SENSEVAL-2: The Swedish Framework

Dimitrios KOKKINAKIS
Språkdata, Göteborg University
Box 200, SE-405 30
Göteborg, Sweden
Dimitrios.Kokkinakis@svenska.gu.se

Jerker JÄRBORG
Språkdata, Göteborg University
Box 200, SE-405 30
Göteborg, Sweden
Jerker.Jaerborg@svenska.gu.se

Yvonne CEDERHOLM
Språkdata, Göteborg University
Box 200, SE-405 30
Göteborg, Sweden
Yvonne.Cederholm@svenska.gu.se

Abstract

In this paper we describe the organisation and results of the SENSEVAL-2 exercise for Swedish. We present some of the experiences we gained by participating as developers and organisers in the exercise. We particularly focus on the choice of the lexical and corpus material, the annotation process, the scoring scheme, the motivations for choosing the lexical-sample branch of the exercise, the participating systems and the official results.

Introduction

Word sense ambiguity is a potential source for errors in human language technology applications, such as Machine Translation, and it is considered as *the* great open problem at the lexical level of Natural Language Processing (NLP). There are, however, several computer programs for automatically determining which sense of a word is being used in a given context, according to a variety of semantic, or defining dictionaries as demonstrated in the SENSEVAL-1 exercise; (Kilgarriff and Palmer, 2000). The purpose of SENSEVAL is to be able to say which programs and methods perform better, which worse, which words, or varieties of language, present particular problems to which programs; when modifications improve performance of systems, and how much and what combinations of modifications are optimal. Specifically for Swedish, we would also like to investigate to what extent sense disambiguation can be accomplished and the potential resources available for the task. We would thus be creating a framework that can be shared both within the exercise and for future evaluation exercises of similar kind, national and international.

1 Choice of Task

Three tasks were identified for SENSEVAL-2, namely: *the lexical-sample*, *the all-words* and *the 'in a system'* tasks. In the lexical sample task, first, we sample the lexicon, then we find instances in context of the sample words and the evaluation is carried out on the sampled instances. In the all-word task a system will be evaluated on its disambiguation performance on every word in the test collection. Finally, in the third type of task, a word sense disambiguation (WSD) system is evaluated on how well it improves the performance of a NL system (MT, IR etc). The reasons we chose the lexical-sample task for Swedish are summarised below:

1. Cost-effectiveness of annotation: it is easier and quicker for the human annotators to sense-tag multiple occurrences of one word at a time, particularly when robust interactive means are utilized (Section 3);
2. The lexical-sample reduces the work of preparing training data since only a subset of the sense inventory is used;
3. More systems can/could (eventually) participate;
4. The all-words task requires access to a full dictionary, which is problematic from the copyright point of view, since industrial partners were also allowed to participate; and, as Kilgarriff and Palmer (2000) noted:
5. Provided that the sample is well chosen, the lexical sample strategy would be more informative about the current strengths and failings of sense disambiguation research than the all-words task.

2 Development Process

In this section we will give a concise description of how the whole exercise (for Swedish) was set up, putting more emphasis on some of the main ingredients of the work, i.e. sampling, resources, annotation and scoring.

A number of likely participants were invited to express their interest and participate in the Swedish SENSEVAL (summer, 2000). A plan for selecting the evaluation material was agreed in Språkdata, and human annotators were set on the task of generating the training and testing material. The material was released to the participants at the end of April 2001 and during the second week of June, 2001 the results were returned for scoring. The Swedish SENSEVAL material was divided into three parts and released in stages:

- **Trial data**: freezing and showing the data formatting conventions (lexicon & corpus);
- **Training data**: the finalised sense inventory and portion of the 'gold standard';
- **Evaluation data**: the rest of the 'gold standard', untagged.

2.1 Dictionary and Corpus

At least three lexical resources were candidates for the Swedish lexicon-sample task. These were the Swedish versions of the WordNet (http://www.ling.lu.se/projects/Swordnet) and the Swedish SIMPLE (http://spraakdata.gu.se/simple/), as well as the Gothenburg Lexical Data Base/semantic Database (GLDB/SDB) (http://spraakdata.gu.se/lb/gldb.html). We chose the GLDB/SDB. The creation of a Swedish version of WordNet, a resource that is extensively used for the semantic annotation of texts in other languages, is under development and had (up to that point) limited coverage, while the SIMPLE lexicon, although available, has limited coverage (in principle it could be used and it is linked to the GLDB/SDB). However, a drawback of the Swedish SIMPLE is that very fine-grained sub-senses are not adequately described (or not described at all) in the material. GLDB/SDB is a generic defining dictionary of 65,000 lemmas available and developed at our department and became the final choice for the lexical inventory (see Allén, 1999[1981] for a description of the model utilized in the dictionary).

For the textual material we chose the Stockholm-Umeå Corpus (SUC), Ejerhed *et al.* (1992). The particular corpus was chosen for three main reasons. It is available to the research community; it is considered the "standard, reference" corpus for contemporary written Swedish; and, third, it is the corpus utilised in the SemTag project (next section).

2.2 Sampling

There is no standard method for sampling the lexical data. However, certain features were considered. These were: frequency, polysemy, part-of-speech and distribution of senses. Words were chosen based not so much on intuition, but rather on their frequency and polysemy. Still, it was hard to find a balance between these two features since high frequency words tend to be monosemous in a corpus, while highly polysemous words tend to have few senses in a corpus. In the case that a word was frequent and polysemous we tried to provide more data (context), than for words that were less frequent. Part-of-speech information was consulted for the decision of choosing more nouns in the sample (highest portion in the GLDB/SDB), than verbs (less than nouns, but more than adjectives in the GLDB/SDB) and adjectives (which are fewer than nouns and verbs in GLDB/SDB). We chose a sample of words where the amount of senses was evenly distributed, i.e. lemmas (dictionary entries) with 2-7 lexemes (senses) and 1-23 cycles (subsenses).

2.3 SemTag

Creating a sense-annotated reference corpus is a laborious task. Therefore, we developed the majority of the test and reference material within an ongoing project highly relevant for our mission, namely SemTag (*Lexikalisk betydelse och användningsbetydelse* – "Lexical Sense and Sense in Context", financed by the *Swedish Council for Research in the Humanities and Social Sciences* (HSFR)); see Järborg (1999). In brief, the purpose of the project is to create a large sample of sense-annotated corpus (several hundreds of thousands of words), which can be used among other things for:

- measuring the performance of automatic methods for WSD;

- testing, in practice and on a large scale, the validity of the lemma-lexeme model implemented in GLDB/SDB;
- the improvement of lexicographic descriptions, and the production of (new and) more fine-grained senses in GLDB/SDB;
- the adjustment of the definitions in GLDB/SDB to better fit the textual use;
- describing new words, not covered by the content of the GLDB/SDB;
- producing material, adequate for training supervised methods to sense disambiguation.

2.4 Corpus/Sense Inventory

Table 1 shows information on the sense inventory, the amount of corpus instances (training/testing) and the distribution of senses and sub-senses (Lexemes/Cycles) in the material for the twenty nouns (N), fifteen verbs (V) and the five adjectives (A). The total amount of training and testing corpus instances was: 8716/1525. The average polysemy in the sample is 3,5/7,6 for lexemes and cycles respectively.

Word	POS	Corpus Instances	Lexemes/ Cycles
barn/1	N	656/115	3/6
betydelse/1	N	295/52	2/1
färg/1	N	110/19	4/11
konst/1	N	77/13	3/6
kraft/1	N	152/27	4/11
kyrka/1	N	154/27	2/3
känsla/1	N	142/25	2/4
ledning/1	N	91/16	4/1
makt/1	N	128/22	3/4
massa/1	N	93/16	6/3
mening/1	N	168/29	4/1
natur/1	N	90/16	3/4
program/1	N	139/24	4/10
rad/1	N	145/25	4/3
rum/1	N	223/39	3/7
scen/1	N	101/17	4/7
tillfälle/1	N	117/20	2/4
uppgift/1	N	174/30	2/3
vatten/1	N	285/50	2/3
ämne/1	N	198/34	4/4
betyda/1	V	198/35	4/4
flytta/1	V	188/33	2/4
fylla/2	V	96/17	4/11
följa/1	V	345/61	5/19
förklara/1	V	169/30	2/9
gälla/1	V	843/148	4/6
handla/1	V	250/44	4/5
höra/1	V	523/92	5/14
måla/1	V	96/16	2/7
skjuta/1	V	79/14	6/15
spela/1	V	267/47	6/23
vänta/1	V	248/43	3/15
växa/1	V	203/36	2/9
öka/1	V	436/77	2/2
öppna/1	V	147/25	4/16
bred/1	A	103/18	3/1
klar/1	A	307/54	4/11
naturlig/1	A	139/24	4/5
stark/1	A	352/62	5/11
öppen	A	189/33	7/21

Table 1. Data for the Swedish Lexical Sample

3 Annotation

The annotation was carried out interactively using a concordance-based interface (developed in SemTag) and which interacts with the corpus and the dictionary; (see http://svenska.gu.se/~svedk/SENSEVAL/images/semtag.gif for a screenshot of this tool). Due to our limited financial resources only two professional lexicographers and a trained Phd student were involved in the tagging process, which was preferred to (untrained) students doing the annotation. High replicability between the human annotators was observed (>95%). The uncertain cases were not used in the training or testing material, while the provided dictionary descriptions for the 40 lemmas were revised (extended and/or modified) prior to their release.

4 Scoring

Prior to SENSEVAL, evaluating WSD performance was based solely on the exact match criterion, which is not consider a "fair" metric, and has a lot of drawbacks (e.g. it does not account for the semantic distance between senses when assigning penalties for incorrect labels, and it does not offer a mechanism to offer partial credit; *cf.* Resnik & Yarowsky (2000)) Instead, in SENSEVAL-2 three scoring policies are adopted:

1. **Fine-grained**: answers must match exactly
2. **Coarse-grained**: answers are mapped to coarse-grained senses and compared to the gold standard tags, also mapped to coarse-grained ones (sense map is required; see below)
3. **Mixed-grained**: if a sense subsumption hierarchy is available, then the mixed-

grained scoring gives some credit to choosing a more coarse-grained sense than the gold standard tag, but not full credit (also using a sense map; see below).

A "sense map" containing a complete list of all sense-ids involved in the evaluation was provided in order to perform the two last types of scoring policies. Each line in the sense map included sense subsumption information and contained a list of the subsumer senses and branching factors.

5 Participants and Results

Five groups showed interest in participating in the Swedish task (eight systems in total). Table 2 provides information for the participating systems, while their average performance is given in Table 3, the score in parenthesis concerns: Verbs/Noun/ Adjectives. All systems returned answers for all instances, thus precision equals recall, all used supervised methods and all systems scored lower on the adjectives and higher on the nouns.

Group (Systems)	Method	Contact Person(s)
Uppsala Univ. (PWE, 3)	TBL-tränade Prolog word experts	T. Lager, N. Zinovjeva
Linköping Univ. (LIU, 1)	Multilevel decision list approach	L. Ahrenberg, M. Merkel, M. Andersson
Göteborg Univ. (Språkdata, 2)	Machine learning & feature overlap	D. Kokkinakis
John Hopkins Univ. (JHU, 1)	---	D. Yarowsky
Maryland Univ. (UMD, 1)	Support vector machine	P. Resnik, J. Stevens, C. Cabezas

Table2. Participants

System	Results	
	Fine-Grained	**Mixed-Grained**
JHU	70,1(63,4/76,9/51,8)	74,7(70,9/79,8/59,5)
PWE-Vote	63,0(58,5/72,7/48,7)	68,6(65,9/75,0/57,9)
Språkdata-ML	62,0(57,8/71,3/48,2)	68,2(66,1/74,9/54,4)
PWE-Simple	61,1(55,4/73,2/43,5)	66,8(63,2/75,7/51,7)
UMD	61,1(56,4/71,4/45,5)	65,6(61,7/73,6/54,3)
LIU	56,5(47,8/71,6/40,8)	61,6(54,7/73,3/49,6)
PWE-Disj	54,0(46,3/67,7/38,4)	60,7(55,3/71,0/47,5)
Språkdata-Overlap	46,0(36,6/57,8/43,1)	55,8(47,8/65,7/53,8)

Table 3. Results. *Overall Precision* followed by precision for *(Verb/Noun/Adective)* instances

Conclusion

The process of WSD is a complex, controversial matter, but relevant for a number of NLP applications. Our contribution to the exercise will eventually sharpen the focus of WSD in Sweden; the material developed in SENSEVAL-2 can be used as benchmark for other researchers that need to measure their system's WSD performance against a concrete reference point (although the dictionary is limited). We think that WSD opens up exciting opportunities for linguistic analysis, contributing with very important information for the assignment of lexical semantic knowledge to polysemous and homonymous content words. The existence of sense ambiguity (polysemy and homonymy) is one of the major problems affecting the usefulness of basic corpus exploration tools. In this respect, we regard WSD as a very important process when it is seen in the context of a wider and deeper NLP system.

Acknowledgements

We would like to thank the *Swedish Council for Research in the Humanities and Social Sciences* (HSFR) for providing financial support for the coordination of the task.

References

Allén S. (1999[1981]). The Lemma-Lexeme Model of the Swedish Lexical Database. *Empirical Semantics*, 376-387. Rieger B. (ed). Bochum. (Reprinted in: Allén S. (1999). *Modersmålet i Fäderneslandet. Ett urval uppsatser under fyrtio år av Sture Allén*, 268-278. Meijerbergs Arkiv 25).

Ejerhed E., Källgren G., Wennstedt G. and Åström M. (1992). *The Linguistic Annotation of the Stockholm-Umeå Corpus project.* Technical Report No. 33, Univ. of Umeå.

Järborg J. (1999). *Lexikon i konfrontation.* Research Reports from the Department of Swedish, Språkdata, GU-ISS-99-6. Available from: http://svenska.gu.se/~svedk/resrapp/konfront.pdf. (In Swedish).

Kilgarriff A. and Palmer M. (2000). Introduction to the Special Issue on SENSEVAL. *Computer and the Humanities*, 00:1-13, Kluwer Acad. Publishers.

Resnik P. and Yarowsky D. (2000). Distinguishing Systems and Distinguishing Senses: New Evaluation Methods for Word Sense Disambiguation. *Natural Language Engineering*, 5(2):113-133, Cambridge.