Parallel Text Processing
Alignment and Use of Translation Corpora

Edited by Jean Véronis

(Université de Provence)

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

Reviewed by
Maria Zimina
(Université de la Sorbonne Nouvelle – Paris 3)

1. Introduction

In the past ten to fifteen years considerable progress has been made in the field of parallel text alignment. The term parallel text itself is now well-established within the computational linguistics community. It refers to texts accompanied by their translations in one or several languages. Aligned texts have proved to be an invaluable source of translation data for terminology banks and bilingual dictionaries. Translation alignment is currently providing the basis for the development of a new generation of tools to assist human translators and to improve the quality and productivity of their work.

The book "Parallel Text Processing: Alignment and Use of Translation Corpora" edited by Jean Véronis, aims at presenting an extensive overview of different research areas in the field of multilingual text alignment. It is the first of its kind to attempt a thorough evaluation and synthesis of recent advances in alignment technology on a large scale. It evolved from the ARCADE project carried out to contribute to the development of a methodology for the evaluation of existing alignment systems (Véronis and Langlais, 1999).

A remarkable example of collaboration of the thirty-four authors having been involved in major alignment research projects in recent years, the book is a collection of high-quality papers presented as individual chapters.

Guidelines for reading and an outline of the book structure are articulated in the preface by Martin Kay and the introductory chapter "From the Rosetta stone to the information society" by Jean Véronis (Chapter 1).

Kay points out that the experience of multilingual text alignment puts forward the lack of empirical knowledge of traditional translation. In this respect, different approaches to parallel text processing described in the book are likely to open new horizons for better understanding of the human translation process. Véronis gives a thorough analysis of ongoing evolution in parallel text processing. His paper aims at providing a survey of the state of the art in multilingual text alignment; it fully succeeds at its goals due to comprehensiveness and broad coverage of most important research issues of the field.

On the whole, inherent diversity of approaches and themes addressed in the book strengthens both a full potential of the alignment technology, catering for the needs of multilingual information society, and a high demand for methodological guidelines and evaluation standard in this field. In consequence, the division of the book into three parts: alignment and methodology (Chapters 2 to 10), applications (Chapters 11 to 15), and resources and evaluation (Chapters 13 to 19) is rather arbitrary. The reader will notice that all three aspects are closely interconnected and are often treated simultaneously. As a matter of fact, the state of the art in parallel text processing has not yet permitted to deal with them separately.

2. Content overview

The first five chapters present different methods used to identify translation correspondences.

In "Pattern Recognition for mapping bitext correspondence" (Chapter 2), D. Melamed suggests an innovative approach to the problem of finding token-level correspondences (bitext maps). Following length correlation alignment strategies by Gale and Church (1991), and Brown, Lai and Mercer (1991), based on internal corpus information, Melamed expands the matching procedure to exploit both dictionary and cognates. The author provides technical insights of search space reduction and filtering techniques increasing the robustness of the method. In Chapter 3 "Multilingual text alignment", M. Simard addresses the question of mapping multilingual translation equivalence. According to the author, aligning multiple versions of texts gives some additional insights in terms of correspondence accuracy. The chapter by Y. Choueka, Eh. S. Conley and I. Dagan, "A comprehensive bilingual word alignment system", describes an innovative alignment system for disparate language pairs (Chapter 4). It contains a detailed description of system architecture and some experimental results regarding the Hebrew-English language pair. The authors propose to convert raw texts to a stream of lemmatized tokens and then use the DK-vec algorithm (Fung and McKeown, 1997) to identify source and target words appearing in the corresponding positions in the two texts. The assumption is made about translation relatedness of these words. An extremely challenging task of alignment and extraction of lexicons is addressed in Chapter 5 by L. Ahrenberg, M. Andersson and M. Merkel "A knowledge-lite approach to word alignment". The system described here will be of interest to researches with diverse backgrounds in natural language processing. It combines statistical measures of co-occurrence with knowledge-lite modules of word categorization, morphological variation, word order, and phrase recognition.

Chapters 6 to 10 are noteworthy for some alternative visions of the concept of text alignment.

In Chapter 6, "From sentences to words and clauses", S. Piperidis, H. Papageorgiou, and S. Boutsis suggest a method enabling identification of clauses in text sentences, and establishment of equivalencies at sentence, clause, noun phrase and word level. The paper emphasizes the idea of significance of clauses as translation units. D. Wu, in his chapter "Bracketing and aligning words and constituents in parallel text using Stochastic Inversion Transduction Grammars", introduces a general formalism for modeling of bilingual sentence pairs and offers a highly innovative algorithm for finding the optimal bilingual parse of a sentence pair. Chapter 8, "The translation network: a model for a fine-grained description of translations" by D. Santos, stands apart from the rest of the papers presented in this section, as it is almost entirely devoted to empirical study of the translation process. Santos critically reviews current translation description models and attempts at interpreting translation through some kind of topology network, mapping from categories specific to the source language into categories particular to the target language. In contrast with the belief of P. Isabelle, who tends to limit the reconstruction of translation correspondences to some passive use of linguistic capability (Isabelle, 1992), Santos puts emphasis on the inherent complexity of this phenomenon. In "Parallel text alignment using crosslingual information retrieval techniques" (Chapter 9),

Ch. Fluhr, F. Bisson, and F. Elkateb demonstrate that parallel text alignment can be successful even if it ignores classical assumptions of sequential order of text units. Indeed, a crosslingual query based on all possible translations of words in a source sentence can help to obtain candidates for alignment among target sentences. The exploration of this method is still at its beginning. For the moment, its chief advantage results in robustness to text desynchronization (omissions, differences in text order etc.). The final methodology chapter "Parallel alignment of structured documents" by L. Romary and P. Bonhomme, extends the notion of alignment to deal with multi-level textual data structured by means of tagging languages, such as SGML. This direction seems to be quite promising due to growing importance of text encoding.

Extraction and use of translation resources contained in parallel corpora are discussed in the following five chapters of the book (Chapters 11 to 15).

P. Fung in "A statistical view on bilingual lexicon extraction: from parallel corpora to non-parallel corpora" draws attention to an under-exploited richness of non-parallel yet comparable corpora currently available in almost any field of knowledge. According to Fung, a rather problematic task of bilingual lexicon extraction from non-parallel corpora can be sorted out by statistical study of context word similarity between candidate terms, representing mutual translation pairs. Chapters 12 and 13 outline different methods of bilingual terminology extraction. The focus is on complex translation units that are particularly difficult to detect automatically. In "Terminology extraction from parallel technical texts", I. Blank develops a method for semi-automatic term identification through sequential tokenization, lemmatization, POS-tagging and pattern matching. The results are presented in the form of a concordance tool suitable for translators and terminologists. The concept of translation memory is widely explored by E. Gaussier, D. Hull and S. Aït-Mokhtar in "Term alignment in use: machine-aided human translation". The authors also present a method for a broader detection of potential correspondences across languages. Syntactic dependencies are explored to delimit potential terminological units. Statistical procedures are then employed to calculate expected alignment frequencies and estimate translation probability for a given pair of units. The paper "Automatic Dictionary Extraction for CLIR" by R. D. Brown, J. G. Carbonell and Y. Yang, points to certain successful applications of automatically generated bilingual dictionaries in Cross-Language Information Retrieval. The reader is presented with an extensive comparative study of different CLIR methods using a bilingual training corpus. The algorithm suggested for dictionary extraction makes use of co-occurrence features, statistical filtering and word order. The last chapter on applications, "Parallel texts in computer-assisted language learning" by J. Nerbonne, demonstrates the value of parallel texts in foreign language learning. From a technical point of view, the basic idea is to use parallel texts glossed to supply information on grammar and dictionary equivalent. The search is performed by lexeme; alternatively inflected forms of words are taken into account. The program provides the user with interactive facilities for looking up information on word use drawn automatically from corpora.

The final section of the book outlines several major projects aimed at the compiling of parallel corpora, standardization and sharing of translation memory resources, as well as evaluation of alignment software (Chapters 17 to 19).

The chapter "Japanese-English aligned bilingual corpora" by H. Isahara and M. Haruno, describes problems of corpus development and presents bilingual corpora compiled in Japan. The reader will find here a comprehensive overview of different facets involved in corpus construction: problems of copyright, computerization and text encoding, automatic alignment tagging, manual post-editing and evaluation. Similar issues are discussed in Chapter 17, "Building a parallel corpus of English/Panjabi" by S. Singh, T. McEnery and P. Baker, in relation to rare word language pairs, such as English/Panjabi. The authors give an impressive coverage of wide-ranging difficulties in corpus construction related to such language pairs:

scarcity of electronic text resources, variations in writing systems, complicated encoding, as a result of heterogeneous text structure, etc.. Nevertheless, it seems that corpus studies are gradually expanding to include more and more of the world's minority languages. Considerable growth of parallel corpora stimulates the demand for unification of resources and standardization of translation memory lookup software. In this respect, the chapter "Sharing of translation memory databases derived from aligned parallel text" by A. K. Melby, deserves special attention as it covers basic features of the translation memory exchange format (TMX). Established in 1997, with participation of leading translation memory software developers, TMX provides a common standard for storage of aligned texts, improving their re-usability in various applications. Finally, guidelines for assessment of alignment software are articulated by J. Véronis and Ph. Langlais in the very last chapter of the book "Evaluation of parallel text alignment systems: the ARCADE project". The authors describe a pioneering evaluation project conducted at the international level that enabled to assess sentence and word alignment accuracy for some major alignment systems. The chapter provides interesting material concerning evaluation protocol. Figures on comparative ranking of systems' performance will convince the reader of important advances in technical aspects of parallel text alignment.

3. Conclusions

The book is highly advised to researchers in computational linguistics, statistics of text, translation, lexicography and terminology. It is also accessible to advanced students of these disciplines, although some mathematical background and programming skills are desirable to go through more technical sections. The introductory chapter is likely to provide guidelines for preparing teaching courses and seminars on parallel corpora and text alignment.

The structure of the book is clearly outlined. Abstracts, keywords and references are provided for all chapters. The reader will be pleased to discover detailed indexes of terms, authors, languages and writing systems. Unfortunately, these anchors are not always sufficient to establish links between individual contributions. This lack of inherent connection between chapters is partially explained by diverse research methodologies suggested by the authors. Nevertheless, it is counterbalanced by common understanding of central issues dealing with such aspects as granularity of text alignment, detection of complex lexical units, storage of equivalencies, parallel text encoding, development of translation memory lookup facilities and elaboration of standards for performance evaluation.

References

Véronis, J., & Langlais, P. (1999). ARCADE: évaluation de systèmes d'alignement de textes multilingues. In K. Chibout, F. Néel, J. Mariani & N. Masson (Eds.), *Ressources et Evaluation en Ingénierie de la Langue* (pp. 77-100): Aupelf-Uref.

Fung, P., & McKeown, K. (1997). A Technical Word and Term Translation Aid using Noisy Parallel Corpora Across Language Groups. *Machine Translation, 12*(1/2), 53-87.

Gale W., & Church K. W. (1991). A program for aligning sentences in bilingual corpora. *Proceedings of the 29th ACL*, Berkeley, CA, 177-184.

Brown, P. Lai, J. C., & Mercer, R. L. (1991). Aligning sentences in parallel corpora. *Proceedings of the 29th ACL*, Berkeley, 169-174.

Isabelle P.(1992). Bi-textual aids for translators. *Proceedings of the Eight Annual Conference of the UW Centre for the New OED and Text Research*, Uni versity of Waterloo, Waterloo, Canada.