Collaborative Machine Translation Service for Scientific texts

Patrik Lambert
University of Le Mans
patrik.lambert@lium.univ-lemans.fr

Jean Senellart
Systran SA
senellart@systran.fr

Laurent Romary
Humboldt Universität Berlin /
INRIA Saclay - Ile de France
laurent.romary@inria.fr

Holger Schwenk
University of Le Mans
holger.schwenk@lium.univ-lemans.fr

Florian Zipser
Humboldt Universität Berlin
f.zipser@gmx.de

Patrice Lopez
Humboldt Universität Berlin /
INRIA Saclay - Ile de France
patrice.lopez@inria.fr

Frédéric Blain
Systran SA /
University of Le Mans
frederic.blain@
lium.univ-lemans.fr

Abstract

French researchers are required to frequently translate into French the description of their work published in English. At the same time, the need for French people to access articles in English, or to international researchers to access theses or papers in French, is incorrectly resolved via the use of generic translation tools. We propose the demonstration of an end-to-end tool integrated in the HAL open archive for enabling efficient translation for scientific texts. This tool can give translation suggestions adapted to the scientific domain, improving by more than 10 points the BLEU score of a generic system. It also provides a post-edition service which captures user post-editing data that can be used to incrementally improve the translations engines. Thus it is helpful for users which need to translate or to access scientific texts.

1 Introduction

Due to the globalisation 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. Finally, international scientists not even consider to look for French publications (for instance PhD theses) because they are not available in their native languages. 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, other communities of users seek 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 other communities.

We propose the demonstration of an end-to-end tool for enabling efficient translation for scientific texts. This system, developed for the COSMAT ANR project,[1] is 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. The tool deals with handling of source document format, generally a pdf file, specialised translation of the content, and user-friendly user-interface allowing to post-edit the output. Behind

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

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

Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, pages 11–15,
Avignon, France, April 23 - 27 2012. ©2012 Association for Computational Linguistics

the scene, the post-editing tool captures user post-editing data which are used to incrementally improve the translations engines. The only equipment required by this demonstration is a computer with an Internet browser installed and an Internet connection.

In this paper, we first describe the complete work-flow from data acquisition to final post-editing. Then we focus on the text extraction procedure. In Section 4, we give details about the translation system. Then in section 5, we present the translation and post-editing interface. We finally give some concluding remarks.

The system will be demonstrated at EACL in his tight integration with the HAL paper deposit system. If the organizers agree, we would like to offer the use of our system during the EACL conference. It would automatically translate all the abstracts of the accepted papers and also offers the possibility to correct the outputs. This resulting data would be made freely available.

2 Complete Processing Work-flow

The entry point for the system are "ready to publish" scientific papers. The goal of our system was to extract content keeping as many meta-information as possible from the document, to translate the content, to allow the user to perform post-editing, and to render the result in a format as close as possible to the source format. To train our system, we collected from the HAL archive more than 40 000 documents in physics and computer science, including articles, PhD theses or research reports (see Section 4). This material was used to train the translation engines and to extract domain bilingual terminology.

The user scenario is the following:

- A user uploads an article in PDF format[3] on the system.

- The document is processed by the open-source Grobid tool (see section 3) to extract the content. The extracted paper is structured in the TEI format where title, authors, references, footnotes, figure captions are identified with a very high accuracy.

- An entity recognition process is performed for markup of domain entities such as: chemical compounds for chemical papers, mathematical formulas, pseudo-code and object references in computer science papers, but also miscellaneous acronyms commonly used in scientific communication.

- Specialised terminology is then recognised using the Termsciences[4] reference terminology database, completed with terminology automatically extracted from the training corpus. The actual translation of the paper is performed using adapted translation as described in Section 4.

- The translation process generates a bilingual TEI format preserving the source structure and integrating the entity annotation, multiple terminology choices when available, and the token alignment between source and target sentences.

- The translation is proposed to the user for post-editing through a rich interactive interface described in Section 5.

- The final version of the document is then archived in TEI format and available for display in HTML using dedicated XSLT style sheets.

3 The Grobid System

Based on state-of-the-art machine learning techniques, Grobid (Lopez, 2009) performs reliable bibliographic data extraction from scholar articles combined with multi-level term extraction. These two types of extraction present synergies and correspond to complementary descriptions of an article.

This tool parses and converts scientific articles in PDF format into a structured TEI document[5] compliant with the good practices developed within the European PEER project (Bretel et al., 2010). Grobid is trained on a set of annotated

[3] The commonly used publishing format is PDF files while authoring format is principally a mix of Microsoft Word file and LaTeX documents using a variety of styles. The originality of our approach is to work on the PDF file and not on these source formats. The rationale being that 1/ the source format is almost never available, 2/ even if we had access to the source format, we would need to implement a filter specific to each individual template required by such or such conference for a good quality content extraction

[4] http://www.termsciences.fr

[5] http://www.tei-c.org

scientific article and can be re-trained to fit templates used for a specific conference or to extract additional fields.

4 Translation of Scientific Texts

The translation system used is a Hybrid Machine Translation (HMT) system from French to English and from English to French, adapted to translate scientific texts in several domains (so far physics and computer science). This system is composed of a statistical engine, coupled with rule-based modules to translate special parts of the text such as mathematical formulas, chemical compounds, pseudo-code, and enriched with domain bilingual terminology (see Section 2). 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. All the PDF files deposited in HAL in computer science and physics were made available to us. These files were then converted to plain text using the Grobid tool, as described in the previous section. We extracted text from all the documents from HAL that were made available to us to train our language model. We built a small parallel corpus from the abstracts of the PhD theses from French universities, which must include both an abstract in French and in English. Table 1 presents statistics of these in-domain data.

The data extracted from HAL were used to adapt a generic system to the scientific literature domain. The generic system was mostly trained on data provided for the shared task of Sixth Workshop on Statistical Machine Translation[6] (WMT 2011), described in Table 2.

Table 3 presents results showing, in the English–French direction, the impact on the statistical engine of introducing the resources extracted from HAL, as well as the impact of domain adaptation techniques. The baseline statistical engine is a standard PBSMT system based on Moses (Koehn et al., 2007) and the SRILM tookit (Stolcke, 2002). Is was trained and tuned only on WMT11 data (out-of-domain). Incorporating the HAL data into the language model and tuning the system on the HAL development set, yielded a gain of more than 7 BLEU points, in both domains (computer science and physics). Including the theses abstracts in the parallel training corpus, a further gain of 2.3 BLEU points is observed for computer science, and 3.1 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 synthesised 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.

Although not shown here, similar results were obtained in the French–English direction. The French–English system is actually slightly better than the English–French one as it is an easier translation direction.

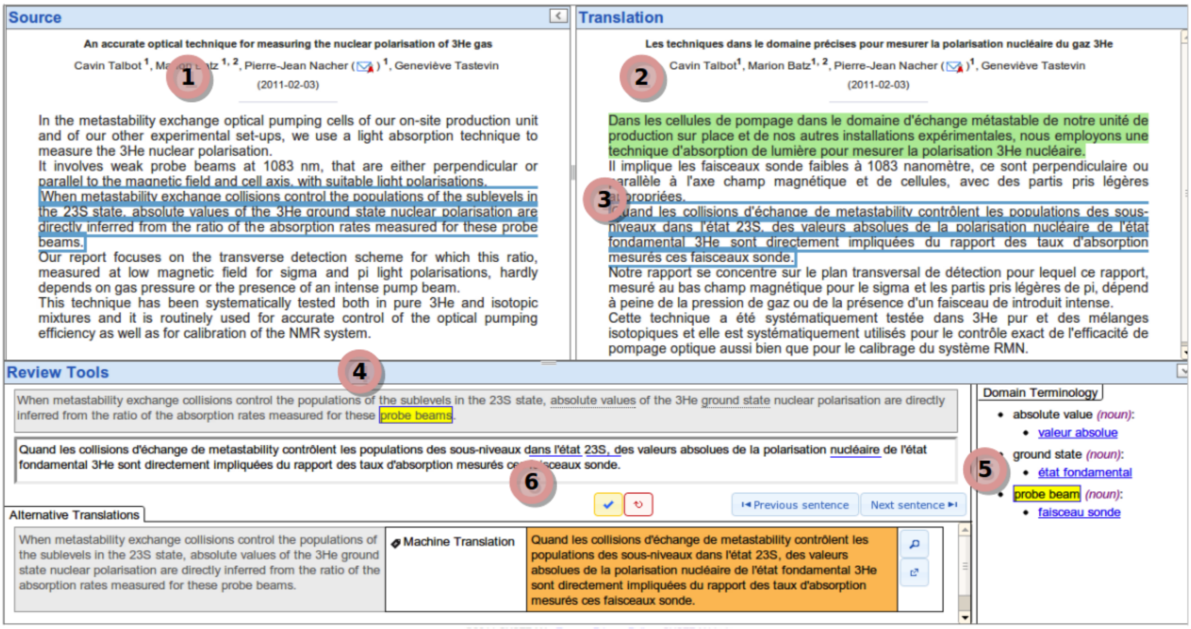
Set	Domain	Lg	Sent.	Words	Vocab.
Parallel data					
Train	cs+phys	En	55.9 k	1.41 M	43.3 k
		Fr	55.9 k	1.63 M	47.9 k
Dev	cs	En	1100	25.8 k	4.6 k
		Fr	1100	28.7 k	5.1 k
	phys	En	1000	26.1 k	5.1 k
		Fr	1000	29.1 k	5.6 k
Test	cs	En	1100	26.1 k	4.6 k
		Fr	1100	29.2 k	5.2 k
	phys	En	1000	25.9 k	5.1 k
		Fr	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 1: 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 computer science (cs) and physics (phys). The following statistics are given for the English (En) and French (Fr) sides (Lg) 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).

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

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

Table 3: 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.



An accurate optical technique for measuring the nuclear polarisation of 3He gas

Cavin Talbot [1], Ma[illegible]tz [1, 2], Pierre-Jean Nacher ( ) [1], Geneviève Tastevin

(2011-02-03)

In the metastability exchange optical pumping cells of our on-site production unit and of our other experimental set-ups, we use a light absorption technique to measure the 3He nuclear polarisation.
It involves weak probe beams at 1083 nm, that are either perpendicular or parallel to the magnetic field and cell axis, with suitable light polarisations.
When metastability exchange collisions control the populations of the sublevels in the 23S state, absolute values of the 3He ground state nuclear polarisation are directly inferred from the ratio of the absorption rates measured for these probe beams.
Our report focuses on the transverse detection scheme for which this ratio, measured at low magnetic field for sigma and pi light polarisations, hardly depends on gas pressure or the presence of an intense pump beam.
This technique has been systematically tested both in pure 3He and isotopic mixtures and it is routinely used for accurate control of the optical pumping efficiency as well as for calibration of the NMR system.

Les techniques dans le domaine précises pour mesurer la polarisation nucléaire du gaz 3He

Cavin Talbot[1], Marion Batz[1, 2], Pierre-Jean Nacher ( )[1], Geneviève Tastevin

(2011-02-03)

Dans les cellules de pompage dans le domaine d'échange métastable de notre unité de production sur place et de nos autres installations expérimentales, nous employons une technique d'absorption de lumière pour mesurer la polarisation 3He nucléaire.
Il implique les faisceaux sonde faibles à 1083 nanomètre, ce sont perpendiculaire ou parallèle à l'axe champ magnétique et de cellules, avec des partis pris légères appropriées.
Quand les collisions d'échange de metastability contrôlent les populations des sous-niveaux dans l'état 23S, des valeurs absolues de la polarisation nucléaire de l'état fondamental 3He sont directement impliquées du rapport des taux d'absorption mesurés ces faisceaux sonde.
Notre rapport se concentre sur le plan transversal de détection pour lequel ce rapport, mesuré au bas champ magnétique pour le sigma et les partis pris légères de pi, dépend à peine de la pression de gaz ou de la présence d'un faisceau de introduit intense.
Cette technique a été systématiquement testée dans 3He pur et des mélanges isotopiques et elle est systématiquement utilisés pour le contrôle exact de l'efficacité de pompage optique aussi bien que pour le calibrage du système RMN.

When metastability exchange collisions control the populations of the sublevels in the 23S state, absolute values of the 3He ground state nuclear polarisation are directly inferred from the ratio of the absorption rates measured for these probe beams.

Quand les collisions d'échange de metastability contrôlent les populations des sous-niveaux dans l'état 23S, des valeurs absolues de la polarisation nucléaire de l'état fondamental 3He sont directement impliquées du rapport des taux d'absorption mesurés ce[illegible]sceaux sonde.

- absolute value *(noun)*:
  - valeur absolue
- ground state *(noun)*:
  - état fondamental
- probe beam *(noun)*:
  - faisceau sonde

When metastability exchange collisions control the populations of the sublevels in the 23S state, absolute values of the 3He ground state nuclear polarisation are directly inferred from the ratio of the absorption rates measured for these probe beams.

Machine Translation: Quand les collisions d'échange de metastability contrôlent les populations des sous-niveaux dans l'état 23S, des valeurs absolues de la polarisation nucléaire de l'état fondamental 3He sont directement impliquées du rapport des taux d'absorption mesurés ces faisceaux sonde.

Figure 1: Translation and post-editing interface.

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

5 Post-editing Interface

The collaborative aspect of the demonstrated machine translation service is based on a post-editing tool, whose interface is shown in Figure 1. This tool provides the following features:

- WYSIWYG display of the source and target texts (Zones 1+2)
- Alignment at the sentence level (Zone 3)
- Zone to review the translation with alignment of source and target terms (Zone 4) and terminology reference (Zone 5)
- Alternative translations (Zone 6)

The tool allows the user to perform sentence level post-editing and records details of post-editing activity, such as keystrokes, terminology selection, actual edits and time log for the complete action.

6 Conclusions and Perspectives

We proposed the demonstration of an end-to-end tool integrated into the HAL archive and enabling efficient translation for scientific texts. This tool consists of a high-accuracy PDF extractor, a hybrid machine translation engine adapted to the scientific domain and a post-edition tool. Thanks to in-domain data collected from HAL, the statistical engine was improved by more than 10 BLEU points with respect to a generic system trained on WMT11 data.

Our system was deployed for a physic conference organised in Paris in Sept 2011. All accepted abstracts were translated into author's native languages (around 70% of them) and proposed for post-editing. The experience was promoted by the organisation committee and 50 scientists volunteered (34 finally performed their post-editing). The same experience will be proposed for authors of the LREC conference. We would like to offer a complete demonstration of the system at EACL. The goal of these experiences is to collect and distribute detailed "post-editing" data for enabling research on this activity.

Acknowledgements

This work has been partially funded by the French Government under the project COSMAT (ANR ANR-09-CORD-004).

References

Foudil Bretel, Patrice Lopez, Maud Medves, Alain Monteil, and Laurent Romary. 2010. Back to meaning – information structuring in the PEER project. In *TEI Conference*, Zadar, Croatie.

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. Association for Computational Linguistics.

Patrice Lopez. 2009. GROBID: Combining automatic bibliographic data recognition and term extraction for scholarship publications. In *Proceedings of ECDL 2009, 13th European Conference on Digital Library*, Corfu, Greece.

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

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