TRANSREAD: Designing a Bilingual Reading Experience with Machine Translation Technologies

François Yvon and **Yong Xu** and **Marianna Apidianaki**
LIMSI, CNRS, Université Paris-Saclay
91 403 Orsay
{yvon,yong,marianna}@limsi.fr

Clément Pillias and **Pierre Cubaud**
CEDRIC, CNAM
2 rue Conté, 75003 PARIS
{clement.pillias,cubaud}@cnam.fr

Abstract

In this paper, we use multilingual Natural Language Processing (NLP) tools to improve the reading experience of parallel texts on mobile devices. Such enterprise poses multiple challenging issues both from the NLP and from the Human Computer Interaction (HCI) perspectives. We discuss these problems, and report on our own solutions, now implemented in a full-fledged bilingual reading device.

1 Introduction

Owing to 15 years of advances in Statistical Machine Translation (SMT), automatically translated texts are nowadays of sufficiently high quality to serve the general public and the translation industry. Contrary to (S)MT which primarily targets readers without any assumed literacy in the source language, the TRANSREAD project studies applications targeting *partially bilingual users*, such as language learners, migrants settling in a new country, inhabitants of multilingual states, editors in the publishing industry and professional translators. Its main goal is to help such users to read texts in the original (source) language, even though a translation might be available in their mother tongue. Bitext processing techniques (Wu, 2010; Tiedemann, 2011) such as cross-lingual alignments at different levels or cross-lingual dictionary access, can facilitate and enrich the reading experience of texts in their original language. Such endeavour poses difficult challenges: it first requires to push existing MT technologies to the limit and to revisit assumptions that are rarely questioned, such as the need to deliver *fully aligned* bitexts, including many-to-many sentence links, and to output *high-precision word and phrase alignments*, even for rare words or gappy multi-word units.

A second challenge is visualisation and interaction design. In fact, most existing interfaces for bilingual reading/writing have targeted specialists of the MT industry, serving purposes such as manual alignment input and visualisation (Smith and Jahry, 2000; Germann, 2008; Gilmanov et al., 2014; Steele and Specia, 2015), MT tracing and debugging (DeNeefe et al., 2005; Weese and Callison-Burch, 2010), MT quality assessment (Federmann, 2012; Chatzitheodorou, 2013; Girardi et al., 2014) or MT post-edition (Aziz et al., 2012). By contrast, our aim is not just to visualize the translation or bilingual correspondences, but rather to enable a smooth and seamless reading experience for the general public. Ebook reading applications typically allow the reader to select a word and to access the corresponding dictionary entry, but applications that exploit the full translation context are much rarer. In *DoppelText*[1], *DuoLir*[2] and *Parallel Text Reader* on iOS, the selection is performed at the sentence level, using alignments. Whatever level is used, this kind of *switch-on-demand* interaction interrupts the flow of reading and can be a source of frustration. Another approach uses *synchronized views*, where the bitext is shown to the reader along with alignment links. In *ParallelBooks*[3], the bitext is synchronized by scrolling and the translation (at the paragraph level) only appears when the user taps

[1] http://www.doppeltext.com/
[2] http://www.duolir.com/
[3] http://www.parallelbooks.com/

Proceedings of NAACL-HLT 2016 (Demonstrations), pages 27–31,
San Diego, California, June 12-17, 2016.

the screen. Synchronized views have the potential to enable a seamless reading of the bitext (Pillias and Cubaud, 2015), at the cost however of a larger screen space. This paper discusses these various challenges and reports on the state of development of our main tool - the bilingual reader. Additional information, including resources and demos, can be found on the project website.[4]

2 Augmenting e-books with alignments

A major requirement of electronic reading devices is their ability to seamlessly reformat and adapt their typesetting, which does not only include the text itself but also other editorial (header, footer), typographic (bold, italic, font sizes and shapes) and dispositional information. An early design choice was to include alignment information as auxiliary source of information to the original electronic book(s). Each side of the bitext is thus stored in the EPUB format,[5] a de facto standard for electronic books, thereby enabling us to take advantage of its inherent ability to encode arbitrary documents and media files, as well as directives regarding their display.

Linguistic annotations are stored in an additional file to the EPUB archive and use an XML-based representation inspired by the XCES format,[6] already proposed in the early 2000s to represent alignment information. References to actual textual units (in the EPUB/html files) are maintained as the complete path from the root of the document to the node containing the unit. Our format for representing annotations is generic enough to represent alignment links at various levels of granularity, as well as other arbitrary information. It relies on two types of basic markups: `<link>` for binary relationships (bilingual links or monolingual co-references), and `<mark>` for unary information concerning one single unit (be it a paragraph, a sentence, a fragment or a word). In our sample file, these tags are used to encode part-of-speech as well as sense disambiguation information. However, only the information related to bilingual alignment links is currently displayed.

[4] `http://transread.limsi.fr`

[5] `http://epubzone.org/epub-3-overview`

[6] `http://www.xces.org/`

3 Challenges of bitext alignments

Our representation of alignment relationships accommodates alignment links at various levels of granularity. Our display currently exploits 5 such levels, based on sentential, sub-sentential and word alignments, which were computed for two short stories by S. Maugham.[7]

3.1 Sentence alignment

Sentence alignment in parallel texts (bitexts) is a well established problem in multilingual NLP (Brown et al., 1991; Gale and Church, 1991; Kay and Röscheisen, 1993), for which a wide array of methods exist (Tiedemann, 2011). The problem is however far from being solved, especially for literary texts where parallelism is less strict than for technical or legal documents (Yu et al., 2012). For this demo, alignments have been produced semi-automatically. The automatic part used techniques presented by (Xu et al., 2015), which implement a multi-pass, coarse-to-fine alignment strategy. The first pass uses very reliable 1:1 alignment links computed using the approach of (Moore, 2002), while the next stages complete this initial partial alignment by including additional correspondences, the probabilities of which are evaluated using a large-scale MaxEnt classifier embarking a very large number of features. Automatic alignments were then manually checked and fixed: for our simple bitexts they were mostly correct, with a link level F-score $\approx 97\%$.

3.2 Word alignments

For this demo, gold word alignments were collected as follows: automatic word alignments were first computed by running the MGiza (Gao and Vogel, 2008) implementation of IBM Model 4 (Brown et al., 1993) in both directions. Alignments in the intersection were checked and corrected following the recommandations of Och and Ney (2003). Even for such simple texts, alignment errors were numerous, with an AER close to 0.17 ('The Promise'), and to 0.19 ('The Verger'). This confirms the intuition that computing high quality word alignments for literary texts might be significantly more difficult than for other text genres. This also calls for improved

[7] 'The verger' and 'The promise', totalling slightly more that 160 sentences each.

techniques for computing confidence measures for word alignments (Huang, 2009): depending on the intended reading context, it might be better to avoid displaying erroneous alignment links.

3.3 Subsentential alignments

The task of designing sound and tractable alignment models is notoriously much harder for groups of words than for words (Marcu and Wong, 2002; DeNero and Klein, 2008). Two main strategies have been explored in the literature: the most common, employed in most SMT systems (Koehn et al., 2007) starts with alignments for isolated words, which are incrementally grown subject to consistency constraints. The alternative way is to start with sentential alignments and adopt a divisive strategy, which yields progressive refinements of an initially holistic pairing; this can be performed exactly under ITG constraints (Wu, 1997); heuristic approaches, capable of handling alignments for arbitrarily long segments have also been proposed in (Lardilleux et al., 2012): both techniques require to evaluate the parallelism of arbitrary chunks. We follow the latter here, also using punctuation marks to select segmentation points. The resulting alignments are deliberately pretty coarse and primarily meant to be used in a contrastive condition for the human tests.

4 A Bilingual Reader

4.1 Design

The current version of the TRANSREAD bilingual reader displays paginated versions of the bitext in parallel views. In Figure 1, the source text is displayed on the right side of the screen and its translation on the left. The user has selected a word in the source version. Touching a word highlights its context on both sides, exploiting the alignment structure in a hierarchical way. Highlighting can be triggered from both versions of the text. The different levels are depicted using bounding boxes, which are pre-computed using the alignment and the HTML graphs. Bounding boxes are not always rectangular, because of word hyphenation for text justification. We have tried to minimize visual overload by simplifying the resulting geometry of the boxes and using a colour scale as background. We use the "natural" theme from (Krause, 2010).

The display size for the bitext can be modified dynamically. When the application starts, the text versions are given an equal space, but the translation on the left can be shrunk so that the reader can concentrate on the original text. Structural highlighting is still functional in this mode, but when the translated text is shrunk to its minimum size, it can only be used as a visual hint.

Our reader is targeted for use on a tablet device. Interactions with the tactile screen of the tablet cause the well-known problem of fingers occluding the touched item. To avoid this issue, a pointer is displayed 1.3cm above the touch position sensed by the system, and its position determines what elements are selected. This *picking* is done by recursively searching the textual elements whose precomputed bounding boxes contain the pointer's position. When this process selects two sibling elements (e.g. two sentences in a paragraph), the system only keeps the one for which the distance from the pointer to the closest bounding box border is greatest. The pointer shape is an upward triangle. When a word is selected, we also display a downward triangle above it attached to the line of text, but following the pointer horizontally. These two triangles enclose the selected word without occluding it.

4.2 Implementation

As readers, we benefit from centuries of high quality printed typesetting. A poor digital typesetting would therefore lower the user experience of our bilingual reader, no matter how rich the interaction may be. But automated typesetting is a complex task, often under-estimated. For instance, justification on small columns, as those produced by bitext presentation, can exhibit "rivers" of white space that are difficult to handle automatically. Our first goal for the implementation was then to select a programming framework that would help for these questions. The Knuth-Plass algorithm[8] was used for justification, along with LaTeX rules[9] for hyphenation. Another important issue was to build a fully reconfigurable software in order to investigate a large design space of interaction for tablets. We have selected the Kivy framework for Python, which en-

[8] As described in `http://defoe.sourceforge.net/folio/knuth-plass.html`

[9] Provided by `http://tug.org/tex-hyphen/`

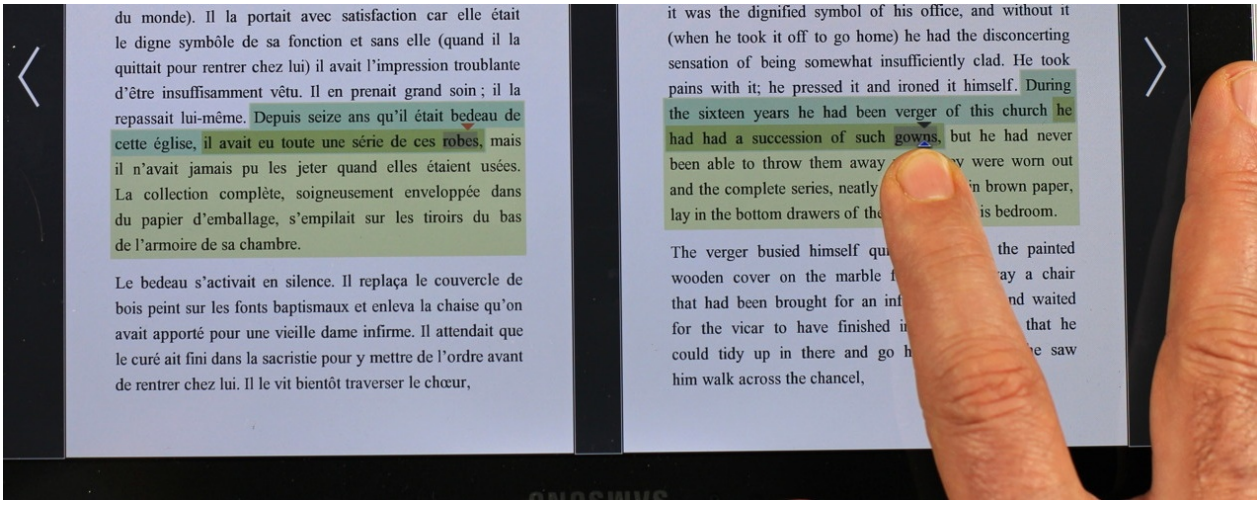

Figure 1: The TRANSREAD bilingual reader application running on tablet

ables cross-platform development for Android or iOS, and GPU-based graphics with OpenGL ES.

5 Perspectives

As reflected in this paper, a top priority is to pursue our efforts towards high-precision alignments, an application where supervised learning techniques could help (Moore, 2005). Additional functionalities in reading are also envisioned, such as an enhanced and non-distracting access to dictionary information for difficult words. Currently, Web Readers and mobile reading devices offer such functionality through a pop-up window presenting the complete dictionary entry. No assistance is however offered to access the right sense in context, which would be especially helpful for polysemous words or when language proficiency is low. In TRANSREAD, we propose to perform this selection automatically. Our word sense disambiguation (WSD) method (Apidianaki and Gong, 2015) exploits word-level alignments to annotate words on both sides of the bitext with the correct senses extracted from BabelNet (Navigli and Ponzetto, 2012). By integrating WSD information in the reader, we will be able to propose definitions, usage examples and Wikipedia entries, as well as synonymous words and semantically correct translations. Our WSD system embeds an alignment-based multi-word expression (MWE) identification mechanism (Marie and Apidianaki, 2015). Such information will serve as part of a smart selection mechanism (Pantel et al., 2014), enabling the system to select appropriate spans and dictionary entries for MWEs found in texts.

An experimental evaluation of the interface general design is currently being conducted. We study, notably, the effect of the depth of the alignment structure on human readers behavior. As short term future work, we shall also investigate other interaction techniques for focus management, such as distortion and 3D views for page turning (Cubaud, 2008). The graphic composition engine developed for the current application already allows such effects. A research agenda should also include long term experiments with real-life reading sessions for a wide range of languages and text difficulties.

6 Conclusion

We have presented a first version of a bilingual reader using NLP and HCI technologies to enhance the reading experience. On our way, some tough challenges had to be overcome, an unexpected issue being the processing of typeset documents which is hardly addressed in the NLP literature. This venture has provided a context where such issues matter, illustrating the benefits of cross-domain research.

Acknowledgments

This work was partly funded by the French "National Research Agency" under project ANR-12-CORD-0015.

References

M. Apidianaki and L. Gong. 2015. LIMSI: Translations as Source of Indirect Supervision for Multilingual All-Words Sense Disambiguation and Entity Linking. In *Proc. SemEval*.

W. Aziz, S. C. M. De Sousa, and L. Specia. 2012. PET: a tool for post-editing and assessing machine translation. In *Proc. LREC*.

P. F. Brown, J. C. Lai, and R. L. Mercer. 1991. Aligning sentences in parallel corpora. In *Proc. ACL*.

P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer. 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2):263–311.

K. Chatzitheodorou. 2013. COSTA MT Evaluation Tool: An open toolkit for human machine translation evaluation. *The Prague Bulletin of Mathematical Linguistics*, 100.

P. Cubaud, 2008. *Digital Libraries*, chapter 3D interaction for digital libraries.

S. DeNeefe, K. Knight, and H. H. Chan. 2005. Interactively Exploring a Machine Translation Model. In *Proc. ACL*.

J. DeNero and D. Klein. 2008. The Complexity of Phrase Alignment Problems. In *Proc. ACL-08:HLT*.

C. Federmann. 2012. Appraise: an Open-Source Toolkit for Manual Evaluation of MT output. *The Prague Bulletin of Mathematical Linguistics*, 98.

W. A. Gale and K. W. Church. 1991. A program for aligning sentences in bilingual corpora. In *Proc. ACL*.

Q. Gao and S. Vogel. 2008. Parallel Implementations of Word Alignment Tool. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*.

U. Germann. 2008. Yawat: Yet Another Word Alignment Tool. In *Proc. ACL-08:HLT (demo)*.

T. Gilmanov, O. Scrivner, and S. Kübler. 2014. SWIFT Aligner, A Multifunctional Tool for Parallel Corpora: Visualization, Word Alignment, and (Morpho)-Syntactic Cross-Language Transfer. In *Proc. LREC*.

C. Girardi, L. Bentivogli, M. A. Farajian, and M. Federico. 2014. MT-EQuAl: a Toolkit for Human Assessment of Machine Translation Output. In *Proc. COLING*.

F. Huang. 2009. Confidence Measure for Word Alignment. In *Proc. ACL-AFNLP*.

M. Kay and M. Röscheisen. 1993. Text-Translation Alignement. *Computational Linguistics*, 19(1).

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proc. ACL (demo)*.

J. Krause, 2010. *Color Index - Revised Edition*.

A. Lardilleux, F. Yvon, and Y. Lepage. 2012. Hierarchical sub-sentential alignment with Anymalign. In *Proc. EAMT*.

D. Marcu and D. Wong. 2002. A Phrase-Based, Joint Probability Model for Statistical Machine Translation. In *Proc. EMNLP*.

B. Marie and M. Apidianaki. 2015. Alignment-based sense selection in METEOR and the RATATOUILLE recipe. In *Proc. WMT*.

R. C. Moore. 2002. Fast and Accurate Sentence Alignment of Bilingual Corpora. In *Proc. of AMTA*, Lecture Notes in Computer Science 2499.

R. C. Moore. 2005. A discriminative framework for bilingual word alignment. In *Proc. EMNLP:HLT*.

R. Navigli and S. P. Ponzetto. 2012. BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network. *Artificial Intelligence*, 193.

F. J. Och and H. Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1).

P. Pantel, M. Gamon, and A. Fuxman. 2014. Smart Selection. In *Proc. ACL*.

C. Pillias and P. Cubaud. 2015. Bilingual Reading Experiences: What They Could Be and How To Design for Them. In *Proc. IFIP Interact*.

N. A. Smith and M. E. Jahry. 2000. Cairo: An alignment visualization tool. In *Proc. LREC*.

D. Steele and L. Specia. 2015. WA-Continuum: Visualising Word Alignments across Multiple Parallel Sentences Simultaneously. In *Proc. ACL-IJCNLP*.

J. Tiedemann. 2011. *Bitext Alignment*. Number 14 in Synthesis Lectures on Human Language Technologies. Morgan and Claypool publishers.

J. Weese and C. Callison-Burch. 2010. Visualizing Data Structures in Parsing-Based Machine Translation. *The Prague Bulletin of Mathematical Linguistics*, 93.

D. Wu. 1997. Stochastic Inversion Transduction Grammar and Bilingual Parsing of Parallel Corpora. *Computational Linguistics*, 23(3).

D. Wu. 2010. Alignment. In *CRC Handbook of Natural Language Processsing*, number 16.

Y. Xu, A. Max, and F. Yvon. 2015. Sentence alignment for literary texts. *Linguistic Issues in Language Technology*, 12(6).

Q. Yu, A. Max, and F. Yvon. 2012. Aligning Bilingual Literary Works: a Pilot Study. In *Proc. NAACL-HLT Workshop on Computational Linguistics for Literature*.