Users and Data: The Two Neglected Children of Bilingual Natural Language Processing Research

Philippe Langlais
RALI-DIRO
Université de Montréal
CP. 6128 Succ. Centre Ville
H3C 3J7 Montréal, Québec, Canada
felipe@iro.umontreal.ca

Abstract

Despite numerous studies devoted to mining parallel material from bilingual data, we have yet to see the resulting technologies wholeheartedly adopted by professional translators and terminologists alike. I argue that this state of affairs is mainly due to two factors: the emphasis published authors put on models (even though data is as important), and the conspicuous lack of concern for actual end-users.

1 Introduction

Parallel corpora (documents collections that are translations of one another) are the bread and butter of machine translation (MT). Solutions have been proposed for mining parallel texts found on the Web (Chen and Nie, 2000; Resnik and Smith, 2003), and for aligning sentences in parallel documents (Gale and Church, 1993), leading to so-called "bitexts". It then becomes possible to align words in parallel sentence pairs, in an unsupervised way (Brown et al., 1993).

Because parallel data is relatively rare, researchers have turned to exploiting comparable corpora, e.g. news articles in different languages covering the same event. Sharoff et al. (2013) thoroughly examine this topic. It is noteworthy that researchers know quite well how to identify parallel sentences in a comparable corpus (Munteanu and Marcu, 2005), and can then use "tried and true" procedures for extracting bilingual lexicons from such a resource (Rapp, 1995; Fung, 1995; Mikolov et al., 2013).

Being able to benefit from both parallel and comparable data is quite an accomplishment from a scientific point of view, and progress is still being made on the task. In contrast, and frustratingly, the technologies that professional translators are adopting continue to rely mainly on sentence-based translation memories. I do not mean to say that other technologies are not being used. For instance, translation agencies are increasingly integrating machine translation into their workflow, but this is mostly driven by cost reduction, and not by a genuine interest in MT on the part of translators, who remain unconvinced.

I submit that this limited adoption of new resources and technologies is due to the conjunction of two factors: the overall lack of concern for actual users, and the clear preference of the research community for the study of models at the cost of research on data. Of course, improvements on models have the potential to impact users. Notably, recent studies (Bentivogli et al., 2016; Isabelle et al., 2017) confirm that neural MT (Sutskever et al., 2014; Cho et al., 2014; Bahdanau et al., 2014) significantly reduces errors, therefore requiring less post-editing. However, better ways of efficiently acquiring and organizing data equally matters.

As for end-users, more interest in their day-to-day concerns should lead to a better adoption of the technologies we develop, which in turn would reveal scientific challenges we had never thought of before. One example of a project I have been involved in is the (at that time pioneering) effort to develop an interactive translation engine named TransType (Foster et al., 1997) in which a translator interacts iteratively with a translation engine in order to produce a translation. After multiple rounds of development, we had several translators beta-test our prototype (Langlais et al., 2002), and we realized that the keystroke saving rate used to measure the improvements brought about by TransType was *not* correlated with the user's productivity gains. This led us to devise a user model that we could not have foreseen at the beginning of the project (Foster et al., 2002). See (González-

Proceedings of the 10th Workshop on Building and Using Comparable Corpora, pages 1–5,
Vancouver, Canada, August 3, 2017.

Rubio et al., 2012) for further developments along these lines.

Doing research in a vacuum certainly facilitates progress. For instance, in recent years we have witnessed a tremendous interest in embedding methods for extracting bilingual lexicons, thanks to the pioneering work of (Mikolov et al., 2013). It is nowadays a standard procedure to measure the quality of embedding representations on what is called the bilingual lexicon induction (BIL) task. One popular evaluation protocol, initially proposed in (Mikolov et al., 2013) consists in identifying the translation of the last 1000 words of the 6000 most frequent words in the training material. However, for many language pairs of interest, existing bilingual lexicons already list the translations of frequent (and less frequent) words. In fact, in (Jakubina and Langlais, 2017), we show that the accuracy of embedding-based methods when translating rare words — which arguably is a test case of better use to end users — is less than 2% at rank 1. I must make it clear at this point that I am excited by embedding methods and their potential to improve the current state of the art. I am merely saying that the way we evaluate these methods does not reflect their true usefulness.

The purpose of this presentation is to pinpoint a number of challenges I feel are worth being reinvestigated. They belong to two categories: understanding better how to acquire and organize (bilingual) data, and better exploiting existing resources, with an emphasis on more representative test cases. This list is not exhaustive, and emanates from the needs expressed by some of the industry professionals I have been discussing with, and from the opinions I have been forming over time when reading (exciting) publications in my field.

2 Overlooked Issues

2.1 Data Acquisition

Finding parallel documents over the Web has been studied early by (Chen and Nie, 2000; Resnik and Smith, 2003). Those systems (and others like them) perform resource alignment by examining their URLs. Since this superficial information is sometimes misleading, they also use other features such as length ratios, lexicon overlap or HTML structure mapping. As noted in (Buck and Koehn, 2016), efforts in gathering parallel data have been mostly ad-hoc and limited in scale. I believe these limitations stem precisely from the fact that we are more concerned by models than data in the academia. Interestingly, the bilingual alignment document shared task at WMT 2016 is a very sensible attempt to promote research to find solutions to the aforementioned problem. I hope this is a rallying first step, fostering a new interest in striking a compromise between efficiency and effectiveness, in the spirit of (Ture et al., 2011).

Conventional wisdom tells us that parallel data is (comparatively) rare, therefore there is a need for mining comparable corpora. Munteanu and Marcu (2005) show that cross-lingual information retrieval coupled with a filter on the publication date of the news offer an efficient way of gathering comparable news data over the Web. Smith et al. (2010) demonstrate that language inter-linked article pairs in Wikipedia offer valuable comparable data. Still, I am not aware of any large-scale and systematic way of **mining comparable data over the Web**.

Gathering **domain-specific bilingual corpora** (parallel and comparable) is a related issue that has many practical benefits, but which I feel is neglected. Compiling domain-specific monolingual data is difficult enough (Groc and Tannier, 2014), in a multilingual setting, it is even more complex to begin to agree on best practices. See (Azoulay, 2017) for a recent attempt and (Morin et al., 2010) for evidences that the quantity of texts acquired should not be the only concern.

2.2 Data Organization

Large-scale acquisition efforts conducted over the Web involve at some point an effort to distinguish parallel data from comparable or even unrelated data. A similar situation arises in institutions that produce documents in multiple languages without necessarily keeping track of which documents are parallel or comparable, and with what level of quality. A typical example of this are news agencies.

The classification of (Fung and Cheung, 2004) is very useful to qualify the kind of bilingual data we are dealing with, as are measures of the comparability of a corpus (Li and Gaussier, 2010; Babych and Hartley, 2014). I think more efforts should be invested in **estimating the quality of a bilingual corpus** (parallel or comparable). It could prove useful for instance when choosing the appropriate extraction technique for a given pair of documents. For example, we could select a

monotonous sentence-alignment if the documents are near-parallel, and Cartesian product-based approaches if the documents are merely comparable.

Produced texts are increasingly becoming multilingual, through various processes that are not all known. While the overused parliamentary Hansard debates are created by a well-known process, for many collections, the **genesis of a document** is simply not known. This poses exciting challenges that have been partially addressed, among which detecting that a text has been produced by translation (whether it be automatic or not) (Carter and Inkpen, 2012; Arase and Zhou, 2013). This feature might impact applications such as plagiarism detection (Ceska et al., 2008).

2.3 Parallel Material Extraction

Having a collection of parallel and comparable corpora available allows for extracting translation units. Sentences have been the focus of much research, and we know rather well how to align sentences in a parallel corpus. While aligning legislative texts and the like is more or less a solved problem (Langlais et al., 1998), **aligning literary texts** is still very challenging (Xu et al., 2015). Goutte et al. (2012) report that statistical MT is robust with respect to noise in sentence alignment. At the same time, Lamraoui and Langlais (2013) show that carefully aligning sentences in a collection as well structured as Europarl (Koehn, 2005) leads to (slight) increases in performance. These somehow contradictory results warrant further investigations.

At the other end of the spectrum of units, we typically seek to align words. So-called IBM models (Brown et al., 1993) are popular generative models that can be learnt in order to extract word pairs in parallel data. Still, **identifying multiword expressions and their translations** remains an actively studied[1] and challenging task. In particular, Isabelle et al. (2017) have observed that idioms are poorly handled by neural machine translation.

Aligning units in a comparable corpus remains a challenge as well. Recognizing sentences that are translations of one another in a comparable corpus has been studied early by (Munteanu and Marcu, 2005), but advances in **embedding methods** might improve the current state of the art. We have participated in this year's BUCC shared task on parallel sentence extraction from comparable corpora with such an approach (Grégoire and Langlais, 2017), and I expect this research avenue to gain in popularity. With the exception of (Kumano and Tokunaga, 2007) and (Quirk et al., 2007), we lack a **generative model of a comparable corpus** that would allow to capture parts of documents that are aligned in a principled way, whatever the granularity (paragraphs, sentences, expressions, words or even subwords).

For progress in extraction to be meaningful, we should pay attention to the way we measure it: Not all units are equally important. For instance, pairs of compositional units are not worth being collected (and therefore evaluated). Likewise, mining sentence pairs in which n-grams have already been seen massively is likely not very helpful. We believe that the community should share a number of benchmarks that are representative of specific uses. Ultimately, this should involve users because they know best what matters to them.

3 Discussion

Progress in acquiring bilingual collections of texts, organizing them into a meaningful repository, and extracting knowledge from it are three avenues that are clearly overlapping. Many of those aspects have received attention by many researchers, and have been the focus of dedicated projects, such as ACCURAT (Skadia et al., 2010).

Still, our (or at least my) understanding of how to efficiently mine bilingual material for a specific use is deficient. I believe one reason for this is that our community is more versed in elaborating models and evaluating them in a vacuum, whereas I think data is definitely part of the game, and we should work on better ways of evaluating our technology. This presentation will be punctuated by a number of studies conducted at RALI, some of which involve real-life users.

[1] The MWE workshop is at its 13th edition.

References

Yuki Arase and Ming Zhou. 2013. Machine translation detection from monolingual web-text. In *51st Annual Meeting of the Association for Computational Linguistics*. Sofia, Bulgaria, pages 1597–1607.

Daphnée Azoulay. 2017. Frame-based knowledge representation using large specialized corpora. In *AAAI Spring Symposium on Computational Construction Grammar and Natural Language Understanding*. Stanford University, CA, pages 119–226.

Bogdan Babych and Anthony Hartley. 2014. Meta-evaluation of comparability metrics using parallel corpora. *CoRR* abs/1404.3759.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. In *ICLR 2015*.

Luisa Bentivogli, Arianna Bisazza, Mauro Cettolo, and Marcello Federico. 2016. Neural versus phrase-based machine translation quality: a case study. *CoRR* abs/1608.04631.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation. *Computational Linguistics* 19(2):263–313.

Christian Buck and Philipp Koehn. 2016. Findings of the WMT 2016 Bilingual Document Alignment Shared Task. In *Proceedings of the First Conference on Machine Translation*. pages 554–563.

Dave Carter and Diana Inkpen. 2012. Searching for poor quality machine translated text : learning the difference between human writing and machine translations. In *25th Canadian Conference on Artificial Intelligence*. Toronto, Canada, pages 49–60.

Zdenek Ceska, Michal Toman, and Karel Jezek. 2008. Multilingual plagiarism detection. In *Proceedings of the 13th International Conference on Artificial Intelligence: Methodology, Systems, and Applications*. AIMSA '08, pages 83–92.

Jiang Chen and Jian-Yun Nie. 2000. Parallel web text mining for cross-language ir. In *Content-Based Multimedia Information Access - Volume 1*. RIAO '00, pages 62–77.

Kyunghyun Cho, Bart van Merriënboer, Çalar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*. pages 1724–1734.

George Foster, Pierre Isabelle, and Pierre Plamondon. 1997. Target-text mediated interactive machine translation. *Machine Translation* 12(1):175–194.

George Foster, Philippe Langlais, and Guy Lapalme. 2002. User-friendly text prediction for translators. In *Empirical Methods in Natural Language Processing*. Philadelphia.

Pascale Fung. 1995. Compiling bilingual lexicon entries from a non-parallel english-chinese corpus. In *3rd Workshop on Very Large Corpora*. pages 173–183.

Pascale Fung and Percy Cheung. 2004. Multi-level bootstrapping for extracting parallel sentences from a quasi-comparable corpus. In *Proceedings of the 20th international conference on Computational Linguistics*.

William A. Gale and Kenneth W. Church. 1993. A program for aligning sentences in bilingual corpora. *Computational linguistics* 19(1):75102.

Jesús González-Rubio, Daniel Ortiz-Martínez, and Francisco Casacuberta. 2012. Active learning for interactive machine translation. In *13th Conference of the European Chapter of the Association for Computational Linguistics*. pages 245–254.

Cyril Goutte, Marine Carpuat, and George Foster. 2012. The impact of sentence alignment errors on phrase-based machine translation performance. In *10th AMTA*.

Françis Grégoire and Philippe Langlais. 2017. Bucc 2017 shared task: a first attempt toward a deep learning framework for identifying parallel sentences in comparable corpora. In *BUCC Workshop*. Vancouver, Canada. Shared task paper.

Clément De Groc and Xavier Tannier. 2014. Evaluating Web-as-corpus Topical Document Retrieval with an Index of the OpenDirectory. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*. Reykjavk, Iceland.

P. Isabelle, C. Cherry, and G. Foster. 2017. A Challenge Set Approach to Evaluating Machine Translation. *ArXiv e-prints* .

Laurent Jakubina and Philippe Langlais. 2017. Reranking translation candidates produced by several bilingual word similarity sources. In *15th Conference of the European Chapter of the Association for Computational Linguitics*. volume 2, Short Papers, pages 605–611.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *tenth Machine Translation Summit*. pages 79–86.

Tadashi Kumano and Hideki Tanaka Takenobu Tokunaga. 2007. Extracting phrasal alignments from comparable corpora by using joint probability smt model. In *11th International Conference on Theoretical and Methodological Issues in Machine Translation*.

Fethi Lamraoui and Philippe Langlais. 2013. Yet another fast, robust and open source sentence aligner. time to reconsider sentence alignment? In *XIV Machine Translation Summit*. Nice, France, pages 77–84.

Philippe Langlais, Guy Lapalme, and Marie Loranger. 2002. Transtype: Development-Evaluation Cycles to Boost Translators' Productivity. *Machine Translation, Kluwer Academic Publishers,* 17:77–98.

Philippe Langlais, Michel Simard, and Jean Véronis. 1998. Methods and Practical Issues in Evaluating Alignment Techniques. In *36th Annual Meeting of the Association for Computational Linguistics (ACL) and 17th International Conference on Computational Linguistic (COLING)*. Montreal, Canada.

Bo Li and Éric Gaussier. 2010. Improving corpus comparability for bilingual lexicon extraction from comparable corpora. In *Proceedings of the 23rd International Conference on Computational Linguistics*. page 644652.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168* .

Emmanuel Morin, Béatrice Daille, Kyo Kageura, and Koichi Takeuchi. 2010. Brains, not Brawn: The Use of 'Smart" Comparable Corpora in Bilingual Terminology Mining. *ACM Transactions on Speech and Language Processing* 7(1):1–23.

Dragos S. Munteanu and Daniel Marcu. 2005. Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics* 31(4):477504.

Chris Quirk, Raghavendra Udupa, and Arul Menezes. 2007. Generative models of noisy translations with applications to parallel fragment extraction. In *In Proceedings of MT Summit XI, European Association for Machine Translation*.

Reinhard Rapp. 1995. Identifying word translations in non-parallel texts. In *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*. pages 320–322.

Philip Resnik and Noah A. Smith. 2003. The web as a parallel corpus. *Computational Linguistics* 29(3):349380.

Serge Sharoff, Reinhard Rapp, and Pierre Zweigenbaum. 2013. *Overviewing Important Aspects of the Last Twenty Years of Research in Comparable Corpora*, Springer, pages 1–17.

Inguna Skadia, Andrejs Vasijevs, Raivis Skadi, Robert Gaizauskas, and Dan Tufi. 2010. Analysis and evaluation of comparable corpora for under resourced areas of machine translation. In *Proceedings of the 3rd Workshop on Building and Using Comparable Corpora. Applications of Parallel and Comparable Corpora in Natural Language Engineering and the Humanities*. pages 6–14.

Jason R. Smith, Chris Quirk, and Kristina Toutanova. 2010. Extracting parallel sentences from comparable corpora using document level alignment. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. page 403411.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*. pages 3104–3112.

Ferhan Ture, Tamer Elsayed, and Jimmy Lin. 2011. No free lunch: brute force vs. locality-sensitive hashing for cross-lingual pairwise similarity. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*. pages 943–952.

Yong Xu, Aurlien Max, and François Yvon. 2015. Sentence alignment for literary texts. *Linguistic Issues in Language Technology* 12(6):1–29.