Mach Translat
DOI 10.1007/s10590-012-9124-2

BOOK REVIEW

Alexander Clark, Chris Fox and Shalom Lappin (eds): Handbook of computational linguistics and natural language processing

Wiley-Blackwell, 2010, xxii + 775 pp, hardbound, ISBN 978-1-4051-5581-6

Pratyush Banerjee

Received: 28 January 2012 / Accepted: 3 March 2012

Over the past few decades, the fields of Computational Linguistics (CL) and Natural Language Processing (NLP) have come a long way from being sub-fields of formal linguistics and artificial intelligence (AI) to full-fledged research areas in their own accord. Based on the solid foundation of the substantial theoretical research, considerable applied research in these areas has also rendered them extremely relevant to large scale industrial applications. The *Handbook of Computational Linguistics and Natural Language Processing* aims at providing a detailed account of the different major research areas in CL and NLP. Alongside introducing different topics, this book also features a detailed overview of state-of-the-art research in each area. Compiled with the objective of serving as an useful reference for graduate students and researchers of computer science, linguistics, mathematics and philosophy, this book encapsulates a wide range of multidisciplinary theoretical topics and techniques in the field of CL and NLP. The *Handbook* has been edited by Alexander Clark from the Department of Computer Science, University of London, Chris Fox from the School of Computer Science and Electronic Engineering, University of Essex, and Shalom Lappin, professor of Computational Linguistics at King's College, London.

The objective of this review is to provide a comprehensive outline of the different research areas, techniques and applications covered in this book, focussing specifically on the area of machine translation (MT). Overall, the book comprises 22 different chapters, each dealing with a specific theory, technique, application area, or specific application in the field of CL and NLP. This collection of chapters has been grouped under four different sections: *Part I—Formal Foundations*, *Part II—Current Methods*, *Part III—Domains of Application* and *Part IV—Applications*. Part I has four

P. Banerjee (✉)
Centre for Next Generation Localization, School of Computing, Dublin City University, Dublin, Ireland
e-mail: pbanerjee@computing.dcu.ie

Published online: 18 March 2012

different chapters, each laying out the theoretical foundations of a different aspect of CL and NLP. Part II comprises seven different chapters and is divided into three sub-sections. The first sub-section, comprising five chapters, describes popular techniques in machine learning (ML) and their subsequent application to NLP tasks. The second and the third sub-sections, each comprising a single chapter, deal with the topics of corpus annotation and evaluation of NLP systems respectively. Part III showcases different application areas for CL and NLP techniques in six different chapters. Finally, Part IV constitutes five chapters each detailing out a specific real-life application of the different techniques in the area. The book concludes with a comprehensive reference list, followed by author and subject indices. Most of the relevant details about MT are packed into the single chapter titled *Machine Translation* (Chapter 19, Part IV) authored by Andy Way. Apart from this, some of the remaining chapters deal with theory and techniques directly used in MT. Furthermore, parts of other different chapters are also relevant to the context of different MT applications and paradigms. Although the actual MT chapter occurs quite late (section IV) of the book, it is primarily to make the reader accustomed to the different technologies used by MT, which are previously covered. In this review, we initially focus on Chapter 19 and gradually move on to the remaining chapters of the book according to their relevance to the MT context.

In Chapter 19, Andy Way provides a comprehensive overview of the different technologies and paradigms involved in MT. What sets this particular chapter apart from the other available overviews of MT (Somers 2003; Hutchins 2003) is the inclusion of the state-of-the-art techniques and paradigms that are still actively being researched and developed in the community. The chapter starts with a detailed introduction of phrase-based statistical machine translation (PBSMT), currently the dominant paradigm in MT (Koehn et al. 2003). Every individual step involved in the development of a standard PBSMT system, ranging from training data selection, to pre-processing, corpus clean-up, word alignment, and to actual training of the translation and language models, tuning, decoding and postprocessing, is covered in great detail in the first section of the chapter. The evolution of statistical machine translation (SMT) from simple IBM word-aligned models to state-of-the-art PBSMT and a comparison of the generative and discriminative models in word alignment also form a part of this section. A discussion on the evaluation of SMT models, mentioning the commonly used metrics for the purpose, concludes the first section of the chapter. In Section 2, different paradigms of MT are discussed, ranging from the conventional rule-based MT (RBMT) and example-based MT (EBMT) to the more recently developed hierarchical, tree-based models and hybrid methods. Section 4 briefly mentions different commonly used MT applications such as *online MT systems*, *translation memory tools*, *spoken language translation* and *sign language translation*. The final section of this chapter focusses on the different areas of MT research carried out by Prof. Way's group at Dublin City University, focussing primarily on the inclusion of syntactic and morphological information in the log-linear framework of SMT. The chapter concludes with a set of general future directions for MT research and a useful list of publications pertaining to the core components in MT. One of the strong points of this chapter is its coverage and organization which clearly presents the current state of MT research to the reader. The use of examples and diagrams is also helpful in understanding the core concepts, particularly for the beginners in the area. Although the competing

paradigms of MT (RBMT, EBMT, hierarchical SMT, etc.) are not explained in detail, the *Further Reading* section at the end of the chapter points the readers to the appropriate publications in the area. However, while the chapter mentions RBMT systems like OpenLogos (Barreiro et al. 2011) and Apertium (Forcada et al. 2011), it misses out on a few more recently successful EBMT/RBMT systems like Cunei (Phillips 2011) and CMU-EBMT (Brown 2011).

As mentioned in Chapter 19, statistical language modelling (SLM) plays a significant role in machine translation, especially in SMT. Apart from being used in MT, SLM is used in numerous NLP applications such as automatic speech recognition, spelling correction, etc. Chapter 3 by Ciprian Chelba introduces the basic hypotheses and techniques involved in the task of statistical language modelling. The chapter starts with the introduction of most commonly used SLM in NLP: the *n-gram* language model. It further explores the relationship of these models with Markov systems and introduces the concepts of *perplexity* and *entropy* used to measure the predictive power of language models. Later in the chapter, the author introduces a sophisticated paradigm of structured language models aimed at capturing the complex dependencies inherent in natural language based on probabilistic context-free grammar (PCFG). The chapter concludes by providing promising future directions in language modelling research and a brief mention of the issues of scalability and domain adaptation for language modelling. Overall, this chapter is a concise introduction to the different aspects in language modelling and its effects on specific NLP tasks like SMT. This chapter also mentions a number of generic language modelling toolkits like SRILM (Stolcke 2002), CMU toolkit (Clarkson and Rosenfeld 1997) and MITLM (Hsu and Glass 2008) useful for language modelling experiments. However IRSTLM (Federico et al. 2008), a commonly used language modelling toolkit, specialized for MT is not mentioned in the context.

Chapter 5 by Robert Malouf presents the highly influential machine learning technique of maximum entropy (MaxEnt). MaxEnt is a general-purpose machine learning method used for classification and prediction and has been widely used in different NLP tasks like sentence boundary detection, part-of-speech tagging, parse selection and, of course, MT. MaxEnt establishes the basis of log-linear models on which state-of-the-art SMT decoders such as *Moses* (Koehn et al. 2003), the most commonly used SMT decoder, are based. This chapter begins with the presentation of the *maximum entropy principle* and an account of parameter estimation techniques. Finally Malouf presents three different applications to highlight the effectiveness of this technique, one of which is a case in MT. The first example concerns with translation of French phrases to English and is essentially modelled as a classification problem rather than an MT task. Another application compares MaxEnt to another advanced modelling technique—hidden Markov models (HMM) for the task of part-of-speech (POS) tagging. Although POS-tagging is often used in RBMT this has not been explicitly mentioned in the *Handbook*.

In addition to SMT itself, the use of grammatical information present in the parallel texts used to train SMT systems has led to the development of competing paradigms in MT research. Some of these approaches rely heavily on parsing techniques to encode the lexical information. In NLP, parsing refers to a set of techniques and theories which allow automatic analysis of the underlying syntactic structures in a natural language

sentence. The relevance and complexity of this task makes it one of the most important and highly researched areas in both CL and NLP. Hence, two separate chapters have been dedicated to the theory and techniques of parsing in this book. Chapter 4 by Mark-Jan Nederhoff and Giorgio Satta introduces the formal foundations of parsing and how it relates to the problem of recognition and representation of the syntactic structures. Both lexicalized and non-lexicalized context-free grammars (CFGs) and the corresponding tabular parsing algorithms are discussed in significant depth, followed by the introduction of Probabilistic context-free (PCFG) parsing. In the later sections, dependency grammar and mildly context-sensitive tree-adjoining grammars (TAGs) and their associated parsing mechanisms are also discussed. The chapter concludes with a discussion of synchronous context-free grammars (SCFG) and their applicability in the context of MT. The focus of Chapter 13 by Stephen Clark is on the probabilistic analysis of sentences in a corpus using supervised learning mechanisms. Starting with generative parsing models like PCFG, the chapter goes into the analysis of the parsing, ranking and optimization algorithms, before moving on to discriminative models of parsing. A detailed description of combinatory categorical grammar (CCG) parsing and its efficiency and accuracy concludes the chapter. Both these chapters provide the reader with a balanced view of the theoretical and practical aspects of parsing. However, since parsing is heavily based on grammar formalisms, it might be better for beginners in the area to go through chapter 1 first. This chapter, titled *Formal Language Theory*, by Shuly Wintner, provides an ideal introduction into the formal classification of languages and their respective expressiveness. Starting from the simple regular languages, and moving on to the more complex context-free and mildly context-sensitive languages, this chapter covers the essential concepts in formal language theory.

Chapter 14 by John A. Goldsmith presents the concepts of *Segmentation* and *Morphology* widely considered to be the first steps in classical NLP. *Segmentation* refers to the task of converting raw text into a sequence of linguistically relevant units for further processing, while *Morphology* refers to the study of surface forms of words in natural language and to the information which could be extracted from such forms. The chapter starts with an introduction of different aspects of natural language, providing clear context and definitions for each. The author then looks into unsupervised approaches of word segmentation and methods of lexicon construction for both space-delimited and unsegmented languages. The latter part of the chapter covers the unsupervised learning of morphology and the use of finite-state transducers in morphological induction. For a discussion on the concepts of finite-state transducers, readers are referred to Chapter 1. Word segmentation, especially for graphically unsegmented languages (Chinese, Japanese, Korean, Arabic, etc.), forms an important part of pre-processing in nearly all forms of MT. Specifically for SMT and EBMT, prior tokenization of the training data is crucial for translation quality. While morphological information is inevitable for developing RBMT systems, EBMT systems often utilize such information for better template matching (Güvenir and Cicekli 1998). SMT uses morphological features as factors in *factored translation models* (Koehn and Hoang 2007) wherein translations operate on more general representations of words (like lemma or part-of-speech), instead of just word surface-forms.

In Chapter 6, Walter Daelemans and Antal van den Bosch review a widely used machine learning model known as *memory-based learning* (MBL). The basic idea is to measure the distance between feature vectors of stored training data and those of 'test' events using similarity metrics of different sorts. The variation in distance measures is further used to create classification of the input data. Modified and extended versions of MBL and their corresponding applications in specific types of NLP tasks like morphophonology, text analysis, dialogue and discourse are discussed in significant detail throughout the chapter. The EBMT paradigm is closely related to MBL, and indeed an implementation of EBMT based on MBL (van den Bosch et al. 2007) is also briefly described in this chapter. In SMT, MBL techniques have been used in domain adaptation to select supplementary training material from out-of-domain data 'similar' to in-domain training data (Wu et al. 2008).

The problem of *Natural Language Generation* (NLG) has been covered in Chapter 20 by Ehud Reiter. The chapter starts with a high-level perspective on the complexity of NLG, followed by the different kind of lexical, stylistic and organizational choices the system has to make throughout the process of NLG. Finally, some of the current NLG systems are summarized along with a discussion on the role of NLG in different NLP applications. Although NLG techniques are used for sentence generation in RBMT systems (Langkilde and Knight 1998), sentence generation for the purpose of MT is not explicitly covered in this chapter.

Some of the chapters in this book cover certain NLP techniques which are not directly related to MT but are often used for pre- and postprocessing of the training data for quality improvements or evaluation. Chapter 15, titled *Computational Semantics*, by Chris Fox, covers the theories of meaning and semantic representation of languages with a focus on corpus-based semantics and its applications. Chapter 18, titled *Information Extraction* by Ralph Grishman discusses the problems of extracting specific information from text, highlighting name, entity, relation and event extraction as the primary tasks in this area. Chapter 21, titled *Discourse Processing*, by Ruslan Mitkov, introduces the computational processing of discourse and the associated theories and methods for the task, focussing on discourse element extraction and anaphora resolution. While Chapters 15, 18, and 21 deal with very specific problems in NLP, a few of the remaining chapters address more general but extremely relevant issues in NLP. Chapter 10, titled *Linguistic Annotation*, by Martha Palmer and Nianwen Xu, discusses the issue of corpus annotation and its effects for supervised learning tasks and evaluation methods. One of the most important issues covered by the chapter, especially in the context of MT is the issue of inter-annotator agreement. SMT is often evaluated using reference translations produced by multiple human translators, thereby making the issue relevant to MT evaluation. Further information about general evaluation techniques are presented in Chapter 11, titled *Evaluation of NLP Systems* by Philip Resnik and Jimmy Lin. This chapter identifies the general difference between *intrinsic* evaluation, required to assess the performance of a particular process, and *extrinsic* evaluation, required to measure the effect of the same process in a larger engineering task, but does not cover the specific techniques used in the extrinsic or intrinsic evaluation of MT or MT components.

Apart from the chapters already discussed, this book also contains a few other chapters on specific application areas of NLP like *Speech Recognition* (Chapter 12

by Steve Renals and Thomas Hain), *Dialogue Management* (Chapter 16 by Jonathan Ginzburg and Raquel Fernández), *Computational Psycholinguistics* (Chapter 17 by Mathew W. Crocker) and *Question Answering* (Chapter 22 by Bonnie Webber and Nick Webb). Specific machine learning techniques like *Decision Trees* (Chapter 7 by Helmut Schmid), *Artificial Neural Networks* (Chapter 9 by James B. Henderson) and *Unsupervised Grammar Induction* (Chapter 8 by Alexander Clark and Shalom Lappin) are also included. In the first part of the book, Chapter 2 by Ian Pratt-Hartmann introduces the notion of computational complexity presenting a series of important complexity results for different language classes and NLP tasks.

Altogether, this *Handbook* covers a wide variety of topics in NLP and CL and, as discussed, is of particular use to researchers in the field of MT. On a more general note, graduate students or novice researchers can utilise this book as a comprehensive starting point for their area of interest within NLP or CL. Although some advanced modelling techniques (HMMs, conditional random fields) are not covered in detail, sufficient pointers have been provided throughout the book to more advanced material. In some chapters, outlines of specific algorithms, along with examples and, in some cases, detailed discussion on open-source toolkits also make the *Handbook* a handy reference when it comes to developing prototype systems. All in all, this is very well compiled book, which effectively balances the width and depth of theories and applications in two very diverse yet closely related fields of language research.

References

Barreiro A, Scott B, Kasper W, Kiefer B (2011) OpenLogos rule-based machine translation: philosophy, model, resources and customization. Mach Transl 25(2):107–126

Brown RD (2011) The CMU-EBMT machine translation system. Mach Transl 25(2):179–195

Clarkson P, Rosenfeld R (1997) Statistical language modeling using the CMU-Cambridge Toolkit. In: ESCA EUROSPEECH 1997, pp 2707–2710

Federico M, Bertoldi N, Cettolo M (2008) IRSTLM: an open source toolkit for handling large scale language models. In: Interspeech 2008: 9th annual conference of the international speech communication association, pp 1618–1621

Forcada ML, Ginestí-Rosell M, Nordfalk J, O'Regan J, Ortiz-Rojas S, Pérez-Ortiz JA, Sánchez-Martínez F, Ramírez-Sánchez G, Tyers FM (2011) Apertium: a free/open-source platform for rule-based machine translation. Mach Transl 25(2):127–144

Güvenir HA, Cicekli I (1998) Learning translation templates from examples. Inf Syst 23(6):353–363

Hsu B-JP, Glass J (2008) Iterative language model estimation: efficient data structure & algorithms. In: Interspeech, pp 841–844

Hutchins WJ (2003) Machine translation: general overview. In: Mitkov R (ed) The Oxford handbook of computational linguistics. Oxford University Press, Oxford, pp 501–511

Koehn P, Hoang H (2007) Factored translation models. In: Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL), pp 868–876

Koehn P, Och FJ, Marcu D (2003) Statistical phrase-based translation. In: HLT-NAACL 2003: conference combining Human Language Technology conference series and the North American Chapter of the Association for Computational Linguistics conference series, pp 48–54

Langkilde I, Knight K (1998) The practical value of N-grams in generation. In: Proceedings of the ninth international workshop on natural language generation, pp 248–255

Phillips A (2011) Cunei: open-source machine translation with relevance-based models of each translation instance. Mach Transl 25(2):1–17

Somers H (2003) Machine translation: latest developments. In: Mitkov R (ed) The Oxford handbook of computational linguistics. Oxford University Press, Oxford, pp 512–528

Stolcke A (2002) SRILM—an extensible language modeling toolkit. In: ICSLP 2002, Interspeech 2002: 7th international conference on spoken language processing, pp 901–904

van den Bosch A, Stroppa N, Way A (2007) A memory-based classification approach to marker-based EBMT. In: Proceedings of the METIS-II workshop on new approaches to machine translation, pp 63–72

Wu H, Wang H, Zong C (2008) Domain adaptation for statistical machine translation with domain dictionary and monolingual corpora. In: Proceedings of the 22nd international conference on computational linguistics, vol 1, COLING '08, pp 993–1000