Parallel Text Processing: Alignment and Use of Translation Corpora

Jean Véronis (editor)
(Université de Provence)

Dordrecht: Kluwer Academic Publishers (Text, speech and language technology series, edited by Nancy Ide and Jean Véronis, volume 13), 2000, xxiii+402 pp; hardbound, ISBN 0-7923-6546-1, $160.00, £99.00, Dfl 300.00

Reviewed by
Philip Resnik
University of Maryland

1. Introduction

One can't help but be fascinated by two sentences in parallel translation, the selfsame meaning diffused, distributed, diverging across alternative expressions. In his *Le Ton beau de Marot: In Praise of the Music of Language*, Douglas Hofstadter goes to the extreme of inventing or examining dozens upon dozens of alternative English translations of a single 18-line French poem, reveling in the nuances of content and form. Articles on machine translation technology in the popular press place the obligatory mangled proverb, obtained by automatic translation and then automatic back-translation, against its original. Closer to home, students of syntactic theory examine word-by-word glosses of sentences in unfamiliar languages in order to see uncovered there the building blocks—morpheme, inflectional marking—that recombine into a fluid expression using familiar words.

By all rights, *Parallel Text Processing* should be merely a technical volume, and this should be merely a technical review. But editor Véronis has chosen to leave until the end the technical inspiration for the book, namely, the ARCADE evaluation exercise for sentence- and word-level alignment systems. At the start of the book, instead, he places a tiny gem of a preface written by Martin Kay. In this preface, Kay makes two significant observations about corpora. First, he notes that what makes current statistical methods tick is that they use knowledge about language as a proxy for world knowledge; second, he argues that aligned texts are a rich source of knowledge about language, knowledge placed there by human translators.

To be sure, Kay expresses doubt about the adequacy of corpora to serve in place of world knowledge where the tough problems are concerned: "More substantial successes in these enterprises [such as high-quality automatic translation] will require a sharper image of the world than any that can be made out simply from the statistics of language use" (page xvii). I myself am more optimistic, because as our field moves from its statistical renaissance into the next phase, I think we are seeing a return to semantic issues that may help provide the "sharper image of the world" Kay believes necessary. I do not mean by this a renewed focus on semantic theory per se, but rather a return to questions of meaning and world knowledge accompanied by the insights of quantitative, corpus-based approaches. To take a few examples, stochastic parsing has moved beyond constituency analyses to syntactic dependencies, which are closer to

underlying thematic relationships (Hajič et al. 1998); large-scale lexical resources are being built with more elaborated semantic underpinnings (Fillmore, Wooters, and Baker 2001; Vossen 1998), and the productive annotate–train–evaluate paradigm is finding application in meaning-oriented tasks ranging from named entity identification and coreference (Harman and Chinchor 2001) to word sense disambiguation (Cotton et al. 2001).

2. Individual Contributions

Whether or not one shares my optimism for the long-term future of semantic depth in corpus-based approaches, it is clear that, as Kay puts it, "The unparalleled richness of aligned texts for a great number of purposes is clear for anyone to see" (page xvii). Véronis begins the collection with "From the Rosetta stone to the information society: A survey of parallel text processing," a very nice survey of how parallel texts are processed and used, laying out clearly the book's organization into major sections on techniques for alignment, applications of parallel texts, and corpus resources and evaluation. Starting in this first chapter by Véronis, points of consensus quickly emerge in alignment methodology, notably principles of lexical anchoring (e.g., dictionary-based word pairs or cognate pairs) and of length correlation (taking advantage of the tendency of short units to translate into short, medium into medium, long into long), with these two principles serving individually or in combination as the basis for establishing correspondences.

The first five methodology chapters lay out alternative approaches taking advantage of multiple sources of information within various levels of analysis. Melamed's chapter, "Pattern recognition for mapping bitext correspondence," lays out a powerful and general geometric approach to identifying token-level correspondences that takes advantage of both the general length correlation and a general matching predicate that can exploit dictionary- or cognate-based lexical anchors as available. Simard, in "Multilingual text alignment: Aligning three or more versions of a text," focuses on sentence-level alignment, adapting techniques from molecular biology. A key innovation here is the exploration of how transitivity of word-level translations in *multilingual* parallel texts can help identify correspondences missed in the component bilingual corpora. Ahrenberg, Andersson, and Merkel, in "A knowledge-lite approach to word alignment," emphasize a modular combination of resources, with the ultimate aim of obtaining a nonprobabilistic translation lexicon from parallel corpus links; the chapter demonstrates portability from English-Swedish to English-French language pairs. The chapter by Choueka, Conley, and Dagan, "A comprehensive bilingual word alignment system—application to disparate languages: Hebrew and English," is distinguished by its focus on relevant properties of the languages, notably issues of morphology and lemmatization in a Semitic language. The chapter is likely to be of interest to the growing segment of the community looking at the English-Arabic language pair. Piperidis, Papageorgiou, and Boutsis, in "From sentences to words and clauses," look at multiple levels of alignment—words, noun phrases, clauses, and sentences—with a focus on machine-assisted translation of English-Greek.

The remaining four alignment methodology chapters explore radically different visions of what it means for two texts to be aligned. Wu, in "Bracketing and aligning words and constituents in parallel text using Stochastic Inversion Transduction Grammars," presents a stochastic grammar formalism that characterizes sentence pairs, and associated algorithms, making it possible to take a generative or language-modeling view of parallel texts. The chapter offers a clear exposition covering bracketing, phrasal alignment, and word alignment. Fluhr, Bisson, and Elkateb, in "Parallel text alignment

using crosslingual information retrieval techniques," view the sentence alignment process as one in which the goal is to match sentences with their translations, much as queries are matched with relevant documents in many retrieval systems based on a vector space notion of semantic proximity. The resulting technique is naturally quite robust in the face of gaps or order variations. In "Parallel alignment of structured documents," Romary and Bonhomme view documents as multilevel structures and take advantage of structure information in alignment. The chapter includes an interesting description of relevant aspects of the Text Encoding Initiative (TEI) guidelines for annotating document structure. Although the chapter by Santos, "The translation network: A model for a fine-grained description of translations," accompanies in this section other chapters concerning alignment methods, the main emphasis of the chapter is a detailed corpus study rather than alignment techniques. The empirical evidence leads Santos to challenge standard assumptions about parallel texts and alignment and to propose an alternative conception of translation inspired by the monolingual notion of aspectual coercion.

The five chapters on applications of parallel text and alignment illustrate significant diversity, and at the same time they all draw attention to fundamental problems concerning the matching of units of meaning at the right granularity. Fung, in "A statistical view on bilingual lexicon extraction: From parallel corpora to non-parallel corpora," raises the question of how to deal with bilingual corpora containing noise, as might be the case when sentence-level units are not cleanly aligned. In order to handle corpora containing comparable documents rather than parallel translations, Fung suggests a vector space approach similar to that of Fluhr, Bisson, and Elkateb but at the level of word contexts rather than sentences. Blank's chapter, "Terminology extraction from parallel technical texts," differs from some of the preceding work in its focus on a particular domain (patent documentation) and technical terminology. Monolingual terminology identification is combined with semiautomatic tools for matching terms; French and German illustrate problems encountered when trying to match multiword versus compounded terms. Gaussier, Hull, and Aït-Mokhtar, in "Term alignment in use: Machine-aided human translation," present methods for term extraction and alignment as well as innovative translation memory techniques that permit multilevel matching. Brown, Carbonell, and Yang, in "Automatic dictionary extraction for cross-language information retrieval," extract a bilingual dictionary using a simple technique based on co-occurrence frequency, use the resulting entries to improve alignments, then extract an improved dictionary; they present comparative results in cross-language retrieval experiments. Nerbonne's chapter, "Parallel texts in computer-assisted language learning," provides background on language learning and presents a tool to assist Dutch students of French. The tool approximates automatic glossing of French sentences in Dutch and permits ready access to parallel examples, relying heavily on the ability to treat morphological variants as equivalents.

The final section of the book concerns efforts to support the community by means of parallel corpora, resource sharing, and common evaluation. Isahara and Haruno's chapter, "Japanese–English aligned bilingual corpora," is really a combination of a proposed alignment technique and a corpus-building project. The chapter makes clear how difficult it can be, owing to availability, quality, and copyright issues, to produce a parallel corpus of reasonable size, even for better-studied languages such as English and Japanese. Singh, McEnery, and Baker's chapter, "Building a parallel corpus of English/Panjabi," strongly reinforces the message of the previous chapter. For modern Panjabi, a representative of Indic languages, the authors after a great deal of work found a collection of children's bedtime stories with English translations, tracked down the author for permission, and ultimately typed the Panjabi text in by hand because

it was unavailable as electronic text. In the chapter by Melby, "Sharing of translation memory databases derived from aligned parallel text," the author describes in detail the translation memory exchange (TMX) format, developed by a standards group in order to facilitate the sharing of translation memory databases. Véronis and Langlais, in "Evaluation of parallel text alignment systems: The ARCADE project," describe the common evaluation of parallel text alignment systems. They report greater than 98% sentence alignment accuracy for systems on "normal" texts; word alignment was evaluated nonexhaustively and yielded accuracy estimates in the vicinity of 75%.

3. Conclusion

Parallel Text Processing succeeds admirably at its goals and will be of use to a wide range of people. One of the book's primary goals is to address a wide range of topics and to cut across communities, and it does a surprisingly good job of balancing introductory and historical material with technical substance. The overall coherent organization is supported further by the abstracts and keywords at the front of each chapter. I can easily imagine using this book both to introduce ideas in a graduate seminar and as a reference for research.

The book should be quite accessible to graduate students and researchers in computational linguistics and to computer scientists. Linguists, translators, or those with less mathematical background might want to brush up on their understanding of basic probability theory and to perhaps get a brief tutorial on dynamic programming from a friendly computer scientist. Even skimming the more technical sections, however, they are likely to be rewarded by interesting material. All readers will be grateful that the editor chose to include separate and fairly thorough indexes for terms, authors, and languages and writing systems.

References

Cotton, Scott, Phil Edmonds, Adam Kilgarriff, and Martha Palmer, editors. 2001. *SENSEVAL-2: Second International Workshop on Evaluating Word Sense Disambiguation Systems*, Toulouse, France, July. ACL SIGLEX. http://www.sle.sharp.co.uk/senseval2/.

Fillmore, Charles J., Charles Wooters, and Collin F. Baker. 2001. Building a large lexical databank which provides deep semantics. In *Proceedings of the Pacific Asian Conference on Language, Information and Computation*, Hong Kong.

Hajič, Jan, Eric Brill, Michael Collins, Barbora Hladká, Douglas Jones, Cynthia Kuo, Lance Ramshaw, Oren Schwartz, Christoph Tillmann, and Daniel Zeman. 1998. Core natural language processing technology applicable to multiple languages: Final report. Technical Report, Center for Language and Speech Processing, Johns Hopkins University, Baltimore. http://www.clsp.jhu.edu/ws98/projects/nlp/report/.

Harman, Donna and Nancy Chinchor. 2001. Message understanding conference proceedings. http://www.itl.nist.gov/iaui/894.02/related_projects/muc/ proceedings/proceedings_index.html

Hofstadter, Douglas R. 1997. *Le Ton beau de Marot: In Praise of the Music of Language.* Basic Books.

Vossen, Piek. 1998. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks.* Kluwer Academic Publishers, Dordrecht.

Philip Resnik is an assistant professor at the University of Maryland in the Department of Linguistics and the Institute for Advanced Computer Studies (UMIACS). He has developed STRAND, a system for automatically finding parallel texts on the Web, and is working on linguistically informed statistical methods for machine translation. Resnik's address is: Department of Linguistics, 1401 Marie Mount Hall, University of Maryland, College Park, MD 20742; e-mail: resnik@umiacs.umd.edu.