Theory and Applications of Natural Language Processing

Aims and Scope

The field of Natural Language Processing (NLP) has expanded explosively over the past decade: growing bodies of available data, novel fields of applications, emerging areas and new connections to neighboring fields have all led to increasing output and to diversification of research.

"Theory and Applications of Natural Language Processing" is a series of volumes dedicated to selected topics in NLP and Language Technology. It focuses on the most recent advances in all areas of the computational modeling and processing of speech and text across languages and domains. Due to the rapid pace of development, the diversity of approaches and application scenarios are scattered in an ever-growing mass of conference proceedings, making entry into the field difficult for both students and potential users. Volumes in the series facilitate this first step and can be used as a teaching aid, advanced-level information resource or a point of reference.

The series encourages the submission of research monographs, contributed volumes and surveys, lecture notes and textbooks covering research frontiers on all relevant topics, offering a platform for the rapid publication of cutting-edge research as well as for comprehensive monographs that cover the full range of research on specific problem areas.

The topics include applications of NLP techniques to gain insights into the use and functioning of language, as well as the use of language technology in applications that enable communication, knowledge management and discovery such as natural language generation, information retrieval, question-answering, machine translation, localization and related fields.

The books are available in printed and electronic (e-book) form:

* Downloadable on your PC, e-reader or iPad
* Enhanced by Electronic Supplementary Material, such as algorithms, demonstrations, software, images and videos
* Available online within an extensive network of academic and corporate R&D libraries worldwide
* Never out of print thanks to innovative print-on-demand services
* Competitively priced print editions for eBook customers thanks to MyCopy service http://www.springer.com/librarians/e-content/mycopy

For further volumes:
http://www.springer.com/series/8899

Chris Biemann

Structure Discovery in Natural Language

Foreword by Antal van den Bosch

Springer

Chris Biemann
Computer Science Department
Technische Universität Darmstadt
Hochschulstr. 10
64289 Darmstadt
Germany

Foreword by
Antal van den Bosch
Centre for Language Studies
Radboud University Nijmegen
P.O. Box 9103
6500 HD Nijmegen
The Netherlands

ISSN 2192-032X e-ISSN 2192-0338
ISBN 978-3-642-25922-7 e-ISBN 978-3-642-25923-4
DOI 10.1007/978-3-642-25923-4
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2011944202

Mathematics Subject Classification (2010): 68T50, 91F20, 05C82, 62H30, 68T05

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Dedicated to my parents, who unsupervised me in a good way

Foreword

Few topics in computational linguistics catch the imagination more than unsupervised language learning. 'Unsupervised' has a magical ring to it. Who would not want to have a system do all the work automatically, and knowledge-free? I suspect it is a common experience of researchers in the field, typically occurring while staring out of a window taking a sip of coffee, to feel a sudden exhilaration: seeing the possibility of circumventing the vexing bottleneck of annotated data gathering with an unsupervised learning algorithm and a lot of unannotated data.

Let us pause here for a bit of exegesis. What does it mean for computational language learning to be *unsupervised*? As just suggested, it involves learning some yet unspecified language processing task on the basis of *unannotated* linguistic data. This in turn begs the question what it means for linguistic data to be unannotated; a reasonable answer would be that unannotated data only consists of linguistic surface elements: sounds or letters, and is devoid of any abstract linguistic elements. This means that any linguistic theory that assumes abstract linguistic elements, be it part-of-speech tags, syllabic roots, or syntactic dependencies, will not be playing any role in unsupervised language learning. This is quite a provocative proposition, and another cause for the rebellious allure of unsupervised learning.

The word *unsupervised*, to continue my exegesis, strengthens the anti-authoritarian connotation even more, but also triggers the question what or whose supervision is thrown overboard. Are we talking about the poor linguist whose wise lessons are ignored? This indeed seems to be the answer suggested by this book, where within a few pages you will read that "Unsupervised means that no labelled training material is provided as input. The machine is exposed to language only, without being told what its output should look like." (this book, p. 2). This is linked to the concept of *knowledge-free*, which is taken to mean "that no knowledge about the specific language, such as e.g. word order constraints or a list of personal pronouns, is given to the system." (this book, p. 2).

Thus, unsupervised language learning is learning from unlabeled data, where labels denote abstract linguistic notions. This aligns well with the parallel meaning attributed to unsupervised learning in the field of machine learning: learning without classification labels. But is the distinction between abstract and surface linguistic elements always clear-cut? Counter to the usual assumption, I argue it is not. Consider, for example, the function word *the*. What is the difference of saying this word carries a syntactic function, and saying that the word *is* a syntactic function? In *the book*, the word *the* marks the beginning of a noun phrase. In *we book*, the word *we* signals a main verb coming up that refers to the first person plural.

If we accept the proposition that linguistic elements such as letters and words can be seen as labels themselves, it is easy to see that supervised machine learning algorithms could be applied to classify strings of linguistic elements into other elements or strings of elements. Learning is still knowledge-free and devoid of linguistic abstractions, but is it unsupervised? If it is not unsupervised in the machine learning sense, can it still be called unsupervised in the no-linguistic-abstractions sense? I would like to disagree, and instead call this type of learning *autosupervised*.

Autosupervised language learning does not occur in an obscure corner of pathological language task definitions. Rather, it is the type of learning that occurs in most present-day statistical text-to-text processing systems: machine translation, paraphrasing, and spelling correction. Even n-gram language models with back-off smoothing can be considered to be the product of a simple self-supervised learning procedure producing a decision list or tree that predicts the next word given the previous $n-1$ words.

Allow me to continue, at least in this Foreword, the use of the word *autosupervised* where normally you would have read *unsupervised*.

Our understanding and command of autosupervised language learning, though as a scientific endeavour still rather young and perhaps just out of its infancy, has deepened over the past two decades. Its potential has in fact become one of the key research questions in computational linguistics. We can now build on the shoulders of pioneers such as Hinrich Schütze, Steve Finch and Nick Chater, Ramon Ferrer-i-Cancho, Chris Manning, Dan Klein, and Alex Clark, and before you lies an important next step in this increasing body of work.

While in computational linguistics the topic is close to reaching the grail status that topics such as machine translation have, it is stereotypical for the field to be largely oblivious to theories from other fields that work with the same idea of data-driven discovery of models of language. These are not *computational* models, but the articles and books in which they are described provide a wealth of inspiration, also in hindsight, for the development of computational models of autosupervised language learning. Fortunately, the better work in the area does acknowledge its roots in 20th-century linguistics, with proponents such as J.R. Firth and Zellig Harris, and occasionally also points to the work of present-day usage-based linguists such as Robert Langacker, William Croft, and Adele Goldberg, and the developmental psychologist Michael Tomasello.

It is important to realize that the boundary between non-computational models of autosupervised language learning and their computational counterparts is thin, and could become void if both sides would work with the same concepts and formalizations. Usage-based linguistic theories have been occasionally criticized for not being entirely formal, but I am convinced it is only a matter of time before this gap will be bridged, thanks to work from both sides of the divide in overlap areas such as corpus linguistics. On the one side one finds studies such as the collostructional analysis work of Stefanowitsch and Gries [229]. On the other side one finds the type of work in autosupervised language learning exemplified by this book.

What would be the formal basis that would connect and equalize work from both sides of the divide? Autosupervised language learning methods have tended to build on light formalizations that make use of simple spaces and metrics (vector spaces, bags of words, *n*-gram models). Despite their sobering simplicity and their complete implicitness, these models are known to harbour the incredible strength that fuels the world's leading search engines, speech recognizers and machine translation systems. Usage-based linguistics, on the other hand, assumes structures (with names such as constructions, collostructions, complex lexical items) of mildly higher complexity: they may have gaps, required and optional elements, and relations between these elements that signal inequality (e.g. dependence). Where bags of words and the Markovian assumption have no answer to these requirements, graph theory does. The book you are now reading describes the building of a machine that starts with this assumption.

This book is built around the concept of an autosupervised structure discovery machine that discovers structure in language data, and can do so iteratively, so that it can discover structure in structured data. It is shown to grasp language identification, part-of-speech tagging, and lexical substitution to levels that rival supervised approaches. To leave sufficient suspense, I trust you will be thrilled to read how.

Nijmegen, September 2011 *Antal van den Bosch*

Preface

After 60 years of attempts to implement natural language competence in machines, there is still no automatic language processing system that comes even close to human language performance.

The fields of Computational Linguistics and Natural Language Processing predominantly sought to teach machines a variety of subtasks of language understanding either by explicitly stating processing rules or by providing annotations they should learn to reproduce. In contrast to this, *human* language acquisition largely happens in an unsupervised way — the mere exposure to numerous language samples triggers acquisition processes that imprint the generalisation and abstraction needed for understanding and speaking that language.

Exactly this strategy is pursued in this work: rather than telling machines how to process language, one instructs them how to discover structural regularities in text corpora. Shifting the workload from specifying rule-based systems or manually annotating text to creating processes that employ and utilise structure in language, one builds an inventory of mechanisms that — once they have been verified on a number of datasets and applications — are universal in a way that allows their application to unseen data with similar structure. This enormous alleviation of what is called the "acquisition bottleneck of language processing" gives rise to a unified treatment of language data and provides accelerated access to this part of our cultural memory.

Now that computing power and storage capacities have reached a sufficient level for this undertaking, we for the first time find ourselves able to leave the bulk of the work to machines and to overcome data sparseness by simply processing larger batches of data.

In Chapter 1, the *Structure Discovery* paradigm for Natural Language Processing is introduced. This is a framework for learning structural regularities from large samples of text data, and for making these regularities explicit by introducing them in the data via self-annotation. In contrast to the predominant paradigms, Structure Discovery involves neither language-specific knowledge nor supervision and is therefore independent of language, domain and data representation. Working in this paradigm instead means establishing procedures that operate on raw language material and iteratively enrich the data by using the annotations of previously applied Structure Discovery processes. Structure Discovery is motivated and justified by discussing this paradigm along Chomsky's levels of adequacy for linguistic theories. Further, the vision of the complete Structure Discovery Machine is outlined: a series of processes that make it possible to analyse language data by proceeding from the generic to the specific. Here, abstractions of previous processes are used to discover and annotate even higher abstractions. Aiming solely at identifying structure, the effectiveness of these processes is judged by their utility for other processes that access their annotations and by measuring their contribution in application-based settings. A data-driven approach is also advocated on the side of defining these applications, proposing crowdsourcing and user logs as means to widen the data acquisition bottleneck.

Since graphs are used as a natural and intuitive representation for language data in this work, Chapter 2 provides basic definitions of graph theory. As graphs based on natural language data often exhibit scale-free degree distributions and the Small World property, a number of random graph models that also produce these characteristics are reviewed and contrasted along global properties of their generated graphs. These include power-law exponents approximating the degree distributions, average shortest path length, clustering coefficient and transitivity.

When defining discovery procedures for language data, it is crucial to be aware of quantitative language universals. In Chapter 3, Zipf's law and other quantitative distributions following power laws are measured for text corpora of different languages. The notion of word co-occurrence leads to co-occurrence graphs, which belong to the class of scale-free Small World networks. The examination of their characteristics and their comparison to the random graph models as discussed in Chapter 2 reveals that none of the existing models can produce graphs with degree distributions found in word co-occurrence networks.

For this a generative model is needed, which accounts for the property of language being a time-linear sequence of symbols, among other things. Since previous random text models fail to explain a number of characteristics and distributions of natural language, a new random text model is developed, which introduces the notion of sentences in a random text and generates sequences of words with a higher probability, the more often they have been generated before. A comparison with natural language text reveals that this model successfully explains a number of distributions and local word order restrictions in a fully emergent way. Also, the co-occurrence graphs of its random corpora comply with the characteristics of their natural language counterparts. Due to its simplicity, it provides a plausible explanation for the origin of these language universals without assuming any notion of syntax or semantics.

In order to discover structure in an unsupervised way, language items have to be related via similarity measures. Clustering methods serve as a means to group them into clusters, which realises abstraction and generalisation. Chapter 4 reviews clustering in general and graph clustering in particular. A new algorithm, Chinese Whispers graph partitioning, is described and evaluated in detail. At the cost of being non-deterministic and formally not converging, this randomised and parameter-free algorithm is very efficient and particularly suited for Small World graphs. This allows its application to graphs of several million vertices and edges, which is intractable for most other graph clustering algorithms. Chinese Whispers is parameter-free and finds the number of parts on its own, making brittle tuning obsolete. Modifications for quasi-determinism and possibilities for obtaining a hierarchical clustering instead of a flat partition are discussed and exemplified. Throughout this work, Chinese Whispers is used to solve a number of language processing tasks.

Chapters 5–7 constitute the practical part of this work: Structure Discovery processes for Natural Language Processing using graph representations.

First, a solution for sorting multilingual corpora into monolingual parts is presented in Chapter 5, involving the partitioning of a multilingual word co-occurrence graph. The method has shown to be robust against a skewed distribution of the sizes of monolingual parts and is able to distinguish between all but the most similar language pairs. Performance levels comparable to trained language identification are obtained without providing training material or a preset number of involved languages.

In Chapter 6, an unsupervised part-of-speech tagger is constructed, which induces word classes from a text corpus and uses these categories to assign word classes to all tokens in the text. In contrast to previous attempts, the method introduced here is capable of building significantly larger lexicons, which results in higher text coverage and therefore more consistent tagging. The tagger is evaluated against manually tagged corpora and tested in an application-based way. The results of these experiments suggest that the benefits of using this unsupervised tagger or a traditional supervised tagger are equal for most applications, rendering unnecessary the tremendous annotation efforts involved in creating a tagger for a new language or domain.

The problem of word sense ambiguity is discussed in detail in Chapter 7. A Structure Discovery process is set up, which is used as a feature to successfully improve a supervised word sense disambiguation system. On this basis, a high-precision system for automatically providing lexical substitutions is constructed.

The conclusion in Chapter 8 may be summarised as follows: Unsupervised and knowledge-free Natural Language Processing in the Structure Discovery paradigm has proven to be successful and capable of producing a processing quality equal to that of conventional systems, assuming that sufficient raw text can be provided for the target language or domain. It is therefore not only a viable alternative for languages with scarce annotated resources, but also overcomes the acquisition bottleneck of language processing for new tasks and applications.

Darmstadt, November 2011 *Chris Biemann*

Contents

Acronyms

Lists of abbreviations used frequently in this volume:

BA	Barabási-Albert Model: a process to generate scale-free graphs
BNC	British National Corpus: a collection of 100 million tokens of British English from different genres
CL	Computational Linguistics: the research area of building linguistic applications with computers
CRF	Conditional Random Field: a supervised machine learning classifier, commonly used for sequence tagging
CW	Chinese Whispers graph clustering algorithm, introduced in this work
DM	Dorogovtsev-Mendes Model: a process to generate scale-free small world graphs with two power law regimes
EP	Entropy Precision: a measure to compare clusterings
ER	Erdős-Rényi Model: a random graph model
F	F-measure: The harmonic mean between precision P and recall R
LCC	Leipzig Corpora Collection: a collection of plain text corpora of standardized size for a large number of languages
LDA	Latent Dirichlet Allocation: a generative clustering algorithm of the topic model family
LSA	Latent Semantic Analysis: a vector space transformation method based on Singular Value Decomposition
MCL	Markov Chain Clustering: a graph clustering method based on random walks
MFS	Most Frequent Sense: strategy of assigning the most frequent sense in WSD, commonly used as a baseline system
MI	Mutual Information: a measure for the dependence between random variables
NER	Named Entity Recognition: the task of finding names in natural language text
NLP	Natural Language Processing: the research area of building systems that can process natural language material

nMI	normalised Mutual Information: the normalised variant of MI
OOV	Out Of Vocabulary rate: the percentage of tokens in a text not known to the model
P	Precision: the number of correct answers divided by the number of total answers
POS	Parts Of Speech: syntactic word classes like verb, noun, pronoun
R	Recall: the number of correct answers given divided by the number of correct answers possible
SD	Structure Discovery: the process of finding regularities in data and annotating them back into the data for later processes
SDM	Structure Discovery Machine: a stack of Structure Discovery processes
ST	Steyvers-Tenenbaum model: a process to generate scale-free small world graphs
SVD	Singular Value Decomposition: a matrix factorisation
SWG	Small World Graph: a graph with a high clustering coefficient and short average path lengths
TWSI	Turk bootstrap Word Sense Inventory: an alternative word sense inventory based on crowdsourcing
WS	Watts-Strogatz-model: a process to generate small world graphs
WSI	Word Sense Induction: task of identifying different meanings of a word
WSD	Word Sense Disambiguation: the assignment of one out of several possible word meanings for a word in context