"Automatic Extension of Feature-based Semantic Lexicons via Bootstrapping Algorithms"

by
Richard Socher

in Partial Fulfillment of the Requirements for the Degree of
Bachelor of Science in Computer Science

at the University Leipzig
Faculty of Mathematics and Computer Science
April 23, 2007

Thesis Supervisors:
Prof. Dr. Gerhard Heyer and Dipl.-Inf. Christian Biemann

Abstract

This work investigates and improves a bootstrapping approach which permits to extend high quality lexical resources with the help of large corpora. The emphasis lies on the extraction of lexical-semantic information and word meaning, which are fundamental components for advanced applications such as semantic parsing, information retrieval or summarizing textual information. ... In the last chapter an outlook with suggestions for further improvements and extensions is given and a novel approach which combines genetic algorithms and bootstrapping is outlined.

Acknowledgements

Thanks to Prof. ...

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

Leipzig, April 23, 2007 _____

Richard Socher

Contents

List of Figures

List of Tables

Chapter 1.

Introduction

A Word that breathes distinctly
Has not the power to die
Cohesive as the Spirit
It may expire if He -
'Made Flesh and dwelt among us'
Could condescension be
Like this consent of Language
This loved Philology.
- *Emily Dickinson* -

Where is the information I need in this huge pile of data? This question will be asked more and more in our information society and the only way to solve it on a large scale is to process the data automatically. Since most often this data is unstructured and available only in natural language, we need to understand the inner structure of natural languages and employ the tools of computer science to effectively extract the information we need.

For this task to be successful a good lexicon is needed. Many fields of advanced natural language processing, such as information retrieval (IR), word sense disambiguation or semantic web applications are based on lexical information. Since the manual creation of high quality lexical resources is very expensive and time consuming, there is a need for automatic or semi-automatic tools to create these. As of now no large high-quality lexical-semantic resources are available for German.

1.1. Corpus Linguistics

1.1.1. Definition

Corpus linguistics is a sub domain of computer linguistics and thus both share the same goals. The main difference lies in the statistical methods of the former which are mostly quantitative linguistic analyses of real world text, so called corpora. A corpus is a text in written or spoken form. In this setting it is available in an electronic form.

1.2. General Definition of a Lexicon

A lexicon is a database that contains orthographic words and certain corresponding information for each word. Depending on the type of lexicon these can be related to the: morphology, syntax or semantics of the word. One word entry can have different part of speech (POS) tags, the word *intimate* for example can be an adjective or a verb. Depending on this POS tag, inflectional information can be different. One syntactically specified word can again represent different semantic concepts. The word *school*, which is a noun could represent an institution or a building for example.

Besides these three domains, a lexicon can also save information extracted from corpora, such as the frequency of an orthographical word, or of its word senses. Huge semantic lexicons are a fundamental basic for many advanced fields in computer linguistics.

... The following figure shows three semantic sorts that inherit their features from left to right: *con-potag* $\supset$ *animate-object* $\supset$ *animal-object*:

$$\begin{bmatrix} \textbf{con-} & \textbf{potag} \\ SORT & d \\ AXIAL & + \\ GEOGR & - \\ INFO & - \\ POTAG & + \end{bmatrix}$$

Figure 1.1.: Features of the semantic sorts *con-potag* $\supset$ *animate-object* $\supset$ *animal-object*

Chapter 2.

Bootstrapping Algorithm

The term bootstrapping is used in numerous scientific fields such as biology, law, electronics, statistics and linguistics. In NLP it refers to a process where few available information (often in the form of seed words) is used as a basis to create more information with the help of a corpus. The emphasis in this approach lies on the extraction of lexical-semantic information and word meaning, based on the *Distributional Hypothesis* and the conclusion that semantic similarity is a function over global contexts. In other words similar words appear in similar contexts. In the presented experiments nouns are classified through their modifying adjectives or the verbs whose object or subject the noun is.

Section 2.1 outlines other works in the area of bootstrapping for semantic lexicons. ... They are explained in 2.2.

2.1. Other bootstrapping methods for the acquisition of semantic lexicons

Early bootstrapping works such as [**?**] were based merely on a couple of seed words and the simple context of one word left and one word right of the seed noun. This approach found new nouns in the given seed word category. The results were sorted by hand.

...

2.2. Pre-processing Steps for Input Data

Scheme 2.1 shows the discussed pre-processing steps for a better overview. Basically the process starts with a flat corpus which is semantically parsed. Then a tool from the Fernuniversität Hagen is applied to extract certain relations. For each relation the following actions are taken:

1. create pairs in the form: *bootword* - *noun* (see section **??** for explanation)
2. delete personal pronouns
3. extract significant co-occurrences

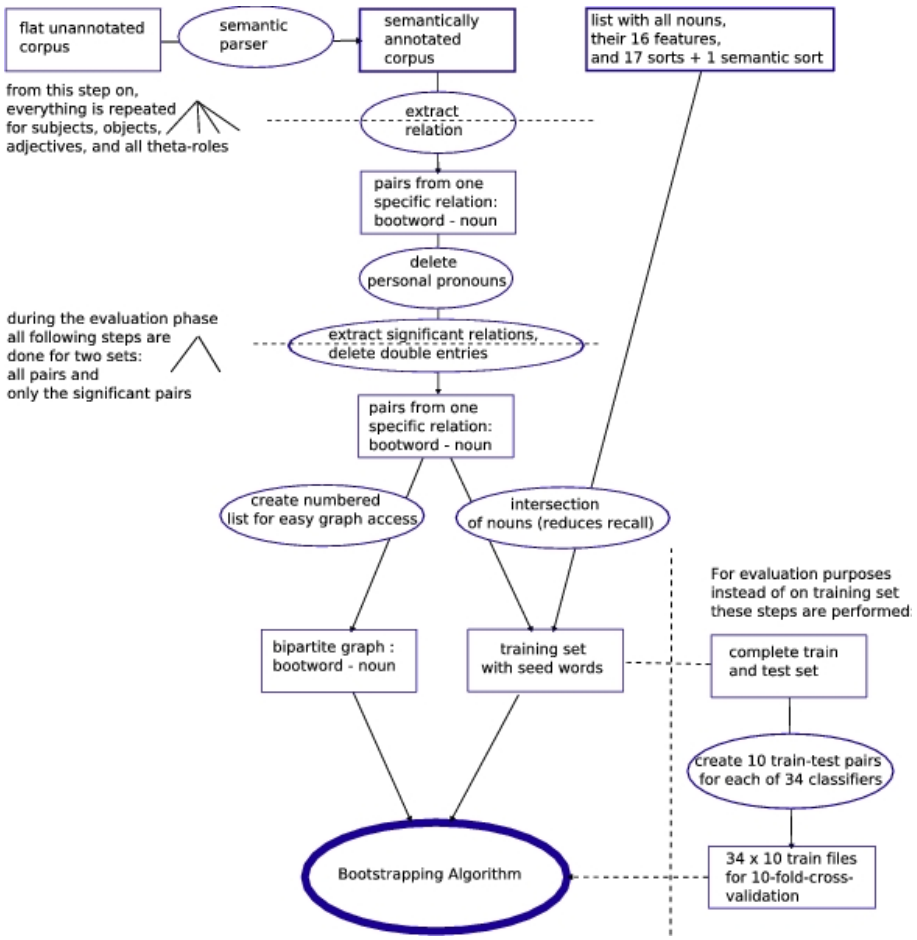

Figure 2.1.: Scheme with pre-processing steps for bootstrapping

4. a set with all co-occurrences (which can appear twice) is saved and one with only unique and significant co-occurrences, for each set the following measures are taken:

   a) one final input of the algorithm is a numbered bipartite graph with the co-occurrences, the numbers are just for easy access via a library

   b) an intersection with the noun database is performed, that means in the training set, only these nouns appear that are actually inside of the test or train set

   c) For each of the 34 semantic types (semantic sort, features and ontological sorts) train and test pairs are created for the bootstrapping process utilizes only one type at the time:

      i. based on the 10-fold-cross-validation method 10 pairs for training and testing are created (see below for explanation)

      ii. each of these pairs together with the graph of the co-occurrences forms the input for the bootstrapping process.

... To illustrate how the extraction from the annotated corpus is performed an example query for the tool netarx from the Fernuniversität Hagen is shown. Netarx is the tool which extracts specific elements from semantic nets. Queries are written in a Prolog style language. The variable x_1 is a specialization of the verb of the sentence, a certain situation. The method of this situation is x_2. If this method is described by a noun, the verb and this noun are printed out: ...

```
'((subs ?x1 ?verb) (meth ?x1 ?x2) ((sub pred) ?x2 ?noun)
(cat ?noun n))' '?verb ?noun'
```

Figure 2.2.: A netarx query for the theta-role *method*

Fortunately there are no problems with polysemous nouns since the morphological parser also disambiguated between the different readings of a word and added a number for each meaning.

One real example is shown in figure 2.3, notice that words are in their basic form and numbered with their corresponding meaning.

```
"kompromißlos.1.1" "einhandkatamarane.1.1"
"akzeptabel.1.1" "lebensdauer.1.1"
"bekannt.1.1" "polit-punk-band.3.1"
"eigen.1.1" "angabe.1.1"
"multipel.1.1" "orgasmus.1.1"
"entscheidend.1.1" "charakter.1.1"
"besonder.1.1" "platz.1.1"
"öffentlich.1.1" "telephonnetz.1.1"
"norwegisch.1.1" "popband.3.1"
```

Figure 2.3.: Example with extracted adjective-noun pairs, showing numbers for each possible polysemous meaning

...

For bigrams the table 2.1 can be used to find the significance level based on the $-2\log\lambda$ value. $P(correct\ rejection\ of\ H_0)$ refers to the probability of a correct rejection of the null hypothesis which states that the bigram occurred by chance.

Log-Likelihood Value	1.32	1.64	2.07	2.71	3.84	5.02	5.41	6.63	7.88	9.14	10.83
$P(correct\ rejection\ of\ H_0)$	0.75	0.80	0.85	0.9	0.95	0.975	0.98	0.99	0.995	0.9975	0.999
$P(false\ rejection\ of\ H_0)$	0.25	0.20	0.15	0.10	0.05	0.025	0.02	0.01	0.005	0.0025	0.001

Table 2.1.: Table for finding the significance level based on log-likelihood ratio

Chapter 3.

Experiments, Results and Interpretation

This chapter covers all experiments on the algorithm. Numerous runs with different parameters are performed to gain the best results. Each type of relation is separately analyzed and compared with other types.

... To get a picture of the consequences of a bias that is too high, table 3.1 also shows precision for the smaller positive class (labelled with 'Precision:+'). Because the feature *spatial* occurs more often positively than negatively the smaller negative class is used for calculation. The overall mean of smaller positive classes is at 73.4%. The mean in the table does not consider the size of the classes, but instead uses the precision values of all features evenly. Overall the precision for positive classes is high enough to use the results, though some features are not suitable for this method. For *method* for example precision is only at 1.0, because the recall is 0, thus no new nouns were falsely assigned the positive class.

Characteristic	Number+	Number-	Bias △	Precision	Precision:+	Recall	F-score
spatial-	4629	4333	0.5165	0.8609	0.8393	0.3625	0.5904
animal	408	8540	0.9544	0.9786	0.7948	0.5234	0.7587
geogr	262	8686	0.9707	0.9406	0.1549	0.4825	0.7145
method	24	8924	0.9973	0.9955	1.0000	0.5391	0.7764
Mean	-	-	0.8092	0.9055	0.5310	0.4352	0.6642

Table 3.1.: Adjectives: features, distribution, bias and bootstrapping results

With the exception of *artif* the precision is above 82%. Notice also that the maximum f-score value is at 0.77. ...

Chapter 4.

Improvements and a Novel Approach

This chapter introduces several improvements of this method. Some are based on the used lexicon, others are generally applicable on other lexicons as well. Furthermore suggestions for practical use and a novel approach of genetic meta-bootstrapping are provided.

4.1. Combination of Characteristics to Complex Semantic Sorts

Any lexicon that defines complex semantic sorts or classes can be extended with the presented method. However the used method is not yet optimal.

Chapter 5.

Conclusion

This thesis explores a statistical bootstrapping approach for learning semantic characteristics for nouns. In particular novel verbal relations have been explored. These outperform previous experiments that bootstrapped solely on adjective-noun co-occurrences. Relations between verbs and their nominal object show especially high quality results with precision above 95% and recall at about 50%. If outcomes of bootstrapping on different relations are combined by only using nouns which were assigned the same class in the separate runs, the results can be seen as almost certain, with a precision of 99% for all binary semantic characteristics. This way recall values of the different characteristics vary between 20% and 40%. Thus some semantic characteristics can be assigned for up to 50,000 new nouns. ...

Appendix A.

Extra Tables

Characteristic	Prec. adj	Prec. obj	Prec. subj	Recall adj	Recall obj	Recall subj	Prec. 3Same	Prec. 3Same pos.	Prec. 2Same	Prec. 2Same pos.	Recall 3Same	Recall 2Same
human:+	0.9205	0.9633	0.9588	0.4154	0.4707	0.4233	0.9979	0.9952	0.9820	0.9492	0.2553	0.4520
geogr:-	0.9406	0.9525	0.9382	0.4825	0.5098	0.4704	0.9907	0.5000	0.9744	0.4176	0.3446	0.5010
spatial:+	0.8611	0.9814	0.9628	0.3625	0.4703	0.4357	0.9992	0.9985	0.9855	0.9858	0.2223	0.4417
legper:+	0.9154	0.9657	0.9631	0.4103	0.4716	0.4258	0.9986	0.9955	0.9834	0.9512	0.2519	0.4526
sort:d+	0.8513	0.9516	0.9343	0.3584	0.4532	0.4186	0.9903	0.9880	0.9667	0.9651	0.2113	0.4268
sort:na-	0.9792	0.9791	0.9882	0.5306	0.5370	0.5045	0.9888	1.0000	0.9831	1.0000	0.4144	0.5295

Table A.1.: Combining results of the three main relations, parameter combination: 2-5

Bibliography

[AAA05] *Proceedings of the 20th National Conference on Artificial Intelligence*, Pittsburgh, 2005.

[BO05] C. Biemann and R. Osswald. Automatische erweiterung eines semantikbasierten lexikons durch bootstrapping auf großen korpora. In B. Schröder B. Fisseni, H.-C. Schmitz and P. Wagner, editors, *Sprachtechnologie, mobile Kommunikation und linguistische Ressourcen Beiträge zur GLDV-Tagung 2005 in Bonn*, pages 15–27, Frankfurt am Main, 2005. Peter Lang.

[Cur04] J. Curran. *From Distributional to Semantic Similarity*. PhD thesis, Edinburgh, 2004.

[Eng88] U. Engel. *Deutsche Grammatik*. Julius Groos Verlag, Heidelberg, 1988.

[Fil68] Charles J. Fillmore. *Universals in Linguistic Theory*, chapter The case for case, pages 1–90. Holt, Rinehart & Winston, New York, 1968.

[KW01] Claudia Kunze and Andreas Wagner. Anwendungsperspektiven des GermaNet, eines lexikalisch-semantischen Netzes für das Deutsche. In Bernhard Schröder Ingrid Lemberg and Angelika Storrer, editors, *Chancen und Perspektiven computergestützter Lexikographie*, volume 107 of *Lexicographica Series Maior*, pages 229–246. Niemeyer, Tübingen, Germany, 2001.

[MBF+90] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. Miller. Five papers on WordNet. *Special Issue of the International Journal of Lexicography*, 1990.

[PL05] Sebastian Padó and Mirella Lapata. Cross-lingual bootstrapping for semantic lexicons: The case of framenet. In *Proceedings of the 20th National Conference on Artificial Intelligence* [AAA05], pages 1087–1092.