Using a Lexical Database for Domain Determination, partial Disambiguation and Dictionary Expansion

Sven-Göran Malmgren, Christian Sjögreen
Department of Swedish Language
Box 200
405 30 GÖTEBORG
SWEDEN
malmgren@svenska.gu.se, christian.sjogreen@svenska.gu.se

Abstract

In this paper, we report on three ways of exploiting the information available in the biggest defining dictionary (and the corresponding lexical database) on modern Swedish. First, domain classification of unknown texts is performed, based on the domain labels in the dictionary/database. Then, partial disambiguation is performed, based on the assumption that ambiguous words that *possibly* belong to the relevant domain, *probably* belong to this domain, given the domain classification. Third, assuming that non-analysable words in the text that are not names possibly belong to the relevant domain, we investigated the possibility of expanding the dictionary automatically in this way. The experiments described in the paper tend to support the hypotheses underlying point 2 and point 3, but of course there are problems.

While point 1 and point 2 have been discussed in several papers on computational linguistics, point 3 seems to have received somewhat less attention, both in lexicographical contexts and in language technology.

1. Introduction

The biggest defining dictionary of Modern Swedish, *Nationalencyklopedins ordbok* (NEO; see Malmgren 1992, 2002) comprises about 62,000 lemmas and about 70,000 lexemes (numbered senses). In addition, there are about 25,000 marked sub-senses. It is based on a lexical database (the *G[öteborg]LDB*) with a number of information categories that are not available in the printed version of the dictionary. Some of these categories have been used for Language Technology purposes, especially lemmatization and automatic disambiguation (see, e.g. Dura 1998, Kokkinakis 2001, Kokkinakis et al 2001). But there are several other categories, e.g. valency information, that have not yet been employed for Language Technology purposes, although their usefulness in this context seems obvious.

In this paper, we shall describe some recent applications of *subject field* or *domain* information in the NEO/GLDB in the context of Language Technology. The aim is threefold. First, automatic domain classification of unknown texts is, of course, valuable in itself. Second, if you know the (likely) domain of an unknown text, you have a better chance to disambiguate several of the ambiguous words of the text. Third, the domain classification offers a possibility of automatic expansion of the dictionary vocabulary, since at least some of the text words not found in the dictionary are likely to belong to the same domain as the text. (For Swedish, this is especially true of compounds; see below.)

To the best of our knowledge, there are not very many studies of this kind with a *lexicographic* approach. On the other hand, domain categorization and domain-based disambiguation make up a relatively common topic in the context of computational linguistics, at least in recent years. In Stevenson & Wilks (2003) a useful survey of word sense disambiguation through 40 years, domain information plays a very limited role. But in the project SENSEVAL (evaluation of sense disambiguators; see, e.g., Kilgariff 1998), some of the participating groups (see esp. Buitelaar 2001, Magnini et al 2001) applied this method of disambiguation. Magnini & Cavaglià 2000 (and later) report from experiments with automatic domain classification based on the WordNet taxonomy enriched by domain codes. As regards Swedish, we are not aware of any previous study of this kind, using domain information from a machine-readable dictionary.

There are about 100 different domain labels in the NEO/GLDB database, e.g. *med(icine)*, *sport(s)*, *mus(ic)*, *math(ematics)*, and *astron(omy)*. For instance, more than 3000 lexemes are marked *med(icine)* in the database. Most of these labels are not available in the printed dictionary. It would be pointless, e.g. in the articles **illness** or **cancer**, to tell the dictionary user that these words belong to the subject field of medicine, since it follows immediately from the definition. The total number of subject field labels is about 90,000, that is, about 1.5 pro lemma on an average. The labels do not make any difference between technical terms in a narrow sense and more general words belonging to the same subject field. For instance, a disease may be denoted by a technical (medical) term as well as a general word. In these cases both are labelled *med.* It should also be noticed that the dictionary is not a dictionary on LSP, even though it contains several technical terms from many different fields. But one should not expect to find terms of highly specialized language in the dictionary.

2. Method

The pilot experiments described in this paper are carried out on non-technical (newspaper) texts belonging to three different subject fields. It is quite possible, however, that the subject field information in the NEO/GLDB would also yield good results if applied to more technical texts, since even in highly technical texts concerning a certain subject field there are probably many semi-technical and non-technical words belonging this subject field (and thus possible to look up in the NEO/GLDB). Several of the unknown words (i.e. those not found in the dictionary) in such a text are likely to be technical terms belonging to the subject field in question.

In order to apply the subject field information of the NEO dictionary to a non-edited text, a lemmatizer is necessary. There are some good lemmatizers available for Swedish (see e.g. Karlsson 1992). Some of them give all possible analyses of ambiguous text-words, while others give unambiguous analyses by means of some heuristic method, with a precision rate of at best 95%. In this paper, the first kind of lemmatizer is relevant, since we will try to disambiguate several of the text-words by means of the domain categorization (a somewhat more 'intelligent' kind of disambiguation).

As is well known, Swedish is a compounding language. In any unknown text, there are compounds not found in any dictionary. Normally, they can be relatively safely analysed by a good lemmatizer. In this study, we will first refrain from analysing words not found in the dictionary (they will, in the first analysis, be regarded as unknown words). Later on they will be examined, being words possibly belonging to the relevant subject fields and thus offering a possibility of expanding the lexicon.

A very simple – invented – English example will illustrate the basic idea. Suppose we have the short English text *he plays the saxophone in a cajun band* and an English lexical database with domain labels. In this database, the word *saxophone* will certainly be marked *mus(ic)* and hardly anything else, and so one of the words of our text is unambiguously marked as *musical.* Besides, there are two ambiguous words, *play* and *band.* They can both belong to the domain of music, but also to other domains. Let us suppose that both words, according to the lexical database, can belong to five different subject fields, out of which music is one. Finally, let us suppose that we do not find the word *cajun* in our dictionary.

This means that there are one certain and two possible words belonging to the subject field *music* in our text. We now give the domain *music* one point for the unambiguous word *saxophone* and 1/5 point for each of the ambiguous words *play* and *band* (since the probability is – from the computer's point of view – 1/5 that these ambiguous words are used in a musical sense). It follows that no other subject field than music can achieve more than 2/5 points, while the musical field achieves 12/5 points. Thus, it should be a good guess – again, from the computer's point of view – that our 'text' is about music. The very simple 'mathematics' may of course be elaborated in different ways; for instance, the relative frequency of the musical and non-musical uses of the words *play* and *band* in authentic texts may be utilized (cf. Magnini et al 2000). For Swedish, this is not yet possible, since there are so far only very small semantically tagged Swedish corpora.

This is the first step of our analysis. Now, having decided that the text is about music, it should be a relatively plausible guess that the two possibly musical words are indeed musical. This hypothesis turns out to be correct. Finally, there may be a good chance that the 'unknown' word *cajun* denotes something with a musical connection. This, too, happens to be correct.

The point, then, is threefold: first we point out the most likely subject area of the text; after that, we can – hopefully, with a reasonable amount of certainty – disambiguate all text-words of which we could say at the beginning that they *possibly* – and only possibly – belong to this subject field; and finally, there is, at least in favourable cases, a reasonable chance that the non-analysable words belong to the domain in question.

Now, this example was a made-up one; reality is something else. Among other things, it has been pointed out that many texts are not homogeneous from the domain point of view (Krovetz 1998). You could easily imagine articles where the medical domain dominates some part of the text and the military domain some other. Part of the problem may be solved by looking not only at the 'best-scoring' domain but also at the second best-scoring one. If the scores of the two best domains are relatively even, two 'winners' might be appointed rather than one (cf. below).

3. Determination of the domains of the authentic texts

We will now examine three randomly selected Swedish texts, one from the legal domain, one from the medical domain and one from the military domain. From the reader's point of view, there is no doubt whatsoever that the texts belong to these domains; among other things, it is indicated by the respective headlines. The following scores were obtained for the three texts (the scores of the three 'best' domains are given in each case):

Legal text		Medical text		Military text	
legal	**45.9**	**med.**	**29.8**	**mil.**	**68.2**
soc(iety)	**36.8**	**traffic**	**8.0**	**med.**	**24.5**
sport	**12.6**	**games**	**7.2**	**sport**	**12.6**

Table 1: The scores of the three 'best' domain labels in the three texts

As is readily seen, the 'correct' domains obtain much higher scores than all other domains in the medical and the military text. In the legal text, the legal domain is challenged by the 'society' domain. As a matter of fact, the text contains much 'society' material. Perhaps the conclusion to be drawn in this case is that the text is a mixture of a legal text and a 'society' text. Tentatively, it might be proposed that the ratio between the best and the second-best domain (this could be called the BDI, or the 'best domain index') should be at least 2 in order to allow an unambiguous domain categorization of the text.
In the military text, the second-best domain lays far behind the best, but it is indeed a very natural second-best domain!

4. Partial disambiguation of the texts

We now turn to disambiguation. For the sake of illustration, we will first take a look at a passage from the military text (see Appendix 1; the domain labels from the NEO/GLDB are given below the corresponding words, and words not found in the dictionary are marked with an asterix, *). There are four words in this passage that are labelled 'mil' but have other labels as well; namely *strategi*, *civilbefolkning*, *öppen*, and *säkerhet*. The guess that 'mil' is the appropriate label turns out to be correct in three of these cases, but not in the fourth: there is no military use of *öppen* ('open') in the text. This might be an indication that the kind of disambiguation proposed here is a bit risky when applied to adjectives (and possibly verbs).
We now turn to the complete texts. The result is as follows:

	correctly disambi-guated words	*incorrectly disam-biguated words*	*precision (%)*
legal	**38**	**9**	**81**
medical	**4**	**1**	**80**
military	**38**	**7**	**84**

Table 2: The results of the automatic disambiguation of the three texts

The result seems to be comparable to the results of similar experiments (on a larger scale). The precision is a bit higher than the precision of methods based on text similarities, but the recall rate is of course much lower, since many ambiguous words in the texts do not, in any of their senses, belong to the 'best' domains. (Cf. Magnini et al 2001.)

5. Expanding the lexicon

We now turn to the point that is of most interest from a *lexicographical* point of view. In the passage from the military text (see Appendix 1), there are two words that are unanalysed since they were not found in the NEO dictionary, namely *krigshandlingar* and *krigsinsats*. By means of a lemmatizer, they can be unambiguously divided into two simple elements occurring in the dictionary, *krig(s)-handling* and *krig(s)-insats*. Since one of the elements (*krig* 'war') has the label 'mil.', we can relatively safely guess that the compounds as a whole belong to the military domain. This is by no means a trivial point; the domain labelling of the compounds must be preceded by domain categorization of the whole text. Cf., for instance, the compounds *operations-bas* (base of operation) and *operations-bord* (operating-table) where the first element can be both medical and military. Automatic domain categorization is possible only if we know the domain of the text where they occur.

If we look at *all* compounds in the military text that are missing in the dictionary, but where one of the elements is labelled 'mil.' (possibly together with other labels), it turns out that the tentative 'mil' analysis of these compounds is always correct. The same holds for the other two texts. In other words, this seems to be a rather safe, and powerful, way of automatically incorporating many more words into the dictionary. Of course, next time we perform domain classification of unedited texts, the system can use these new words, and so on; thus, the system is, in a way, self-improving. It should be noted, however, that this method is applicable only to typically compounding languages, like Swedish. For languages like English, where new terms are normally multi-word lexical units, it would not be efficient (cf. Jacquemin & Bourigault 2003).

Again, of course the point is not that the dictionary is expanded with lots of randomly chosen compounds, which would be very easy. The point is, rather, that the new compounds are labelled ('mil.', 'med.', etc.).

Finally, what about the words that do not occur in the dictionary and also can not be analysed as compounds (like *cajun* in the 'English' example above)? The natural hypothesis is that there is a good chance that these words are – probably relatively technical – words belonging to the domain in question. Apart from names, there is only one such case in our texts, *epidural* in the medical text. In this case, the hypothesis turns out to be correct; *epidural* is a medical word. But of course the hypothesis must be tested on much larger text materials.

6. References

Buitelaar, P. **2001. 'The SENSEVAL-2 Panel on Domains, Topics and Senses' in:** ***Proceedings of SENSEVAL-2. Second International Workshop on Evaluating Word Sense Disambiguation Systems, Held /.../ 5–6 July 2001.*** **New Brunswick, NJ.**

Dura, E. **1998. *Parsing words.* Göteborg.**

Jacquemin, C. and Bourigault, D. **2003. 'Term Exctraction and Automatic Indexing' in R. Mitkov (ed.), *The Oxford Handbook of Computational Linguistics.* Oxford: Oxford University Press.**

Karlsson, F. **1992. 'SWETWOL: a comprehensive morphological Parser for Swedish' in *Nordic Journal of Linguistics* 15.**

Kilgariff, A. **1998. 'SENSEVAL: An Exercise in Evaluating Word Sense Disambiguation Programs' in Th. Fontenelle, Ph. Hiligsmann, A. Michiels, A. Moulin and S. Theissen** (eds.), *Actes EURALEX'98. Papers submitted to the Eighth EURALEX International Congress on Lexicography in Liège, Belgium.* Liège.

Kokkinakis, D. 2001. 'The Språkdata-ML System as used for SENSEVAL 2' in *Proceedings of SENSEVAL-2.*

Kokkinakis, D., Järborg, J. and Cederholm, Y. 2001. 'SENSEVAL-2: The Swedish Framework' in *Proceedings of SENSEVAL-2.*

Krovetz, R. 1998. More than one Sense per Discourse. NEC Research Memorandum.

Magnini, B. and Strapparava, C. 2000. 'Integrating Subject Field Codes into WordNet' in *Proceedings of LREC-2000, Second International Conference on Language Sources and Evaluation, Athens Greece.* Athens.

Magnini, B., Strapparava, C., Pezzulo, G. and Gliozzo, A. 2001. 'Using Domain Iinformation for Word Sense Disambiguation' in *Proceedings of SENSEVAL-2.*

Malmgren, S.-G. 1992. 'From Svensk ordbok ('A Dictionary of Swedish') to Nationalencyklopedins ordbok ('The Dictionary of the National Encyclopedia')' in H. Tommola, K. Varantola, T. Salmi-Tolonen and J. Schopp (eds.), *EURALEX'92 Proceedings I–II. Papers submitted to the 5th EURALEX International Congress on Lexicography in Tampere, Finland.* Tampere: University of Tampere.

Malmgren, S.-G. 2002. 'Lexicography in The Nordic countries: traditions and recent developments' in A. Braasch & C. Povlsen (eds.), *Proceedings of the Tenth EURALEX International Congress, Copenhagen, Denmark, August 13–17, 2002.* Copenhagen:CST.

Stevenson, M. & Wilks, Y. 2003. 'Word-Sense Disambiguation' in R. Mitkov (ed.), *The Oxford Handbook of Computational Linguistics.* Oxford: Oxford University Press.

Appendix 1: A passage from the military text

Kanske		kan	man	redan	ana		en	ny	sorts		krigföring,	
											mil.	
Possibly		**can**	**you**	**already**	**imagine**		**a**	**new**	**kind of**		**warfare,**	

En	ny	sorts	strategi.		En	strategi		som	bygger	på	en	väl
			mil.			**mil.**						
			Games			**games**						
a	**new**	**kind of**	**strategy.**			**A**	**strategy**		**based on**			
	a	**well**										

genomförd		kombination		av	ytterst		häftig		krigsinsats			och
		math.					**meteorol.**		*			
		sport					**psychol.**					
		games										
performed		**combination**		**of**	**extremely**		**intensive**		**war efforts**			**and**

en noga		planerad	psykologisk		krigföring,		riktad	mot	Iraks	soldater		
			psychol.		**mil.**					**mil.**		
A carefully		**planned**		**psychological**		**warfare,**			**aimed**	**at**	**Iraq's**	
	soldiers											

och	civilbefolkning.			En	effektiv		växelverkan		mellan		den	
	mil.				**econ.**		**phys.**					
	soc(iety)					**phys.**		**fil**				
and	**civilian**	**population.**		**An**	**efficient**	**interaction**			**between**		**the**	

förlamande		dödsskräck		krigshandlingarna			skapar	och	propagandans			
psychol.		*		*					**soc.**			
med.									**massmedia**			
paralysing		**fear of death**		**the acts of war**			**create**	**and**	**the propaganda's**			

lockande		erbjudande		om	en	öppen	dörr	till	säkerhet			
		econ.				**mil.**	**home**		**mil.**			
						soc.			**soc.**			
						sport			**psychol.**			
						(etc.)			**econ.**			
tempting		**offer**		**of**	**an**	**open**	**door**	**to**	**security**			

och	trygghet.											
	psychol.											
and	**safety**											