An Aligned French-Chinese corpus of 10K segments from university educational material

Ruslan Kalitvianski **Lingxiao Wang** **Valérie Bellynck** **Christian Boitet**
LIG-GETALP, Bâtiment IMAG, 700 av. Centrale,
CS 40700, 38058 Grenoble cedex 9, France
firstname.lastname@imag.fr

Abstract

This paper describes a corpus of nearly 10K French-Chinese aligned segments, produced by post-editing machine translated computer science courseware. This corpus was built from 2013 to 2016 within the MACAU project, by native Chinese students. The quality, as judged by native speakers, is adequate for understanding (far better than by reading only the original French) and for getting better marks. This corpus is annotated at segment-level by a self-assessed quality score. It has been directly used as supplemental training data to build a statistical machine translation system dedicated to that sub-language, and can be used to extract the specific bilingual terminology. To our knowledge, it is the first corpus of this kind to be released.

1 Introduction

The ongoing MACAU project, started in 2012 at the University of Grenoble, aims at providing multilingual access to course material taught at the university (Kalitvianski et al, 2015). It is motivated by the fact that many foreign students struggle with understanding material taught in French, and have to spend extra time in dictionary lookup and translation to fully comprehend the meaning.
The MACAU platform[1] is designed to create multilingual versions of initially monolingual course material by producing machine translations into the desired language, and by providing an interface that allows readers to post-edit these translations, segment by segment, until the desired level of quality is achieved.

A direct by-product of this activity is a bilingual corpus of post-edited sentences, constituting full courses, exercises and so on, concerning several fields of theoretical and practical computer science. Such a corpus could be employed as supplemental data for training a custom machine translation system. It can also serve for extraction of domain-specific lexicon.

In this paper we describe the data, provide corpus statistics, and delineate potential uses for the corpus.

2 The MACAU corpus

In this section we describe the MACAU project within which this corpus was constructed, and give the corpus' characteristics.

2.1 The MACAU project

The MACAU project has been ongoing since 2012. Its purpose is to help foreign students access educational material produced by the university in their native tongues, as those are the ones they understand best.

This is achieved by post-editing machine-translated documents, segment by segment. The post-edition is done via the iMAG web interface (Boitet et al, 2008). An iMAG is an interactive multilingual access gateway, which allows its users to visit a web page in the language of their choice while preserving its layout.

Pages are automatically segmented into translation units, typically sentences or titles. Segments are substituted by either a machine translation output if the segment is not found in the dedicated translation memory, or by the best post-edition available if the segment has been post-edited. Users can contribute corrections directly on the page by hovering the mouse pointer over the segment they desire to

[1] Currently migrating to macau.imag.fr

Proceedings of the 3rd Workshop on Natural Language Processing Techniques for Educational Applications,
pages 117–121, Osaka, Japan, December 12 2016.

correct, which makes a post-editing palette appear. The quality of a post-edition is explicitly "self-assessed" by the post-editor through a score in [0..20]. That score can later be revised, for example by other Chinese students using the Chinese version to study. Note that the interface allows to see the target (Chinese) and source (French) versions to appear side by side, so that, while learning some topic in computer science, Chinese students can also progress in French.

Post-editing through the iMAG interface is typically 3 times faster than translation from scratch (15-20 mins vs. 1 hour per standard page of 250 words). Also, for the post-editor, such an interface has the benefit of allowing post-editors to see the segment within its context.

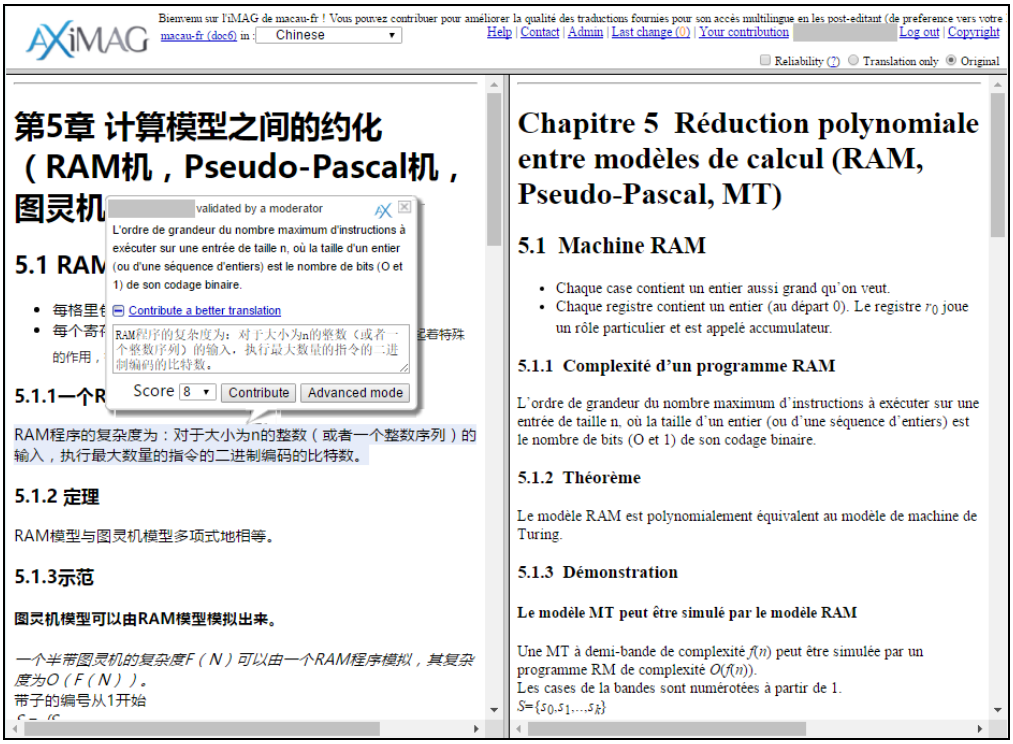

Figure 1: a chapter of a course on computational complexity, in bilingual view, with a post-editing window displayed over a segment.

For the MACAU project, the source documents are provided by teachers and also by students, and cover subjects concerning bachelor and master-level computer science. Table 1 below describes the quantity and the subject matters included in the corpus.

Subject matter	Content type	Pages (html)
Introduction to Propositional and First-Order Logic	Full book	45
C programming	Teacher lectures	14
Object-oriented programming	Teacher lectures	13
Computational Complexity	Lecture notes	13
Human-Machine interaction	Teacher lectures	7
Formal Languages and Parsing	Teacher lectures, hand-outs	5
Modelling of digital systems	Exam paper	2
AI and automatic planning	Exam paper	2
Introduction to Ergonomics	Student report	1

Table 1. Current status of the MACAU platform

Post-editing was performed by three Chinese-speaking university students, selected for their knowledge of the courses. Two were masters' students, one was a third year bachelors'. All have been taught the subject in French.

Students were selected on the basis of their knowledge of Chinese, their familiarity with the subject matters, and their interest in the task. They were explained the purpose of the task, and practiced on training documents before post-editing those that are included in the corpus. They received some monetary compensation for their work (as interns).

The students were asked to post-edit machine-generated translations segment by segment, through the iMAG interface, until an acceptable Chinese formulation of each segment was obtained. They were also told that the priority was not literary quality, but rather understandability. A group of two other native Chinese speakers subsequently verified the correctness a subset of randomly selected translations.

2.2 Corpus characteristics

The corpus is a collection of 9662 aligned French-Chinese segments, extracted from courseware in HTML. This corpus is cleaned of all HTML markup, however it does contain other non-linguistic elements, such as mathematical and logical formulas.

The segmentation was produced automatically and has not been corrected manually, therefore some segments correspond to fragments of sentences, and, more rarely, to two sentences fused together. Primary translations were obtained automatically via Google Translate.

The average source segment length is ~72 characters, or about 11 French words, and the median length is 53 characters. 25% of the source segments are less than 26 characters long and another 25% are over 100 characters long. Moreover, the corpus contains 108860 words, of which 8819 are unique French tokens.

The corpus initially contained many redundancies, but has been substantially cleaned. The remaining few source redundancies differ by their Chinese translations. The quality of the segments, as judged by bilingual readers, is considered adequate for understanding.

La complexité d'un programme pseudo-Pascal est l'ordre de grandeur du nombre d'instructions élémentaires à exécuter sur une entrée de taille n.	伪帕斯卡程序的复杂性是基本指令的数量级上的大小为 n 条输入运行。
Question: Sont-ils décidables dans le modèle de calcul déterministe?	问题：他们是能用确定性计算解决的问题吗？
(a ⇒ b)∧(b ⇒ c)∧¬(a ⇒ c) est insatisfaisable.	(a ⇒ b)⋀(b ⇒ c)⋀¬(a ⇒ c)是不可满足的。
Voici la référence principale, et son résumé, qui nous semble tout à fait clair.	这里是主要的参考和总结，这似乎是相当清楚的。
Ces trois formules n'ont pas de variables libres.	这三个公式没有自由变量。

Table 2. Examples of segments from the corpus

This corpus is now available on GitHub[2].

3 Building a specialized MT system for that sublanguage

One possible use of this corpus is the training of a specialized MT system for educational documents.

3.1 Motivations and method

We are interested in increasing the usage quality of machine translation systems. We measure usage quality as a function of post-edition times related to an estimate of the human translation time, which by default is assumed to be 60 minutes per standard 250 word page.

$$Q = 1 - (2/100 \times \frac{Tpe_{total}(for\ the\ task)}{Thum_{estim}(for\ the\ task)} \times Thum_{std_{page}}(\mathrm{mn})) \quad (1)$$

Formula 1: A measure for the usage quality of a MT system.

For example, Q = 40% is Tpe_{total} = 30 mn/p (8/20), and Q = 90% if Tpe_{total} = 5 mn/p (18/20).

This corpus has been used by Wang (2015) as supplemental data for training a specialized Moses (Koehn et al. 2007) probabilistic machine translation system through incremental training, yielding better usage quality than a generalistic PMT system.

[2] https://github.com/macau-getalp/macau

3.2 Usage of Moses incremental training

When new training data is available, a way of adding it to an existing model is incremental training. It is an iterative process that avoids the time-consuming retraining of a new model from scratch[3].
The V_0 of the system was trained on 100K bilingual segments from the MultiUN corpus (Eisele et al, 2010). Batches of 5000 segments taken from several in-domain corpora were iteratively added, including a raw form of this corpus that contained 16000 unfiltered segments.

3.3 Evolution of post-editing times

After 16 iterations, results show that the incremental training method reduces post-edition times, in a short amount of time (16 iterations, about 90 hours of computation, without ever recompiling everything). This system yields a usage quality of 70%, with 15 mins/std_page, better than Google Translate.

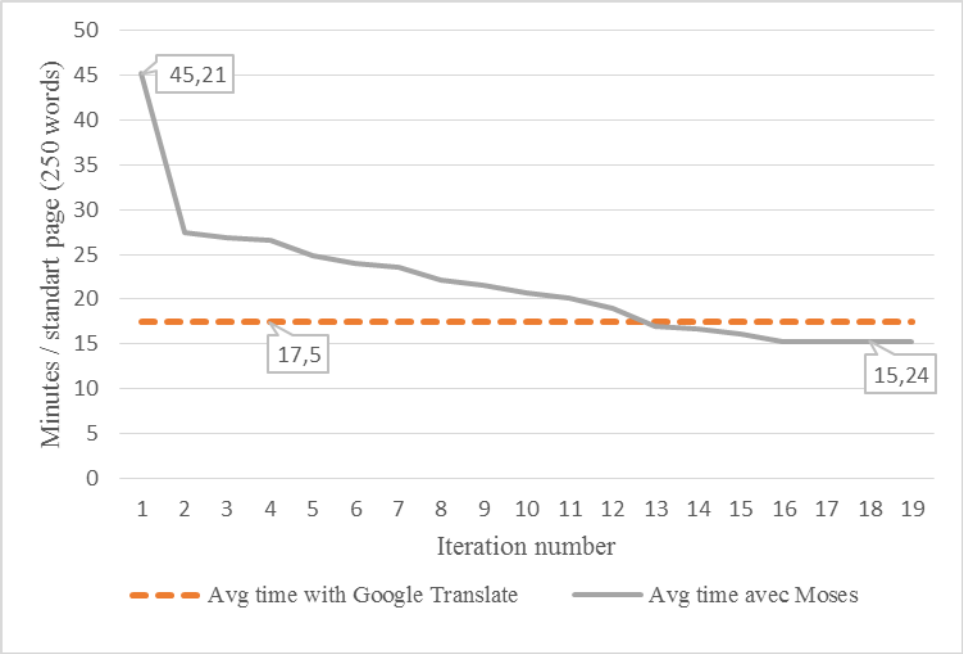

Figure 2: observed reduction in post-edition times after incremental training

The BLEU score (Papineni et al. 2002) improved as well, going from 13.8% after the first iteration to 48.3% after the last one.

Conclusion

We have presented a bilingual parallel corpus of nearly 10K aligned French-Chinese segments, produced over three years in the course of the MACAU project. This corpus is released under a free license and will be periodically updated as new post-editions become available. To our knowledge, this is the first corpus of this kind to be published

The multilingual access platform being open to everyone, this corpus can be extended by anyone by post-editing either pre-existing or newly uploaded documents, something that we encourage.

Although a large-scale evaluation of the usefulness of the platform will be carried out in the near future, we have already observed that the process of post-editing improves understanding and exam grades. An example of this is a student whose exam grade rose from 2.5/20 to 11/20 after a month of post-editing material related to the subject. Undoubtedly, several factors were at play, but this appears to be an interesting avenue of investigation.

[3] The details of the incremental training process are described here:
http://www.statmt.org/moses/?n=Advanced.Incremental

Acknowledgements

The authors would like to express gratitude to the *PédagoTICE* initiative, as well as to Pr. Marie-Christine Rousset, Guillaume Huard and Pascal Lafourcade for their assistance.

References

Christian Boitet, Cong-Phap Huyhn, Hong-Thai Nguyen and Valérie Bellynck. 2010. *The iMAG concept: multilingual access gateway to an elected Web sites with incremental quality increase through collaborative post-edition of MT pretranslations.* In Proceedings of TALN-2010, 8 p.

Ruslan Kalitvianski, Valérie Bellynck and Christian Boitet. 2015. *Multilingual Access to Educational Material through Contributive Post-editing of MT Pretranslations by Foreign Students*; In Proceedings of ICWL 2015, 10 p.

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, Annual* Meeting of the Association for Computational Linguistics (ACL), demonstration session, Prague, Czech Republic, June 2007.

Lingxiao Wang. 2015. *Outils et environnements pour l'amélioration incrémentale, la post-édition contributive et l'évaluation continue de systèmes de TA. Application à la TA français-chinois*. PhD dissertation, Université de Grenoble.

Andreas Eisele and Yu Chen. 2010. *MultiUN: A Multilingual Corpus from United Nation Documents.* In the Proceedings of the Seventh conference on International Language Resources and Evaluation, European Language Resources Association (ELRA), Pages 2868-2872

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. *BLEU: a method for automatic evaluation of machine translation.* ACL-2002: 40th Annual meeting of the Association for Computational Linguistics. pp. 311–318.