraged in several papers (Lagoudaki, 2007; Zanettin, 2002).

The forerunner of using the web as a source for collecting parallel corpora is Philip Resnik and his system STRAND, described since the late nineties in a series of papers, the last one being Resnik & Smith (2003), where the core STRAND system is explained as well with several improvements comparing to the original core version. The main idea behind this approach is to find webpages that exhibit a parallel structure at the level of URL and/or page composition, and that could be mutual translations. In practice this was done relying on some advanced options of the AltaVista search engine, which allowed to find out whether a page contains links to different language versions of that document contained in the same website. The retrieved pages were then subject to a candidate pairs detection task that has been carried out with several strategies, combining automatic language identification, URL matching, average document lengths and other content-based similarity measures to detect pairs of pages even when they do not present similarity just at the level of structure. Other systems, developed independently from STRAND but employing similar approaches, have seen the birth in the same period: PTminer (Ma & Liberman, 1999), BITS (Chen & Nie, 2000), PTI (Chen *et al.*, 2004). A similar and more recent implementation of these strategies is described in Mohler & Mihalcea (2008), who presented a system called Babylon, developed with the purpose to find parallel texts for under-resourced languages. This was done by crawling the web using Google Search APIs on a set of seed words in Quechua, and then looking for Spanish language counterpart pages or even checking whether there are portions of parallel text within a single page. Another project that tried to overcome the lack of publicly available parallel corpora for certain languages (in this case English, Latvian, Lithuanian and Romanian) is ACCURAT (Pinnis *et al.*, 2012), which

found a possible solution in the exploitation of comparable corpora, with the development of tools for the alignment of comparable documents and extraction of parallel sentences and bilingual mapping of translated terminology.

So the two main steps to collect a parallel corpus from the web can be summarised as

1. crawl the web to find potential multilingual pages or websites, and
2. locate and pair translated pages in the two languages (whose textual content will be later extracted and aligned).

The above mentioned papers describe strategies which perform both passages but a number of other works focus only on the second step, assuming a list of bilingual pages or websites has already been obtained. Some of them consider the extraction of bilingual content from dynamic content websites: Fry (2005) proposes when possible to exploit the RSS syndication format, and a similar strategy is proposed by Tsvetkov & Wintner (2010) with their system called PCB (Parallel Corpora Builder), which crawls news sites with dynamic content to obtain an English-Hebrew parallel corpus. Another paper is Almeida & Simões (2010), that describes a system called GWB (GetWebBitext) using as starting point the crawl of specific websites from a set of keywords in L1 fed into an implementation of Yahoo! APIs, then trying to detect corresponding URLs in L2, in a similar way to the STRAND approach but implementing a strategy that avoids the need to download and parse HTML files from the web (as it happened in the previously mentioned STRAND-alike approaches).

A very recent alternative to the simple STRAND-alike models is Paradocs, developed by Patry & Langlais (2011), which does not rely on file or URL naming informations, but rather works entirely with content-based features (numerical entities and hapax words), and it is a language neutral system.

The majority of these tools are not released for public use, so similar strategies need to be reimplemented from scratch. But some developers have rather decided to share their resources and make them publicly available, and this is the case of Bitextor (Esplà-Gomis & Forcada, 2010). Its authors remark on the importance

and usefulness of extracting parallel corpora from the web in particular for corpus-based machine translation like the statistical approach, but possibly for rule-based MT systems as well, being parallel data a possible source of texts where to extract translation rules. Based on several strategies above described in the previous papers (URL comparison, content comparison, text length etc.), they compiled an open source system able to extract bitexts from a given website, outputting them in TMX format containing segments aligned at the sentence level.

2.2.4 Issues of the WAC approach

Previously implemented approaches to the "web as parallel corpus" have been reviewed in the previous paragraph. As shown it is possible to collect parallel corpora from the web, but there are some shortcomings that should be considered as well - some of them more general "web as corpus" issues, some others more specific about the collection of parallel data:

- The Internet definitely provides a large amount of easy to access language data, but overall its content is unstable, widely unknown and in continuous change. This means that, despite the advantage of having an ever-growing, easy-to-access resource, there are sites and pages appearing and disappearing all the time (Wattam *et al.*, 2012), making difficult (if not impossible) to replicate experiments involving information retrieval from the web. Moreover the extremely wide nature of what the web contains means a lack of consistency in terms of formats (e.g. the use of different text encodings), and the fact that many texts are not necessarily correct from a linguistic or grammatical perspective (not to talk about the amount of spam and automatically generated text) can be a serious problem - whereas the purpose of the research is not to study peculiarities of the Internet language itself;

- The availability of search engines APIs provides an easy way to crawl the web through major search engines, but their behaviour is actually unknown since the nature of so-retrieved pages depends on how they have been ranked by search engine - see the case of paid ads. Also the behaviour of search engines themselves is not consistent, being subject to continuous changes

/updates, and it is quite unlikely to get the same search results on the same set of keywords, presenting again problems of repeatability. Moreover this kind of services provided by Google, Yahoo!, Bing etc. are not always available: on the contrary they are in many cases turned into paid services after a period of free availability, in a way that developers have either to purchase them or reimplement URL collection applications based on which search engine offers this kind of service for free at some point;

- The issue about copyright: the debate about how ethical (and/or legal) can be the reproduction of documents retrieved from the web has never come to an end or a proper solution, probably because of the several variables involved: national copyright laws deals with this matter in different ways (and countries with out-to-date legislations may not even contemplate digital data), the concept of fair use, which can be subject to a number of interpretations, and in general the fact that there is no clear and shared judgement on (the paradoxical situation of) having such a large amount of easily reproducible digital data and the fact that in theory each single file on the Internet belongs exclusively to its author (Hemming & Lassi, 2003) - unless otherwise stated in the case of using open source licenses (GNU or Creative Commons) and public domain material;

- Another problem is the availability of data in particular languages (and language pairs): while the web definitely contains a large amount of texts in high-density languages (English above all, but also Spanish, French, Chinese), documents written in medium- and low-density languages may be available in remarkably smaller quantity. This restricts the choice of data even more when it comes to parallel documents, since the stress in not on the availability of data in one single language but on specific language pairs;

- Retrieval of parallel data from the web may be difficult also considering that there are no standards for the construction of multilingual webisites: there may be recurring strategies - especially thanks to the spread use of templates in various content management systems - but in general every webmaster is free to organise the content of a bilingual or multilingual

website the way he prefer. So there is an high degree of inconsistency in the structure of multilingual site, and it may be hard to pair and align the parallel content of many of them with any of the strategies mentioned in the previous paragraph. This way there are many possible and useful data that would contribute to a parallel corpus, but are not going to just because their bilingual content is difficult or even impossible to retrieve and pair.

In general, all these issues contribute to the sparsity of assorted data and the lack of parallel corpora of very large size, if compared to the average sizes of currently available monolingual corpora. Being intrinsic to the nature of the web itself, some of these problems have no real solution, while others can be somehow faced. For example, talking about the copyright issue, while sharing copyrighted data retrieved from the web is not strictly legal, sharing lists of paired URLs is: this is what Philip Resnik did with the URLs of pages he collected with STRAND[1]. But the webpage where he published this list was last updated June 2002 and these collections are now more than 10 years old, so most of the URLs are no more available, at least via direct link: web archiving resources like *The Wayback Machine* by the *Internet Archive* may solve the problem but it depends whether a copy of a page has been archived or not.

Also a limitation in most of the above mentioned papers is that very few of them perform an analysis of their results, mainly coming from those papers discussing the retrieval of parallel text from specific websites. So there is a stronger focus on the mechanisms to obtain parallel data from the web rather than what is possible to get with those methods. This is one of the aspects that this project proposal wants instead to further explore.

2.3 Genre, domain and document dissimilarity

Identifying the nature of texts is a major matter in corpus linguistics, as it can have several useful practical applications. As said in Sharoff (2007, 1), having documents classified into categories "can be used for various purposes, such as improving the relevance of information retrieval or selecting more appropriate

[1] http://www.umiacs.umd.edu/~resnik/strand/

language models in POS tagging, parsing, machine translation, or in word sense disambiguation".

Several document classification methods have been developed by different research groups, however it seems that two terms have become, among others, the object of particular interest: *genre* and *domain*, in the entire group of sometimes unclear and overlapping terminologies variously used by linguists or language scholars to define textual classes. Let us consider this in detail, in particular the two following sections will review 1) the background of studies on text variability, and 2) the main techniques to automatically understand the composition of a collection of documents.

2.3.1 Background of studies on text variation

Genre studies and text categorization issues have been object of investigation for decades (Adamzik, 1995), and they became a matter of great interest in corpus linguistics especially with regards to the possibility of performing automatic classification of large amounts of documents (especially when collected with unsupervised techniques). A number of approaches have been proposed during the years and the community has struggled to agree on a common way of defining classes. This is no surprise, since classification of documents can follow different strategies (e.g. supervised vs. unsupervised methods) and for different purposes. But still with the proliferation of different taxonomies came a certain amount of confusion about the basic definition of classification concepts themselves.

Lee (2001) explored this matter with his study on the BNC. Above all he mentions the work of Douglas Biber, since he introduced an analysis that distinguishes several textual dimensions, using quantitative methods and corpus analyses throughout a series of works, a relatively recent one being Biber & Conrad (2009). Lee discusses Biber's distinction between *genre* - seen as a categorization that relies on variables that are external to the text itself (e.g. audience or purpose) - and *text type* - given by internal criteria (Biber, 1988).

Talking about *register*, Lee seems to take from the analysis of Biber (1988) the close connection between this concept and genre, saying that they are

> two different points of view covering the same ground [...] with *register* being used when we are talking about lexico-grammatical and discoursal-semantic patterns associated with situations (i.e., linguistic patterns), and *genre* being used when we are talking about memberships of culturally-recognisable categories.
> (Lee, 2001, 46)

but, given that genre is more about whole texts while register regards language situations at the paragraph level, he concludes that

> I prefer to use the term genre to describe groups of texts collected and compiled for corpora or corpus-based studies. *(ibid.)*

Lee then considers the term *text type*, defining it as so vague and elusive that it can mean anything, and so not very useful for classification purposes. He clarifies also terms like *topic*, defining it as "what the text is about", and *domain*, described as the subject field. He says that "*genre* is the level of text categorisation which is theoretically and pedagogically most useful and most practical to work with, although classification by *domain* is important as well" (Lee, 2001, 37). He recognizes the importance of domain in building a balanced corpus, because "*domain* was probably the most important criterion used to ensure a wide-enough coverage of a variety of texts"(*ibid.*, 53).

A classification into domains can be quite intuitive and useful from a linguistic point of view (and also easy to perform in unsupervised ways with data-driven approaches, see next section). But in several occasions most of the attention is given to the concept of genre, as it seems to be slightly more problematic: Kim & Ross (2010, 146), talk about the "elusive nature of genre" even though "there is a shallow agreement that genre is a concept that can be used to categorise documents by structure and function" (*ibid.*, 153).

While *domain* seems to be more intuitive and easy to define[1], *genre* is not such an easy concept, and so the choice of genre classes has appeared to be quite disparate in the various contributions. However, it seems that some common trends can be identified: one direction may be the adoption of a classification

[1] And, to a certain extent, overlapping with the concept of *topic*.

based on "look'n'feel" labels, reflecting the practical use of a text (recipe, review, faq, blog post, academic paper etc.), as suggested by Sharoff (2010, 169). Even though intuitive, relying on the somehow shared definitions of text genre in a society and relying on inference and interpretation (Waller, 1987, 148), this kind of classification can be problematic in several ways: it assumes the existence of a stable and broadly shared palette of genres, but the use of a same label can change given different contexts such as cultural shifts between different languages and cultures (hence the significance for translation). Also when dealing with corpora from the web it is even more difficult to come up with a fixed set of genres, with the content of the Internet continuously changing and so with new rising and evolving typologies of text. For example, see how the concept of *blog* moved from being originally mainly related to define personal online diaries from being nowadays more a content management system format - and in a way moving from being a *domain* lookalike category to a *genre* one.

Another trend is to choose broader, functional classes as categories. The advantage of this approach is to work with concepts that correspond to major aims of communicative production, present in old and new textual forms without the problems of arbitrariness given by the look'n'feel approach. For example texts falling under the label 'instruction' can be internet FAQs as much as more traditional tool manuals. And still a classification like this can be very useful: texts sharing the same communicative intention most probably share similar sentence structures, particular uses of verbs and tenses, etc. - and so becoming useful when selecting training data with the same communication purpose of a particular document.

Automatic genre/domain identification turns out to be necessary to perform when dealing with large-scale classifications of hundreds, thousands or millions words corpora (furthermore, a systematic automatic classification guarantees consistency and replicability in a way that its human counterpart cannot give). This leads to the question how to automatically recognize and group texts under these classes. In the next section we are going to talk about the practical techniques used to discriminate documents belonging to different text types.

2.3.2 How to measure text variability

The machine learning paradigm can be considered one of the most prominent methods for text classification. Sebastiani (2002) gives an overview on the previous decades when a remarkable growth of interest on the possibilities of automatic classification of texts emerged, mainly due to technical progresses about availability of texts in digital format and their management. He describes how text categorization has moved during the '80s and '90s from *knowledge engineering* to *machine learning*, mentioning then the different choices that can be made about categorization itself (single- vs. multi-label, hard vs. ranking categorization) and its applications (document organization, text filtering, word sense disambiguation and categorization of web pages).

The field of study about topic modeling, well explained in Steyvers & Griffiths (2006), can provide a great help to understand the composition of a corpus. Topic models (like Latent Dirichlet allocation) are able to statistically analyse large collections of unlabelled texts, connecting words that occur together or in similar contexts and then creating clusters with these groups of words (topics/domains). This way it is possible to get an idea of the composition of a corpus from the content of the corpus itself, and obtain suggestions about how to organize categories.

Machine learning plays an important role also when the focus is on the detection of genres:

> The approach dominating automatic genre identification research is based on supervised machine learning, where each document is represented like a vector of features (a.k.a. the vector space approach), and a supervised algorithm (e.g. Support Vector Machine or Naive Bayes) automatically builds a genre classification model by "learning" from how a set of features "behave" in exemplar documents. (Santini *et al.*, 2010, 18)

Talking about these features:

> Many different feature sets have been tried out to date, e.g. function words, character n-grams, Parts of Speech (PoS) tags, PoS tags trigrams, Bag of Words (BoW), or syntactic chunks. *(ibid.)*

About the choice of the features to be selected for text classification purposes, it is worth to mention (Forsyth & Holmes, 1996), who underlines how crucial the selection of suitable linguistic markers is to get a working discriminant method. The suggestion of these authors is the employment of textual feature finding techniques that minimally depend on human judgement and relying more on data driven analysis. The authors support this hypothesis with an analysis of five systems of this kind.

So it seems that a reliable technique to understand the composition of a collection of texts is to learn from the data themselves the features that can help understanding their topics/domains and genres. Clustering is definitely a good starting point to retrieve keywords to detect topic/domains, but also detection of genres can be performed on the basis of recurrent linguistic patterns (i.e. POS trigrams or punctuation statistics) recognized as specific for particular genres. These patterns could be *flexigrams*, a term defining extended n-grams which are e.g. any combination of two words in a span of four (2/4gram), three in a span of six (3/6gram) etc. The use of this particular kind of n-grams for feature finding would be justified with the "the tendency for speakers and writers, as well as listeners and readers, to work with chunks of language rather than isolated words" (Forsyth & Sharoff, 2011). They can be used to build document dissimilarity matrices in order to locate the closest documents to a specific one, as will be shown in the next chapter.

2.3.3 Text varieties and MT

The main contributions about the issue of dependency on text typologies in SMT concern **domain adaptation**, which is defined as the way to face "the problem that arises when the data distribution in our test domain is different from that in our training domain" (Jiang, 2008). And SMT is one of the situations where this problem can arise: as previously shown there is a lack of assortment when it comes to choose which data to use to train SMT systems. Several parallel corpora

are freely available (Europarl, JRC-Acquis etc.), but most of them belong to very specific communicative contexts (like parliamentary proceedings) and so their applicability and usefulness when translating texts related to a different field may be limited. But they can still be exploited as training data, implementing them with portions of "in-domain" parallel data to tune an SMT system towards specific translation purposes.

During the Second Workshop in Statistical Machine Translation (ACL 2007), one of the main topics was precisely the employment of domain adaptation in a specific SMT task. Two papers explore the use of mixture modeling in SMT: Civera & Juan (2007) explore the capabilities and drawbacks of domain adaptation, employing a new mixture version of the Hidden Markov Model (HMM) then used to generate topic-dependent Viterbi alignments for a translation task with Moses. Also Foster & Kuhn (2007) describe various approaches to the mixture adaptation model[1] and, this is interesting for us, decomposing the training corpus into genres, because for the authors "this is the simplest way to exploit heterogeneous training material for adaptation" (*ibid.*, 130). Koehn & Schroeder (2007) show different experiments of domain adaptation on the same baseline SMT task (e.g. training it only on out-of-domain or in-domain or combined training data, different combinations of language models etc.). The results of these experiments demonstrate the importance of the in-domain language model to get good performances.

Many other papers focus instead on the exploitation of monolingual data used to get, via automatic translations, synthetic in-domain bilingual corpora: Wu *et al.* (2008) describe a situation where in-domain bilingual data do not exist, so out-of-domain parallel corpora are used to train a baseline system then implemented, as said, with in-domain automatically-translated monolingual corpora and in-domain translation dictionaries; Bertoldi & Federico (2009) conduct an experiment with similar approach, again synthesizing a bilingual corpus by translating monolingual adaptation data, and highlighting, as done by the above

[1] Cross-domain versus dynamic adaptation, linear versus loglinear mixtures, language and translation model adaptation, various text distance metrics, different ways of converting distance metrics into weights, and granularity of the source unit being adapted to.

mentioned (Koehn & Schroeder, 2007), that a key role to get improvements is played by the language model adaptation.

Another notable contribution is the one of Haque *et al.* (2009), that not only explore domain adaptation in SMT combining large out-of-domain and scarce in-domain training data, but also introduce the employment of clustering to extract sentences from out-of-domain data that are more similar to in-domain data, combining them with in-domain data themselves into a unified translation model. The idea of extracting phrase pairs that are supposed to be more relevant to the domain of interest from out-of-domain data is also shown in Foster *et al.* (2010), where the relevance to the target domain is delimited by discriminative instance weighting on phrase pairs similarity and their similarity or not to general language (even if most probably talking in terms of 'general language' is not completely correct, as it is used in this paper to define out-of-domain data in contrast to the specific features of in-domain language).

Niehues & Waibel (2010) propose an approach where factored translation models are used to integrate domain knowledge into a SMT system introducing a corpus identifier as additional factor; the result is a model that allows to understand if a phrase pair belongs to a certain domain. Daumé III & Jagarlamudi (2011) show how to face the problem of out-of-domain terms in an SMT system by mining unseen words, in particular employing an approach based in canonical correlation analysis, previously shown in Haghighi *et al.* (2008). The language pairs obtained this way are then integrated into Moses.

Seeing the several approaches that can be used to integrate small in-domain data with big out-of-domain quantities of parallel texts, it appears that some of them have tried to maximise the usefulness of domain adaptation employing some of the text classification strategies we have talked about in the previous section. Notably we have the case of Foster & Kuhn (2007), that opts for a classification into genres. Here the problem is that the term *genre* is used as-is, without defining what is meant with it, or even how the training corpus has been divided into genres. Or we have the case of Haque *et al.* (2009), where clustering is performed to find sentences that are more similar to in-domain data. In this case the description of the procedure employed is much more defined, but there

is not even one example about the result of these clusterings talking about the nature of the clusters themselves.

In general, the overall situation does not show a widespread use of text classification techniques, and much attention is given to the term 'domain' rather than any other. A recent paper by Gavrila & Vertan (2014) points out that very little attention is given to the concept of genre, and most importantly that using as training data texts having the same domain to certain documents that need to be translated but different genre may have an impact on the quality of SMT.

2.3.4 Less data for MT

The previous section reported some of the main strategies adopted to tune SMT systems towards specific translation purposes, and it has been pointed out how in general not much attention is given to this matter under the point of view of the domain/genre aspect of the texts involved. Another thing that has not been object of broad attention in this field is the chance to use small amounts of data when performing SMT training.

However some interest has been shown towards this possibility, like (Popovic & Ney, 2006) which pointed out the difficulties of dealing with the limited availability of large bilingual corpora for each required domain, along with the time and effort required to process them, while "small corpora have certain advantages like low memory and time requirements for the training of a translation system, the possibility of manual corrections and even manual creation". Some examples of the use of small amounts of training data are shown, comparing translations made from different size of Spanish-English translations (1K, 13K and 1.3M sentences task-specific corpora) and Serbian-English (0.2K and 2.6K sentences). These experiments show that, although the translation quality increasing with the size of the training corpus, understandable translations can be obtained from smaller task-specific training data as well, and if conventional dictionaries, phrasal books and morpho-syntactic knowledge are available these can be integrated to further improve the performances of these translations. Suggested uses of these translations are document classification or multilingual information retrieval.

Moore & Lewis (2010) point out the advantages of using accurately selected training data instead of large amounts of non-domain-specific data, in particular their approach is focused on the creation of language models: having already a small language model in the domain of interest, sentences from a non-domain-specific source are selected based on comparing cross-entropy having as reference the in-domain language model. Results are positive and this method has proven to be beneficial also in terms of reducing the amount of training data without a loss in terms of translation quality.

Axelrod *et al.* (2011) use a similar approach to the previous one, again focusing on the possibility of extracting sentences from a large non-domain-specific corpus that are most relevant to a target domain. This method is implemented as domain adaptation of an existing SMT system but the authors also explore the use of small amounts of training data. Their results show that "a domain adapted system comprising two phrase tables trained on a total of 180k sentences outperformed the standard multi-model system which was trained on 12 million sentences".

Orland (2013) interestingly further tests this approach on 14 different domains, and again the intelligent selection of subsamples of training data using ranking techniques proven to be successful in some cases, in particular he found that "for some domains, a significantly smaller amount of intelligently selected training data can yield BLEU scores nearly as high or higher than when all the available training data is used", and that "some domains, with fairly limited vocabulary and language variation, are better candidates for the technique described in this work than other domains".

These works have shown how, while the main trend is still the domain adaptation of publicly available corpora, the possibility of employing small amounts of parallel data is being explored as well, proving that it is worth to further explore this direction, which is the aim of the research here presented. In particular the approach here proposed wants to put a bigger focus on understanding the nature of candidate training data in order to sample very small amounts of appropriate training data. In these previously mentioned papers experiments have been conducted mainly on Europarl etc. In the next chapter an analysis of the English-Italian portion of Europarl is presented, together with an explanation

of the reasons which led to the creation of new corpora for the purposes of this study.

Chapter 3

Methodology of targeted corpus collection and analysis

3.1 Preliminary studies

3.1.1 Choice of tools and resources

As previously said, the rule "more data is better data" is usually applied in a situation of scarceness of publicly available, assorted parallel corpora. The main aim of this research is to investigate instead the possibility to maximise the use of small-sized training sets selected accordingly to the best match between training data and text(s) that need to be translated (and so doing also to explore the language variability in parallel data for SMT). So the first step is to get a corpus containing parallel texts belonging to a reasonably varied quantity of topics/genres.

The choice of the right corpus for such purposes depends on several factors, first of all the languages involved. The experiments for this project have been conducted on the English-Italian language pair because of the author's knowledge of both languages. Talking about the availability of parallel data, this language pair may be not as well resourced as e.g. English-French or English-Spanish, but there are still some corpora which provide a fairly large amount of data for this language pair.

Another crucial point when choosing the right corpus for the above mentioned purposes is assortment: we need a certain degree of text variability in order to make sense of the sampling procedure. Let us consider Europarl: while its size for the English-Italian language pair would ensure a remarkable quantity of parallel data for machine translation experiments[1], its text content needs to be analysed in order to understand how varied it is. The next section shows an analysis conducted on English-Italian Europarl, with some document dissimilarity sampling experiments.

3.1.2 Topic modeling analysis on Europarl

Europarl has been described in section 2.2.2, and its importance and usefulness as one of the main resources in the field of SMT has been repeatedly pointed out. One if its advantages is the legal and free provision of linguistic material with all the official languages used in the European Parliament (including our language pair of interest), in the form of monolingual corpora or parallel corpora with English-based pairs. This explains why Europarl has been considered as the first choice for the research here described, but it was necessary to verify whether or not its content was actually suitable for the purposes of this project. Europarl is made by transcriptions of proceedings of the European parliament, so it is reasonable to suppose a certain degree of repetitiveness in terms of register of the utterances pronounced by the members of the Parliament and the content of communications and matters discussed during the sessions.

In order to do that, Mallet (McCallum, 2002) has been employed to perform an unsupervised topic modeling analysis on the English side of the English-Italian Europarl corpus. Among the main customisations Mallet offers the possibility to choose the number of topics, the number of words per topic and use hyperparameter optimisation (i.e. allowing some topics to be more prominent than others, if it looks like some of them are). After several iterations, trying to avoid the generation of too many near-duplicate topics (i.e. topics appearing to be too

[1] 1,909,115 paired sentences, with a total of 47,402,927 Italian words and 49,666,692 English words (Europarl version 7).

similar), the adequate number of topics settled to 20. Results are shown in table 3.1.

N.	Keyword clusters	Assigned labels
1	report language parliament committee proposal mr amendments council rapporteur amendment position speaker support view legal community proposals proposed principle	*?*
2	european europe president union mr people parliament citizens treaty political today presidency time eu constitution member states future world	*?*
3	international people peace situation war aid united union military european resolution president security government support humanitarian mr country iraq	*Warfare*
4	european states member rights report data protection eu legal justice terrorism immigration law asylum citizens countries people union crime	*Immigration*
5	health research people programme european diseases tobacco drugs states disease member human public europe information patients care framework treatment	*Medical-addictions*
6	directive food products health environment safety proposal animal protection animals amendments consumer legislation waste water environmental consumers commission substances	*Food production*
7	economic euro financial european growth monetary bank stability crisis economy policy central states market currency pact countries markets investment	*Eurozone economy*
8	countries trade development world eu european agreement union developing economic international africa wto china aid cooperation negotiations agreements global	*Global market*
9	israel education european palestinian cultural people culture programme sport israeli young peace languages middle support palestinians europe media training	*Middle east*
10	council european union policy countries presidency mr president rights parliament political states office common agreement security enlargement summit process	*?*
11	commission mr president member european important time make commissioner made work states debate point parliament question council fact clear	*?*
12	transport safety european road air proposal tourism traffic rail report maritime directive europe sector mr sea environmental states passengers	*Transports*

13	budget commission parliament financial european funds committee programme policy year eur budgetary money report support council fund aid court	*Finance*
14	parliament mr vote report president committee amendment group european members procedure amendments house minutes rules mrs resolution voting rapporteur	*?*
15	report social women european policy eu states member employment people development work support voted europe economic writing union regions	*?*
16	report european union mr parliament policy council countries committee social economic europe community people president agreement treaty question employment	*?*
17	rights human people country president democracy political situation resolution government freedom china death democratic european eu mr respect elections	*Human rights*
18	eu european union turkey russia countries accession country ukraine russian negotiations turkish political relations enlargement romania report cooperation region	*Eastern Europe*
19	fisheries fishing agricultural policy sector report production proposal farmers support agriculture rural market measures european aid regions commission reform	*Primary sector*
20	directive market services european proposal report member competition states parliament internal companies legal workers public rights legislation regulation protection	*Legal*

Table 3.1: Topics and word clusters in English side of English-Italian Europarl.

Labels have been arbitrarily assigned based on human judgement, after having intuitively guessed what each topic is about according to the meaning of keyworkds for each group. Clusters with a *?* label do not denote any particular unbalance towards topics as specific as the other ones. They appear quite similar each other, mainly including terminology about the European parliament debates activity, such as *president*, *report*, *parliament*, *council* etc.

Also having a look at the distribution of single documents across the topics[1], it appears that not many documents show a strong unbalance towards the main

[1]Mallet outputs every single document as being assigned to different topics in different probability.

topic they have been assigned to, meaning that the affiliation of single documents to a specific topic is weak. One of the main reasons for this is probably the fact that sessions of the European parliament deal with different topics at one time; this means that this not a case where a corpus document always corresponds to a single argument.

It can be concluded that, although this experiment has led to a better understand of the composition of Europarl (at least of the portion of English Europarl aligned with its Italian counterpart), it has demonstrated that this corpus is for several reasons (too homogeneous, more topics per document) not the best choice to be used for the extraction of subsamples for SMT.

3.1.3 Document similarity experiment

The above reported analysis confirmed that Europarl may not be the best resource to be used in subsampling experiments where ideally a certain degree of text variability is needed. However, an attempt of selecting a subsample of training data for a specific translation task from Europarl has been made anyway. The purpose of this experiment was to check whether the use of similarity measures could be useful to extract a portion of documents which are recognised to be the most similar to a specific document, and use them as training data for SMT.

One of the easiest ways to measure the degree of similarity between two documents is using **cosine**. This is done by mapping words from the interested documents into vectors, then the cosine of the angle between the vector of the test document to be machine-translated and every vector of all the documents in Europarl is singularly calculated in order to compute their distance. Cosine measures are between 0 and 1, where values towards zero mean a lower similarity between the test document and a possible document to be used as training data to translate it. So the highest values can be chosen as (possibly) the most suitable to be used as training data to translate that specific document.

The test document has been randomly selected from the Internet, it is a journal article about a controversy in the Catholic church, dated 10 September 2009. It was written in Italian and provided with English human translation

(which has been used as benchmark for automatic MT evaluation); its size is around 1350 words per language on 47 sentence pairs[1].

This is the pipeline in detail:

1. Cosine similarity has been computed between the test doc and each one of the 6,216 Europarl files (on the English side of the EN-IT subcorpus);
2. Files have been sorted according to the cosine similarity score and the first 500 have been selected;
3. The result of this selection has been used to train a SMT system in Moses;
4. The obtained parameter file has been used to translate the test document (in the English-to-Italian direction);
5. Steps 3 and 4 have been repeated substituting the training data selected with cosine similarity with 500 other documents randomly extracted from Europarl, in order to obtain a term of comparison and validate the quality of the resulting translations;
6. MT evaluation of the quality of these translations has been carried out by using BLEU;
7. The whole above described process has been repeated in the opposite language direction (Italian-to-English).

The outcome of this analysis is presented in Table 3.2.

[1] http://chiesa.espresso.repubblica.it/articolo/1339977 and http://chiesa.espresso.repubblica.it/articolo/1339977?eng=y for the original documents.

Direction	Training set	Specs	BLEU score
IT>EN	500 most similar	303,615 sentence pairs ∼7M words per lang	27.5
IT>EN	500 random	117,973 sentence pairs ∼2M words per lang	26.1
EN>IT	500 most similar	303,615 sentence pairs ∼7M words per lang	26.5
EN>IT	500 random	117,973 sentence pairs ∼2M words per lang	23.3

Table 3.2: Results of the cosine similarity subsampling experiment with Europarl

Results confirmed that a selection of similar documents to a specific text that need to be translated gives better translation results compared to the same quantity of documents selected randomly - at least according to the automatic MT evaluation (with an improvement of 1.4 BLEU points for the IT>EN direction and 3.3 for EN>IT). However on the other hand this improvement (especially on the IT>EN direction) is counterbalanced by the remarkably smaller quantity of random data (less than half the size of the *ad hoc* training sets). This is most probably due to the fact that cosine similarity calculates the similarity between two documents as a single whole feature, without distinguishing between the several aspects that make a texts similar to another one, like terminology, sentence structure, grammar, length etc. So chances are that training data possibly appropriate and useful in terms of genre or domain are discarded from this kind of selection maybe because their size differs from the one of the considered text.

3.1.4 Observations

In the two previous paragraphs the outcome of a topic modeling analysis of the English side of English-Italian Europarl and a document dissimilarity subsampling experiment for SMT have been presented.

The topic modeling analysis ended up being useful to understand the composition of Europarl and to decide whether to use this corpus in the SMT subsampling experiments or not. Even though a certain variety of topics is present in Europarl,

it also emerged that its textual content is very homogeneous and repetitive - not surprisingly since the corpus is composed exclusively by transcriptions of parliament proceedings. Another technical problem is the fact that single documents in Europarl are presented in the form of sessions (sometimes split in more than one file), where the object of the debate can be more than a single one, which is not very compatible with the requirement of having a consistent labelling for each possible candidate training text. For these reasons Europarl is unsuitable for a research which aims at exploring text variability on different axes of text variability.

However this subsampling experiment has shown that it is possible to employ a document similarity measure to select the most suitable training data for a specific task: even just the employment of a simple measure like cosine similarity on the 500 most similar texts to a specific document has shown an increase of BLEU score compared to a same quantity (in number of single documents) of randomly selected training texts. However the limits of this success can be noticed looking at the size of the random sample, which is much smaller (in number of words) than the *ad hoc* subsample.

All this led to two main decisions: firstly, not to further employ Europarl in the main project experiments because ideally a corpus containing a wider assortment of text variety across both the domain and genre dimensions would be employed. A possible alternative solution is to collect a new parallel corpus from the web, as shown in 3.2.

Secondly, the use of document similarity measures to subsample training data for document-specific SMT has been proven a useful strategy. However the limits of cosine similarity have been pointed out, so it is necessary to explore the use of another measure which may be more suitable for our purpose of finding the most useful training data for a particular task. One possibility is to use a (dis)similarity measure strategy that makes use of specific linguistic features (see 3.4).

3.2 Parallel corpus collection from the web

In section 3.1 a series of experiments on Europarl has shown the unsuitability of this corpus for SMT subsampling experiments, leading to the decision of col-

lecting a brand new parallel corpus for the study described in this thesis. This section reports in detail the procedures followed to collect this new corpus, reimplementing some of the strategies described in 2.2.3.

In 3.2.1 a first experiment to collect a multilingual corpus from a specific website, exploiting the functionality of RSS syndication, is reported. In 3.2.2 the method eventually used to collect the new English-Italian parallel corpus is described.

3.2.1 The "RSS method"

As shown in in 3.1 it is possible to adopt different strategies to collect parallel corpora from the web. According to Fry (2005) the collection of parallel documents from dynamic websites providing RSS syndication of their content is possible. This strategy is quite simple to perform, since it mainly depends on two factors:

1. The extraction of articles (in one of the languages of the considered website) from the RSS stream - essentially an XML file, and

2. The identification of their equivalents in the second language - which can be performed consistently to every article once located the link to the other language version in a single one, since they all come from the same website and have an identical HTML structure.

This can be easily reimplemented (and customised) from scratch employing Unix tools and scripts (Fry mentions `rss2email`, `procmail`, `grep` and `wget`).

This strategy has been tried on *Presseurop.eu*, a "news website publishing a daily selection of articles chosen from more than 200 international news titles, then translated into ten languages (English, German, French, Spanish, Romanian, Italian, Portuguese, Dutch, Polish and Czech)"[1]. The pipeline consisted in the following steps:

[1] `http://www.presseurop.eu/en/about`. The website ceased to publish updates on December 2013 due to budgetary reasons, but the project was resumed by volunteers on May 2014 with the name *Voxeurop.eu*. The corpus collection here described was performed before the closure of Presseurop.

1. After downloading the RSS page from the website[1], only those URLs that actually contain news article pages have been extracted from the HTML code of this page.

2. This RSS page was continuously updated with links to the last 50 articles published on the website. In order to get more links, John Fry decided to set up a pipeline of regular expressions which allowed to collect 329 articles through 5 weeks of RSS feed subscription via e-mail. However this method can be time consuming if one wants to collect a larger amount of data, so he enriched his corpus crawling past versions of his previously considered RSS feeds via archived versions of those same RSS page. Similarly the *Wayback Machine* provided by the *Internet Archive*[2] has been used with the purpose to try to collect more pages contained in the *Presseurop.eu* website. This way 97 available previous versions of the *Presseurop.eu* RSS page have been recovered. URLs containing articles from each one of them have been extracted and added to the original list of 50 links. After sorting and eliminating duplicates a list of 1920 URLs of articles has been obtained.

3. As previously said, *Presseurop.eu* published articles collected from several national online newspapers and magazines, providing translations in the other 9 languages for each one of them. Following a brief analysis of one of these pages, a Bash script has been written in order to extract links of these translations from English versions of every article and output them in a tab-delimited format.

4. The last step consisted in the corpus collection process itself: plain text content of each URL was downloaded with jusText (Pomikalek, 2011) and single *English-other language* parallel corpora were aligned with Hunalign (Varga *et al.*, 2005) (See Steps 3 and 4 in 3.2.2 for a more detailed description of these two tools).

[1] http://www.presseurop.eu/en/feed/rss/all.xml.
[2] http://archive.org/web/.

The final result is a multilingual corpus of 1920 single documents, translated in ten different languages and provided as a series of bilingual corpora aligned to the English side of the corpus.

Corpus	N. docs	N. sentences	Eng words	Other lang words
En-Cz	1920	24,941	473,242	380,241
En-De	1920	23,182	451,546	386,434
En-Es	1920	24,076	477,003	496,629
En-Fr	1920	26,375	515,130	520,511
En-It	1920	23,285	477,243	437,836
En-Nl	1920	26,452	486,594	497,725
En-Pt	1920	26,748	508,753	501,730
En-Ro	1920	26,189	500,139	484,382

Table 3.3: Composition of the Presseurop multilingual corpus.

This corpus has been collected to see whether the overall text variability in the documents provided by this website is wide enough for the main research purpose of this project (i.e. the study of extraction of subsamples for training SMT systems based on text variety criteria). The employment of the RSS strategy - and the multilingual corpus collected following this method - demonstrated its validity and practice as a convenient method to collect parallel corpora from the web (in particular from a specific website and, where possible, an extended multilingual corpus). But again it is necessary to check whether these corpora provides a reasonably wide variety of different typologies of documents in order to be used in the following SMT experiments. Texts have been extracted from a news website, so while it is reasonable to expect a varied coverage in terms of topics[1] the variability in terms of genres should be much more restricted around the journal/magazine article type.

Once again topic modeling analysis has been used to understand the composition of a corpus, repeating what was previously done with Europarl: unsupervised

[1]The website categorises its content under six different subjects: Politics, Society, Economy, Science & the Environment, Culture & Ideas, Europe & the World.

clustering of 20 topics feeding the English side of the *Presseurop.eu* corpus into Mallet. Table 3.4 reports the result of this analysis:

N.	Keyword clusters	Assigned labels
1	city belgium belgian flemish de town local residents capital art tourists region mayor village museum million building cities cultural	*Belgium*
2	energy nuclear power oil companies industry gas food production europe plants sea million green waste company year electricity farmers	*Energy-environment*
3	world history war language culture society years past today cultural life book church century modern english revolution human communist	*Culture*
4	german germany merkel berlin angela chancellor germans sarkozy die der federal zeitung minister summit spiegel euro nicolas franco press	*France-Germany*
5	romania romanian roma police court law bulgaria corruption state justice data secret bucharest information internet crime rom reports bulgarian	*Romania-Bulgaria*
6	russian russia war nato kosovo military eu serbia country soviet moscow croatia president defence foreign belarus years bosnia putin	*Russia-Yugoslavia*
7	people don countries greece country germany euro eu money german agree uk understand make politicians good system years greeks	*Germany-Greece*
8	people time back years world long day country don good put year end days ago make europe place things	*?*
9	italy turkey italian libya border arab turkish mediterranean gaddafi migrants berlusconi foreign refugees europe military north france eu africa	*Italy-Libya*
10	sweden police danish film swedish movement protest streets people denmark demonstrations protests demonstrators demonstration october youth football daily programme	*Sweden-Denmark*
11	europe european eu countries union political states economic crisis national world policy power time future leaders france common euro	*EU crisis*
12	spanish spain portugal portuguese el madrid daily lisbon pa blico de jos mundo zapatero la crisis notes barcelona paper	*Spain-Portugal*

13	greece euro debt greek crisis banks financial eurozone bank austerity markets government economic country billion economy tax countries money	*Greek crisis*
14	european eu daily commission brussels government member states time country notes parliament minister council reports issue points europe state	*Europarl*
15	people years country work young social year workers million euros market home number labour job jobs children population working	*Job policies*
16	irish ireland british britain uk london scotland cameron independent times daily dublin english david guardian government headlines independence news	*UK-Ireland*
17	party government political hungarian hungary elections minister parliament prime vote left parties country democracy power media law democratic majority	*Hungary*
18	french france sarkozy le paris president nicolas international lib ration chief jean american italian left head mario monde hollande	*France*
19	polish poland warsaw gazeta china chinese poles wyborcza czech daily rzeczpospolita smoking dziennik prawna notes europe beijing ski explains	*Poland-Czech-China*
20	dutch immigration immigrants people women netherlands society country wilders debate muslim party muslims social islam integration live pvv anti	*Netherlands-Islam*

Table 3.4: Topic modeling analysis on presseurop.eu.

The results of this analysis shows that most topics correspond to news related to specific countries or zones of Europe (or specific matters involving a specific country or groups of countries), but some of them are more general (8, 11, 14, 15). However the situation is overall similar to the one we met with Europarl: there could be a certain degree of text variability in terms of topics/domains (even though not much), but since all the documents come from a single website containing translated versions of journal and magazine articles there is a strong unbalance towards a journalistic style of texts. At this point the solution would be the collection of a parallel webcorpus composed from texts coming from different websites, as shown in the next sections.

3.2.2 The BiTextCaT pipeline: steps and tools

This chapter so far reported the first steps conducted in the direction of finding a corpus suitable for subsampling SMT experiments: in 3.1 the employment of an already existing resource like Europarl has been considered, since this corpus has the advantage of being a freely available resource providing a reasonably large amount of data. Its content has been evaluated through topic modeling analysis and tested with an SMT subsampling experiment based on cosine similarity, and this led to the decision of not adopting it because a corpus containing a wider variety of texts across different text variability axes would be preferable: even though Europarl has proven to contain a certain amount of different subjects, the very codified verbal interaction employed in parliament sessions prevents the presence of a big variability in terms of communicative situations.

This led to consider the possibility of building a new corpus from the web. Thus previous attempts have been considered (reported in section 2.2) and in particular the RSS method has been reimplemented. This strategy has proven to be easy to perform, providing a practical way to collect and pair multilingual texts from the Internet (as shown in section 3.2.1). But collecting multilingual documents from a single website led (not surprisingly) to a similar situation to Europarl, i.e. the lack of variety of documents on different axes of text variability, which is an essential requirement for the purposes of this research.

Even though it was quite predictable, through these two attempts we verified that getting parallel data from one single source is not suitable for the purpose of studying text variation (and its impact on SMT). In order to compile a bilingual corpus made out of material available on the Internet which would satisfy the required criteria, the most appropriate solution is probably to perform automatic crawls on the web and obtain documents coming from a variety of multilingual websites. In order to do that, advanced functionalities provided by search engines have been employed to locate webpages that potentially belong to multilingual websites (see section 2.2.3).

This strategy mainly consists in 4 passages: 1) Mining the web to locate candidate pages in L2 (the second language), 2) retrieval of page counterparts in L1, 3) pages content download and 4) sentence alignment. The following

paragraphs describe in detail every step of this pipeline, which have been used to build an Italian-English (and later a German-English) parallel corpus.

Step 1: URL collection

This first step involved the location of translated pages on the Internet. In order to do that, the original pipeline by (Resnik & Smith, 2003) used to rely on Altavista APIs, which are now discontinued, so the method had to be reimplemented with other search engine APIs publicly available at the moment of the corpus construction (similarly to what was made by Mohler & Mihalcea (2008) with Google). At the moment of this corpus collection (first trimester of 2012) Microsoft Bing was available. Its search engine API have been employed through its implementation in the *UrlCollector.jar* module which is part of BootCaT (Baroni & Bernardini, 2004), a freely available toolkit containing a series of scripts to collect and process monolingual corpora from the web. *UrlCollector.jar* requires as input a plain text file containing a list of words, one per line (or more than one in case of n-grams), which are then used as search terms into the search engine, and gives as output a list of URLs containing the top results (default settings are 10 results per query). So this script allows to automatise the process of collecting results of queries on a search engine in a way that it would be otherwise difficult to implement[1].

This strategy to discover and collect URLs from bilingual websites follows some suggestions from previous attempts in this direction: assuming that national top level domains are expected to have sites in the language of respective countries (Ma & Liberman, 1999) and that a translated text is usually located along with its original version in the same website (Chen & Nie, 2000). For example the search for English language content in an Italian website it is likely to lead to texts that could most likely be the English translation[2] of an Italian

[1] Even though not impossible, for example the use of the command line web browser Lynx has been considered as an alternative.

[2] As shown in Atwell *et al.* (2007), even though American English is expected to be the dominating variety of text content on the web, apparently British English is widely present as well, but this difference is noticeable in a small amount of cases. There are no studies in particular on whether the British or American variety is prevalent in English webpages of Italian webistes.

content contained in the same site. So Italian national top level domains have been crawled looking for this specific type of pages. This was done by using the advanced search options of search engines APIs, provided as operators to be typed in together with search words[1]. The operator `site:` has been used with the `.it` option to restrict the search to Italian websites only, and the operator `inanchor:` with the options `en`, `eng`, `english_version` to find pages with these language specific recognizers in their URLs. The maximum number of URLs per query has been set to 50 (the maximum allowed) and the 1000 tuples from ukWaC (Ferraresi, 2007) have been used as search queries. This list is made by random combinations of "basic words and mid-frequency words collected from other corpora"[2] (in this case the BNC) and it is suitable for the collection of a general-purpose corpus (in the sense of being varied in its content on several axes of text variability). These are the first 20 lines of the list:

```
grey gently
drawing totally
path eating
watching explanation
dealt lack
radical organised
relationships studied
gets accused
conservative hoping
realise increasing
unions pure
culture stories
violence cottage
noise glass
tape easily
gate flowers
```

[1] A full list of these operators for Bing is available at http://msdn.microsoft.com/en-us/library/ff795620.aspx.

[2] http://wacky.sslmit.unibo.it/doku.php?id=seed_words_and_tuples.

```
choose lake
bottom accommodation
colleagues article
corner network
```

Firstly the string `inanchor:eng site:it` has been added to each line of this list, then the search with *UrlCollector.jar* on this file has been performed.

Then this search has been repeated with `en`, and `english_version`, in order to increase the chances of finding more multilingual websites, using different language identifiers. This way three lists of URLs have been obtained, one for each considered language tag, and they have been merged and alphabetically sorted and cleaned from duplicates. The next step was the retrieval of the other language counterparts, which is described in the next section.

Step 2: Retrieval of corresponding pages in Italian

A list of URLs of webpages in English collected from Italian websites has been obtained from crawling the Internet, and chances are that a certain amount of these URLs are likely to contain English translations of Italian pages coming from their website of origin. At this point it was necessary to retrieve the Italian language counterparts of each URL, having as result pairs of URLs whose content should be the same, but in two different languages. This was the most challenging task since there is no standard for the construction of multilingual sites and webmasters structure the bilingual contents in many different ways. For this reason previous works suggested several strategies to perform the current task: looking for parent or sibling pages (defined by Resnik & Smith (2003) respectively as "a page that contains hypertext links to different-language versions of a document" and "a page in one language that itself contains a link to a version of the same page in another language"), automatic language identification, URL substitution rules, page content and/or structure matching etc.

Reimplementing all those strategies from scratch would have been not only very time-consuming but also the creation of a pipeline for the collection of multilingual corpora from the web is not the primary subject of this thesis. So, even

though this corpus collection process is crucial because a bilingual resource with certain features (i.e. a certain degree of text variability) was needed in order to conduct our SMT subsampling experiments, not all the possible, previously strategies existing in literature have been recreated, only those which were more likely to fit this pipeline and make it work properly. Since the previous step was mainly based on the extraction of language identifiers in URLs, presumably that the most useful strategy would be 1) try to change the current language identifier with the most likely correspondents in the other language and 2) test the new URL in order to validate its existence.

An example of how this strategy works is this:

`http://sitename.it/?lang=en > http://sitename.it/?lang=it`

In order to do that (and to maximise the possibility to retrieve as many candidates as possible), a manual analysis of the URLs has been performed to understand what are the most recurring language identifiers in URLs and how to get the translated page in the other language - for example while for most of them the substitution of the English tag with the Italian one is enough, in some cases the deletion of the language tag had to be performed instead. So it was necessary to go through the URL list in order to identify the most common and widespread language tags as they appear in multilingual website URLs (and how they change from a language to another). Based on this analysis some generalisations have been made and included in a script performing an "if-then-else" reiteration where every URL from the list created in Step 1 is subject to these substitutions, tested in order to verify the existence of its counterpart in the other language - or marked as "can't find it" if the test is unsuccessful after trying all the possible combinations. Results have been stored in a tab-delimited text file alongside the original URLs.

At the end of this process all the "can't find it" lines have been discarded and only those pairs that the script managed to retrieve have been kept. During this passage the URL list shortened considerably, the reasons being both the actual absence of a counterpart in Italian for many texts (i.e. the document was published on a Italian website but only in English without being the translation

of any Italian content) or the impossibility to automatically retrieve the Italian version due to the structure of websites. Apparently this strategy is more likely to work with pages coming from dynamic websites, which are more consistent in the use of language identifiers in their URLs (like the example above).

This step has been the most challenging, and its main drawback is probably the fact that it required a lot of manual checks in order to ensure the retrieval of actual parallel webpages. It is not the purpose of this thesis to go more in depth on this direction, but for future attempts at further developing the current task in this direction is advisable.

Step 3: Plain text extraction

The extraction of the main plain text content from webpages - removing all the irrelevant material like boilerplate (menus, sidebars, embedded content etc.) and non-textual elements - is notably one of the main practical issues when compiling a corpus from the web (see Baroni *et al.* (2008)). JusText (Pomikalek, 2011) has been already mentioned in section 3.2.1, it was employed when collecting the multilingual Presseurop corpus with the RSS method. It accepts as input an URL and processes its content, deciding whether a chunk of text contained in a particular webpage has to be considered as part of its relevant main content or not, based on a series of customisable heuristics (length, quantity of stopwords, etc.). JusText has been used on this occasion as well, and in order to keep a consistent download of the content of translated pages (the possibility to customise the selection document by document was unfeasible due to the high volume of data), the default settings with only the language customisation have been used.

JusText has been included in a script able to collect each document saving them in plain text format in two folders, one for each language, where translated documents have been evenly named. At the end of this process the result is a parallel corpus aligned at the document level.

Step 4: Sentence alignment

At this point the corpus collection stage is completed, but most of the tasks which a parallel corpus can be used in (including SMT) require it to be aligned

at the sentence level. Even though this is the last step of the process of making a parallel corpus it is one of the most important, since an incorrect sentence alignment would further lead to translation faults. In particular an essential pre-requirement is to have a correct sentence split, in order to have one sentence per line, which is then paired with its second language counterpart (or counterparts in case the alignment is asymmetric).

Sentence split has been performed using the NLTK package Punkt (Kiss & Strunk, 2006) and then documents have been aligned at the sentence level with Hunalign (Varga *et al.*, 2005).

3.2.3 Final corpora

This pipeline has been performed on the `.it` top level domain, to obtain an English-Italian corpus. However it would have been interesting to test it also on a different language pair. So the whole pipeline has been repeated on the `.at` and `.de` top level domains and substituting Italian language URL identifiers with German ones in order to build an English-German parallel webcorpus. This corpus has not been used in the final SMT experiments but its content has been analysed and results about are reported in 3.3 along with the ones about the English-Italian corpus for the sake of completeness.

Table 3.5 shows the number of documents, sentences and words of the final parallel corpora. Data obtained from the Austrian and German national top level domain are presented separately.

Corpus	**N. docs**	**N. sentences**	**Eng words**	**Other lang words**
ENGITA	2,932	109,156	2,213,599	2,172,191
ENGER (at)	710	30,542	455,586	404,240
ENGER (de)	5,009	22,2186	3,352,269	3,017,548

Table 3.5: Composition of the Italian-English and German-English parallel corpora.

3.3 Text classification: Topics/Domains

3.3.1 Understanding the composition of our corpora

Two parallel corpora have been compiled using the strategy described in the previous section. They have been built in a semi-supervised way on a list of random words combinations (which was created with the purpose of being used for the collection "general-purpose" corpora). This means that there was little control on the final composition these corpora, the main limitation being the behaviour of the search engine employed for their collection, i.e. it is not possible to know exactly how Bing makes certain pages appear as top results. So at this stage it is possible to guess that these corpora cover different web varieties, but a deeper analysis of their content give a clearer idea of the kind of bilingual documents the web contains at least the Italian-English and German-English webspace explored through the national top level domains of Italy, Germany and Austria.

3.3.2 Probabilistic topic modeling

Again Mallet has been used to perform a topic modeling analysis on the English side of the corpora. Just as reminder, the output of Mallet are clusters of words that frequently occur together in the corpus they have been extracted from, and based on them it is possible to assign a label to each cluster. After several attempts this time the number of 10 topics has emerged as the most informative.

N.	Keyword clusters	Assigned labels
1	music fashion art world work film style time years italian great young collection love show york model year life	*Fashion and lifestyle*
2	italian italy international years development european world year university public economic research social group system company cultural state national	*Economy-statuses*
3	hotel sea area city day wine centre rooms di visit offers island km rome located room beautiful park place	*Accommodations*
4	church god pope life world faith vatican catholic benedict christ holy xvi time council religious jesus great cardinal people	*Catholicism*
5	inter time team don good back ll people content season world made great purposes make marketing prices find didn	*Sport-football*
6	production engine car pic water produced system high version pics products quality product made company realized hp cc models	*Motors-Mechanics*
7	data information order site user website page personal photo set details time web system version access code list software	*Software*
8	century church di city ancient town built san museum roman st art building castle back important part area works	*Monuments-Attractions*
9	goku body stars earth time dragon people energy star moon sun vegeta planet ball years children year treatment power	*Kids entertainment*
10	china islam muslim samir violence beijing religion party today bishops country fight corruption challenge church khalil dissident maoism children	*World-Religions*

Table 3.6: Topic modeling analysis on Engita.

Recall that this topic modeling tool automatically creates clusters of words that frequently occur together in a given corpus. And when collecting a corpus from the Internet relying on search engines APIs it is likely that a certain quantity of pages comes from recurring websites (mainly because those websites have an high ranking on search engines). So it is no surprise that some of the topics presented in table 3.6 originated mainly (in certain cases exclusively) from specific websites: Topic 1, with words suggesting that a portion of this corpus is made of pages about fashion/lifestyle, most likely originated from 209 documents coming from the Italian online version of the magazine *Vogue*. Similar examples are Topic 4 (with 172 documents coming from a journalistic blog about the Catholic

church), topic 5 (125 texts from the Inter Football Club website), topic 10 (with the remarkably large number of 452 documents coming from a news missionary website). Some topics appear to correspond to specific websites even with smaller amount of pages, like topic 9 (from a website called *Dragon Ball Arena*) and another example could be topic 6, which may be due do the presence of 27 pages coming from an amateur website about vintage cars from Easter Europe, but we also have the presence of other pages about motors.

The remaining topics instead collect contributions from different sources of the same kind: Topic 3 looks quite homogeneously originated from webistes of hotels and Topic 8 comes from several pages describing artistic heritage, monuments, art exhibitions and places of cultural and tourist interest. Topic 7 collects contributions from pages related to IT services of various nature (tutorial, electronic specifications etc.) and Topic 2 is the least intuitively clear, but it is related to pages about economy, statuses of organisation, academic regulations etc., in general collecting official documents.

N.	Keyword clusters	Assigned labels
1	linz art ars electronica world city space center exhibition architecture works page work festival design time project artists year	*Arts-tourism*
2	austria austrian university european research international students information europe vienna years eu development countries public march english year application	*University*
3	hotel ski austria salzburg city family holiday mountain rooms area enjoy located offers winter lake vienna offer children day	*Accommodations*
4	roma anthem national people groups time written group languages life called song language mat world great family term maximoff	*?*
5	lga backplate compatible gigabyte ga asus mainboard noctua mounting support contact kit amd preinstalled msi ds asrock tattoo socket	*IT hardware*
6	file chiles chile sharity windows stereoscopic server left hot species video directory user decoder flavour install orange files player	*Software*
7	market domain information data system trading software exchange time number company service application mail free nic registration exaa energy	*Software 2*
8	society marcuse social technology political means production art culture change order time work labor state existing result technological capitalism	*Politics-society*
9	vienna professor music year philharmonic concert austria century austrian world st upper orchestra opera sch musical years festival	*Music*
10	fan cooler high system noctua low mounting nh cpu fans noise control pressure mm mainboards speed quality performance pwm	*IT hardware 2*

Table 3.7: Topic modeling analysis on Enger.at.

The more limited size of the Austrian corpus produces some redundancy in the topic modeling analysis: we have two very similar topics (6 and 7) about IT software and other two about IT hardware (5 and 10). Topics 1 and 9 appears to be related to the world of arts, with the first one collecting pages on events like exhibitions or art festivals and the second one more focused on music. Topic 2 originates from pages of academic nature, with university regulations etc., while Topic 3 recalls pages from hotels and other kind of tourist accommodations and services. Topic 8 is originated from pages containing cultural essays, and Topic 4 was difficult to determinate, but we can say it collects texts of cultural interests (essays, pamphlets).

N.	Keyword clusters	Assigned labels
1	data file software version system check information user files windows time server open program source work application test support	*Software*
2	company business companies production products market energy management development technology quality industry germany services customer service gmbh systems german	*Business*
3	power high control system time sound output range frequency low audio signal level voltage light input space digital supply	*Electronic*
4	berlin film art world german years theatre work exhibition festival works history year time music international museum war germany	*Culture-turism*
5	team world time year swing place pilots flying competition show top flight glider aachen good km pilot great german	*Flights*
6	research german germany university education international federal training students project information study programme development institute funding science cooperation universities	*University*
7	hotel city area day time find site free tour people park station place visit information offer rooms room water	*Accommodations*
8	mm high made engine car system weight design parts model air special performance front cm case side steel vehicle	*Motor-mechanics*
9	music piano sound people time digital life body world work playing action musical make sounds grand kawai touch works	*Music*
10	countries country migration law people migrants eu germany population european number states government social policy economic criminal legal immigration	*Law-society*

Table 3.8: Topic modeling analysis on Enger.de.

In the corpus collected from the Germany top level domain we can see topics similar to the ones originated from the previous two analysis: software and electronics (Topics 1 and 3), hotels and tourist services (Topic 7), cultural events and attractions and music (Topics 4 and 7), academic pages (Topic 9). Topic 8 is originated from pages related to the car-motor industry, Topic 2 from corporate and industry pages and Topic 10 from articles about business.

3.3.3 General conclusions

Even though these three corpora differs in relation to certain features - Engita and both Enger because of the different languages, Enger.at and Enger.de because of the different size and country of origin - it is possible to notice some recurring topics: tourist accommodations, art and culture, IT and corporate pages. These topics allowed to understand what kind of documents are contained in these corpora, and the nature of the websites they come from. This way what can be reasonably considered some of the main areas of the Italian-English and German-English bilingual webspaces have been located. They correspond to certain kinds of activities, entities and services which target a wider audience than just their native, i.e. tourism or the offer of other kind of services, from education to private and corporate business.

Also this analysis allowed us to understand whether these corpora collected from the web contain a certain degree of variability in terms of text types to justify their use in a SMT subsampling experiment.

3.4 Document dissimilarity analysis

The topic modeling analysis was useful to understand what kind of documents compose these new corpora - and in general it has provided an overview of the most common typologies of multilingual websites in Italian-English and German-English. The main purpose of this project is test the possibility of extracting subsamples of candidate training data for SMT from parallel corpora according to the nature of each single document one wants to translate. In section 3.1.3 a subsampling experiment employing a similarity measure has been reported, showing that this may be a possible path to follow, but the limitations of cosine similarity have been pointed out as well. So we decided to use similarity measures taking into account certain linguistic features that possibly would help as well, as shown in the following sections of this chapter.

3.4.1 Flexigrams-based analysis: the Teaboat suite

Classification of documents into genres is usually performed based on various linguistic features, such as POS-trigrams (Sharoff, 2010). According to Forsyth & Sharoff (2011) "speakers and writers tend to work with chunks of language rather than isolated words", and this led to the decision of rather trying an "enhanced version" of normal n-grams. Some of the functionalities contained in the Teaboat suite (Forsyth & Sharoff, 2014), a toolkit originally designed for terminology extraction from comparable corpora based on flexigrams, has been employed for this purpose.

The word "flexigram" refers to a particular type of lexical bundle, defined as "a sequence n/m-gram where n is a group of tokens occurring within a segment of m tokens, where m n " (*ibid.*). Flexigrams are a useful linguistic feature to capture the variation of language in situations like

thank you mister John
thank you very much
thank all of **you**
I say **thank you**

where the two words *thank* and *you* are variably located in chunks of four words, and not necessarily in contiguous positions. This is a 2/4 flexigram combination.

The original purpose of Teaboat was to analyse the content of comparable corpora and extract translated words and multiword expression from them, creating intra-corpus dissimilarity matrices in order to carry out such task. But then dealing with multiword expressions is a major matter also in machine translation, in particular with phrase-based SMT where segments of words are the basic unit of the translation process. So it was appropriate to employ Teaboat to discriminate documents in Engita and Enger, creating document dissimilarity matrices based on different flexigrams combinations. The possibility to measure the distance between all the documents in Engita and Enger provides a complete view of how documents are spread (even with a graphic output in a bidimensional space). And, given a document one wants to translate, it can be included in the

corpus and understand what are its most similar documents by performing this analysis; then subsample and use them to train ad-hoc SMT systems.

3.4.2 Dissimilarity matrices generation

The document dissimilarity analysis has been performed using the scripts contained in the Teaboat toolkit on the English side of Engita and Enger, trying three different basic combinations of flexigrams: 1/1 (single words), 2/4 (bigrams in spans of 4 words), 3/6 (trigrams in spans of 6 words). As an example, below it is reported how the whole pipeline worked with 2/4 flexigrams on Enger.at. The steps to create an intra-corpus dissimilarity matrix are the following:

1. First of all Teaboat needs a metafile which allows the program to look into our data (in specific a lemmatised version of the English side of Enger.at, in a folder containing one file per document), and this metafile can be generated with a dedicated script. Teaboat provides also a script to remove duplicates and near-duplicates. Duplicates have already been discarded during the corpus collection stage, but the dedicated Teaboat script has been run on the corpus as well, in order to get rid of the near-duplicates. The number of texts dropped from 711 to 678 after this passage. According to the manual "the default similarity scoring function uses Pearson's reciprocal-rank similarity measure, on character n-grams, defaulting to a gramsize of 4".

2. The second step involves the identification of flexigrams, in what was originally referred as the *feature finding* stage. Clearly the number of combinations of groups of two words in spans of four may be huge, so the dedicated script "saves only the most frequently occurring N flexigrams from each document, where N is the rounded square root of the number of tokens in that document". The program reads as input the metafile created in the previous passage and outputs a list of the most frequent flexigrams in the considered corpus, sorted and provided with some statistics.

Flexigram	Span size	Rank	Raw frequency	% rate
('the', 'of')	4	1	6970	1.56200
('of', 'the')	4	2	5771	1.29330
('in', 'the')	4	4	3698	0.82873
('the', 'the')	4	3	3774	0.84576
('to', 'the')	4	5	3027	0.67836
('and', 'the')	4	7	2683	0.60127
('be', 'the')	4	6	2906	0.65124
('the', 'be')	4	8	2512	0.56295
('the', 'and')	4	9	1868	0.41862
('be', 'to')	4	10	1626	0.36439
('be', 'a')	4	11	1530	0.34288
('a', 'of')	4	13	1464	0.32809
('of', 'and')	4	12	1481	0.33190
('be', 'in')	4	14	1407	0.31531
('for', 'the')	4	18	1202	0.26937
('on', 'the')	4	17	1291	0.28932
('at', 'the')	4	21	1138	0.25503
('the', 'in')	4	19	1154	0.25861
('and', 'be')	4	22	1101	0.24674
('in', 'of')	4	23	1077	0.24136

Table 3.9: first 20 results of the feature finding analysis.

In table 3.9 the content of the columns is 1) flexigram, 2) span size, 3) rank based on raw frequencies, 4) raw frequency of each flexigram in the corpus, 5) percentage rate of the occurrence of each flexigram in the corpus.

3. The last step involves the creation of the dissimilarity matrix. The dedicated script takes as input the flexigrams found in the previous passage and applies them to our corpus, working in self-test mode, i.e. the script computes dissimilarities of each document of a given corpus against each other document of the corpus itself. The output is a tab-delimited format text file where each column is a text with the numeric values quantifying the dissimilarities between that text and each other document in the corpus.

   The dissimilarity is computed using the inverse tetrachoric correlation coefficient, estimated according to Karl Pearson's formula (Upton & Cook, 2008) and so explained in (Forsyth & Sharoff, 2014):

   > $d(p,q) = 1 - \sin\left(\frac{\pi}{2} \times \left(\sqrt{ad} - \sqrt{bc}\right) \div \left(\sqrt{ad} + \sqrt{bc}\right)\right)$ where a, b, c, d are counts in a fourfold table constructed by reference to the median values in the vectors [of rates of flexigram occurrences] such that a is the number of times both values exceed their median, b is the number of time the first value exceeds its median while the second does not, c is the number of times the second value exceeds its median while the first does not and d is the number of times neither value exceeds its median. (In fact, all four counts were incremented by 1 as an attentuation factor to avoid zero cell counts).

3.4.3 Graphic representation

The final stage of this document dissimilarity analysis consists in creating a visual representation of the dissimilarity matrix in R (R Development Core Team, 2008). This step is not strictly necessary, but having a graphic version of the dissimilarity matrix is helpful to better understand the nature of our corpora.

So, keeping as example Enger.at, a graphic visualisation of the corpus has been obtained computing multi-dimensional scaling (isoMDS) in R. The final

output was a graph where we can see documents of the corpus as dots, which are mutually distant from each other according to their (dis)similarity.

An heatscatter to highlight the density zones of the graph. In addition, centroids for the topics (coming from the previous analysis with Mallet) have been included, so the above mentioned topic modeling analysis and the flexigrams-based graph (in its dotted scatterplot visualisation option) have been combined together by mapping the 10 topics on the heatscatter graph. This was made by measuring the mean position of their assigned documents and generating the plot.

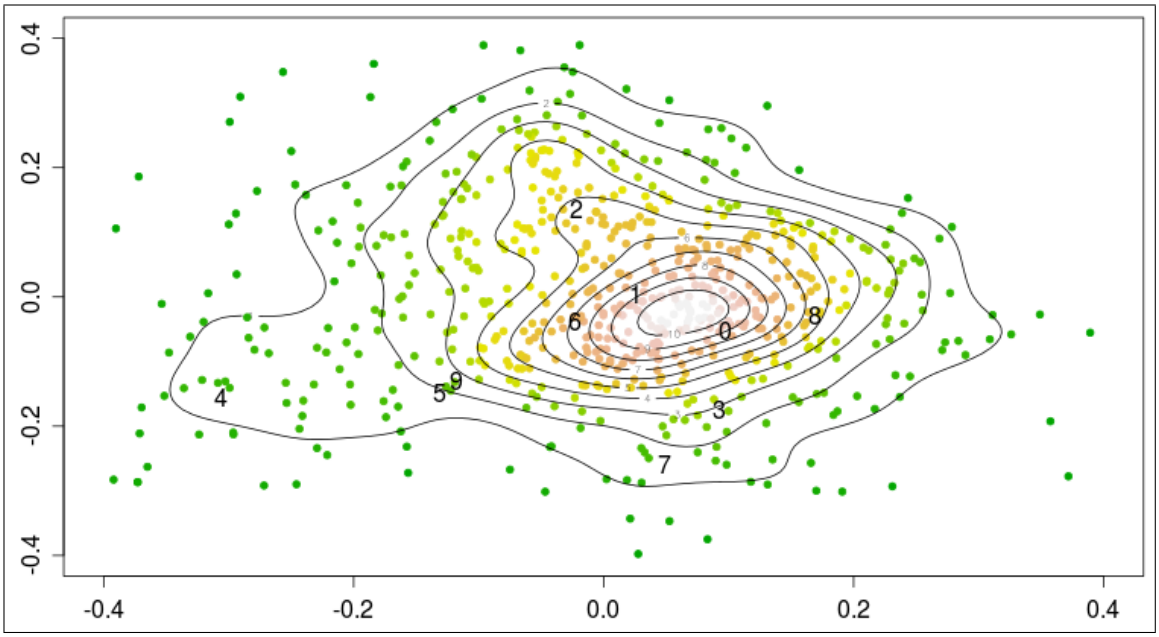

Figure 3.1: The final visualisation graph for Enger.at.

In figure 3.1 numbers correspond to centroids of Mallet categories and, following a manual analysis of their content, the two analyses appeared to suitably overlap in this representation: for example topic 5 (computer software) and 9 (computer hardware), which reasonably share similar areas of action, appear close on the graph generated by the flexigrams. This vicinity is useful when selecting documents for SMT since it enlarges the choice of training data to documents belonging not only to the considered topic, but also to a similar one.

3.5 Conclusions

In this chapter the identification of corpora suitable for the purpose of extracting subsamples for SMT, and the analysis of their content, has been illustrated. As a result English-Italian and German-Italian bilingual corpora have been collected from the web, and their content has been analysed in order to understand their composition and at the same time a methodology to measure dissimilarity between documents has been established. This strategy will be employed for subsampling experiments, which are reported in Chapter 4.

Chapter 4

Use of Focused Data in SMT

4.1 Experimental set up

In the previous chapter the collection of general-purpose bilingual corpora has been shown, together with the topic modeling analysis of the content of the obtained corpora and the generation of intra-corpus dissimilarity matrices with their graphic representation. In this chapter, we present the employment of such dissimilarity matrix analysis to generate ad-hoc subsamples, and their employment in SMT to translate a specific document. This pipeline has been tested using as starting point three documents, then it has been extended to their ten most similar documents (for a total of 33 translations for each language direction). The resulting translations have then been evaluated via automatic MT metrics.

4.1.1 The subsample-translate pipeline

It is possible to summarise the proposed subsampling pipeline in 5 steps:

1. Having a document in English to be translated in Italian (the **test doc**) and a parallel corpus English-Italian, rather than using the whole content of this corpus only a portion of it is going to be employed as training data to build an SMT system. In particular a selection the most similar documents to the test document (the **subsample**);

2. The similarity between the test doc and the documents belonging to the subsample is given by the proximity of these other documents to the test doc in the dissimilarity matrix, previously generated as described in 3.4.2. So - if not already present, as in the case of the three test docs here considered - the test doc has to be included in the whole corpus, with all the possible candidate training documents, and a dissimilarity matrix is generated[1];

3. The test doc is located in a vector space among all the possible candidate training documents, and it is possible to visualise it surrounded by them, with the most similar closer to it, in the R-generated plot. A selection of the 500 most similar documents to the test doc is made (graphically this would be like drawing a circle around the test doc);

4. This subsample is then extracted from the parallel corpus (both languages) and used in Moses: the content of the subsample is lowercased and tokensied, English and Italian sides of the subsample are used to build the translation model, the Italian side only to train the language model;

5. The resulting translation system is then used to machine-translate the test document from English to Italian.

As explained in 3.4.2, flexigrams have been chosen for their flexibility to capture language variability, and for that reason they may be a good feature for document discrimination since they take into account non necessarily contiguous multiword expression combinations. The aim of this research is also to try to understand which flexigram combination may be the most suitable and effective for the subsampling task: the employment of flexigrams-based analyses is here justified by the assumption that speech is not only given by single words but also multiword expressions/phrases (e.g. in SMT itself the use of approaches based on the use of phrases rather than single words is an established method), but it is not known whether it is worthwhile to employ multiword combinations rather than just staying with 1/1 flexigrams (i.e. unigrams, single words) or using a more traditional method like cosine similarity (as shown in section 3.1.3) for this

[1]These passages are done on the English side of the parallel corpus.

task. For this reason the above mentioned 5 steps pipeline is repeated on 2/4, 3/6 and 1/1 flexigrams-generated dissimilarity matrices and using cosine similarity, from English to Italian. Everything is performed on the other language direction as well, from Italian to English.

These subsample translations are based on the 500 documents most similar to the test doc extracted from the Engita corpus[1]. In order to have terms of comparison to evaluate the performances of these subsampling-based translations, the test docs have been translated also with the whole content of the Engita parallel corpus (removing every time the test document from the training data) and with 500 randomly selected documents (ensuring the test doc was not ending up in the random selection). One random selection has been performed for each subsampling experiment.

In total, 16 translation systems are created for each test doc:

- three models, generated on the subsample of 500 most similar documents from 1/1, 2/4 and 3/6 flexigrams-based dissimilarity matrices, from English to Italian;
- three models, generated on the subsample of 500 most similar documents from 2/4, 3/6 and 1/1 flexigrams-based dissimilarity matrices, from Italian to English;
- three models, generated on 500 randomly selected documents from 2/4, 3/6 and 1/1 flexigrams-based dissimilarity matrices, from English to to Italian;
- three models, generated on 500 randomly selected documents from 2/4, 3/6 and 1/1 flexigrams-based dissimilarity matrices, from Italian to English;
- one model, generated on the subsample of 500 most similar documents according to cosine similarity, from English to Italian;

[1]Recall the choice of limiting the subsampling to the 500 most similar documents was made thinking of building translation systems on very small corpora, simulating a "worst case" scenario. But choosing 500 documents rather than another quantity is completely subjective, thinking of a quantity of documents that may be small but big enough to ensure an acceptable minimum amount of training data.

- one model, generated on the subsample of 500 most similar documents according to cosine similarity, from Italian to English;
- one model, generated on the whole content of Engita corpus, from English to Italian;
- one model, generated on the whole content of Engita corpus, from Italian to English;

The translations obtained from the employment of these 16 translation systems are then evaluated as reported in 4.1.3, while the next section provides a description of the test docs on which these experiments are tested.

4.1.2 Choice of test documents

In this section a description of the three test docs employed in the experiments is provided. All of them were randomly selected from the Engita corpus: this means they are originated from the web, already provided both in Italian and English and aligned at the sentence level.

CONCOR .
http://www.concordiahotel.it
http://www.concordiahotel.it/en

This text comes from the homepage of a hotel, as such its main purpose is promotional.

Hotel Concordia in Rome: Stylish and Warm Welcome! The Concordia Hotel in Rome is located in an attractive 18th century building, set in the spellbounding heart of Rome . In the spring and summer, breakfast on the roof garden is a definite plus. Soak up the atmosphere in the nearby Spanish Steps , visit the Keats-Shelley House and throw your coin in the Trevi Fountain . Or head along to the Villa Borghese to see the water clock which dates back to the 19th century. From the Pincio terrace overlooking Piazza del Popolo , stop to admire the stunning views of Rome, the twin churches, Santa Maria dei Miracoli and Santa Maria di Montesanto, as well as Saint Peter's Dome . With the boutiques on Via Condotti , the nearby Via del Corso and the glossy Via Veneto , minutes away, it is hard to resist shopping in Rome. The Concordia in Rome is a well loved and traditional hotel, providing great value accommodation , and staff dedicated in providing exceptional service to meet your demands.

Table 4.1: English text of *concor*.

ARCHIM .

http://archimede.imss.fi.it/kircher/indice.html

http://archimede.imss.fi.it/kircher/index.html

This text comes from a completely different context comparing to the first test doc: it belongs to a museum website, in particular a subsection dedicated to a research project. This page has the purpose of providing a presentation and a description of the project.

The Athanasius Kircher correspondence project was created with the goal of making the manuscript correspondence of Athanasius Kircher available on the Internet. The project was commenced through the collaboration of the Istituto e Museo di Storia della Scienza in Florence, the Pontifical Gregorian University in Rome and the European University Institute in Fiesole, under the direction of Michael John Gorman and Nick Wilding. Since September 2000, the project has been rehoused at Stanford University. A new searchable version of the correspondence, using Luna Insight software has been developed at Stanford, and is now available to researchers. Comments on the new interface are very welcome. During his lifetime, the Jesuit polymath Athanasius Kircher (1602-1680) was widely regarded as the physical embodiment of all the learning of his age. A refugee from war-torn Germany, Kircher arrived in Rome just after Galileo's condemnation, where he was heralded as possessing the secret of deciphering hieroglyphics. He wrote over thirty separate works dealing with subjects ranging from optics to music, from Egyptology to magnetism. He invented a universal language scheme, attacked the possibility of alchemical transmutation and devised a host of remarkable pneumatic, hydraulic, catoptric and magnetic machines, which he displayed to visitors to his famous museum , housed in the Jesuit Collegio Romano. His books, lavishly illustrated volumes destined for Baroque princes with a love of the curious and exotic, are permeated with a strong element of the Hermetic philosophy of the Renaissance, synthesised with the Christianised Aristotelianism of the Jesuit order to which Kircher belonged. Kircher had over 760 correspondents, including scientists, physicians, Jesuit missionaries, two Holy Roman Emperors, popes, and potentates throughout the globe.

Table 4.2: First ten lines of English text of *archim*.

ABSTRA .

http://www.abstract.it/portfolio/intranet/honda-intranet?set_language=it

http://www.abstract.it/portfolio/intranet/honda-intranet?set_language=

en[1]

This texts comes from an IT company website. As part of their portfolio they showcase their work with a description of specific projects (including the needs of the client, the outcome etc.). This text is then descriptive in a similar way to ARCHIM but being ABSTRa of a more commercial nature it contains elements typical of the promotional communication.

Sezioni Honda intranet
The need
The intranet Honda Italy is the result of a deep study by the Abstract's team.
It fully meets the needs expressed by the customer: a simple but at the same time articulated website, so that gave its users the chance to navigate between appointments, meetings, documents, projects and communications quickly, respecting at the same time, the policies of permissions internal to the company.
The peculiarity of the intranet Honda Italy lies in the processes that led to the creation of the final product.Above the portal there is a structured study made with innovative tools that gave birth to a result not only welcome but highly responsive to the requirements expressed by the user.
The challenge of the intranet is always something new, every company has a complex structure of power and an extensive and almost never linear organization.
The biggest is the company the more complex become the decision-making processes and the levels of security.
The portal
In this case the level of complexity is high because of the number of users, documents and reports run by each department, everybody wished that the intranet met a certain requirement without taking into account the needs of different departments or, in cases of collaboration between them, often creating situations of conflict or overlap.
Not to mention that every intranet has a tremendous impact on work organization and every company should be ready to bear the costs of this impact.
An intranet, often change not only the processes but also the same approach to work and remains a work in progress to be developed over time through experience.

Table 4.3: First ten lines of English text of *abstra*.

[1]These links are not working anymore.

Table 4.4 shows details about the three test documents. Even though they share the fact of being texts produced with the purpose to be published on the web, they differs under many points of view: length, communicative purpose and style, possibly posing different challenges when automatically translating them from a language to another.

Test doc	*Sentences*	*English words*	*Italian words*
CONCOR	10	190	195
ARCHIM	56	1858	1775
ABSTRA	27	726	734

Table 4.4: Details about the three considered test docs.

4.1.3 MT evaluation set up

Several automatic MT evaluation metrics have been employed in order to assess the performances of the SMT subsampling strategy. All of them rely on the use of a human translation of each test doc as benchmark for comparison but they play different roles in the evaluation: BLEU (Papineni *et al.*, 2002), METEOR (Lavie & Denkowski, 2009), added to Translation Error Rate (Snover *et al.*, 2009) and Soft Cardinality (Jimenez *et al.*, 2012).

In particular Translation Error Rate (TER) is an error metric which measures the number of corrections needed to transform the sentences of a machine-translated text into a human reference translation. TER is defined as "the minimum number of edits needed to change an hypothesis so that it exactly matches one of the references, normalized by the average length of the reference" (*cit.*), so a lower score means less edits to perform and so a better translation (compared to the provided human translation).

Soft Cardinality (SC) was originally designed to measure monolingual sentence similarity but it has been employed for the evaluation of MT output as well[1]. While usual similarity scores measure similarity in a crisp manner (i.e. either two elements like words are identical or not), SC takes into account the fact that similar words contribute less to the dissimilarity of two sets than completely

[1] http://www.ttc-project.eu/images/stories/TTC_D7.2.pdf, par. 5.2, page 13.

different words. This is achieved by dividing words in smaller n-grams and then comparing two texts based on such n-grams rather than whole words.

4.2 Results of automatic MT evaluation

4.2.1 First results

The whole subsample-translate-evaluate pipeline on the three documents on both language directions has been performed. A first set of experiments was run using flexigrams from a lemmatised version of the English side of our corpus; results were quite unsatisfactory since in no case the translations based on subsamples outperfomed the ones built using the whole corpus.

Chances are that generating flexigrams without restrictions of any sort were not resulting into a strong discrimination between different text types, i.e. a large number of highly frequent flexigrams are made by combinations of stopwords which may be too generic to generate proper document dissimilarity.

So it has been decided to restrict the amount of stopwords in the three considered flexigrams combinations as following:

- No stopwords for unigrams;
- Up to one stowords for 2/4 flexigrams;
- Up to two stowords for 3/6 flexigrams.

Dissimilarity matrices has been generated based on this new flexigrams subsample extractions and the three document were translated and evaluated.

The following tables contain the results of the automatic evaluation. The two highest BLEU scores for each evaluation task have been highlighted.

Concor EN>IT	**BLEU**	**METEOR**	**TER**	**SC**
Whole corpus	0.1300	0.3709	0.7230	0.4906
Subsample Cosim	0.2117	0.4683	0.6923	0.6113
Subsample 2/4	0.2184	0.4805	0.6820	0.6003
Subsample 3/6	0.1318	0.3652	0.7384	0.4794
Subsample 1/1	0.1910	0.4331	0.6923	0.5434
Random 1	0.1837	0.4230	0.6974	0.5374
Random 2	0.2072	0.4562	0.7128	0.5991
Random 3	0.1849	0.4585	0.6923	0.6079
Concor IT>EN	**BLEU**	**METEOR**	**TER**	**SC**
Whole corpus	0.1545	0.2696	0.6789	0.4004
Subsample Cosim	0.2315	0.3240	0.6631	0.5995
Subsample 2/4	0.1983	0.3191	0.6894	0.5017
Subsample 3/6	0.0963	0.2586	0.7789	0.3795
Subsample 1/1	0.1961	0.3002	0.6473	0.4724
Random 1	0.1307	0.2852	0.7105	0.4068
Random 2	0.1411	0.2993	0.7210	0.4370
Random 3	0.1462	0.2901	0.7526	0.4094

Table 4.5: Automatic MT evaluation on translations of CONCOR.

Archim EN>IT	**BLEU**	**METEOR**	**TER**	**SC**
Whole corpus	0.1147	0.2730	0.7671	0.3240
Subsample Cosim	0.1228	0.3055	0.7325	0.3969
Subsample 2/4	0.1330	0.2806	0.7525	0.3143
Subsample 3/6	0.1513	0.3268	0.7491	0.3914
Subsample 1/1	0.1124	0.2758	0.7749	0.3102
Random 1	0.1716	0.3368	0.7255	0.3812
Random 2	0.4497	0.5868	0.4792	0.6062
Random 3	0.1535	0.3337	0.7558	0.3982
Archim IT>EN	**BLEU**	**METEOR**	**TER**	**SC**
Whole corpus	0.1072	0.2045	0.7350	0.3263
Subsample Cosim	0.1443	0.2463	0.7061	0.3916
Subsample 2/4	0.1357	0.2000	0.6995	0.3270
Subsample 3/6	0.1773	0.2638	0.7150	0.4292
Subsample 1/1	0.1351	0.2140	0.7454	0.3335
Random 1	0.1739	0.2542	0.7098	0.4019
Random 2	0.4508	0.3791	0.4701	0.6322
Random 3	0.1791	0.2770	0.7241	0.4488

Table 4.6: Automatic MT evaluation on translations of ARCHIM.

Abstra EN>IT	**BLEU**	**METEOR**	**TER**	**SC**
Whole corpus	0.0949	0.3138	0.7179	0.4718
Subsample Cosim	0.1872	0.4160	0.6335	0.5337
Subsample 2/4	0.0948	0.3062	0.7316	0.4021
Subsample 3/6	0.1689	0.4032	0.6825	0.5028
Subsample 1/1	0.1627	0.3710	0.6798	0.5064
Random 1	0.1954	0.4350	0.6362	0.5922
Random 2	0.1720	0.4208	0.6689	0.5934
Random 3	0.1723	0.4033	0.6662	0.5368
Abstra IT>EN	**BLEU**	**METEOR**	**TER**	**SC**
Whole corpus	0.1279	0.2368	0.6524	0.3827
Subsample Cosim	0.2262	0.3087	0.5702	0.5761
Subsample 2/4	0.0913	0.2093	0.6936	0.3570
Subsample 3/6	0.2156	0.3107	0.6016	0.5027
Subsample 1/1	0.1885	0.2900	0.6112	0.4924
Random 1	0.2419	0.3236	0.6002	0.5240
Random 2	0.1994	0.3244	0.6332	0.5358
Random 3	0.2275	0.3154	0.6057	0.4994

Table 4.7: Automatic MT evaluation on translations of ABSTRA.

In general, it is possible to observe that sets made by 500 documents outperforms the sets made using the whole corpus. So limiting the presence of stopwords actually does contribute to the selection of better subsamples but in different ways: results for CONCOR shows that the subsamples based on cosine similarity and the ones made by 2/4 flexigrams-based dissimilarity matrix gave the best results on all the four considered automatic MT evaluation metrics, for both language directions. TER is the only exception, since it scored exactly the same for the English-to-Italian direction on Subsample Cosim and Subsample 1/1 and for the Italian-to-English direction, where the Subsample 1/1 outperformed both Subsample Cosim and Subsample 2/4. Also please note how the third best result for English-to-Italian translations is given by a random selection (Random 2).

Results for ARCHIM shows instead that the subsamples (both Cosim and the flexigram-based ones) still give better results than the whole corpus but this time the 3/6 selection prevails, for both language directions (with the exception of Subsample 2/4 with TER for Italian-to-English). However the ad-hoc subsamples are not the ones giving the best scores overall: in fact the 500 random-based translations for ARCHIM in general give better scores than all the 500 ad-hoc subsample-based translations, in particular the random selection based on 3/6 flexigrams remarkably outperforms the best ad-hoc subsample translation on both language directions. Table 4.8 shows the first ten lines of this outlier (on the left) compared to the first ten lines of the best subsample-based translation (Cosim).

Line	Random 2	Subsample Cosim
1	This work, it should be clear on the expression as a preliminary description of the project mirato to make un'edition of Athanasius Kircher's correspondence of be available on the Internet.	This work it intendersi as a description of the preliminary mirato to ensure un'edition of the letters of Athanasius Kircher consultabile the internet.
2	The first phase of ongoing project is already complete, thanks to the cooperation of ' Institute and the museum of history of science of Florence, the Institute of ' universitario of European Fiesole and of the pontifical gregorian university in Rome.	" project are already been completata, with the collaboration of ' and national museum of the history of the science of Florence, of ' universitario European institute of Fiesole and pontificia gregoriana university of Rome.
3	It is available for the research online a sperimentale version of the database containing the, Kircher's correspondence preserved in the archives of the gregorian university, connected to the images of scannerizzate manuscripts.	It is available for the research an online database of sperimentale version, with the letters of kircher conserved in the archives of gregoriana university, connected to the images scannerizzate of the manuscripts.
4	The database will be ampliato prossimamente with the addition of kircher's letters present in other, and is with further resources of research.	The database will be ampliato prossimamente with the addition of the letters of Kircher preserved in other archives and further resources to research.
5	In the course of his life, l' eclettico jesuit Athanasius Kircher (1602-1680), by many considered as the personificazione of the entire sapere dell'time.	In his life, l'eclettico gesuita Athanasius Kircher (1602-1680) for considered as the personificazione of all to dell'period.
6	A refugee from war-torn germany, Kircher arrived in Rome just after Galileo's condemnation, where he possessing the secret of deciphering of geroglifici.	Fuggito of germany dilaniata from the war, Kircher he Rome, shortly after the by Galileo, preceduto from fama possedere of the secret of decifrazione of geroglifici.
7	He wrote, more than 30 works, trattando arguments that spaziano dall'optics to music, dall'egittologia the magnetismo.	He more than 30, trattando argomenti that spaziano dall'ottica to the music, dall'egittologia to magnetismo.
8	Invented a one of the universal language, attacked discussion l'hypothesis of alchemical transmutation and devised a host of remarkable pneumatic, hydraulic, catottriche and magnetic machines, which he displayed to visitors to his famous museum, housed in the jesuit collegio romano.	Invent a pattern of universal language, pose in discussione l'ipotesi of trasmutazioni alchemiche and progett a straordinarie machines pneumatiche, idrauliche, catottriche and magnetiche, espose in the museum, ospitato in collegio " gesuita.
9	His books, volumes explanatory illustrati recorded in principles barocchi lovers of everything that was stravagante and esotico, are permeated by a philosophy ermetica the renaissance, synthesised with l'aristotelismo cristianizzato typical dell'jesuit order, which Kircher belonged.	His books , volumes riccamente illustrati destinati to the barocchi amanti of all it stravagante and esotico, are deeply permeati from the philosophy ermetica rinascimentale, fusa with l' aristotelismo cristianizzato typical dell'order gesuita, which Kircher apparteneva.
10	Kircher had over 760 correspondents, including scientists, physicians, jesuit missionaries, two holy roman emperors, popes, and potentates throughout the world.	Kircher he had more 760 corrispondenti, among which scienziati, medici, missionari gesuiti, two imperatori of " sacred impero, and potentati popes of the whole world.

Table 4.8: Random 2 translation for ARCHIM compared to Subsample Cosim translation.

It is possible to observe how both translation contain similar kind of translation errors, such as unseen words and misplaced punctuation. However while several words in both sets may have not been translated because of their high specificity and so unlikely to be present in the training set - such as *egittologia* (Egyptology), *ermetica* (hermetic), *catiottriche* (catoptric), or entity names like instututions - some other words ended up being correctly translated by Random 2 but not Subset Cosim: *complete* vs. *completata* (line 2), *jesuit* vs. *gesuita* (line 5), *refugee* vs. *fuggito* and *war-torn* vs. *dilaniata from the war* (line 6), *hypothesis* vs. *ipotesi* (line 7), *permeated* vs. *permeati* (line 9), *scientist phisicians missionaries empereors* vs. *scienziati medici missionari imperatori* (line 10). Matching words yield higher scores on most MT evaluation metrics, so this is probably the reason why this random sample outperformed every other translation. However it is not clear why this translation system, built on a randomly collected sample of parallel documents from Engita, ended up having a training set that ended up being much better than ad-hoc subsets. This may be due to randomness.

ABSTRA gives a different picture again: the 3/6 subsamples generally give the best performance over the use of the whole corpus, the 2/4 and the 1/1 subsamples, with the exception of TER and SC scores for the English-to-Italian translations which are slightly better; but they are outperfomed by other systems, in particular Subsample Cosim and Random 1 (with Random 1 performing betther than Subsample Cosim) offer the two highest scores for the English-to-Italian translations, while Random 1 and Random 3 outperform all the ad-hoc subsamples for the Italian-to-English translations.

So on the one hand it is possible to observe how it is not necessary to employ the whole corpus to obtain better translations, but on the other hand every one of these three sets of experiments provide different results, with flexigrams-based translations outperformed by cosine similarity and random samples widely performing better than both flexigrams-based and cosine similarity-based selections. This is a negative result, however the outcome of this analysis may be not statistically significant, for example due to the reduced size of the translated documents. This is the reason why some further analyses have been performed as described in the next paragraph.

4.2.2 Further analysis

In order to validate the results from the previous analysis, in particular those regarding the two best translations for each test set, statistical significance tests have been computed using the bootstrap method (Koehn, 2004). This strategies relies on the generation of BLEU scores repeatedly calculated (e.g. 1000 times) on resamplings of the sentences of the same set that needs to be evaluated, and dropping the top 25 and bottom 25 BLEU scores in order to obtain a 95% confidence interval and have more reliable scores. Tables 4.9, 4.10 and 4.11 show the results of this analysis.

Concor	**EN>IT**		**IT>EN**	
	Subs. Cosim	Subs. 2/4	Subs. Cosim	Subs. 2/4
Original	0.2117	0.2184	0.2315	0.1983
Actual	0.2078	0.2138	0.2304	0.1915
95% conf. int.	0.1978	0.1986	0.2264	0.1882

Table 4.9: Bootstrap results for the two best BLEU scores for CONCOR.

Archim	**EN>IT**		**IT>EN**	
	Random 1	Random 2	Random 2	Random 3
Original	0.1716	0.4497	0.4508	0.1791
Actual	0.1223	0.4082	0.4130	0.1448
95% conf. int.	0.1218	0.4052	0.4060	0.1441

Table 4.10: Bootstrap results for the two best BLEU scores for ARCHIM.

Abstra	**EN>IT**		**IT>EN**	
	Subs. Cosim	Random 1	Random 1	Random 3
Original	0.1872	0.1954	0.2419	0.2275
Actual	0.1496	0.1544	0.2119	0.2032
95% conf. int.	0.1501	0.1518	0.2109	0.2012

Table 4.11: Bootstrap results based for the two best BLEU scores for ABSTRA.

This test helps us assessing whether the change in BLUE score across the two best systems for each document (and each language direction) corresponds to a true difference in terms of actual translation quality (and so understanding which of the two systems is better than the other). Scores for CONCOR dropped around 0.01 points for each system, confirming the previous analysis: Subsample 2/4 slightly outperform Subsample Cosim on the English-to-Italian direction; but we have the opposite situation for the Italian-to-English direction, with Subsample Cosim outperforming Subsample 2.4 by 0.0382 points. Scores for ARCHIM dropped more considerably after the significance test (around 0.04 for all systems), but the two outlying results given by the Random 2 systems confirmed their dominant position for both systems. The results are a bit more inconsistent for ABSTRA, with scores dropping around 0.03 points and bein quite stable for Random 1 for both language directions (but to be precise the score for Random 2 for English-to-Italian is 0.0436, and 0.0263 on Random 3 for the Italian-to-English direction). However again in this case the order of the winning translations are confirmed from the original analysis.

As said the test docs evaluated as described in the previous section may be too small to consider the automatic MT evaluation scores calculated on their translations reliable. So another way of getting a more reliable evaluation is to enlarge the test set. In order to do so, the ten most similar documents to each test doc have been included in the test set (and removed from the training set). The following table shows the total sizes of these new evaluation sets of 11 documents (tokenised):

Each translated document has been then evaluated with BLEU, and the average score of each test set + its ten most similar documents have been calculated. Standard deviation has been produced as well, in order to check the amount of variation from the average calculation of the 11 scores, since a certain amount of discrepancy is expected when translating different documents with the same translation system.

At the same time whole 11 documents test sets have been translated and evaluated as single chunks of text. This is due to the fact that BLEU is more reliable when calculated on bigger sizes of text.

Test sets	English Words	Italian Words	Sentences
Concor 2/4 and 10 most similar	2161	2466	126
Concor 3/6 and 10 most similar	6734	7236	307
Concor 1/1 and 10 most similar	5523	5262	235
Archim 2/4 and 10 most similar	19791	21529	756
Archim 3/6 and 10 most similar	15827	15457	496
Archim 1/1 and 10 most similar	8169	7986	408
Abstra 2/4 and 10 most similar	22966	22390	908
Abstra 3/6 and 10 most similar	5352	5453	279
Abstra 1/1 and 10 most similar	13503	13039	551

Table 4.12: Size of extended evaluation sets.

EN>IT	Concor	Doc2	Doc3	Doc4	Doc5	Doc6	Doc7	Doc8	Doc9	Doc10	Doc11	Average	Stdev	Concor+10
Subsample Cosim	0.2141	0.1797	0.1493	0.3538	0.0957	0.0806	0.2118	0.1975	0.3110	0.2094	0.2444	0.2043	0.0214	0.2809
Subsample 2/4	0.2028	0.111	0.3048	0.2661	0.1241	0.2427	0.0122	0.1978	0.0536	0.2952	0.1599	0.1791	0.0303	0.1963
Subsample 3/6	0.135	0.0667	0.27	0.0528	0.1734	0.049	0.1469	0.0647	0.1438	0.1353	0.1665	0.1276	0.0222	0.1191
Subsample 1/1	0.2001	0.1853	0.2484	0.0163	0.1212	0.1776	0.211	0.1312	0.1636	0.2479	0.2091	0.1737	0.0063	0.2078
Random 2/4	0.3247	0.0547	0.263	0.2387	0.1167	0.1206	0.0114	0.2022	0.0679	0.2389	0.0606	0.1544	0.1867	0.1590
Random 2	0.1785	0.0765	0.309	0.1288	0.2436	0.0708	0.2101	0.0833	0.1898	0.1622	0.1477	0.1636	0.0217	0.1534
Random 3	0.1546	0.169	0.2926	0.0338	0.1043	0.1753	0.1871	0.1072	0.1808	0.2125	0.1291	0.1587	0.0180	0.1830

IT>EN	Concor	Doc2	Doc3	Doc4	Doc5	Doc6	Doc7	Doc8	Doc9	Doc10	Doc11	Average	Stdev	Concor+10
Subsample Cosim	0.2287	0.0706	0.1052	0.2608	0.0812	0.0954	0.1743	0.2023	0.1955	0.1603	0.1868	0.1601	0.0296	0.2098
Subsample 2/4	0.2157	0.0896	0.2764	0.2474	0.1515	0.1961	0.0379	0.2081	0.0558	0.2464	0.2764	0.1819	0.0429	0.1959
Subsample 3/6	0.0911	0.0739	0.2616	0.0514	0.1755	0.0848	0.1231	0.0657	0.1655	0.1206	0.1514	0.1240	0.0426	0.1336
Subsample 1/1	0.197	0.2456	0.317	0.0195	0.07	0.1469	0.2141	0.1123	0.1314	0.2085	0.1534	0.1650	0.0308	0.1819
Random 1	0.3003	0.0581	0.2117	0.2558	0.1268	0.1404	0.0361	0.1514	0.0448	0.1977	0.0873	0.1464	0.1506	0.1430
Random 2	0.1385	0.0761	0.2674	0.0991	0.2476	0.0941	0.165	0.0765	0.1767	0.1586	0.1653	0.1513	0.0189	0.1453
Random 3	0.1399	0.2122	0.3143	0.0262	0.083	0.1401	0.2164	0.1017	0.1622	0.2132	0.1078	0.1560	0.0226	0.1848

Table 4.13: BLEU scores of translations of CONCOR and its ten most similar documents.

EN>IT	Archim	Doc2	Doc3	Doc4	Doc5	Doc6	Doc7	Doc8	Doc9	Doc10	Doc11	Average	Stdev	Archim+10
Subsample Cosim	0.1135	0.1186	0.0515	0.058	0.1133	0.1656	0.1392	0.1109	0.1518	0.1105	0.1219	0.1140	0.0059	0.1631
Subsample 2/4	0.1289	0.1213	0.0298	0.1569	0.0779	0.136	0.0614	0.1385	0.1271	0.166	0.1519	0.1177	0.0162	0.1367
Subsample 3/6	0.1516	0.052	0.0969	0.3829	0.2322	0.1726	0.1359	0.2567	0.1335	0.1155	0.1035	0.1666	0.0340	0.1602
Subsample 1/1	0.1127	0.1344	0.0216	0.1732	0.2653	0.0506	0.1316	0.1399	0.0642	0.1209	0.1187	0.1211	0.0042	0.1785
Random 1	0.08	0.0998	0.0465	0.1517	0.0611	0.1287	0.0525	0.1325	0.1185	0.1349	0.1388	0.1040	0.0415	0.1237
Random 2	0.1631	0.0526	0.1172	0.3535	0.2015	0.1971	0.1892	0.2207	0.127	0.1136	0.1401	0.1705	0.0162	0.1758
Random 3	0.1245	0.1388	0.0319	0.0829	0.2652	0.0643	0.1396	0.1851	0.0582	0.191	0.1559	0.1306	0.02220	0.1597

IT>EN	Archim	Doc2	Doc3	Doc4	Doc5	Doc6	Doc7	Doc8	Doc9	Doc10	Doc11	Average	Stdev	Archim+10
Subsample Cosim	0.1293	0.1179	0.0358	0.0552	0.1288	0.2029	0.1678	0.1554	0.1506	0.1452	0.1198	0.1280	0.0067	0.1735
Subsample 2/4	0.1422	0.1124	0.0504	0.2296	0.1124	0.1789	0.0686	0.1525	0.1525	0.1504	0.1622	0.1374	0.0141	0.1667
Subsample 3/6	0.1717	0.0321	0.0903	0.2944	0.2105	0.2149	0.1668	0.2061	0.189	0.1031	0.0818	0.1600	0.0635	0.1817
Subsample 1/1	0.1402	0.1207	0.056	0.155	0.256	0.0517	0.1849	0.1127	0.0403	0.1731	0.1679	0.1325	0.0195	0.1794
Random 1	0.0696	0.088	0.0302	0.1839	0.0751	0.1744	0.0329	0.1141	0.1343	0.1301	0.111	0.1039	0.0292	0.1292
Random 2	0.175	0.0467	0.134	0.3297	0.1887	0.2269	0.2144	0.2105	0.1698	0.1085	0.0801	0.1713	0.0671	0.1927
Random 3	0.1505	0.1257	0.0775	0.0795	0.2832	0.1784	0.2104	0.1433	0.1329	0.1862	0.1323	0.1545	0.0128	0.1699

Table 4.14: BLEU scores of translations of ARCHIM and its ten most similar documents.

EN>IT	Abstra	Doc2	Doc3	Doc4	Doc5	Doc6	Doc7	Doc8	Doc9	Doc10	Doc11	Average	Stdev	Abstra+10
Subsample Cosim	0.1595	0.1148	0.0798	0.1723	0.2178	0.0474	0.2036	0.1924	0.0952	0.0822	0.1323	0.1361	0.0192	0.1967
Subsample 2/4	0.1561	0.1543	0.1432	0.1059	0.0874	0.2079	0.0918	0.1314	0.123	0.1467	0.1625	0.1372	0.0045	0.1400
Subsample 3/6	0.1695	0.102	0.1795	0.1745	0.0752	0.2574	0.1635	0.0599	0.1068	0.1104	0.1791	0.1434	0.0067	0.1536
Subsample 1/1	0.1789	0.3163	0.0615	0.1834	0.149	0.0803	0.12	0.1935	0.2379	0.1525	0.1263	0.1636	0.0371	0.1788
Random 1	0.191	0.1597	0.1681	0.1623	0.2138	0.2538	0.0443	0.19	0.2329	0.1812	0.246	0.1857	0.0388	0.2220
Random 2	0.184	0.0964	0.2027	0.246	0.0982	0.1789	0.1329	0.0992	0.1218	0.1428	0.1483	0.1501	0.0252	0.1810
Random 3	0.0851	0.0765	0.0633	0.0888	0.0895	0.03	0.0781	0.1521	0.1158	0.0854	0.0823	0.0860	0.0019	0.0934

IT>EN	Abstra	Doc2	Doc3	Doc4	Doc5	Doc6	Doc7	Doc8	Doc9	Doc10	Doc11	Average	Stdev	Abstra+10
Subsample Cosim	0.2288	0.1202	0.1168	0.1702	0.2038	0.0331	0.2132	0.1802	0.0905	0.0978	0.1548	0.1463	0.0523	0.2031
Subsample 2/4	0.1664	0.1171	0.1015	0.1291	0.211	0.2464	0.0478	0.1406	0.1875	0.175	0.1914	0.1558	0.0176	0.1883
Subsample 3/6	0.1915	0.1072	0.1988	0.209	0.076	0.3085	0.1279	0.0498	0.1138	0.153	0.2032	0.1580	0.0082	0.1693
Subsample 1/1	0.2412	0.2969	0.0676	0.1822	0.1622	0.0673	0.1343	0.132	0.2479	0.2799	0.1347	0.1769	0.0753	0.1826
Random 1	0.1716	0.1644	0.1838	0.172	0.2346	0.2321	0.0628	0.1806	0.2021	0.1792	0.2984	0.1892	0.0896	0.2257
Random 2	0.2393	0.0755	0.2006	0.2752	0.0801	0.2945	0.1392	0.0476	0.1114	0.1055	0.1777	0.1587	0.0435	0.1843
Random 3	0.0727	0.0545	0.0327	0.1037	0.0943	0.0665	0.0896	0.141	0.1342	0.2143	0.1034	0.1006	0.0217	0.1034

Table 4.15: BLEU scores of translations of ABSTRA and its ten most similar documents.

Again the two best results for each evaluation session have been highlighted, both in the "average" and "full set" columns. CONCOR favours again the subsamples: the translations made with systems trained on cosine similarity-based subsamplings offer the best results for both the average and full set, for the English-to-Italian drection, followed respectively by 2/4 and 1/1 flexigrams-based subsamplings. For the Italian-to-English direction the picture is slightly different, with Subsample Cosim performing as the best system only for the full set (followed by Subsample 1/1), while the best average score is represented by Subsample 2/4 followed by Subsample 1/1.

ARCHIM instead provides his best subsampling result on 3/6 in the average results for the English-to-Italian translations, but it is outperformed by the Random 2 selection, while the best score for the full set is Subsample 1/1 followed very closely by Random 2. The average and full set results for the Italian-to-English translations are instead consistent, with the Random 2 systems having the highest scores and Subsamples 3/6 as the second best results.

The analysis of ABSTRA shows a consistency between the English-to-Italian and Italian-to-English translations: the best subsampling performances are given by the Random 1 systems both for the average and full set translation, while the second best scores are given by Subsample 1/1 flexigrams-based translations for the average scores and Subsample Cosim for the full set scores.

To sum up, for all the three considered documents the earlier analysis has shown that better results were obtained from 500 subsamples compared to the employment of the whole corpus. This result demonstrates that using all the available data (in this case an Italian-English bilingual corpus with 2,932 words, 109,156 sentence pairs, 2,200,000 words per language - minus the test document) does not necessarily yield better results than a selection of its documents based on a flexigrams-based dissimilarity matrix (2/4 or 3/6, most probably depending on the text type of the test document) or cosine similarity. The size of subsamples may vary since the selection is made on a document level, rather than at the sentence level (as in other approaches), but it can be as small as ten times less than the use of the whole corpus. Also the time and resources used in the training process are remarkably reduced: on a 4x Intel i7-3520M with 8GB RAM computer

the training step in Moses took around 1 hour to generate translation models based on the whole corpus while only an average time of 15 minutes was enough to train SMT systems based on 500 documents sets.

However, in most cases selection based on cosine similarity outperformed flexigrams-based selection, and the fact that in several occasions random sets of 500 documents yielded even better results than the ad-hoc subsamples (with the striking case of the Random 2 outlier for ARCHIM) should be seriously considered as well. These results suggest that even though the selection of subsamples with the flexigrams approach may be advantageous compared to the use of the whole corpus, it may be not the best subsampling. Most probably this means that the flexigrams approach works better with documents belonging only to certain genres/domains. Let us consider CONCOR: the reasons behind the positive results for this test doc may be found in the nature of the document itself, which can be described as belonging to a very specific communication purpose. As shown in section 3.3.2 there is a whole topic dedicated to the text type of hotel sites, which appeared to be a very recognisable kind of multilingual web content. Also a certain number of unseen words in CONCOR did not require to be translated, as they are proper names of places (squares, churches, roads), while most of the other untranslated words are verbs. But still most cosine similarity-based subsampling translations for CONCOR in all the considered analyses performed better than the flexigrams-based subsampling, suggesting that despite its limitations cosine similarity may be still more reliable.

Something similar happened for the other two main test documents, ARCHIM and ABSTRA, but they ended up being a more difficult challenge: looking into the texts themselves (see examples in section 4.1.2) they appear to be made of more complex constructs and they do not crisply belong to one single specific text type the same way of CONCOR. Also results from the MT evalation have shown a confusing situation where very often random sets outpeformed both flexigrams-based and cosine similarity-based subsamplings (which scored overall similarly without showing a particular emergence of one method on the other). So probably the subsampling pipeline as it was presented in this thesis, based on a linguistic feature like flexigrams, works better for certain text types than others, in particular it seems that it works better for narrow domains, but cosine

similarity offered similar if not better results (at least based on automatic MT evaluation results).

This means that, while the use of tiny subsamples as training data against the use of all the available data has been proven useful, the best way to select the subsamples themselves is not clear: flexigrams-based selections have given good results in one case out of three, and in that case it has been outperfomed by cosine similarity systems. ARCHIM and ABSTRA offered a more confusing picture where random sets often outperformed both cosine similarity and flexigram-based systems, which means that in order to translate documents belonging to the text types of ARCHIM and ABSTRA probably require selections based on strategies other than flexigrams-based selections and cosine similarity. This possibility is further discussed in the concluding chapter of this thesis.

Chapter 5

Conclusion

The aim of this thesis was to explore the possibility of obtaining good performances from SMT approaching the problem from two main points of view: 1) by using very small training sets rather than huge quantities of (mostly) out-of-domain data, and 2) getting to know the nature of parallel data under the point of view of their text varieties (above all domain), in order to better understand which documents are the most suitable to be used as training data for specific translation tasks. Previous research has shown that limiting the quantity of training data when building SMT systems can give several advantages, such as the use of fewer computational resources (compared to the use of larger quantities of data), experiencing little or no loss in terms of translation performance, in some cases even better results. Also discriminating between documents belonging to different textual varieties has been previously explored, but the research here presented wanted to further address these two aspects, in particular using even smaller quantities of data and borrowing analysis techniques of textual data from genre/domain studies. These techniques have been used also in order to choose a suitable parallel corpus for the final subsampling experiments, subsequently leading to the decision of creating a new parallel corpus from the web. In order to do so, a pipeline to collect parallel corpora from the web has been set up (based on previous but mostly currently unavailable attempts), and analysis the resulted the situation of the current research on the "web as multilingual corpus" has been addressed as well.

So different aspects of the MT studies and in general of computational linguistics have been touched upon: SMT, genre/domain, document dissimilarity studies, and the "web as (parallel) corpus" approach. In the following sections the nature of this research is briefly reviewed, the main findings are reported, limitations and possible directions of future research are pointed out.

5.1 Main findings

5.1.1 Parallel corpora from the web

A big part of this project has been spent on the collection of parallel corpora from the web, since freely available parallel corpora in the language pair of interest (English-Italian) were not available - apart from Europarl, which ended up not being suitable due to the lack of wide text variability. After reviewing previous attempts at collecting multilingual data from the web, some of them have been considered: the first one was Fry's "RSS method", which allowed to easily collect a multilingual corpus but having the deficiency of working on a single website only, and so bringing again the problem of having a limited amount of text varieties. The second one is Resnik's original pipeline relying on search engines, BiTextCaT is based on it, but implementing the several steps of parallel corpus collection from the web with currently available tools: relying on the performances of Microsoft Bing's publicly available APIs, and further processing with state-of-the-art boilerplate cleaning and sentence aligning tools, three corpora have been collected from the national top level domains of Italy, Austria and Germany.

This pipeline is composed of freely available tools, in particular the *UrlCollector.jar* script contained in BootCaT (URL collection in the first language), jusText (plain text extraction), the NLTK package Punkt (sentence split), Hunalign (sentence alignment), while a substitution rules script has been compiled from scratch for the purpose of retrieving pages in the second language, which is released under GNU Genereal Public License. The corpora obtained with BiTextCaT cannot be made available because the content of single webpages is copyright of their owners but, as Resnik previously did, the list of URLs can be

shared without breaking any copyright[1], and their content can be downloaded following the BiTextCaT guidelines[2].

The size of the resulted parallel corpora is remarkably smaller than the average dimensions of parallel corpora usually employed in SMT: the English-Italian portion of Europarl is 1,909,115 sentence pairs (49,666,692 English words and 47,402,927 Italian words), while Engita is 109,156 sentence pairs (2,213,599 English words and 2,172,191). Getting more data would have required more iterations when performing the retrieval of candidate URLs from the web and, since the second step of the BiTextCaT pipeline (retrieval of corresponding pages in the second language) is currently based on substitution rules of language tags in URLs only (section 3.2.2, Step 2), possibly the implementation of other strategies to retrieve parallel pages. Nevertheless having a small corpus was in line with the idea of challenging normal trends of using large collections of (mostly non relevant) bilingual data. In short, having a small parallel corpus containing a certain degree of text variability was better for the aims of this research project, and this implementation of the "web as parallel corpus" method here presented, provided as an open source toolkit easy to reimplement and to customise accordingly to specific purposes, has proven to be useful for this purpose.

5.1.2 Topic modeling and document dissimilarity

Topic modeling has been extensively used in this research. It has been employed in the first stage of this research when looking for a parallel corpus with certain features, i.e. containing a certain degree of text variability for the subsequent SMT experiments. In particular the topic modeling functionality contained in the tool Mallet has been helpful to understand the composition of Europarl, the first considered candidate, and the webcorpus built downloading some content of the multilingual website Presseurop.eu, leading to the decision of discarding them. Later it has been used to explore the content of Engita and Enger (.at and .de) as well, showing the variety of topics covered in these three corpora. In particular it has been possible to note some recurring topics over the three

[1] http://corpus.leeds.ac.uk/brunez/parcorp.html.
[2] http://corpus.leeds.ac.uk/brunez/bitextcat.html.

corpora (e.g. tourism, industry, art), leading to some insights about the kind of multilingual webisites the Internet contains[1].

So even just the use of topic modeling analysis has been very useful to understand the textual variability contained in the considered corpora, helping to decide whether to use them or not. The possibility to set the number of generated topics in Mallet has been reiterated several times before settling on 10 topics for Engita and Enger, ending up being a suitable amount for such small corpora and allowing the avoidance of redundancy (in this case too many topics/clusters of words referring to the same kind of texts).

Another analysis has been performed in order to discriminate among documents belonging to different text varieties: dissimilarity matrices have been generated in R with multidimensonal scaling (isoMDS) based on the analysis of corpus documents using extended n-grams (flexigrams) as linguistic features. In particular this analysis has been tested on three of them: 2/4, 3/6 and 1/1 (unigrams). These dissimilarity matrices have been output in bidimensional graphic representations, where previously produced topics have been mapped. So combining the findings of the topic modeling with the document dissimilarity analysis (later used for subsampling), provided a good strategy to obtain a complete overview of the composition of a collection of parallel texts.

5.1.3 Less (but focused) data are better data?

The main object of interest of this thesis was to further explore the possibility of using very small training sets for SMT, putting some attention on the text variety of each single document(s) in need of translation. The idea of using less training data has been previously addressed in literature, and one of the purposes of this thesis has been to try to go further in this direction simulating a situation where the starting point is to have already a small amount of data to choose from, then select even smaller subsamples of parallel texts which have been chosen based on certain linguistic features. Awareness of the nature of the texts involved, both the documents that need to be translated and the whole corpus of possible

[1] As previously pointed out in this thesis, all these analyses have been performed on the English side of the corpora.

training data, played an important role in this process: topic modeling gave some idea of the main domains of the available data, while the document dissimilarity analysis provided a way of visualising the distance between all the documents (including the test docs) and an instrument to select subsamples for SMT based on certain linguistic features (flexigrams). The tests have been conducted setting the size of the subsamples to 500 most similar texts to the test documents, using 2/4, 3/6 and unigrams as linguistic features for the dissimilarity matrices and cosine similarity in order to compare the employment of this novel system to subsample training data for SMT to a more traditional dissimilarity method. Random selections of 500 documents and the content of the whole corpus have been used as training data as well in order to have benchmarks to evaluate the performances of the subsamplings. Some restriction on very common stopwords has been necessary to obtain a better discrimination between documents for the flexigram-based subsamplings, significance test has been run on the results of the automatic MT evaluation and the test sets have been extended to the 10 most similar documents to the original test docs in order to get more reliable BLEU scores.

These experiments have proven that using smaller samples of 500 documents actually do provide better performance than using the whole parallel corpus. This result demonstrates how using all the available data does is not only more time and resources consuming but can also give worse results.

But while the use of lesser amounts of data has been proven to be working in all the considered tests, it seems that the subsampling method based on flexigrams does not provide remarkable results: in the case of the first test doc, CONCOR, cosine similarity performed better than the best result given by subsample selections (2/4, for both language directions), while ARCHIM and ABSTRA gave as best performances the ones made by random samples, with subsamples as second best results, and not always the same. This means that the flexigrams-based method may work for certain text types but still it may not be necessarily the one to be employed since a more traditional (and simpler) method like cosine similarity in our case performed better. This limitation is further elaborated in the next section.

5.2 Scope

The final experiments of this projects have shown how the subsampling method based on flexigrams gives a positive outcome in one case out of three, when comparing 500 flexigrams-based subsamples to 500 random selections of training data but not cosine similarity-based subsamples, which gave better results. Also even if only the flexigrams-based subsamples are considered there is no consistency in the best results for each test doc. This means that is not possible to do a generalisation about which is the best feature (2/4, 3/6 or 1/1) to base subsamplings on, and that the subsampling method as it is does not work for every kind of document and anyway it does not provide overall better results than cosine similarity.

Talking about the test documents themselves, it is possible that CONCOR offers a more doable challenge than ARCHIM and ABSTRA, in particular ABSTRA looks like quite an elaborate document to translate with MT. Also CONCOR has the big advantage of being very easy to identify as belonging to a text type having a specific and recognisable topic in the Engita corpus, i.e. hotel webpages. This means that most likely the 500 most similar documents to CONCOR, both from the flexigrams-based and the document similarity analysis, contain a large amount of pages belonging exactly to the same text type of CONCOR, while the 500 most similar documents to ARCHIM and ABSTRA may have not a similar situation. So subsampling training sets (either based on flexigrams as feature to discriminate between documents or on cosine similarity) properly works only in certain cases - possibly with documents belonging, in a similar way to CONCOR, to narrow domains. This means that most probably subsampling works also for other different text varieties, but the selection would be made on other linguistic features than flexigrams or a cosine similarity-based selection, and that would require calibration for each document/text type, in order to decide what is the optimal strategy for each circumstance.

Also in this project everything was performed on extremely small amounts of data. This was made with the purpose of simulating some sort of extreme scenarios of lack of big quantities of parallel data, in order to demonstrate the validity of working with smaller amounts than big corpora of tens, hundreds or

more million of words. This has been useful to demonstrate that this is possibly true, but when it comes to perform automatic MT evaluation results are less reliable when working with such small amounts of text. This is the reason why the best results have been validated using statistical significance and test sets enlarged in the latter stage of the experiments, but still it is possible that results of automatic MT evaluation may be impaired for this reason.

Another thing to consider is that all these analyses and experiments have been made at the document level: whole documents and not single sentences have been used as the basic unit to analyse texts contained in the considered collections, then dissimilarity matrices have been generated discriminating between documents which have been then employed in the SMT experiments. But working at the document level means working with single documents of different length, which may make a big difference both when creating subsamples and at when performing automatic MT evaluation.

5.3 Future research

As just shown, some interesting results have been achieved during this research: the importance of being aware of the nature of the text variety of parallel resources that may be considered to perform MT has been pointed out, and the possibility of using less training data than all the available resources has been shown. This can find practical application in situations like industrial settings, where situations like time constraints and lack of large amounts data in an intended genre/domain are likely to happen.

However the subsampling method here presented has been proven working only in certain circumstances and in general not providing a remarkable improvement on methods based on more traditional sampling selections (i.e. cosine similarity), which means that there is still more to further explore on this path: in this project certain specific features has been chosen to represent discrimination between different types of documents (flexigrams), but the proposed methodology does not depend on these features. As said in the previous section different text types may need subsamplings based on different linguistic features: flexigrams have proven to work better with a text belonging to a narrow domain,

identifiable based on key words (and multi-words) which are usually related to that particular kind of text, i.e. hotel webpages. Documents like the second test document instead, being of a more academic/institutional nature, may need a different kind of features to be identified, for example recurring sentence structures typical of academic texts may suggest the employment of syntactic features instead. So, in order to predict whether the use of certain features would work or not using the subsampling approach here proposed on documents belonging to different text types, it is crucial to consider the adoption of this kind of approach and if so identify which features are most representative of different text types. In order to do that going more in depth with a study about the genres and domains (and their combinations) of bilingual/multilingual documents would be very useful.

Also the experiments here presented were made on the English-Italian language pair and with subsamples of 500 documents, so there is a range of possibilities to explore: for example it would be interesting also to understand how this works with different language pairs. It is likely that pairs of typologically more distant languages would require lager subsamples and different training settings. A more extensive use of BiTextCaT, i.e. reiterating the URL collection, employing a larger set of seed words etc.[1] can lead to the collection of larger quantities of parallel corpora which would allow to select subsamples of different amounts - in particular it may be interesting to try out increasing amounts of training data and monitor improvements (or deteriorations) of the translation quality. Finally, Moses has been employed for these experiments using a less-than-standard "baseline" set up, without tuning, in order to simulate worst-case scenarios, but having enough data it is possible to use this engine in its traditional baseline set up Plus Moses offers an extremely wide variety of settings and parameters that can be adjusted, so it would be interesting to see how to find a good balance between employing a more or less sophisticated set up (e.g. using more language models, weighting optimisation etc.) and the training times.

So there is room for improvement and experimentation when dealing with the employment of small sets of data for document-specific SMT. The project

[1]There is also room for improvements with regards to the pages in the second language, now implemented only as substitution rules at the URL level.

presented in this thesis wanted to explore this direction, which is definitely worth to be further investigated.

Appendix A

Tools and workflow

This appendix collects all the programs, toolkits etc. used in this thesis project, with details about how they have been run for the purposes of this project.

A.1 Bilingual corpus collection

A.1.1 BootCaT

http://bootcat.sslmit.unibo.it

BootCaT is available both as a front-end application and as a collection of command line scripts. In particular the only script used in the BiTextCaT pipeline is *UrlCollector* Java script. The version used in this project is the one contained in BootCaT 0.60, available until 2 July 2012. After this date Bing applied some restrictions on the use of their APIs, making its search engine functionalities still available through its Windows Azure Marketplace (but limiting the amount of retrievable results).

The script itself has not been modified but it has been employed slightly differently from its original usage since URL language parameters have been added to each line of the seed list in order to retrieve pages belonging to multilingual websites (see 3.2.2). The number of results per query has been set to the maximum, 50.

A.1.2 L2 pages Retrieval Script

http://corpus.leeds.ac.uk/brunez/bitextcat.html#l1_url_retrieval

This script has been used to retrieve pages in the second language of interest. It can be easily customised changing the language URL identifiers (`en`, `eng`, `english`) with the ones of the intended language.

A.1.3 jusText

http://code.google.com/p/justext

jusText has been used to retrieve plain text content from webpages, using its standard settings.

A.1.4 Punkt

http://www.nltk.org/api/nltk.tokenize.html?highlight=punkt#module-nltk.tokenize.punkt

This Python module contained in the Natural Language Toolkit (NLTK) has been integrated into a Python script then used to split the sentences of Engita (and Enger).

Punkt uses as reference parameters files trained in the target languages, and the module provides already them for English, Italian and German.

A.1.5 Hunalign

http://mokk.bme.hu/en/resources/hunalign

Hunalign can use a bilingual dictionary as additional parameter to improve the quality of sentence alignments, so it has been provided with Italian-English and German-English dictionaries based on the ones provided with the CAT tool OmegaT[1].

[1] http://www.omegat.org.

A.2 Document analysis and dissimilarity

A.2.1 MALLET

http://mallet.cs.umass.edu/topics.php

MALLET has been used has been used for every topic modeling analysis contained in this thesis. It has always been used with hyperparameter optimisation.

A.2.2 TEABOAT

http://corpus.leeds.ac.uk/tools/teaboat.zip

TEABOAT (Term Equivalent Associator Based On Anchor Texts) is a suite of Python scripts designed to extract and align terminology (words and multiword expressions) from comparable corpora. The feature finding function and the subsequent creation of dissimilarity matrix based on it contained in Teaboat have been used in this project.

In specific three scripts contained in this suite have been employed: *dropdup.py* (discarding duplicates and near-duplicates), *flexlex.py* (extracting flexigrams) and *flexdifs.dat* (generating dissimilarity matrices).

A.2.3 R

http://www.r-project.org

Dissimilarity matrices produced with Teaboat have been further processed in R, in specific using R Studio[1] for an easier plot visualisation. Classic Multidimensional Scaling[2] in particular have been used to generate the dissimilarity matrices.

A.3 Moses

http://statmt.org/moses

[1] http://www.rstudio.com.

[2] http://stat.ethz.ch/R-manual/R-devel/library/stats/html/cmdscale.html.

Moses has been used to train translation models based on subsamplings selected using Teaboat and R with the set up as described in 4.1.1. Moses has been employed in its "vanilla" setting, as described in the section about how to train a baseline system in the Moses official website: http://www.statmt.org/moses/?n=Moses.Baseline (without Tuning). Data have been preprocessed in a format suitable for Moses with the tools provided with Moses itself.

References

Adamzik, K. (1995). *Textsorten – Texttypologie, Eine kommentierte Bibliographie*. Nodus, Münster. 28

Almeida, J.J.a. & Simões, A. (2010). Automatic Parallel Corpora and Bilingual Terminology extraction from Parallel WebSites. *Proceedings of the 3rd Workshop on Building and Using Comparable Corpora, LREC 2010*, 50–55. 24

Atwell, E., Arshad, J., Lai, C.M., Nim, L., Asheghi, N.R., Wang, J. & Washtell, J. (2007). Which English Dominates the World Wide Web, British or American? In *Proceedings of CL2007*, Birmingham. 52

Axelrod, A., He, X. & Gao, J. (2011). Domain Adaptation via Pseudo In-domain Data Selection. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, 355–362, Association for Computational Linguistics, Stroudsburg, PA, USA. 36

Baroni, M. & Bernardini, S. (2004). BootCaT: Bootstrapping corpora and terms from the web. In *Proceedings of the LREC 2004 conference*, vol. 4, 1313–1316, ELRA. 52

Baroni, M. & Bernardini, S. (2006). Wacky! Working papers on the Web as Corpus. GEDIT, Bologna, Italy. 21

Baroni, M., Chantree, F., Kilgarriff, A. & Sharoff, S. (2008). Cleaneval: a Competition for Cleaning Web Pages. In *LREC*, European Language Resources Association. 56

Berger, A.L., Brown, P.F., Della Pietra, S.A., Della Pietra, V.J., Gillett, J.R., Lafferty, J.D., Mercer, R.L., Printz, H. & Ureš, L. (1994). The Candide system for machine translation. In *Proceedings of the workshop on Human Language Technology*, HLT '94, 157–162, Association for Computational Linguistics, Stroudsburg, PA, USA. 16

Bertoldi, N. & Federico, M. (2009). Domain adaptation for statistical machine translation with monolingual resources. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, StatMT '09, 182–189, Association for Computational Linguistics, Stroudsburg, PA, USA. 33

Biber, D. (1988). *Variation across speech and writing*. Cambridge University Press. 28

Biber, D. & Conrad, S. (2009). *Register, Genre, and Style*. Cambridge Textbooks in Linguistics, Cambridge University Press. 28

Bloodgood, M. & Callison-Burch, C. (2010). Bucking the Trend: Large-scale Cost-focused Active Learning for Statistical Machine Translation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, 854–864, Association for Computational Linguistics, Stroudsburg, PA, USA. 10

Braune, F. & Fraser, A. (2010). Improved unsupervised sentence alignment for symmetrical and asymmetrical parallel corpora. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, 81–89, Association for Computational Linguistics, Stroudsburg, PA, USA. 19

Brown, P.F., Della Pietra, V.J., Della Pietra, S.A. & Mercer, R.L. (1993). The mathematics of statistical machine translation: parameter estimation. *Comput. Linguist.*, **19**, 263–311. 17

Brunello, M. (2012). Understanding the composition of parallel corpora from the web. In *Proceedings of the Seventh Web as Corpus Workshop (WAC7)*, 7–13. 2

Chen, J. & Nie, J.Y. (2000). Parallel Web text mining for cross-language information retrieval. In *Recherche d'Informations Assistée par Ordinateur (RIAO)*, 62–77, Paris. 23, 52

Chen, J., Chau, R. & Yeh, C.H. (2004). Discovering parallel text from the World Wide Web. In *Proceedings of the second workshop on Australasian information security, Data Mining and Web Intelligence, and Software Internationalisation - Volume 32*, ACSW Frontiers '04, 157–161, Australian Computer Society, Inc., Darlinghurst, Australia, Australia. 23

Civera, J. & Juan, A. (2007). Domain adaptation in statistical machine translation with mixture modelling. In *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, 177–180, Association for Computational Linguistics, Stroudsburg, PA, USA. 33

Daumé III, H. & Jagarlamudi, J. (2011). Domain Adaptation for Machine Translation by Mining Unseen Words. In *Association for Computational Linguistics*, Portland, OR. 34

Eisele, A. & Chen, Y. (2010). MultiUN: A Multilingual Corpus from United Nation Documents. In N.C.C. Chair), K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner & D. Tapias, eds., *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, European Language Resources Association (ELRA), Valletta, Malta. 22

Elizalde Cecilia, Pouliquen Bruno, Mazenc Christophe, J.G.V.J. (2012). TAPTA4UN: Collaboration on machine translation between the World Intellectual Property Organization and the United Nations. In *Proceedings of the Translating and the Computer Conference*. 11

Esplà-Gomis, M. & Forcada, M.L. (2010). Combining content-based and URL-based heuristics to harvest aligned bitexts from multilingual sites with Bitextor. In *Fourth Machine Translation Marathon Open Source Tools for Machine Translation*. 24

Ferraresi, A. (2007). *Building a very large corpus of English obtained by Web crawling: ukWaC*. Ph.D. thesis, University of Bologna. 53

Forsyth, R. & Holmes, D. (1996). Feature-finding for text classification. *Literary and Linguistic Computing*, **11**, 163–174. 32

Forsyth, R. & Sharoff, S. (2011). From crawled collections to comparable corpora: An approach based on automatic archetype identification. In *Proc. Corpus Linguistics Conference*, Birmingham. 32, 63

Forsyth, R. & Sharoff, S. (2014). Document dissimilarity within and across languages: a benchmarking study. *Literary & Linguistic Computing*, **29(1)**, 6–22. 64, 67

Foster, G. & Kuhn, R. (2007). Mixture-model adaptation for SMT. *Proceedings of the Second Workshop on Statistical Machine Translation*, 128–135. 33, 34

Foster, G., Goutte, C. & Kuhn, R. (2010). Discriminative instance weighting for domain adaptation in statistical machine translation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, 451–459, Association for Computational Linguistics, Stroudsburg, PA, USA. 34

Fry, J. (2005). Assembling a Parallel Corpus from RSS News Feeds. In *Proceedings of the Workshop on Example-Based Machine Translation, MT Summit X*, Phuket, Thailand. 24, 46

Gascó, G., Rocha, M.A., Sanchis-Trilles, G., Andrés-Ferrer, J. & Casacuberta, F. (2012). Does more data always yield better translations? In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, 152–161, Association for Computational Linguistics, Avignon, France. 11

Gavrila, M. & Vertan, C. (2014). Text Genre An Unexplored Parameter in Statistical Machine Translation. In Z. Vetulani & J. Mariani, eds., *Human*

Language Technology Challenges for Computer Science and Linguistics, 456–467, Springer, Pozna, Poland. 35

Haddow, B. & Koehn, P. (2012). Analysing the Effect of Out-of-domain Data on SMT Systems. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, WMT '12, 422–432, Association for Computational Linguistics, Stroudsburg, PA, USA. 11

Haghighi, A., Liang, P., Berg-Kirkpatrick, T. & Klein, D. (2008). Learning Bilingual Lexicons from Monolingual Corpora. In *Proceedings of ACL-08: HLT*, 771–779, Association for Computational Linguistics, Columbus, Ohio. 34

Haque, R., Naskar, S.K., van Genabith, J. & Way, A. (2009). Experiments on Domain Adaptation for English-Hindi SMT. In *In Proceedings of PACLIC 23: the 23rd Pacific Asia Conference on Language, Information and Computation*, 670–677, Hong Kong. 34

Hemming, C. & Lassi, M. (2003). Copyright and the Web as Corpus. 26

Jiang, J. (2008). A Literature Survey on Domain Adaptation of Statistical Classifiers. 32

Jimenez, S., Becerra, C. & Gelbukh, A. (2012). Soft Cardinality: A Parameterized Similarity Function for Text Comparison. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics - Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, SemEval '12, 449–453, Association for Computational Linguistics, Stroudsburg, PA, USA. 76

Kilgarriff, A. & Grefenstette, G. (2003). Introduction to the special issue on the web as corpus. *Computational Linguistics*, **29**, 333–347. 22

Kim, Y. & Ross, S. (2010). Formulating representative features with respect to document genre classification. In *Genres on the web*, Springer. 29

Kiss, T. & Strunk, J. (2006). Unsupervised Multilingual Sentence Boundary Detection. *Comput. Linguist.*, **32**, 485–525. 57

Koehn, P. (2004). Statistical Significance Tests for Machine Translation Evaluation. In D. Lin & D. Wu, eds., *Proceedings of EMNLP 2004*, 388–395, Association for Computational Linguistics, Barcelona, Spain. 84

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, 79–86, AAMT, Phuket. 21

Koehn, P. (2010). *Statistical machine translation*. Cambridge University Press, Cambridge. 9, 16, 19, 21

Koehn, P. & Schroeder, J. (2007). Experiments in domain adaptation for statistical machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, 224–227, Association for Computational Linguistics, Stroudsburg, PA, USA. 33, 34

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A. & Herbst, E. (2007). Moses: open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, 177–180, Association for Computational Linguistics, Stroudsburg, PA, USA. 18, 19

Lagoudaki, E. (2007). Challenges and Possibilities for Extracting Parallel Corpora from the Web The Translators Dream Scenario. In *Proceedings of the X Symposium on Social Communication*, 22–27, Santiago De Cuba. 23

Lavie, A. & Denkowski, M.J. (2009). The Meteor Metric for Automatic Evaluation of Machine Translation. *Machine Translation*, **23**, 105–115. 76

Lee, D.Y. (2001). Genres, registers, text types, domains and styles: clarifying the concepts and navigating a path through the BNC jungle. *Language Learning and Technology*, **5**, 37–72. 28, 29

Ma, X. & Liberman, M. (1999). Bits: A method for bilingual text search over the Web. In *Machine Translation Summit VII*. 23, 52

McCallum, A.K. (2002). MALLET: A Machine Learning for Language Toolkit. 39

Mohler, M. & Mihalcea, R. (2008). Babylon Parallel Text Builder: Gathering Parallel Texts for Low-Density Languages. In *LREC'08*. 23, 52

Moore, R.C. & Lewis, W. (2010). Intelligent Selection of Language Model Training Data. In *Proceedings of the ACL 2010 Conference Short Papers*, ACLShort '10, 220–224, Association for Computational Linguistics, Stroudsburg, PA, USA. 35

Niehues, J. & Waibel, A. (2010). Domain adaptation in statistical machine translation using factored translation models. *EAMT 2010 Proceedings of the 14th Annual conference of the European Association for Machine Translation*. 34

Orland, L. (2013). Intelligent Selection of Translation Model Training Data for Machine Translation with TAUS domain data. 36

Papineni, K., Roukos, S., Ward, T. & 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, 311–318, Association for Computational Linguistics, Stroudsburg, PA, USA. 76

Patry, A. & Langlais, P. (2011). Identifying Parallel Documents from a Large Bilingual Collection of Texts: Application to Parallel Article Extraction in Wikipedia. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*, 87–95, Portland, Oregon. 24

Pinnis, M., Ion, R., Stefanescu, D., Su, F., Skadina, I., Vasiljevs, A. & Babych, B. (2012). ACCURAT Toolkit for Multi-level Alignment and Information Extraction from Comparable Corpora. In *Proceedings of the ACL*

2012 System Demonstrations, ACL '12, 91–96, Association for Computational Linguistics, Stroudsburg, PA, USA. 23

Pomikalek, J. (2011). *Removing Boilerplate and Duplicate Content from Web Corpora*. Ph.D. thesis. 47, 56

Popovic, M. & Ney, H. (2006). Statistical Machine Translation with a Small Amount of Bilingual Training Data. In *LREC 2006*, 25–29, Genoa, Italy. 35

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. 67

Resnik, P. & Smith, N.A. (2003). The Web as a parallel corpus. *Comput. Linguist.*, **29**, 349–380. 23, 52, 54

Santini, M., Mehler, A. & Sharoff, S. (2010). *Genres on the Web: Computational Models and Empirical Studies*. Text, Speech and Language Technology, Springer. 31

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Comput. Surv.*, **34**, 1–47. 31

Sharoff, S. (2007). Classifying Web corpora into domain and genre using automatic feature identification. In *Proc. of Web as Corpus Workshop*, vol. 5, 1–10, Louvain-la-Neuve. 27

Sharoff, S. (2010). In the garden and in the jungle: Comparing genres in the BNC and Internet. In A. Mehler, S. Sharoff & M. Santini, eds., *Genres on the Web: Computational Models and Empirical Studies*, 149–166, Springer, Berlin/New York. 30, 63

Snover, M., Madnani, N., Dorr, B.J. & Schwartz, R. (2009). Fluency, Adequacy, or HTER?: Exploring Different Human Judgments with a Tunable MT Metric. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, StatMT '09, 259–268, Association for Computational Linguistics, Stroudsburg, PA, USA. 76

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufis, D. & Varga, D. (2006). The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006)*, Genoa, Italy. 22

Steyvers, M. & Griffiths, T. (2006). Probabilistic Topic Models. In T. Landauer, D. Mcnamara, S. Dennis & W. Kintsch, eds., *Latent Semantic Analysis: A Road to Meaning.*, Laurence Erlbaum. 31

Tsvetkov, Y. & Wintner, S. (2010). Automatic Acquisition of Parallel Corpora from Websites with Dynamic Content. In *LREC'10*. 24

Upton, G. & Cook, I. (2008). *The Oxford Dictionary of Statistics*. Oxford University Press, Oxford. 67

Varga, D., Németh, L., Halácsy, P., Kornai, A., Trón, V. & Nagy, V. (2005). Parallel corpora for medium density languages. In *Proceedings of the RANLP 2005*, 590–596. 19, 47, 57

Waller, R. (1987). *The typographic contribution to language*. Ph.D. thesis, University of Reading. 30

Wattam, S., Rayson, P. & Berridge, D. (2012). Document Attrition in Web Corpora: an Exploration. In N. Calzolari, K. Choukri, T. Declerck, M.U. Doan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk & S. Piperidis, eds., *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, European Language Resources Association (ELRA), Istanbul, Turkey. 25

Weaver, W. (1955). Translation. In W.N. Locke & A. Booth, eds., *Machine Translation of Languages*, MIT Press, Cambridge, MA, USA. 9, 16

Wu, H., Wang, H. & Zong, C. (2008). Domain adaptation for statistical machine translation with domain dictionary and monolingual corpora. In *Proceedings of the 22nd International Conference on Computational Linguistics -*

Volume 1, COLING '08, 993–1000, Association for Computational Linguistics, Stroudsburg, PA, USA. 33

Zanettin, F. (2002). Corpora in translation practice. In *LREC 2002*, 10–14, Las Palmas de Gran Canaria. 23