Swedish SENSEVAL, a Developer's Perspective

Dimitrios Kokkinakis, Jerker Järborg and Yvonne Cederholm
Språkdata, Göteborg University
Box 200, SE-405 30, Sweden
{First.Last}@svenska.gu.se

Abstract

There are, hopefully, many computer programs for automatically determining which sense of a word is being used in a given context, according to a variety of semantic, defining or other types of dictionaries. SENSe EVALuation (SENSEVAL) is an open, community-based evaluation exercise for Word Sense Disambiguation (WSD) programs, arranged for a second consecutive time. The purpose of the exercise 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. Moreover, not only do we want to know which programs perform best, but also, the developers of a program want to know when modifications improve performance, and how much and what combinations of modifications are optimal.

1. Introduction

According to dictionaries, common words have more than one meaning. Usually, only one of these meanings apply in a given context, either written or spoken. This is no issue for people in their daily interaction with others, but it is a difficult task for computers. The task is of great importance in a number of Natural Language Processing (NLP) applications, such as Machine Translation (MT) or (Cross-Language) Information Retrieval ([CL]IR). Word sense ambiguity is a potential source for errors in such tasks and it is considered as *the* great open problem at the lexical level of 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. SENSe EVALuation (SENSEVAL), Kilgarriff (1998), Kilgarriff & Palmer (2000) is an open, community-based evaluation exercise for Word Sense Disambiguation (WSD) programs arranged for a second consecutive time.

The purpose of the exercise 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. Moreover, not only do we want to know which programs perform best, but also, the developers of a program want to know when modifications improve performance, and how much and what combinations of modifications are optimal. Specifically for Swedish, we would also like to investigate to what extent WSD can be done, the potential resources available for the task and create a framework that can be shared both within SENSEVAL and for future evaluation exercises of similar kind, national and international. SENSEVAL is designed to meet all these needs.

This paper will present some of the experiences we gained by participating as developers and organisers in the SENSEVAL exercise for Swedish. Particularly, the choice of the lexical and textual material, the annotation process, the scoring scheme, and the motivations for choosing the ”lexical-sample” branch of the exercise.

2. Short History

SENSEVAL-1 was the first open evaluation exercise for WSD programs. Three languages (English [18 systems], French [5 systems] and Italian [2 systems]) and a total of 23 research groups participated. SENSEVAL-1 was held in Sussex, UK in 1998. The exercise was conceived at the SIGLEX workshop: "Tagging Text with Lexical Semantics. Why, What and How?" held in 1997 in Washington. A range of Machine Learning algorithms and a variety of lexical resources were utilised. Two important points are worth to be mentioned w.r.t. SENSEVAL-1. One was the fact that by the end of the exercise the purity of the approach was less important than the robustness of the system performance; and second, the discussion created more awareness among the participants of how fundamental the lexicon is to the task.

3. Lexical Sample

Three tasks were identified for SENSEVAL-2, these are: *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 (SENSEVAL-1 was a lexical-sample exercise). 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 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 the evaluation material;
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;
5. Provided that the sample is well chosen, the lexical sample strategy would be more informative about the current strengths and failings of WSD research than the all-words task (Kilgarriff & Palmer (2000)).
1.

Table 1 gives brief information w.r.t. the different languages participating in the lexical-sample part of SENSEVAL-2.

Language	*Amount of Words*	*Available Context*	*Lexicon*	*Format*	*Corpus*	*Sample/ words*
Basque	**40**	5 sents around	Euskal Hiztegia	TEI-SGML	Newspaper	75+15*n*
Chinese	**15**	2-3 sents	???	???	Sinica Corpus	10-60+?
Danish	**100** (50/25/25)	50 tokens	SIMPLE+Nudansk Ordbog	???	Newspaper	75+15*n*
English	???	???	WordNet 1.7	XML	BNC, web, PennTreebank	???
Italian	**100** (50/25/25)	2 sents around	ItalWordNet	XML	Newspaper, Periodical	???
Swedish	**40** (20/15/5)	2 sents around	GLDB/SDB	XML	SUC	843-77+148-13

Table 1. Lexical-sample participants in SENSEVAL-2

4. SENSEVAL-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. resources, sampling, 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 by the end of April, 2001 and the state-of-affairs at this moment (May, 2001) is that the participants are working with the material. During the second week of June, 2001 the results will be available, a two-day workshop will be held in Toulouse, France, devoted to the SENSEVAL-2 exercise. 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.

4.1 Dictionary and Text

At least three lexical resources were candidates for the Swedish lexicon-sample task. These were the Swedish versions of S-WordNet (http://www.ling.lu.se/projects/Swordnet) and SIMPLE (http://spraakdata.gu.se/simple/), and the Gothenburg Lexical Data Base (GLDB/SDB) (http://spraakdata.gu.se/lb/gldb.html). The GLDB/SDB was chosen since the S-WordNet had (up to that point) limited coverage and is also an ongoing project; while SIMPLE, although available, has limited coverage (in principle it could be used since it is linked to GLDB/SDB). GLDB/SDB is a generic defining dictionary of 65,000 entries.

Creating a sense-annotated reference corpus is a laborious task. Therefore, we developed the majority of the test and reference material within an ongoing, highly relevant for our mission project, namely SemTag ('Lexikalisk betydelse och användningsbetydelse' - Lexical Sense and Sense in Context); see Järborg (1999). For the textual material the Stockholm-Umeå Corpus (SUC), Ejerhed *et al.* (1992), was chosen, basically for two reasons. One because it is available to the research community, and, second because it is the corpus utilised in SemTag.

Table 2 shows information on the sense inventory, the amount of corpus instances and the distribution of senses (lexemes) and sub-senses (cycles) in the material.

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

Table 2. Swedish lexical sample (***Training/Testing; total: 8716/1525*)

4.2 Sampling

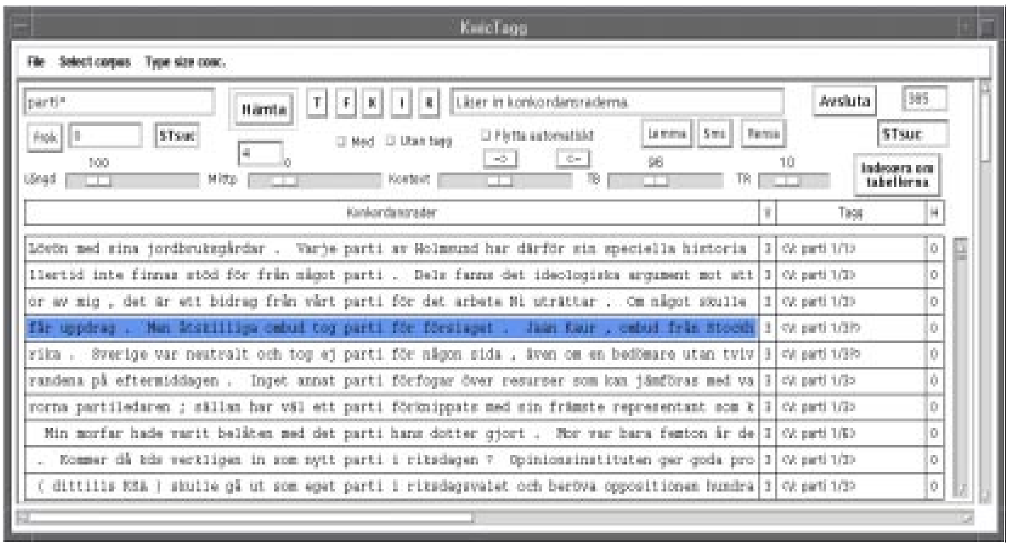
There is no standard method for sampling the lexical data. However, certain features were considered. These were:

Frequency Polysemy Part-of-speech Distribution of senses

Words were chosen based not so much on intuition, but rather on their frequency and polysemy. Still, it is hard to find a balance between these two features since high frequency words tend to be monosemous in a corpus, while high 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 words that were less frequent. Part-of-speech information was accounted for 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 less than nouns and verbs in GLDB/SDB). We chose a sample of words where the amount of senses was evenly distributed, i.e. lemmas with 2-7 senses and 1-23 subsenses.

4.3 Annotation Process

The annotation was carried out interactively using a concordance-based interface, Figure 1. 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. The replicability between those were on the 95% level.



Konkordansrader	#	Tagg	H
Lövön med sina jordbruksgårdar . Varje parti av Holmsund har därför sin speciella historia	3	<X parti 1/1>	0
lliertid inte finnas stöd för från något parti . Dels fanns det ideologiska argument mot att	3	<X parti 1/3>	0
or av mig , det är ett bidrag från vårt parti för det arbete Ni uträttar . Om något skulle	3	<X parti 1/3>	0
får uppdrag . Men åtskilliga ombud tog parti för förslaget . Jaan Kaur , ombud från Stockh	3	<X parti 1/3>	0
rika . Sverige var neutralt och tog ej parti för någon sida , även om en bedömare utan tviv	3	<X parti 1/3>	0
randens på eftermiddagen . Inget annat parti förfogar över resurser som kan jämföras med va	3	<X parti 1/3>	0
rorna partiledaren ; sällan har väl ett parti förknippats med sin främste representant som k	3	<X parti 1/3>	0
Min morfar hade varit belåten med det parti hans dotter gjort . Mor var bara femton år de	3	<X parti 1/6>	0
. Kommer då kds verkligen in som nytt parti i riksdagen ? Opinionsinstituten ger goda pro	3	<X parti 1/3>	0
(dittills KDA) skulle gå ut som eget parti i riksdagsvalet och beröva oppositionen hundra	3	<X parti 1/3>	0

Figure 1. Annotation interface

Since some of the aims with SemTag is to improve the lexicographic descriptions in the GLDB/SDB and test in practice the validity of the lemma-lexeme model implemented, the development of the annotated instances for SENSEVAL-2 gave us a chance to revise our sense inventory and make adjustments and improvements to the descriptions found in the database; i.e. in the form of adding new sub-senses or modifying definitions of senses.

4.4 Interchange and Result Format

The corpus instances and dictionary format was in XML with DTDs provided. An example of a corpus instance (SUC file:AD04_BRV) for the 5th sense of the verb höra here: 'belong' is:

```
<instance id="höra.301"><answer instance="höra.301" senseid="höra_1_5"/>
  <context>Den ämnesdidaktiska forskningen kom igång i Sverige först på 70-talet.
    Geografiundervisningen diskuterades dock redan på 50-talet. Sverige <head>hörde
    </head>  till de ledande nationerna när det gällde den" nya  geografin". Utbytet mellan
    Lund och USA var livligt och Gösta Wennberg som bodde och arbetade i Lund på den
    tiden tog starka intryck. 1964 kom han till Uppsala och blev metodiklektor på lärarhögskolan.
  </context>
 </instance>
```

The systems required to return, for scoring, a one-line-per-answer for each unique corpus reference for the token being tagged and for which they were returning a result. One or more sense-identifiers, optionally associated with a probability measure (see also Section 5), could be attached. The BNF for scoring is:

```
<lexical_sample_answer>    ::= lexical-element instance-id <sense-tag-list>
<sense-tag-list>           ::= <weighted-list> | <unweighted-list>
<weighted-list>            ::= sense-id[/weight] {sense-id[/weight]}
<unweighted-list>          ::= sense-id {sense-id}
<weight>                   ::= INTEGER | positive REAL NUMBER
```

5. Scoring

Prior to SENSEVAL evaluating WSD performance was based on the exact match criterion given by the formula:

%correct=100 × (#exactly matched sense tags/#assigned sense tags)

which is not consider a "fair" metric, and has a lot of drawbacks, such as that it does not account for the semantic distance between senses when assigning penalties for incorrect labels, and that 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" contains a complete list of all sense-ids involved in the evaluation and is necessary for performing the two last types of scoring policies. Each line in the sense map includes sense subsumption information and contains a list of the subsumer senses and branching factors.

6. Participants

Three groups showed interest on participating in the Swedish task:

Group	Method	Contact Person(s)
Uppsala University, Linguistics	*TBL-tränade Prolog Word Experts; (Peewees)*	*Torbjörn Lager* *Natalia Zinovjeva*
Linköping University, Computer & Info. Science	Multilevel Decision List Approach	Lars Ahrenberg, Magnus Merkel Mikael Andersson
Göteborg University, Språkdata	*Machine Learning*	*Dimitrios Kokkinakis***

Table 3. Swedish participants in SENSEVAL-2 (***also in the developer's group*)

7. Conclusions

The process of Word Sense Disambiguation is a complex, controversial matter, but relevant for a number of Natural Language Processing applications. Our contribution to the exercise will eventually sharpen the focus of WSD in Sweden; the material developed in SENSEVAL-2(Swedish) can be used as benchmark for other researchers that need to measure their system's WSD performance against a concrete reference point (although the number of words is rather small). 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 and component when it is seen in the context of a wider and deeper NL processing system.

References

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, In Swedish

Kilgarriff A. (1998), *SENSEVAL: An Exercise in Evaluating Word Sense Disambiguation Programs*. In Proceedings if the 1st LREC, Granada, Spain pp 581-588

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