Corpus-Based Extension of Semantic Lexicons in Large Scale

Dimitrios Kokkinakis, Maria Toporowska Gronostaj and Karin Warmenius
Språkdata, Göteborg University
Box 200, SE-405 30, Sweden
{First.Last}@svenska.gu.se

Abstract

During recent years there has been an increased interest to acquire or extend, on a large-scale, high-quality semantic lexicons. The methodology is usually corpus-driven. It is based on the (re-)use of machine readable resources of various types, and the application of cost effective ways to eliminate the acquistion bottleneck, i.e. derivational morphology, customization of off-the-shelf resources, statistical techniques and shallow parsing. This paper investigates how, and to what extent the flexibility and robustness of a partial parser can be utilized to fully automatically achieve this goal. Our work is based on the observation that members of a semantic group are often surrounded by other members of the same group in text. Given a few category members we use parsed corpora to collect surrounding contexts and try to identify other words that also belong to the same group.

1. Introduction

During recent years there has been an increased interest to use corpus-driven approaches to acquire high quality semantic lexicons on large scale: (Grefenstette (1994); Dorr & Jones (1996); Hearst and Schütze (1996); Takunaga *et al.* (1998); Lin (1998)). This paper investigates the use of a cost-effective way to eliminate the acquisition bottleneck by exploiting the flexibility and robustness of a system based on a partial parser. The parser uses fine-grained syntactic contexts for identifying similarities between words and acquire large quantities of high quality general purpose semantic knowledge. Given a few category members of a semantic group, we investigate whether it is possible to collect appropriate surrounding contexts and identify other words, on a large scale, that also belong to the same semantic group.

Our point of departure is not to acquire the semantic lexicons from scratch, rather, to build on what we have already at our disposal. That is, lexical resources of high quality, manually produced and verified but quantitatively *insufficient* for realistic large-scale tasks. Therefore, we focus our attention to explore and exploit inexpensive methods to progressively enrich the resources with several thousands of new, classified lexical units.

Our work is based on the observation that members of a semantic group are often surrounded by other members of the same group throughout a corpus. By a semantic group, we understand here for instance enumerative and conjunctive phrases of the form: *xa, xb, ..., xc* or *xa and xb, ..., and xc*, where *x* can be any content-poor item, such as determiners and numerals, and *a b c* nouns or names. We further slightly constrain this general observation by searching for particular types of phrases, of particular length and of particular semantic content provided by the available, limited semantic resources. These resources, then, are progressively enriched and applied in a bootstrapping manner back to the phrases extracted from the corpora in order to classify as many as possible of the words that are members of the retrieved phrases. The level of the fine-grained syntactic analysis is made possible through the use of a robust parser developed by Abney (1997) in which Kokkinakis & Johansson Kokkinakis (1999) have developed a large coverage grammar for written Swedish. The semantic lexicons we refer to are the Swedish SIMPLE lexicon (*Semantic Information for Multifunctional Plurilingual Lexica*) and gazeteers of person, location and organization names. Previous experiments in a small scale for Swedish (Kokkinakis *et al.* (2000)) have demonstrated that the task of enriching semantic resources using syntactic information is feasible. Therefore we wanted to investigate to what magnitude this can be done and evaluate, at least for some of the semantic groups, the quality of the acquired semantic units.

2. Related Research

Context similarity plays and important role in word acquisition. The use of syntax for generating

semantic knowledge and ways of measuring semantic similarity based on distributional evidence and syntagmatic relations have been put forward in the literature by many researchers. A common characteristic of almost all approaches is the computation of the semantic similarity between two words on the basis of the extent to which words' average contexts of use overlap.

Our method has similarities to the work by Hearst (1992), who uses lexico-syntactic patterns of the form: 'NP {, NP}* {,} and other NP' for the extraction of hyponymic relations, such as: '*...temples, treasuries, and other important civic buildings'*. However, more influential source of inspiration has been the work described by Grefenstette (1994). He examined an approach to extract corpus-specific semantics using a system, SEXTANT, which processes a text by tagging, partially parsing and creating dependencies between words in phrases extracted. The dependency relations are considered as attributes of the SEXTANT, and they are compared using a weighted Jaccard similarity measure (i.e. Count(attributes shared by x and y)/Count(attributes processed by x or y)) in order to discover words used in a similar manner. A result from this process was a list of similar words for each word in the corpus.

When it comes to the acquisition and extension of name lists, for the benefits of tasks such as named entity recognition, similar approaches are applicable. Stevenson and Gaizauskas (2000), for instance, build categorised lists of names from manually annotated training data, combining various types of filters. The authors claim that a performance measure of 87% f-score on a standard data set is achieved using these corpus-derived lists.

3. Resources

We apply a method, described in a more detail in the next section, uniformally onto two semantic lexicons. The first is the Swedish SIMPLE lexicon, developed within the EU-financed project with the same name. The content and design of the SIMPLE model, applied in 12 European languages, is documented in Lenci *et al.* (1998). The notion of semantic type is central for the SIMPLE model and its ontology. Information on semantic class, domain, argument structure of predicative expressions etc., constitute a relevant part of the semantic type specification.

The Swedish lexicon provides descriptions for 10,000 semantic units (roughly 6,000 words), comprising 7,000 nouns, 2,000 verbs and 1,000 adjectives; this paper will elaborate on the noun part of the lexicon. As a vital part of the different entries' semantic unit is the notion of semantic class whose value is an element in a semantic class list (95 classes) hierarchically structured, e.g. ANIMAL and BUILDING. Ambiguous entries are also denoted as such. For instance, *glas* 'glass' is marked with the classes: AMOUNT, CONTAINER, MATERIAL and UNIT-OF-MEASUREMENT. The second lexicon is a list of frequent proper names: PERSON (4900), LOCATION (4300) and ORGANIZATION (1300). These originate from a previous work in the framework of creating an information extraction system for Swedish (in the EU-financed project AVENTINUS).

4. Methodology

We have experimented with a corpus-driven approach, using a cascaded finite-state syntactic parser (CASS-SWE), based on work done by Kokkinakis & Johansson Kokkinakis (1999), which seems a plausible way of progressively enriching the semantic resources. An advantage of CASS-SWE is its ability to identify, with high accuracy, arbitrarily complex nominal and other types of phrases, a property that we consider here as crucial for aiding the identification of new semantic entries. The method rests on the assumption that words entering into the same syntagmatic relation with other words are perceived as semantically similar. Essentially the approach is as follows:

1. Gather, part-of-speech annotate and parse large corpora (in our case using CASS-SWE, a parser that uses part-of-speech annotated input);
2. From the resulted analyzed forest of chunks, filter out long noun phrases;
3. Filter out knowledge-poor elements, such as determiners and punctuation; and use the lemmatised and normalised content of the extracted phrases;
4. First Pass:

    Measure the overlap between the members of the phrases extracted and the entries in the SIMPLE lexicon and the gazeteers;

If conditions apply, add new categorised entries in the database;
Repeat the previous two steps, until very few or nothing can be matched;

5. Second Pass:
   Apply compound segmentation on the members of the phrases left;
   Check whether they are lexicalised using a defining dictionary, do not use them if they are;
   Repeat the process from step (4) by matching this time the heads with the content of the database;

The bootstrapping mechanism dynamically grows the original lists, so that each iteration produced a larger semantic dictionary.

4.1 Corpora and Part-of-Speech Annotation

The corpora we used consisted of over 42 million tokens. Most of the material was provided by the Swedish Language Bank[1]. We part-of-speech tagged the corpora using Brill's tagger (Brill (1992)) trained on Swedish material, using a very fine-grained tagset[2]. For instance, the noun *jurister* 'lawyers' receives by the tagger the description NCUPNI, which is interpreted as a *common noun, non-neuter, in plural form, nominative case* and *indefinite form*. Note also that a pre-tagger filter recognises a large number of multi-word expressions and compound names of the form 'Los Angeles' and 'Dow Jones'.

4.2 Parsing and Grammar Rules

The parsing process is using CASS-SWE, a flexible parser for Swedish, in which *levels* or *bundles* of rules of very special characteristics and content can be rapidly created. From the already encoded rules in CASS-SWE we extracted two subsets. One, having common nouns (63) and one proper nouns (45) respectively in their Right-Hand-Side. The only requirement we posed was that each extracted phrase should contain at least three members of each respective phrasal group. Knowledge-poor items such as conjunctions and determiners, which are not specific to any category and are common across all phrases, were removed. Phrases containing adjectives such as *andra/annan* 'other' were excluded, since the noun following (oftenly) signalls a higher in the hierarchy concept. As in the example: *skor, tröjor och andra produkter* 'shoes, blouses and other products'. Normalization was performed by using the base form of every common noun in the phrases. The rule subsets were then applied on the corpus. From the large forest of chunks produced, a large number of phrases for each category was extracted. The amount of unique retrieved phrases for the first group were 35,955 and for the second 71,636. Examples of the rules are given below. For clarity, the names of the tags on the RHS have been edited for readability:

Example of Common Noun Rule (F stands for punctuation):

'Rule-CN --> DETERMINER? COM-NOUN (COM-NOUN F)* COM-NOUN CONJ COM-NOUN'

e.g.: *färger, penslar, papper och matsäckar* 'colours, brushes, paper(s) and lunch-boxes'

Example of Proper Noun Rule:

'Rule-NP --> APPOSITION-NOUN? PROP-NOUN+ (F PROP-NOUN)+ CONJ PROP-NOUN+'

e.g.: *Venezuela, Trinidad och Island*

The retrieved phrases could be easily recognised since each level of rules can be indexed with a unique identifier. There is also the possibility to generate the results in a linear format having as content only the part of the phrases we are interested to retrieve and ignore other syntactic, irrelevant in this case, annotations produced by the parser.

4.3 First Pass Overlap

The way we measure the overlap between the members of the phrases extracted and the entries in the SIMPLE and the gazeteers is simply by matching a database with the content of the resources against the content of the phrases. We assume that if at least two of the members of a phrase (a figure arbitrarily taken) are also entries in the lexicon, with the *same* semantic class, and the rest of the phrase members *have not* received a semantic annotation, then there is a strong indication that the

1. The material consists of four are newspaper collections (*press95, press96, press97, press98*) and a collection of contemporary novels (*romii*). For more information on the material visit: http://spraakdata.gu.se/lb.
2. A slightly simplified variant of the tagset can be found in: http://spraakdata.gu.se/lb/parole/.

rest of the members are co-hyponyms, and thus semantically similar with the two already encoded in the lexicon. Accordingly, we annotate them with the same semantic class. For instance the common nouns: *jurist* 'lawyer', *optiker* 'optician' and *läkare* '(medical) doctor' have being manually coded in the (original) SIMPLE lexicon, with the OCCUPATION-AGENT semantic class (*individuals or groups of humans identified according to a role in professional, social or religious disciplines*). Thus, in the extracted noun phrase: *jurister, läkare, optiker, psykologer och sjukgymnaster* (after lemmatisation and removal of the knowledge poor items, the conjunction *och* 'and' and punctuation) the three first nouns will get the OCCUPATION-AGENT label, while the two last, namely 'psychologist' and 'physiotherapist' will also get the same label by the system. This is because they satisfy the condition stated earlier, namely that they have not received a semantic class annotation and the rest of the members of the phrase (at least two) have been assigned the same semantic class.

In case where original members of the lexicon are ambiguous in the same way, that is, they receive same labels, then a new word matched will also receive the ambiguous labelling. For instance, in the phrase: *flaskor, tallrikar, vinglas* 'bottles, plates, wine glasses' the last word is not matched by the lexicon, however the first two are assigned the classes CONTAINER and AMOUNT. Accordingly, the word *vinglas* will be assigned the two semantic classes.

The new items are inserted in the database and the process is repeated from step (4) until nothing else can be matched in the remaining phrases, or ambiguity, multiple, different classes for the members of a phrase, prohibits the continuation of the process. For example in the case of a phrase such as *barn, kvinnor, husdjur och möbler* 'children, women, pets and furniture' nothing will be entered in the database. Since, according to the SIMPLE, *barn* and *kvinnor* will be assigned the class BIO (*classification of human beings according to biological charactersitics, like age, sex, etc.*) and *möbler* the class FURNITURE. Therefore, the unknown to the lexicon word *husdjur* is prohibited from obtaining a semantic class, since two different classes appear within a single phrase.

4.4 Second Pass Overlap

After we tested the method outlined so far, we discovered that a large number of phrases were not used by the system since none or only one of the members of the phrases was covered by the lexicons, either the original or the enriched version. Therefore, we found it compelling to devise a way to deal with these cases by taking account the compounding characteristic of the Swedish language (proper nouns were not treated on the second pass). Over 27,000 common noun phrases were not matched while in 35% of these all content was annotated but no match could be obtained.

The fact now that over 70%, or approximately 80,000, of all the entries in the SAOL (1998) are compound forms casts light onto the need to design effective tools for compound segmentation, as new, casual compounds are created constantly in Swedish. We assume that a considerable number of casual or on the fly created compounds can inherit relevant parts of semantic information provided on their heads by the SIMPLE lexicon and thus, can be easily incorporated in it. In order to restrain automatic incorporation of lexicalised compounds with idiomatic, metaphoric or metonymic meaning, we check whether a compound is included as a separate entry in a defining dictionary (lexicalised). For this purpose we used the GLDB/SO (http://spraakdata.gu.se/lb/gldb.html). If this is the case, the compound is not subjected to automatic inheritance.

Compound segmentation involves identifying grapheme combinations that are not-permitted in non-compound forms in the language, which carry information of potential token boundaries. The heuristic principle for the segmentation is based on producing short *n-gram* character sequences from hundreds of non-compound lemmas, and then generating n-grams that are not part of the lists produced. After manual adjustments and iterative refinement a list of such graphemes has been produced and used for segmentation. Ambiguities are unavoidable, although the heuristic segmentation has been evaluated for high precision; we do not force the system to overgenerate spurious decomposition points.

Examples of n-gram sequences include: ivb, iv|b, e.g. *skriv|bord* 'writing desk' and ngss, ngs|s, e.g. *forsknings|skola* 'research school'; '|' denotes where the segmentation should take place, while *bord* and *skola* are heads on the previous compounds. We apply heuristic decomposition on the members of the phrases left, and run the process from step (4) once again. This time by matching the content of the enriched database with the compounds' heads of the segmented strings (if any) in the remaining

phrases. For instance, in the phrase: *färjor, kryssnings|fartyg, tankers och ro-ro-|fartyg* 'ferries, cruise liners, tankers and ro-ro-vessels' no classes are assignd during the first pass, while during segmentation the second and fourth words' heads get the label VEHICLE since they match the entry *fartyg* 'vessel', and the rest two are matched with the same class since they satisfy the condition stated earlier (no other classes involved and at least two belong to the same one).

5. Evaluation

We will discuss now the results we obtained by applying the previously outlined method on large corpora, both in terms of quantity and quality. Table 1 summarizes the results with respect to the quantity aspects. The first pass was repeated six times. During the second pass the material in the remaining phrases was segmented and the enriched content of the database was matched against the heads of the segmented members of the phrases. This time resulting in fewer entries in the database. This can be explained by the fact that we are rather restrictive during segmentation.

	Original	Pass-1	Pass-2	Total
Common Nouns	2,921	5,110	1,100	9,131
Proper Nouns	10,550	25,700	---	36,250

Table 1. Quantitative acquisition results

Class	Original	New	Wrong/Spurious	Precision
FLOWER	19	26	3	88,5%
PHENOMEN.	36	29	9	69%
ORGANISAT.[NE]	1,300	395	22	94,4%
BIO	46	107	12	88,8%
IDEO	17	74	9	87,8%
VEHICLE	33	118	17	85,6%
APPARATUS	22	27	2	92,6%
GARMENT	25	184	19	89,7%
ILLNESS	38	66	8	87,9%

Table 2. Qualitative acquisition results

The most obvious way to evaluate the results of our technique is by using a gold standard (a human-compiled collection of related words). Since general-purpose thesauristic resources for Swedish are non-existent (there is a current effort to develop a Swedish WordNet at the university of Lund) and we do not have access to machine-readable versions of synonym dictionaries, we carried out two evaluations in another manner. First, we performed a manually qualitative evaluation for a number of semantic groups, based solely on our common sense and judgement, table (2). Precision was simply calculated as the ratio of valid entries to the total produced. Second, we tested a number of words based on the information found in synonym dictionaries for Swedish (Walter (1991); Strömberg (1998)). Two such words were *bil* 'car' and *rederi* 'shipping company', which according to the two dictionaries had 7+8=11 unique and 3+4=6 unique synonyms respectively. We then looked at the classes these words belonged to (according to SIMPLE), namely VEHICLE and AGENCY, to see whether the synonyms in the paper dictionaries occurred in that semantic class. For the word *bil* 8 of the synonyms occurred and 3 did not (*vagn, kärra, åk*); while for *rederi* 1 occurred and 5 did not (*fartygsbolag, linje, skeppsägare, båtbolag, sjöfartsbolag*). Varying figures, from no matches at all, to all matches, were found for a number of other words. Both methods have drawbacks, but seemed to be the closest we can come with respect to quality evaluation, this way, of course, we can (presumably) only evaluate precision. The general conclusion, that we can only partly stipulate on, is that existing synonym dictionaries cover a number of infrequent, sometimes "old-fashioned", terms and seemed not to be up-dated with the contemporary language style. Something that can only be achieved by processing large electronic corpora, which does not seem to be the case for the dictionaries consulted.

5.1 Error Analysis

There were four basic sources of erroneous entries identified. Part-of-speech and lemmatisation

errors; a number of long, enumerative noun phrases with many unknown to the lexicon entries, where two or three (happened) to correctly get the same semantic label, but few the wrong one. This caused the undesired effect of introducing new entries with wrong labels. For instance, in the phrase: *tröja halsduk strumpa underkläder skiva album* 'sweater scarf sock underwear record album' the entries *tröja* and *strumpa* received the label GARMENT the rest no labels, and consequently *halsduk* and *underkläder* achieved the correct label (GARMENT) while *skiva* and *album* the wrong one. A final source of error was polysemy, which also exhibited similar effect as the previous one, and also prohibiting the incorporation of new entries. For instance, in the phrase: *depression ångest spänning* 'depression anxiety excitement' after the first iteration the two first words received by the lexicon the label EMOTION. The third one was also labelled EMOTION (according the discussion in Section 4.3), which is correct according to the specific subsense of that word in that context. Once received that label, later processing of phrases, where *spänning* has another sense, such as: *tryck#*ATTRIBUTE *spänning#*EMOTION? *vibration tyngdkraft#*ATTRIBUTE 'pressure tension vibration gravitation', the already received annotation for *spänning* causes a phrase as this one to be rejected for further processing since two different labels are involved in the phrase and *vibration* cannot get a correct label (*spänning* ought to have the ATTRIBUTE class in this context).

6. Contribution and Further Work

We have presented a simple, quite efficient method to acquire general-purpose semantic knowledge from large corpora. Our main contributions of this paper are: the use of partially parsed corpora for extending semantic lexicons, the application of a unified way to process compounds, while infrequent words are not a major headache as in other (statistically-based) approaches. Both parsing and compounding are of equal importance; through parsing we allow the incorporation of new, mainly non-compound words, through compounding we allow new compounds of existing words. Regarding further work, we have to devise a better way to evaluate the results and decrease the amount of spurious generated entries. Actually, most of them originate from part-of-speech errors and not so much from the competence of the grammar. We will also continue the work in augmenting the rest of the SIMPLE's vocabulary. Lack of semantic resources in electronic form, such as large ontologies for Swedish, prohibits us to make more solid evaluation. The future release of the Swedish WordNet will be considered for such evaluation.

References

Abney S. 1997. Part-of-Speech Tagging and Partial Parsing. *Corpus-Based Methods in Language and Speech Processing*, Young S. and Bloothooft G., eds, Chap. 4, pp. 118 136, Kluwer AP

Brill E. 1992. A Simple Rule-Based Part of Speech Tagger. *3rd Conference on Applied Natural Language Processing (ANLP)*, Trento, Italy

Dorr B. and Jones D. 1996. Acquisition of Semantic Lexicons: Using Word Sense Disambiguation to Improve Precision. *SIGLEX Workshop: "Breadth and Depth of Semantic Lexicons"*, pp. 42-50, Santa Cruz, USA

Grefenstette G. 1994. *Explorations in Automatic Thesaurus Discovery.* Kluwer AP

Hearst M.A. 1992. Automatic Acquisition of Hyponyms from Large Text Copora. *14th COLING*, pp. 539-545, Nantes, France

Hearst M.A. and Schütze H. 1996. Customizing a Lexicon to Better Suit a Computational Task. *Corpus Processing for Lexical Acquistion*, pp. 77-94, Boguraev B. and Pustejovsky J. (eds.). MIT Press

Kokkinakis D. and Johansson Kokkinakis S. 1999. A Cascaded Finite-State Parser for Syntactic Analysis of Swedish. *9th EACL*, pp. 245-248, Bergen, Norway

Kokkinakis D., Toporowska-Gronostaj M. and Warmenius K. 2000. Annotating, Disambiguating \& Automatically Extending the Coverage of the Swedish SIMPLE Lexicon. *2nd Language Resources and Evaluation Conference (LREC).* Athens, Hellas

Lenci A. *et al*. 1998. *SIMPLE WP2, Linguistics Specifications*. Deliverable 2.1, Pisa

Lin D. 1998. Automatic Retrieval and Clustering of Similar Words. *COLING-ACL98*, Montreal, Canada

Sanfilippo A. *et al*. 1999. *Preliminary Recommendations on Lexical Semantic Encoding*. EAGLES LE3-4244, Draft version

SAOL 1998. *Svenska Akademiens Ordlista över Svenska Språket* (The Swedish Academy Word-List). Norstedts & Svenska Akademien

Stevenson M. and Gaizauskas R. 2000. Using Corpus-derived Name Lists for Named Entity Recognition. *6th Conference on Applied Natural Language Processing and First Conference of the North American Chap-*

ter of the Association for Computational Linguistics, pp. 290-296. Seattle
Strömberg A. 1998. *Stora synonymordboken*. Strömbergs Bokförlag AB, Falköping
Takunaga T., Fujii A., Iwayama M., Sakurai N. and Tanaka H. 1997. Extending a Thesaurus by Classifying Words. *"Automatic Information Extraction and Building of Lexical Semantic Resources" Workshop*, Vossen P. (*et al.*) (eds), pp. 16-21, Spain
Walter G. 1991. *Bonniers synonymordbok*. Bonniers