Natural Language Engineering **11** (3): 239–246. © 2005 Cambridge University Press
doi:10.1017/S1351324905003827 Printed in the United Kingdom

Parallel texts

RADA MIHALCEA
Department of Computer Science & Engineering, University of North Texas, P.O. Box 311366, Denton, TX 76203 USA
e-mail: rada@cs.unt.edu

MICHEL SIMARD
Xerox Research Centre Europe, 6, Chemin de Maupertuis, 38240 Meylan, France
e-mail: michel.simard@xrce.xerox.com

(*Received May 1 2004; revised November 30 2004*)

Abstract

Parallel texts[1] have become a vital element for natural language processing. We present a panorama of current research activities related to parallel texts, and offer some thoughts about the future of this rich field of investigation.

1 Introduction

Parallel texts have become a vital element in many areas of natural language processing (NLP). They represent one of the richest and most versatile sources of knowledge for NLP, and have been used successfully not only in problems that are intrinsically multilingual, such as machine translation and cross-lingual information retrieval, but also as an indirect way of attacking "monolingual" problems, for example in semantic and syntactic analysis.

Why have parallel texts proven such a fruitful resource? Most likely because of their ability to represent meaning: the translation of a text in another language can be seen as a semantic representation of that text, which opens the doors to a tremendously large number of language processing applications that operate on such representations. In his famous memorandum from 1949, Warren Weaver wrote: *"When I look at an article in Russian, I say: 'This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode'."* And this is in fact one of the main reasons why parallel texts have proven such a formidable knowledge source: in the absence of alternative "true" semantic representations, parallel texts give us the means to discover the *meaning* of a text, and consequently use it in various ways and for various purposes.

[1] Parallel texts, also known as bilingual corpora, are bodies of text available in two or more languages.

https://doi.org/10.1017/S1351324905003827 Published online by Cambridge University Press

Our goal in this special issue is to bring together articles highlighting the achievements and challenges in building and using parallel texts. We were fortunate to receive a large number of high-quality submissions, covering many different aspects of the domain, ranging from corpus-building and annotation to translation modeling and automation, to knowledge transfer and extraction. We present here a panorama of this vast domain of research and study, and introduce the various contributions that appear in this issue as we go along.

2 Building parallel texts

The amount of text that undergoes translation is staggering. Billions of dollars are spent yearly on translation worldwide. By current standards, this also means billions of translated words. And where there is translation, there are parallel texts. Certainly one of the first challenges for NLP researchers and developers is to access this data, collect it, and make it available in a format suitable for electronic use.

Perhaps the most widely used solution for obtaining a corpus is to negotiate with owners of existing parallel texts the rights to access, use, and disseminate the data. Large, multilingual collections of governmental and institutional texts, such as the Canadian and Hong-Kong Hansards, the Proceedings of the European Parliament, various publications emanating from the United Nations, etc. were initially collected this way. However, the road to fruitful negotiations with public institutions is often long and winding, and must be traveled with patience. As for private collections, such as newswire archives, most publishers seem reluctant to clear such data from copyright protection, even for exclusive research purposes.

Then, there is the Web. As far as anyone can tell, while the Web is still growing at a steady pace, English texts make up a diminishing portion of the lot, which means that other languages are taking up a larger share of the pie (Kilgariff and Grefenstette 2003). How much of this "multilingual Web" is actually made up of parallel text, and how the different language-pairs are represented remain open questions.

Methods for automatically harvesting parallel corpora by crawling the Web, such as those proposed by Resnik and Smith (1999) and others are relatively easy to implement, and can be quite effective in locating pairs of URL's that point to mutual translations. The amount of data that can be quickly collected this way is sometimes impressive: Kraaij, Nie and Simard (2003) report collecting 15 million words of English-French bitext, 3.6 million words English-German, 2.5 million words English-Italian and 19 million words English-Chinese in this way; Resnik and Smith gathered approximately 2 million words of English-Arabic text and possibly over 100 times as much English-Chinese[2]. However, along with this sort of automatic collection inevitably come issues of data quality and uniformity. Furthermore, it must be pointed out that just because something is on the Web does not mean it is in the public domain; therefore, collecting data from the Web does not entirely solve the problems of ownership and copyright.

[2] The STRANDS Database – http://www.umiacs.umd.edu/~resnik/strand/

https://doi.org/10.1017/S1351324905003827 Published online by Cambridge University Press

Yet, the Web represents only a fraction of the phenomenal amounts of parallel texts that lie dormant (or half-asleep) on the hard disks of private and public institutions all over the world. In a growing number of cases, thanks to the interest in translation memory technology, this information is systematically being archived and organized. Of course, the research community cannot truly hope to ever have access to this wealth of knowledge. But as those who own the data discover the value of their assets, and as they realize that researchers are finding ways of using it, there is still hope that someday, it will be fully exploited.

In this issue, Luisa Bentivogli and Emanuele Pianta present the *MultiSemCor* corpus and the details of their methodology for assembling a richly annotated corpus of parallel English and Italian text. Possibly one of the most interesting aspects of their work is the initial method for acquiring the corpus: by translating into Italian an existing annotated corpus in English, and then automatically transferring the annotations. Such an approach may seem surprising in view of the wealth of existing parallel texts. But as the authors point out, "Not only are translators more easily available than linguistic annotators, but translations may be a more flexible and durable kind of annotation". Furthermore, by controlling the translation process itself, they attain better quality alignments and annotations than could be obtained via pre-existing, uncontrolled translations.

3 Processing parallel texts

In their raw, original format, parallel texts would not be of much use for computer applications, if it were not for a variety of tools designed to prepare such corpora. Possibly the most critical phase in this preparation work is the alignment of parallel texts at various degrees of resolution (documents, sentences, words). While the work carried out at the IBM T.J. Watson Research Center in the late eighties and early nineties (Brown, Della Pietra, Della Pietra and Mercer 1993) remains the basic reference when it comes to sentence and word-alignment, arguably the most influential event of recent years for parallel text processing is the work of the 1999 Johns Hopkins Summer Workshop on Statistical Machine Translation, and the decision to make available to the research community the machine translation software that was developed alongside this event. The so-called **Egypt** toolkit (Al-Onaizan, Curin, Jahr, Knight, Laffferty, Melamed, Och, Purdy, Smith and Yarowsky 1999), and its successor component **GIZA++** (Och and Ney2003) implement the required machinery to build and train IBM and other statistical translation models. As a side-effect, they also produce alignments between the words of the parallel texts used to train the model.

Professor Frederick Jelinek, who headed the Continuous Speech Recognition Lab at IBM when these statistical translation models were first being developed, reportedly once said: "Every time I fire a linguist, my performance goes up". Indeed, **GIZA++** aligns words using exclusively statistical information. And yet, there is a growing belief that linguistic annotations and knowledge-based heuristics could play a role in improving word-alignment quality.

https://doi.org/10.1017/S1351324905003827 Published online by Cambridge University Press

The contributions of David Talbot and Jörg Tiedemann to this issue could signal the return of the linguists to the lab. Talbot presents a method that allows a priori information, in the form of alignment constraints, to infiltrate the maximum likelihood parameter estimation at the core of **GIZA++**. His approach allows improvements in both alignment and translation quality, especially when training data is scarce. But more importantly, it opens the door to a well-founded integration of various types of linguistic constraints and annotations in the alignment process.

As for Tiedemann, he shows how an alignment method based on an optimized combination of statistical, linguistic, and heuristic clues can significantly improve over techniques for word alignment that are solely based on statistical information. In particular, he proposes the use of a genetic algorithm to implement an evolutionary procedure that optimizes the combination of various knowledge sources for word alignment, with clear improvements observed in the alignment of an English–Swedish corpus as compared to the best baseline alignment produced by statistical methods.

4 Exploiting parallel texts

Machine translation is perhaps the most voracious consumer of parallel texts. Church and Mercer once pointed out that "more data are better data" (Church and Mercer1993). And indeed, most (if not all) existing studies on the relationship between the amount of available training data and the ensuing quality of data-driven machine translations seem to support this claim. If this is encouraging for language pairs for which large quantities of parallel texts are readily available (e.g. English-French and English-Chinese), it comes as bad news for most other language pairs, for which such data are available only in small quantities or even lacking completely. Yet, there is a growing interest in data-driven methods that attempt to make better use of smaller amounts of data. In their contribution to this special issue, Andy Way and Nano Gough take on a long debate in the machine translation community, and compare how statistical and example-based MT systems perform, when exposed to varying amounts of parallel text. The results of their experiments, if sometimes surprising, are no less instructive.

If mechanically going from one language to the other is a research field almost as old as the computer itself, a much more recent trend follows the idea of using existing translations as a *bridge* to carry linguistic knowledge from one language to the other. Following the seminal work of Yarowsky and Ngai (2001), considerable work has recently been devoted to cross-lingual natural language processing applications in which semantic, syntactic, or discourse annotations, usually attached to text in a well-studied language, are transfered to one that has traditionally received less attention.

The work of Bentivogli and Pianta, mentioned above, follows precisely this line, as does that of Rebecca Hwa and colleagues, who show here how syntactic dependencies can be effectively transferred across parallel texts. In work centered around the idea of creating treebanks with minimal human intervention, they demonstrate that parallel texts represent a useful resource for creating syntactic annotations

https://doi.org/10.1017/S1351324905003827 Published online by Cambridge University Press

in several languages, resulting into automatically annotated corpora that can be used to train syntactic parsers. An interesting aspect of their work is a clear evaluation of how much mileage one can gain by adding a small amount of language-specific information: manual correction rules requiring less than one person-month are shown to effectively account for the limitations of the automatic dependency projection, and significantly improve the performance of the syntactic parser trained on the noisy annotations.

Yet another trend that has emerged in recent years is the more ambitious idea of automatically *discovering* knowledge via parallel text. For instance, Church and Gale (1991), and then later Resnik and Yarowsky (1999), have shown how word sense distinctions can be automatically learned using the translations of an ambiguous word in a second language. But the most direct application of this idea is perhaps observed in automatic lexicon extraction. Of course, bilingual lexicons can be useful for machine translation, cross-language information retrieval and other NLP applications, but they are also of direct use to human translators and terminologists, who must deal with the constant evolution of terminology in specialized domains and general language.

Magnus Sahlgren and Jussi Karlgren present here an intriguing new method to automatically extract lexical equivalents from parallel text. Their approach builds upon the familiar idea of linking word-pairs based on the similarity of their distribution in parallel corpora (Kay and Röscheisen 1988; Fung and Church 1994), but relies on a clever dimension reduction scheme, in the spirit of the *Random Projection* method (Papadimitriou, Raghavan, Tamaki and Vempala 1998), to overcome complexity and robustness issues: high-dimensionality vectors, describing the contexts within which individual words appear in the corpus, are projected into a much lower-dimensionality space through a random projection. In turn, this makes it possible to process very large quantities of parallel text very efficiently, both in terms of computation time and space.

5 Perspectives on the future of parallel texts

Will machine translation continue to be the main consumer of parallel texts? Probably so, but most likely in ways different than what we have seen so far.

In the last two decades, we have witnessed significant improvements in a number of important text processing problems, through approaches that rely (exclusively, in some cases) on large amounts of textual data as sources of knowledge. Much of this recent progress is the result of major advances in statistical natural language learning, and more generally on the application of emerging machine learning techniques to natural language. But recently, a shift in this tendency has been noticed. Although data-driven approaches have undoubtedly revolutionized the field of natural language processing, many researchers seem to agree that, in the future, progress is more likely to come from a better understanding of deeper linguistic aspects, rather than from better machine learning algorithms or more data.

This trend is particularly apparent in research in machine translation, where more and more emphasis is being placed on richer knowledge representations for

https://doi.org/10.1017/S1351324905003827 Published online by Cambridge University Press

improved automatic translation. Under this new tendency of placing focus on deeper linguistic knowledge, rather than pure data-based statistical modeling, the role of parallel texts in machine translation is likely to also face a shift.

So far, parallel texts have been used in data-driven approaches to machine translation mostly as means of learning lexical or phrasal equivalences. But as we learn how to effectively use deeper linguistic knowledge in performing complex NLP tasks, the role of parallel texts as a bridge to transfer knowledge between languages and as a source of knowledge about language structure and meaning are likely to take more importance. This last direction, in particular, is most promising. Parallel data, be it multilingual or not, constitutes multiple views or perspectives on the same information. What makes it so rich is precisely the fact that it is never entirely parallel. Approaches to word-sense disambiguation that rely on parallel texts exploit these differences directly to infer meaning. Parallel bracketing methods such as those proposed by Wu (1997) again rely on differences in word-order to discover structure in different languages. It is these "orthogonalities" in parallel data that we must learn to explore and exploit, as they hold the clues to answer many questions. Every additional language brings along a new perspective, thus multiplying these orthogonalities, and alongside the opportunities for resolving new problems in NLP.

Acknowledgements

We thank all the authors who responded enthusiastically to the call for papers for this special issue, for all the hard work that went into their submissions. We received an impressive number of articles (36), out of which only a small number could be retained. We are also grateful to the members of the review committee, whose excellent work was essential for selecting the articles that make up this issue, and who have done a wonderful job providing feedback to the authors. Many thanks go to Elliott Macklovitch for his insightful comments on an earlier draft of this paper. We owe him the analogy between Warren Weaver's quote and the view of parallel texts seen as semantic representations. Finally, we would like to thank John Tait and the NLE editorial board for their trust and their help in putting together this special issue.

Review Committee

Lars Ahrenberg (Linköping U.), Susan Armstrong (ISSCO), Michael Barlow (Rice U.), William Byrne (JHU), Chris Callison-Burch (U. Edinburgh), Francisco Casacuberta (UP València), Violetta Cavalli-Sforza (CMU), Jiangping Chen (U. North Texas), Ken Church (Microsoft), Silviu Cucerzan (Microsoft), Ido Dagan (Bar-Ilan U.), Jason Eisner (JHU), George Foster (NRC Canada), Pascale Fung (HKUST Hong Kong), Eric Gaussier (XRCE), Ulrich Germann (U. Toronto), Daniel Gildea (U. Rochester), John Goldsmith (U. Chicago), Julio Gonzalo (UNED), Cyril Goutte (XRCE), Gregory Grefenstette (CEA France), Eduard Hovy (USC/ISI), Pierre Isabelle (NRC Canada), Hitoshi Iida (Tokyo U.), Philipp Koehn (U. Edinburgh),

https://doi.org/10.1017/S1351324905003827 Published online by Cambridge University Press

Wessel Kraaij (TNO/TPD Netherlands), Shankar Kumar (JHU), Philippe Langlais (U. Montréal), Alon Lavie (CMU), Elizabeth Liddy (Syracuse U.), Elliot Macklovitch (U. Montréal), Robert Moore (Microsoft), Dan Melamed (NYU), Ruslan Mitkov (U. Wolverhampton), Hermann Ney (RWTH Aachen), Hwee Tou Ng (NUS Singapore), Jian-Yun Nie (U. Montréal), Franz Och (Google), Kemal Oflazer (Sabancı U.), Martha Palmer (U. Pennsylvania), Kishore Papineni (IBM), Ted Pedersen (U. Minnesota, Duluth), Jessie Pinkham (U. Chicago), Andrei Popescu-Belis (ISSCO/TIM/ETI U.Geneva), Dragomir Radev (U. Michigan), Florence Reeder (MITRE), Philip Resnik (U. Maryland), Antonio Ribeiro (ECJRC), Charles Schafer (JHU), Harold Somers (UMIST), Hideki Tanaka (ATR SLT), Arturo Trujillo (Canon CRE), Jean Véronis (U. Provence), Clare Voss (ARL), Andy Way (Dublin City U.), Yorick Wilks (U. Sheffield), Dekai Wu (HKUST Hong Kong), Kenji Yamada (ISI)

Additional Reviewers

Katarina Probst (CMU), Evgeny Matusov (RWTH Aachen), Nicola Ueffing (RWTH Aachen), Martin Volk (Stockholm U.)

References

Al-Onaizan, Y., Curin, J., Jahr, M., Knight, K., Lafferty, J., Melamed, D., Och, F.-J., Purdy, D., Smith, N. H. and Yarowsky, D. (1999) Statistical Machine Translation - Final Report, JHU Workshop 1999. Technical report, Johns Hopkins University.

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

Church, K. and Gale, W. (1991) Concordances for Parallel Text. *Seventh Annual Conference of the UW Centre for the New OED and Text Research*, Oxford, UK.

Church, K. and Mercer, R. (1993) Introduction to the Special Issue on Computational Linguistics Using Large Corpora. *Computational Linguistics*, **19**(1): 1–24.

Fung, P. and Church, K. W. (1994) K-Vec: A New Approach for Aligning Parallel Texts. *Proceedings of the 15th International Conference on Computational Linguistics (COLING-94)*, pp. 1096–1102. Kyoto, Japan.

Kay, M. and Röscheisen, M. (1988) Text-translation Alignment. Technical report, Xerox Palo Alto Research Centre.

Kilgariff, A. and Grefenstette, G. (2003) Introduction to the Special Issue on the Web as Corpus. *Computation Linguistics*, **29**(3): 333–348.

Kraaij, W., Nie, J.-Y. and Simard, M. (2003) Embedding Web-based Statistical Translation Models in Cross-Language Information Retrieval. *Computational Linguistics*, **29**(3): 381–419.

Och, F. J. and Ney, H. (2003) A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, **29**(1): 19–51.

Papadimitriou, C. H., Raghavan, P., Tamaki, H. and Vempala, S. (1998) Latent Semantic Indexing: A Probabilistic Analysis. *Proceedings of the 17th ACM Symposium on the Principles of Database Systems*, pp. 159–168. ACM Press.

Resnik, P. and Yarowsky, D. (1999) Distinguishing systems and distinguishing senses: new evaluation methods for word sense disambiguation. *Natural Language Engineering*, **5**(2): 113–134.

https://doi.org/10.1017/S1351324905003827 Published online by Cambridge University Press

Resnik, P. (1999) Mining the Web for bilingual text. *Proceedings of 37th Annual Meeting of the Association for Computational Linguistics (ACL 1999)*, College Park, MD.

Wu, D. (1997) Stochastic Inversion Transduction Grammars and Bilingual Parsing of Parallel Corpora. *Computational Linguistics*, **23**(3): 377–404.

Yarowsky, D. and Ngai, G. 2001. Inducing multilingual pos taggers and np bracketers via robust projection across aligned corpora. *Proceedings of North American Chapter of the Association for Computational Linguistics NAACL-2001*, pp. 200–207.

https://doi.org/10.1017/S1351324905003827 Published online by Cambridge University Press