SMT for restricted sublanguage in CAT tool context at the European Parliament

Najeh Hajlaoui
European Parliament
Luxembourg

ABSTRACT

This paper shows that it is possible to efficiently develop Statistical Machine Translation (SMT) systems that are useful for a specific type of sublanguage in real context of use even when excluding the exact Translation Memory (TM) matches from the test set in order to be integrated in CAT "Computer Aided Translation" tools. It means that the included part is quite different from the existing translations and consequently harder to translate even for an SMT system trained on the same translation data.

Because we believe on the proximity of sublanguages even though it is still hard to practically define the sublanguage notion, we are proposing on the framework of the MT@EP project at the Directorate General for Translation (DG TRAD) of the European Parliament (EP) to develop SMT systems specific for each EP Parliamentary Committee optimised for restricted sublanguages and constrained by the EP's particular translation requirements.

Sublanguage-specific systems provide better results than generic systems for EP domains showing a very significant quality improvement (5-25% of BLEU score), mainly due to the EP context specificity and to the proximity of sublanguages. This approach is also satisfactory for pairs of under-resourced languages, such as the Slavic families and German.

1. Previous work

In general, a sublanguage is a subset of the language (Harris, 1970) identified with a particular semantic domain or a linked family of domains (Kittredge, 1978), (Kittredge, 1982). In our previous research (Hajlaoui, 2008), we showed that, despite the great distance between mother languages (e.g. Arabic and French) (Hajlaoui, Daoud, & Boitet, 2008), the two correspondent sublanguages are very near one to another as shown in Figure 1. It was a new illustration of Kittredge's analysis (Kittredge, 1993), (Kittredge & Lehrberger, 1982).

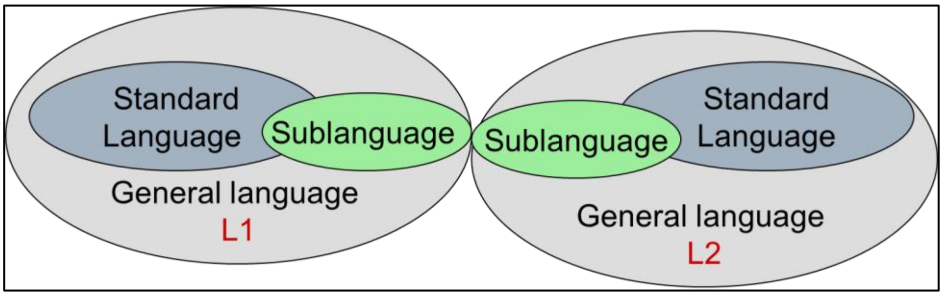

Figure 1: Sublanguages are very near one to another

We also showed that SMT system works very well for small sublanguages with a very small training corpus (less than 10 000 words) (Hajlaoui & Boitet, 2008). This proves that, in the case of very small sublanguages, SMT may be of sufficient quality, starting from a corpus 100 to 500 smaller than for the general language. We are proving in this work the validity of this approach in real context of use clarifying and answering some related questions. We describe also the type of resources we need, mainly Thematic Translation Memory (TTM) presenting some promising results.

The issue consists to do further research on the type of sublanguages for which it is possible to develop efficient (useful) SMT systems in the context of CAT tools.

In the following section, we are testing our conjuncture which consists that SMT systems work very well for domain sublanguages using small training corpus.

2. SMT applied on sublanguages

Assuming that a sublanguage is a subset of the language identified with a particular semantic (family of) domain, in our EP context, health, environment, economy, etc. seems to constitute the restricted sublanguages we are looking for.

In the context of existing applications developed in the DG TRAD, one of the constraints to take into account concerning the use of MT in the EP workflows is to take the ad-hoc vocabulary to translate EP documents (amendments, laws, etc.). Our objective is to help EP translators by reusing an existing base of Translation Memory data to better translate unmatched sentences. Our technical choice involves automatic selection of data to resolve problems of context and quality.

The corpus must obviously reach a critical size to allow reliable statistical treatment. SMT approach works very well for restricted domains with little or no human revision, for example the rules-based TAUM-METEO system is purposely developed for the weather service in Canada to provide weather forecasts in French and English.

Based on some statistical information, we know that environment (ENVI), economy (ECON), and Control of budget (CONT) are ones of the main domains in terms of number of documents treated at the EP. Consequently, we built SMT systems for those domains using the Moses decoder (Koehn, et al., 2007) with the phrase-based factored translation models (Koehn & Hieu, 2007) to mainly translate from English to French. The language models for French were 3-gram ones over each training domain data using the IRSTLM toolkit (Federico, 2008). We used

Minimum Error Rate Training (MERT) (Och, 2003) to optimize the systems. The domain data is randomly split to three sets (training, tuning, and testing).

The following tables (Table 1, Table 2 and Table 3) show BLEU, METEOR and TER scores of our English-to-French SMT systems build specifically for three domains announced above comparing them to the existing "Generic" systems and to "Google" MT systems. The scores are computed on tokenized, truecased text, using the MultEval tool version 0.5.1 (Clark, 2011). The BLEU score show that the improvement can reach 25% of a "Specific" system over the "Generic" system and much more (about 29%) over "Google" system depending on the domain.

MT systems	Training set (Nb. Sent)	Tuning set (Nb. Sent)	Testing set (Nb. Sent)	BLEU	METEOR	TER
Specific	49003	1020	1020	65.7	75.7	29.9
Generic	NA	NA	1020	40.6	56.8	45.0
Google	NA	NA	1020	36.1	53.9	47.4

Table 1: BLEU, METEOR and TER English-French scores for the CONT (Control of budget) domain

MT systems	Training set (Nb. Sent)	Tuning set (Nb. Sent)	Testing set (Nb. Sent)	BLEU	METEOR	TER
Specific	106736	2207	2207	60.4	72.6	32.6
Generic	NA	NA	2207	44.7	61.8	43.6
Google	NA	NA	2207	43.3	61.2	42.3

Table 2: BLEU, METEOR and TER English-French scores for the ENVI (Environment) domain

MT systems	Training set (Nb. Sent)	Tuning set (Nb. Sent)	Testing set (Nb. Sent)	BLEU	METEOR	TER
Specific	101669	936	936	58.6	70.9	35.8
Generic	NA	NA	936	41.5	57.8	45
Google	NA	NA	936	34.2	52.1	49.7

Table 3: BLEU, METEOR and TER Englis-French scores for the ECON (Economy) domain

The first results are promising. They showed a general improvement of 5%-25% of BLEU score over generic MT systems depending on the domain and on the test set. They don't concern only English-to-French but the approach is also satisfactory for pairs of under-resourced languages, such as the Finno-Ugric or Slavic families and German (tested for English-to-German, English-to-Estonian and English-to-Bulgarian). It is mainly due to the lexical convergence which is the main characteristic of restricted sublanguage. It is also due to the EP context specificity and to the proximity of sublanguages.

Contrary to the huge volume of data used to develop generic SMT systems[1], the training data used to develop specific systems are very small[2]. However the choice of the data sets is very important. In fact, a specific sublanguage training set avoids the introduction of out-domain vocabulary and a representative test set is more relevant than a single EP document. Based on a single EP document the translation result cannot be generalized since it depends on the matching chance of that document with the training data. Consequently, it is very important to take a representative test set of the domain data.

In order to integrate the specific SMT service in CAT tool, we would like in the next section to see whether this approach is still working when we exclude the part of the test set which can be translate with Translation Memory (full matching).

3. SMT in CAT tool context

In order to combine SMT with Translation Memories (TM), we would like to exclude the part having exact TM matches from the test set keeping as possible the fact that the way to select a test set for a specific system is a bit different from the generic system case. The test set should be in-domain and it should be representative of the domain to be able to provide a general conclusion. We called this part of sentences to be excluded from the test set "natural overlap" in order to make difference between it and the "artificial overlap": the "natural overlap" is the basic function in the SMT approach, which might be important in the case of a restricted sublanguage (small domain) due to the lexical convergence and the limitation of the vocabulary; it is one of the main features of a given sublanguage. While an "artificial overlap" consists to include the test set or a part of it in the training set which is of course forbidden.

As defined, the "natural overlap" is the part of the test set[3] which have an exact TM matches. Consequently, in order to detect the "natural overlap" called also "lexical convergence", four cases can be distinguished.

- Same source but different target
- Same source and same target
- Same target but different source
- Different source and different target

In our actual experiments, we defined it as having the same source and the same target because it happens that even with exact TM matches, users need to post-edit the translation as shows the following French-to-English example for research domain.

- Source: il fait de la recherche.
- TM source: il fait de la recherche.
- TM target: he is doing search.
- Reference: he is doing research.

[1] For instance, it is around 20 millions of sentences: 380 millions of words for English and 415 millions of words for French.

[2] In general, it is less than 100 000 sentences.

[3] It might happen even though the domain data are randomly divided into three data sets.

By updating CONT domain data to train the specific system and taking a bigger test set, we obtained a 23.4% improvement of BLEU scores of the EP Specific system over the Generic system for the full test set (3678 sentences[4]). The BLEU score went from 46.5% to 69.9% with the specific system as show in Table 4. In this experiment, we used a larger test set to restrict it to only 2017 sentences by excluding the part having the same source and the same target.

By excluding sentences having exact matches with translation memories, we might generally reduce the representativity of the domain in the test set but we still have an improvement over a generic system 16.5% of BLEU score for the CONT domain (English-to-French). The BLEU score went from 47.2% to 63.7% with the specific system.

MT system	Training set (Nb. sent)	Tuning set (Nb. sent)	Testing set (Nb. sent)	BLEU	METEOR	TER
Specific	61185	1226	3678	69.9	79.6	24.8
Generic	NA	NA	3678	46.5	62.6	41.3
Specific	61185	1226	2017	63.7	75.3	30
Generic	NA	NA	2017	47.2	63.3	40.6

Table 4: Third comparison between specific and generic system results

This improvement on MT quality can be converted to monetary benefit. It is important to demonstrate by extrapolation that a huge number of words will be better translated showing at the end an important benefit by reducing the post-editing time.

4. Conclusion and perspectives

We showed that it is possible to efficiently build SMT systems that are useful for specific sublanguages. In the actual context of use, we excluded the exact TM matches from the test set since TM is prioritised to MT to keep the same translation terminology. Because our specific system is built based a local data selected for a given specific domain, we think that it might properly respect also the same terminology as the TM. It is one of the challenges that might be tested for the future work.

The performance of a specific system is proportional to the coverage of the domain. The coverage is usually reached after a certain size of training data. In the next experiments, we will try to automatically detect the domain coverage using some previous experiences as well.

We are developing in a first step a limited number of specific domain systems for some language pairs such as EN-FR, EN-DE, etc. Then, they will be hosted to allow human evaluation involving EP users. We will mainly measure the post-editing time spent by users, which is the

[4] Average length equal to 27 words.

main indicator of MT quality improvement. We should note that human factors like correlation, instability over time should be taken into account during the human evaluation campaign.

The sublanguage specific systems will be integrated in the EP translation workflows to improve Translation Memory results offering in priority previous human translations. To more prioritize Translation Memory, the development of an algorithm to translate only unmatched segments with MT is in progress. It include all the sentences that have a higher match score (e.g. between 82% and 99%).

References

Clark, J. C. (2011). Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. *In Proceedings of ACL-HLT 2011 (46th Annual Meeting of the ACL: Human Language Technologies.* Portland, OR.

Federico, M. N. (2008). IRSTLM: an open source toolkit for handling large scale language models. *In Proceedings of Interspeech.* Brisbane, Australia.

Hajlaoui, N. (2008). *Multilinguïsation de systèmes de e-commerce traitant des énoncés spontanés en langue naturelle.* Grenoble, France: Thèse. Université Joseph Fourier.

Hajlaoui, N., & Boitet, C. (2008). TA statistique à petits corpus pour de petits sous-langages. . *Proc. ToTh 2008 :Conférence sur la Terminologie & Ontologie : Théories et Applications*, (p. 15). France - Annecy, .

Hajlaoui, N., Daoud, D., & Boitet, C. (2008). Methods for porting NL-based restricted e-commerce systems into other languages. *Proc. LREC 2008 "Language Resources and Evaluation Conference"* , (p. 7). Maroc Marrakech.

Harris, Z. (1970). Mathematical structures of language. *Mathematical Gazette. Vol. 54(388)*, 173-174.

Kittredge, R. (1978). Textual cohesion within sublanguages: implications for automatic analysis and synthesis. *Proc. Coling-78. Vol. 1/1. August 14-18.* Bergen, Norvège.

Kittredge, R. (1982). Variation and Homogeneity of Sublanguages. in Sublanguage - Studies of Language in Restricted Semantic Domains., (p. 20). Walter de Gruyter. Berlin / New York.

Kittredge, R. (1993). Sublanguage Analysis for Natural Language Processing. *Proc. First Symposium on Natural Language Processing.*, (pp. 69-83.). Thailand, Bangkok.

Kittredge, R., & Lehrberger, J. (1982). *Sublanguage - Studies of language in restricted semantic domain.* Walter de Gruyter. Berlin / New York.

Koehn, P. (2005). A Parallel Corpus for Statistical Machine Translation. *Machine Translation Summit.*

Koehn, P., & Hieu, H. (2007). Factored Translation Models. *In Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing (EMNLP) and Computational Natural Language Learning (CONLL)*, (pp. 868-876). Prague, Czech Republic.

Koehn, P., Hieu, H., Alexandra, B., Chris, C.-B., Marcello, F., Nicola, B., et al. (2007). Moses: Open Source Toolkit for Statistical Machine Translation. *In Proceedings of 45th Annual Meeting of the Association for Computational Linguistics (ACL), Demonstration Session*, (pp. 177-180). Prague, Czech Republic.

Och, F. J. (2003). Minimum Error Rate Training in Statistical Machine Translation. *In Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL)*, (pp. 160-167). Sapporo, Japan.