Mach Translat (2011) 25:377–381
DOI 10.1007/s10590-011-9117-6

# Nitin Indurkhya and Fred J. Damerau (eds): Handbook of Natural Language Processing (second edition)

## CRC Press, Boca Raton, 2010, xxxiii+678 pp, Hardbound, ISBN 978-1-4200-8592-1

**Sandipan Dandapat**

Received: 5 September 2011 / Accepted: 12 October 2011 / Published online: 26 October 2011


The second edition of the *Handbook of Natural Language Processing* provides detailed coverage of the techniques and applications of current natural language processing (NLP). This edition has removed outdated material and upgrades and expands some of the chapters from the earlier version of the handbook (Dale et al. 2000). This handbook also covers some emerging areas of recent NLP, such as sentiment analysis, web distance and word similarity. This edition of the handbook was compiled by Nitin Indurkhya, a researcher at the University of New South Wales, and late text processing pioneer Fred J. Damerau of the IBM T.J. Watson Research Centre.

A review of this book has already been published by Jochen Leidner (2011) which focuses on the overview of the topics covered therein compared with other available related handbooks. In this review, I aim to provide a more detailed outline of the different research areas covered in the book, focusing particularly on the area of machine translation (MT). The book has 26 chapters in three different sections, namely, *Classical Approach* (Part I), *Empirical and Statistical Approach* (Part II) and *Applications* (Part III). The first part of the book has six chapters. The organization of these chapters is based on the chronological flow pattern of standard processing stages of NLP, typically found in pipelined rule-based MT (RBMT) architectures. The subsequent 17 chapters in the second part of the book follow a similar organizational pattern. However, in this review, I will primarily focus on the chapters directly related to MT. I will also cover MT-related topics detailed in other chapters of this book.

There are primarily two chapters about MT in this book, namely, *Statistical Machine Translation* (SMT) (Chapter 17), by Abraham Ittycheriah, and *Chinese Machine Translation* (Chapter 18), by Pascale Fung. Along with these two chapters dealing directly with MT, there is a separate chapter on *Alignment* (Chapter 16), by Dekai Wu.

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

Alignment is an essential technique at the heart of SMT systems. It is therefore convenient to read Chapter 16 first as it describes different strategies for bitext alignment in general. This chapter defines the concepts of alignment keeping in mind its uses and applications. The chapter focuses on alignment at three different levels: at sentence level; at character, word and phrase level; and finally at the level of tree structures. All the above alignment techniques are described in detail with the corresponding algorithms and examples. A strong point of Chapter 16 is that it describes the algorithms in detail with mathematical derivations and with many relevant examples.

Chapter 17 is the main chapter which deals with the different components of SMT systems. The chapter first gives a brief introduction to the SMT approach. It then describes the individual components of an SMT system, i.e. language models, parallel corpora, word alignment, different translation models (IBM model, phrase-based system, syntax-based MT using a hierarchical phrase-based MT system (*Hiero*) and direct translation model). Finally it touches on the search strategy used in SMT decoders. Given that MT is a huge area of NLP standing in its own right, this chapter is disappointingly short. People with specific interest to SMT will find a more detailed description in the textbook by Koehn (2010).

Chapter 18 describing Chinese MT is the last chapter pertaining to MT. Chinese is the most widely spoken language in the world and a lot of effort has been put into Chinese MT. Thus, this chapter will be of interest to many readers. Readers would be able to visualize the theoretical concepts described in Chapter 17 with real application scenarios in Chapter 18. A complete chapter devoted to the application perspective will equip readers with strategies to develop MT systems for other languages. First, the chapter talks about different pre-processing issues for Chinese, such as word segmentation, part-of-speech (POS) tagging and chunking. Here the author also describes maximum-entropy-based Chinese word segmentation and translation-driven word segmentation. The subsequent three sections briefly describe phrase-based SMT, example-based MT (EBMT) and RBMT for Chinese. In the next section, the author describes semantic-based SMT and the interlingua approach. Finally, the last section focuses on the different applications of Chinese MT, namely, term translation and named-entity translation, Chinese spoken language translation, and cross-language information retrieval (IR) using MT.

All in all, the area of SMT is broadly covered in the book. Though there is some detail missing in Chapter 17 when it talks about SMT approaches which might cause difficulty for beginners. The book does not cover RBMT and EBMT though there is still some work going on in these areas. Some such recent successful systems include Apertium (Forcada et al. 2011), Cunei (Phillips 2011), CMU-EBMT (Brown 2011), OpenLogos (Barreiro et al. 2011) and OpenMaTrEx (Dandapat et al. 2010) which are not mentioned in the book. Many hybrid platforms for MT also draw on RBMT and EBMT as well as SMT. Readers working in the area or those looking to research the area of RBMT and EBMT might be disappointed with the coverage of the book.

Other than on MT, the book covers a wide range of NLP approaches and applications. These primarily include the areas of text processing (Chapter 2), morphology (Chapter 3), POS tagging (Chapter 10), multiword expressions (MWEs) (Chapter 12), parsing (Chapters 4, 8 and 11) word sense disambiguation (WSD) (Chapter 14) and natural language generation (NLG, Chapters 6, 22 and 23). The above areas are

covered with significant detail in the book with different classical, empirical and statistical approaches. These topics are the backbone of any RBMT system and often of interest to those working in this area. Also, SMT frameworks use a large number of pre-/post-processing tools based on these components.

Text processing and morphological analysis are the first two steps of classical NLP. The book covers these topics in two separate chapters: *Text Preprocessing* (Chapter 2), by David D. Palmer, and *Lexical Analysis* (Chapter 3), by Andrew Hippisley. *Text Preprocessing* deals with converting a raw text into a well-defined sequence of linguistically meaningful units. This chapter focuses on the challenges of text processing (e.g. character set dependency, language dependency, corpus dependency and application dependency), tokenization (both space-delimited and unsegmented languages) and sentence segmentation. The approaches for Chinese and Japanese tokenization are given as an example of unsegmented language tokenization. The subsequent chapter (Chapter 3) talks about lexical analysis of text. This chapter focuses on computational morphology. First, the chapter introduces the finite state *morphonology*[1] with detailed examples. Starting with simple inflectional morphology, the author also describes difficulties in morphology (isomorphism and contiguity problem) and lexical analysis. Finally, the author discusses the well adopted paradigm-based approach to morphology. The chapter is very well organized; however, it uses some linguistic terms (e.g. *non-ablauting verbs*, *lemma's canonical form*) without defining them, which may cause a problem for beginners to this area. On the other hand, people with a linguistic background might have difficulties in understanding finite-state morphology (in Section 2) without being introduced to finite state automata (not described in the book).

Chapter 10, by Tunga Güngör, deals with issues and methodologies in POS tagging. The bulk of the chapter describes different approaches to POS tagging namely, rule-based tagging, statistical tagging (based on different graphical models) and combination tagging (an effective way of using different tagging techniques using co-training or voting strategies). Finally, the chapter describes language-specific tagging (Chinese and Korean) and morpho-syntactic tagging for highly inflected languages. POS tagging is an essential component to build a RBMT system. In addition, POS tagging is also used in SMT (e.g. source context modelling) and EBMT (to find closely-matching examples). The readers of this journal might be interested in reading the specific work on training a POS tagger to build an MT system (Sánchez-Martínez et al. 2008) which is absent from this book.

The concept of MWEs is introduced in Chapter 12, by Timothy Baldwin and Su Nam Kim. MWEs generally need special treatment during translation as the structure and meaning (semantics) of MWEs can not be derived directly from their component words. A detailed linguistic description on the properties of MWEs is given in the chapter. It summarizes research issues on identification and extraction of MWEs from text without providing a detailed account of existing computational methodologies.

WSD is described in Chapter 14, by David Yarowsky. WSD is widely used in different IR and MT techniques which are described in Section 3 of the chapter. The

[1] The more widely used term for this is *morphophonology* which refers to the interaction of word formation (morphology) with the sound system of language (phonology).

chapter starts with some early approaches to WSD which is followed by different supervised machine learning approaches for WSD and feature selection for the same. Some lightly supervised (e.g. WSD via word-class disambiguation, hierarchical class models and iterative bootstrapping) and unsupervised algorithms (e.g. hierarchical clustering and randomize algorithm *Buckshot*) are also described in the chapter.

The book has three chapters on parsing, namely, *Syntactic Parsing* (Chapter 4), by Peter Ljunglöf and Mats Wirén, *Treebank Annotation* (Chapter 8), by Eva Hajičová et al., and *Statistical Parsing* (Chapter 11), by Joakim Nivre. The *Syntactic Parsing* part presents the classical approaches to parsing and presents a wide range of techniques that can be used to parse a natural-language sentence. The chapter on *Treebank Annotation* describes the syntactic annotation of corpora, which is used for statistical parsing. Readers will benefit more by reading the parsing chapters as they are ordered in the book, so that they will read the work on treebank annotation (Chapter 8) before statistical parsing (Chapter 11). Chapter 11 describes the techniques for statistical parsing. First the chapter briefly introduces the concepts of statistical parsing in terms of syntactic representation with examples from constituent structure and dependency structure, statistical models and evaluation metrics for parsing. In the next section, the author defines the framework of probabilistic context free grammars with its use in statistical parsing models and the algorithms for learning probabilistic grammar rules and estimating parse structure. The subsequent two sections focus on two different learning methods for statistical parsing: generative models and discriminative models using supervised machine learning techniques. It also introduces semi-supervised and unsupervised learning techniques for statistical parsing. Altogether, the chapter broadly covers the different approaches of statistical parsing, supplies a significant number of references for further reading and provides references to different past and ongoing research directions of statistical parsing. As a result, the chapter makes very good reading for students beginning in their research on statistical parsing. However, undergraduate students (with no background in NLP) might face some difficulty, especially students with no background in machine learning.

There are three chapters in this book which deal with NLG. One of the chapters is included in the section about classical approaches to NLP, namely Chapter 6 (*Natural Language Generation*), by David D. McDonald, which mainly surveys works in the field of NLG. The other two chapters on NLG are in the applications section. These two chapters are *Report Generation* (Chapter 22), by Leo Wanner, and *Emerging Applications of Natural Language Generation in Information Visualization*, *Education and Health Care* (Chapter 23), by Barbara Di Eugenio and Nancy L. Green. Chapter 6 starts by pointing out the differences between NLG and natural language comprehension. Then the chapter focuses on different components for language generation. Furthermore, it describes different representations required to produce fluent multisentential and multi-paragraph text, built around linguistic realisation. Readers of this journal might be more interested in NLG in the context of MT, especially sentence generation in a rule-based system or in the use of a language model in an SMT system for fluent translation. However, this book does not address the issue of sentence generation in the context of MT. Chapter 23 shows three applications of NLG. Starting with the techniques used for multimedia presentation generation, the chapter illustrates the issues of a multimedia presentation generation system by showing an example from

different systems over the past two decades. Another application of NLG is shown in Chapter 22 which is *Report Generation*. The main focus here lies in the process of generating reports from input (usually structured data—numerical time series or formal content) to text planning in order to produce a more informative report. The chapter also briefly describes linguistic realisation of the report generation. Finally, 10 different sample report generators are illustrated.

Apart from the aforementioned chapters, the handbook also includes chapters for foundational background (e.g. *Corpus Creation* (Chapter 7), by Richard Xiao, and *Fundamental Statistical Techniques* (Chapter 9), by Tong Zhang). One chapter in the book also touches upon the overview of modern speech recognition (Chapter 15). The book also has separate chapters on IR (Chapter 19), information extraction (Chapter 21), question answering (Chapter 20), ontology construction (Chapter 24) and biomedical NLP (Chapter 25).

All in all, the handbook covers the wide area of NLP and its applications. This will essentially help researchers and graduate students to access starting-point material for a particular area of interest. The handbook also covers the associated algorithms with examples which will help to develop prototype systems. Other than some minor typographic errors,[2] the handbook is error-free. All together, this is a high quality compilation of up-to-date theories and applications of NLP.

## References

Barreiro A, Scott B, Kasper W, Kiefer B (2011) OpenLogos 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

Dale R, Moisl H, Somers H (2000) Handbook of natural language processing (1st edn.). Marcel Dekker, New York

Dandapat S, Forcada ML, Groves D, Penkale S, Tinsley J, Way A (2010) OpenMaTrEx: a free/open-source marker-driven example-based machine translation system. In: Proceedings of the 7th international conference on natural language processing (IceTAL 2010), Reykjavík, Iceland, pp 121–126

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

Koehn P (2010) Statistical machine translation. Cambridge University Press, Cambridge

Leidner JL (2011) Handbook of natural language processing (second edition) Nitin Indurkhya and Fred J. Damerau (editors) (University of New South Wales; IBM Thomas J. Watson Research Centre) Boca Raton, FL: CRC Press, 2010, xxxiii+678 pp; hardbound, ISBN 978-1-4200-8592-1, $99.95. Comput Linguist 37(2):395–397

Phillips AB (2011) Cunei: open-source machine translation with relevance-based models of each translation instance. Mach Transl 25(2):161–177

Sánchez-Martínez F, Pérez-Ortiz JA, Forcada ML (2008) Using target-language information to train part-of-speech taggers for machine translation. Mach Transl 39(1–2):29–66

[2] A list can be found in http://www.computing.dcu.ie/~sdandapat/typos.html.