*P.D.NIMSARKAR
Professor of Linguistics (Former)
Dept. of Linguistics, Foreign and Indian Languages
RTM NAGPUR UNIVERSITY
NAGPUR

**K.DASARADHI
Research Scholar
RTM NAGPUR UNIVERSITY
NAGPUR

USAGE OF STATISTICAL MACHINE TRANSLATION IN TEXTUAL TRANSLATION

Abstract

The aim of this paper 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. 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. Discriminating between documents belonging to different textual varieties has been previously explored, but the present paper 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 sub-sampling 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 presentation on the web as 'multilingual corpus' has been addressed as well.

Key Words: analysis, collection, elements, evaluation, information, language, sentence

Introduction

The baseline approach of SMT (Statistical Machine Translation) is based on the analysis of probability distributions of segments contained in collections of bilingual texts in

the two languages of interest selected for the purpose. 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). 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 de ne 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.

SMT Vs RBMT

SMT can provide several advantages compared to traditional RBMT (Rule based Machine translation). However there are several open issues in SMT that still need to find a proper solution.

Two requirements are important in order to get a good translation from SMT: having a good coordination between the training data provided to the system and a specific text that needs to be translated in 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.

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. (Koehn 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 rst two language pairs and 40 million words for the latter. 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 and Gargantua 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 (Varga 45).

Parallel Texts

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 bi-texts, while in the translation industry the concept of translation memory (TM) is more specific and defines bi-texts 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.

Parallel corpus is instead used to de ne a large collection of bilingual texts, which can be provided in different formats depending on their size, purpose etc.

Current 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 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 (Baroni and Bernardini 21).

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 as a paid service, with fees in the order of thousands 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 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, 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. The texts contained in this multilingual corpus are provided both as monolingual with detailed XML markup 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.

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 is a corpus collecting the complete body of European Union law applicable to the member states, available in 22 European languages (Steinberger 22).

Talking about other proceedings of parliamentary debates, another well known parallel corpus is the Canadian Hansard, an English-French corpus containing debates from the

Canadian Parliament and another corpus made by institutional documents from the United Nations is MultiUN (Eisele & Chen 22).

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.

Collecting Corpora from Internet

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. (Kilgarri and Grefenstette 29).

This definitely 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 (Zanettin 23).

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 where the core Strand system is explained as well with several improvements comparing to the original core version (Resnik and Smith 54).

The main idea behind this approach is to find webpages that exhibit a parallel structure at the level of URL 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 (Resnik & Smith 54).

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 and Liberman 52).

A similar and more recent implementation of these strategies is described in Mohler & Mihalcea 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 accurate 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 (Pinnis 23).

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

1.crawl the web to nd potential multilingual pages or websites

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 46).

RSS syndication format and a similar strategy is proposed by Tsvetkov and Wintner with their system called PCB (Parallel Corpora Builder), which crawls news sites with dynamic content to obtain an English-Hebrew parallel corpus. Another concept is that describes a system called GWB (Get Web Bitext) 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 les from the web (as it happened in the previously mentioned Strand-alike approaches) (Almeida and Simoes 56).

A very recent alternative to the simple Strand-alike models is Paradocs, developed by which does not rely on le or URL naming informations, but rather works entirely with content-based features (numerical entities and hapax words), and it is a language neutral system (Patry and Langlais 62)

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 (Espla-Gomis and Forcada 24). 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 bi-texts from a given website, outputting them in TMX format containing segments aligned at the sentence level.

Genre studies and Text Categorization

Genre studies and text categorization issues have been object of investigation for decades (Adamzik 28) 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 un-supervised 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.

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 and Ross 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 categorize documents by structure and function".

While domain seems to be more intuitive and easy to de ne[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 based on look'n'feel" labels, reflecting the practical use of a text (recipe, review, Faq, blog post, academic paper etc.), (Sharo 169).

Even though intuitive, relying on the somehow shared definitions of text genre in a society and relying on inference and interpretation such 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). 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.(Waller 148).

Measuring Text Variability

The machine learning paradigm can be considered one of the most prominent methods for text classification (Sebastiani 34).

He 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 avail-ability 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 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 (Tsvetkov and Winter 24).

Typologies in SMT

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 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 (Jiang 32).

Conclusion

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 16).

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 16). SMT system is 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. 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 commercializing SMT systems, like Language Weaver (founded in 2002 and acquired in 2010 by SDL). 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.

Work Cited

Adamzik, K. Textsorten, *Texttypologie, Eine kommentierte Bibliographie*. Nodus, Munster. 1995

Almeida, J.J.a. & Simoes, A. *Automatic Parallel Corpora and Bilin-gual Terminology extraction from Parallel WebSites. Proceedings of the 3rd Workshop on Building and Using Comparable Corpora*, LREC. 2010

Baroni, M. & Bernardini, S. Wacky, 'Working papers on the Web as Corpus'. GEDIT, Bologna, Italy. 2006

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. & Ures, L. *The Candide system for machine translation. In Proceedings of the workshop on Human Language Technology*, HLT, Association for Computational Linguistics, Stroudsburg, PA, USA.1994

Brown, P.F., Della Pietra, V.J., Della Pietra, S.A. & Mercer, R.L. 'The mathematics of statistical machine translation: parameter estimation.' Comput. Linguist., 2003

Chen, J. & Nie, J.Y. 'Parallel Web text mining for cross-language infor-mation retrieval. In Recherche d'Informations Assistee par Ordinateur ' (RIAO), 2000

Chen, J., Chau, R. & Yeh, C.H. 'Discovering parallel text from the World Wide Web',.in Proceedings of the second workshop on Australasian in-formation security, Data Mining and Web Intelligence, and Software Interna-tionalisation - Volume 32 , ACSW Frontiers '04, Australian Computer Society, Inc., Darlinghurst, Australia, Australia. 2004

Eisele, A. & Chen, Y. 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. 2010

Espla-Gomis, M. & Forcada, M.L. 'Combining content-based and URL-based heuristics to harvest aligned bi-texts from multilingual sites with Bitextor' in Fourth Machine Translation Marathon Open Source Tools for Ma-chine Translation. 2010

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

Jiang, J. A Literature Survey on Domain Adaptation of Statistical Clas-si ers. 32, 2008

Kilgarriff, A. & Grefenstette, G. 'Introduction to the special issue on the web as corpus'. Computational Linguistics, 2003

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

Koehn, P. Europarl: 'A Parallel Corpus for Statistical Machine Translation', in Conference Proceedings: the tenth Machine Translation Summit, 2005

Koehn, P. *Statistical Machine Translation*. Cambridge University Press, Cambridge. 2010

Ma, X. & Liberman, M. Bits: 'A method for bilingual text search over the Web', in Machine Translation Summit VII. 1999

Patry, A. & Langlais, P. '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 Com-parable Corpora: Comparable Corpora and the Web, 87{95, Portland, Oregon. 2011

Pinnis, M., Ion, R., Stefanescu, D., Su, F., Skadina, I., Vasiljevs, A. & Babych, B. 'Accurate Toolkit for Multi-level Alignment and Information Extraction from Comparable Corpora' in Proceedings of the ACL 2012 System Demonstrations, ACL , Association for Computational Linguistics, Stroudsburg, PA, USA. 2012

Resnik, P. & Smith, N.A. The Web as a parallel corpus. Comput. Linguist., 2003

Sebastiani, F. *Machine Learning in Automated Text Categorization*. ACM Comput. Surv., 2002

Sharoff, S. 'In the garden and in the jungle: Comparing genres in the BNC and internet', in A. Mehler, S. Sharo & M. Santini, eds., *Genres on the Web: Computational Models and Empirical Studies*, Springer, Berlin/New York. 2010

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufis, D. & Varga, D. 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. 2006

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

Tsvetkov, Y. & Wintner, S. 'Automatic Acquisition of Parallel Cor-pora from Websites with Dynamic Content', in LREC'10, 24, 2010

Varga, D., Nemeth, L., Halacsy, P., Kornai, A., Tron, V. & Nagy, V. 'Parallel corpora for medium density languages', in *Proceedings of the RANLP*. 2005

Waller, R. *The typographic Contribution to Language*. Ph.D. thesis, University of Reading. 1987

Weaver, W. , *Machine Translation of Languages*, MIT Press, Cambridge, MA, USA.1955

Zanettin, F. ‘Corpora in translation practice’.in LREC, 10{14, Las Palmas de Gran Canaria}. 2002