Unsupervised and Knowledge-free Natural Language Processing in the Structure Discovery Paradigm

Dissertation Summary
November 19, 2007

Christian Biemann
University of Leipzig

Abstract

After almost 60 years of attempts to implement natural language competence on 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 the machine a variety of subtasks of language understanding either by explicitly stating processing rules or by providing annotations the machine 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 learn 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 being verified on a number of datasets and applications – are universal in a way that allows their execution on 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 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 lan-

guage, domain and data representation. Working in this paradigm rather means to set up discovery 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 sketched: A series of processes that allow analysing language data by proceeding from the generic to the specific. At this, abstractions of previous processes are used to discover and annotate even higher abstractions. Aiming solely to identify 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.

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 built from 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 among other things a time-linear sequence of symbols. 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 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.

For discovering 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 cost of being non-deterministic and formally not converging, this randomised and parameter-free algorithm is very efficient and especially 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.

Chapter 5 constitutes the practical part of this work: Structure Discovery processes for Natural Language Processing using graph representations.

First, a solution for sorting apart multilingual corpora into monolingual parts is presented, 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.

Next, 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. Results of these experiments suggest that the benefit of using this unsupervised tagger or a traditional supervised tagger is equal for most applications, rendering unnecessary the tremendous annotation efforts for creating a tagger for a new language or domain.

A Structure Discovery process for word sense induction and disambiguation is briefly discussed and illustrated.

The conclusion in Chapter 6 is summarised as follows: Unsupervised and knowledge-free Natural Language Processing in the Structure Discovery paradigm has shown to be successful and capable of producing a processing quality equal to systems that are built in a traditional way, if just 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 might also overcome the acquisition bottleneck of language processing for new tasks and applications.