Automatic Translation of Scientific Documents in the HAL Archive

Patrik Lambert*, Holger Schwenk*, Frédéric Blain*†

* LIUM, University of Le Mans
Avenue Laennec, 72085 Le Mans cedex, France
† SYSTRAN SA, Paris, France
Name.Surname@lium.univ-lemans.fr

Abstract

This paper describes the development of a statistical machine translation system between French and English for scientific papers. This system will be closely integrated into the French HAL open archive, a collection of more than 100.000 scientific papers. We describe the creation of in-domain parallel and monolingual corpora, the development of a domain specific translation system with the created resources, and its adaptation using monolingual resources only. These techniques allowed us to improve a generic system by more than 10 BLEU points.

Keywords: domain adaptation, machine translation, scientific literature

1. Introduction

Due to the globalization of research, the English language is today the universal language of scientific communication. In France, regulations require the use of the French language in progress reports, academic dissertations, manuscripts, and French is the official educational language of the country. This situation forces researchers to frequently translate their own articles, lectures, presentations, reports, and abstracts between English and French. In addition, students and the general public are also challenged by language, when it comes to find published articles in English or to understand these articles.

This problem, incorrectly resolved through the use of generic translation tools, actually reveals an interesting generic problem where a community of specialists are regularly performing translations tasks on a very limited domain. At the same time, another community of users seeks translations for the same type of documents. Without appropriate tools, the expertise and time spent for translation activity by the first community is lost and do not benefit to translation requests of the second community. From the international point of view, we see the reverse problem – where specialists are simply not considering French papers because they are missing language expertise. We can therefore list at least three types of actors:

1. French scientists – writing natively in French, and good enough in English. These scientists are passing a noticeable amount of time for translating their own publication from English to French or vice-versa. Their translation effort could be leveraged for the other actors through appropriate work environment.

2. French public looking for publications only available in English and are using online translation tools not appropriate for such translations.

3. International scientists not even considering to look for French publications (for instance PhD theses) because they are not available in their native languages. These actors are specialists in the field and possible English native speaker and could use interactive translation tools.

Building efficient translation tools for scientific texts which are automatically adapted to various scientific areas is the challenge of the French National project COSMAT.[1] These systems will be closely integrated into the HAL open archive[2], a multidisciplinary open-access archive which was created in 2006 to archive publications from all the French scientific community. In addition to provide the automatic translations, the interface will also allow to post-edit the output, which itself will be then used to improve the translations engines.

2. Task Description

In this paper we describe the development of a phrase-based statistical machine translation system (PBSMT) for the translation of scientific documents in several domains from French to English and from English to French. These documents in the HAL archive may be scientific research papers (published or not), PhD theses (contained in the connected TEL archive), scientific reports, etc., coming from public or private teaching and research institutions in France or abroad. HAL contains more than 100,000 documents from about 30 scientific domains. For development and evaluation of our PBSMT system, we focused on two of the 30 domains, namely computer science and physics. Large amounts of monolingual and parallel data are available to train a SMT system between French and English, but not in the scientific domain. In order to improve the performance of our translation system in this task, we extracted in-domain monolingual and parallel data from the HAL archive. This process is explained in the following, before describing out-of-domain resources also used as training data.

[1] http://www.cosmat.fr

[2] http://hal.archives-ouvertes.fr/?langue=en

Corpus	Sentences	Selected			Random		
		Words	OOV (%)	Single (%)	Words	OOV (%)	Single (%)
Source: fr							
dev info	1100	28334	0.86	58.44	29517	0.62	58.11
dev phys	1000	28626	0.87	60.35	29805	0.60	59.26
test info	1100	28704	0.88	57.58	29107	0.64	59.15
test phys	1000	28277	0.91	59.73	28656	0.59	59.69
Source: en							
dev info	1100	25767	0.87	57.64	25670	0.66	57.61
dev phys	1000	26021	1.02	57.93	26579	0.64	57.23
test info	1100	26135	0.89	56.24	25546	0.70	55.84
test phys	1000	25859	1.12	58.21	25482	0.61	57.45

Table 1: Number of sentences, words, OOV words and singletons for the development and test data: (left) chosen randomly in a subset selected according to IBM 1 model scores, and (right) development and test data selected totally randomly.

Set	Domain	Lang.	Sent.	Words	Vocab.
Parallel data					
Train	cs+phys	En	55.9 k	1.41 M	43.3 k
		Fr	55.9 k	1.63 M	47.9 k
Dev	cs	En	1100	25.8 k	4.6 k
		Fr	1100	28.7 k	5.1 k
	phys	En	1000	26.1 k	5.1 k
		Fr	1000	29.1 k	5.6 k
Test	cs	En	1100	26.1 k	4.6 k
		Fr	1100	29.2 k	5.2 k
	phys	En	1000	25.9 k	5.1 k
		Fr	1000	28.8 k	5.5 k
Monolingual data					
Train	cs	En	2.5 M	54 M	457 k
		Fr	761 k	19 M	274 k
	phys	En	2.1 M	50 M	646 k
		Fr	662 k	17 M	292 k

Table 2: Basic statistics for the parallel training, development, and test data sets extracted from thesis abstracts contained in HAL, as well as monolingual data extracted from all documents in HAL, in the two domains of focus: computer science (cs), and physics (phys). The following statistics are given for the English (en) and French (fr) sides of the corpus: the number of sentences, the number of running words (after tokenisation) and the number of words in the vocabulary (M and k stand for millions and thousands, respectively).

Extraction of In-domain Resources

The HAL pdf files corresponding to the computer science and physics domains were made available to us. The pdf files were then converted to plain text (via the TEI format) using the Grobid[3] open-source converter. The documents in HAL are nearly exclusively monolingual, but the thesis from French universities must include both an abstract in French and in English. Although in some cases the two abstracts may not be strictly parallel translations or may contain translation errors, our experiments show that these abstracts turned out to be useful parallel data.

The abstracts were first aligned at the sentence level. Then training, development and test data were selected in the following way. To avoid including incorrectly aligned sentence pairs in the development and test data, the selection was performed based on the cost of the IBM Model 1 (Brown et al., 1993) for each sentence pair. The development and test data sets were chosen at random within the subset of sentence pairs whose IBM 1 score satisfies a criterion. About 50k running words for each of the computer science and physics domains were selected. The rest of the data was used as training set. The statistics of these parallel data sets are given in Table 2.

In the case of the development and test data, since more frequent words give a smaller contribution to the IBM 1 model cost, this selection method may introduce a bias and favor sentences easier to translate. To check this, we compared the number of out-of-vocabulary (OOV) words and the number of singletons in our selected sets and in randomly selected sets with the same number of sentence pairs (see Table 1). The vocabulary taken into account for the OOV count was the vocabulary of a corpus containing Europarl, News-Commentary, 10^{9}_{2} data (see the description of out-of-domain data below), as well as the theses abstracts. As the counts in Table 1 show, the number of OOV or singleton words was not higher in our selected development and test sets, suggesting that the selected set was not significantly easier to translate than a randomly selected set.

We also extracted (with Grobid) text from all documents deposited in HAL in computer science and physics to train our language model (see the statistics of these monolingual data in Table 2).

Out-of-domain Training Data Used

The data extracted from HAL was used to adapt to the scientific literature domain a generic system mostly trained on data provided for the shared task of Sixth Workshop on Statistical Machine Translation[4] (WMT 2011). These data are described in Table 3.

[3] http://grobid.no-ip.org

[4] http://www.statmt.org/wmt11/translation-task.html

Source	Translation Model	Language Model	Tuning Domain	CS		PHYS	
				words (M)	Bleu	words (M)	Bleu
En	wmt11	wmt11	wmt11	371	27.3	371	27.1
En	wmt11	wmt11	hal	371	28.4	371	28.3
En	wmt11	wmt11+hal	hal	371	36.0	371	36.2
En	hal	wmt11+hal	hal	1.4	37.2	1.4	38.8
En	wmt11+hal	wmt11+hal	hal	287	38.3	287	39.3
En	wmt11+hal+adapted	wmt11+hal	hal	299	38.8	307	40.0

Table 4: Results (BLEU score) for the English–French systems. The type of parallel data used to train the translation model or language model are indicated, as well as the set (in-domain or out-of-domain) used to tune the models. Finally, the number of words in the parallel corpus and the BLEU score on the in-domain test set are indicated for each domain: computer science and physics.

Corpus	English	French
Bitexts:		
Europarl	50.5M	54.4M
News Commentary	2.9M	3.3M
Crawled (10^9 bitexts)	667M	794M
Development data:		
newstest2009	65k	73k
newstest2010	62k	71k
Monolingual data:		
LDC Gigaword	4.1G	920M
Crawled news	2.6G	612M

Table 3: Out-of-domain development and training data used (number of words after tokenisation).

The parallel out-of-domain data used were the Europarl corpus (European Parliament proceedings), the News-commentary corpus (quality commentary articles about the news) and a selection[5] of the French–English 10^9 corpus (mostly crawled from bilingual Internet sites).

The monolingual out-of-domain data used to train our language model were the monolingual version of the bitexts, the News corpus provided at WMT 2011 (crawled from the web) and LDC's Gigaword collection.

Finally, we employed the development data provided at WMT 2011 to compare the tuning of the system with out-of-domain or in-domain data.

3. Translation Adaptation Experiments

3.1. Concatenation of Parallel Corpora

In this section we evaluate the impact of introducing the resources extracted from HAL into a generic baseline system. The results are reported in Table 4. The baseline system is a standard PBSMT system based on Moses (Koehn et al., 2007) and SRILM (Stolcke, 2002) and trained and tuned only on WMT11 data (out-of-domain). We first train SMT models by just concatenating the in- and out-of-domain parallel corpora.

Tuning the translation model[6] on our HAL development set, we gained more than a BLEU point. Incorporating the HAL data into the language model and tuning the system on the HAL development set[7], the gain is more than 7 BLEU points, in both domains (computer science and physics). Using only in-domain data (1.4 million words of theses abstracts) as parallel corpus, the gain is 1.2 BLEU point in computer science and 2.6 BLEU points in physics with respect to using 371 million words of out-of-domain data. Concatenating the theses abstracts and the out-of-domain data in the parallel training corpus, a further gain of 1.1 BLEU points is observed for computer science, and 0.5 points for physics.

The last experiment performed aims at increasing the amount of in-domain parallel texts by translating automatically in-domain monolingual data, as suggested by Schwenk (2008). The synthesized bitext does not bring new words into the system, but increases the probability of in-domain bilingual phrases. By adding a synthetic bitext of 12 million words to the parallel training data, we observed a gain of 0.5 BLEU point for computer science, and 0.7 points for physics.

3.2. Weighting In- and Out-of-domain Translation Models

Instead of simply concatenating the parallel in- and out-of-domain parallel corpora, we used them to train distinct phrase tables. The weights of the corresponding translation models in the SMT log-linear combination were optimised via mininum error-rate training (MERT (Och, 2003)). This method is similar to that of Koehn and Schroeder (2007). The results are presented in Table 5. For all systems in Table 5, the translation direction is English–French, the language model is that referred to as "wmt11+hal" in Table 4 and the set used for tuning is the in-domain set. Thus the

[5] We applied the same two filters as Lambert et al. (2010) to select this subset. The first one is a lexical filter based on the IBM model 1 cost of each side of a sentence pair given the other side, normalised with respect to both sentence lengths. This filter was trained on a corpus composed of Europarl, News-commentary, and United Nations bitexts. The other filter is an n-gram language model cost of the target sentence, normalised with respect to its length. This filter was trained with all monolingual resources available except the 10^9 data.

[6] We re-used the previous language model, which had been built via linear interpolation of the models trained for the different corpora involved, tuning the interpolation coefficients on the wmt11 development set.

[7] this time both the translation model coefficients and language model interpolation coefficients were tuned on the HAL development set.

Translation Model	CS	PHYS
wmt11+hal concatenated	38.3	39.3
wmt11+hal two PT	37.3	39.1
wmt11+hal two PT with back-off	36.8	38.7

Table 5: BLEU score results of phrase-table combinations.

first line of Table 5, which refers to the baseline with only one phrase table trained on concatenated in-domain and out-of-domain parallel data, refers to the same system as the penultimate line of Table 4. The two remaining lines of Table 5 refer to systems with two phrase tables (one trained on WMT'11 data and the other one with HAL data). The difference lies in the way decoding is performed. In the first case ("two PT"), the phrase pairs belonging to one or the other table are exploited independently. This means that the translation options for a given phrase pair are collected in both tables, with their respective scores. If a phrase pair is present in both tables, it will thus appear twice in the set of translation options, with a different score, and translation hypotheses will be formed with each score. In the second case ("two PT with back-off"), the out-of-domain table is used only as back-off of the in-domain table. The out-of-domain table phrase pairs are only used to translate phrases which are unknown in the in-domain table. In both cases, the reordering model was build from the concatenation of in-domain and out-of-domain corpora. As show the results of our experiments in Table 5, two independent phrase tables worked better than having the out-of-domain table as back-off of the in-domain one. However, no improvement with respect to the baseline trained on concatenated parallel corpora was achieved. This may be due to the increased instability of the MERT process with 5 additional coefficients to optimise (Foster and Kuhn, 2009).

4. Conclusion

This paper described ongoing work to develop a statistical machine translation system to translate scientific texts between French and English. We started with a state-of-the-art system optimized for the news domain (WMT'11 evaluation). Domain specific parallel data for training, development and test was automatically extracted from bilingual abstracts of French theses. We plan to make these corpora freely available to stimulate research on the translation of scientific texts. We also performed an adaptation of the translation model using monolingual data only. All these techniques lead to an improvement of the BLEU score of more than 10 BLEU points.

We tried decoding with two distinct phrase tables trained respectively on the in-domain and out-of-domain data, but did not achieve any improvement with this set-up.

In the future, we will work on a more fine-grained adaptation of the language and translation models to sub-domains, for instance data bases, networking, AI, etc instead of one generic domain "computer science". The translation systems will be smoothly integrated into the HAL archive, delivering quick translations of abstracts and whole PDF documents. The end user will also have the possibilities to post-edit the automatic translations. These corrections will be used to improve the models of our system.

Acknowledgments

This work has been partially funded by the French Government under the project COSMAT (ANR ANR-09-CORD-004). We would like to thank Patrice Lopez and Laurent Romary for the help with extracting the data from the HAL archive.

5. References

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

George Foster and Roland Kuhn. 2009. Stabilizing minimum error rate training. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 242–249, Athens, Greece.

Philipp Koehn and Josh Schroeder. 2007. Experiments in domain adaptation for statistical machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 224–227, Prague, Czech Republic.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proc. of the 45th Annual Meeting of the Association for Computational Linguistics (Demo and Poster Sessions)*, pages 177–180, Prague, Czech Republic, June.

Patrik Lambert, Sadaf Abdul-Rauf, and Holger Schwenk. 2010. LIUM SMT machine translation system for WMT 2010. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 121–126, Uppsala, Sweden.

Franz J. Och. 2003. Minimum error rate training in statistical machine translation. In *Proc. of the 41th Annual Meeting of the Association for Computational Linguistics*, pages 160–167.

Holger Schwenk. 2008. Investigations on large-scale lightly-supervised training for statistical machine translation. In *IWSLT*, pages 182–189.

Andreas Stolcke. 2002. SRILM: an extensible language modeling toolkit. In *Proc. of the Int. Conf. on Spoken Language Processing*, pages 901–904, Denver, CO.