Lemma selection in domain specific computational lexica – some specific problems

Sussi Olsen

Center for Sprogteknologi
Njalsgade 80, DK-2300, Denmark
Phone: +45 35 32 90 90
Fax: +45 35 32 90 89
e-mail: sussi@cst.dk
URL: www.cst.dk

Abstract

This paper describes the lemma selection process of a Danish computational lexicon, the STO project, for domain specific language and focuses on some specific problems encountered during the lemma selection process. After a short introduction to the STO project and an explanation of why the lemmas are selected from a corpus and not chosen from existing dictionaries, the lemma selection process for domain specific language is described in detail. The purpose is to make the lemma selection process as automatic as possible but a manual examination of the final candidate lemma lists is inevitable. The lemmas found in the corpora are compared to a list of lemmas of general language, sorting out lemmas already encoded in the database. Words that have already been encoded as general language words but that are also found with another meaning and perhaps another syntactic behaviour in a specific domain should be kept on a list and the paper describes how this is done. The recognition of borrowed words the spelling of which have not been established constitutes a big problem to the automatic lemma selection process. The paper gives some examples of this problem and describes how the STO project tries to solve it.

1. Introduction

The Danish STO project, SprogTeknologisk Ordbase, (i.e. Lexical database for language technology) (see Braasch et al. 1998,) is a national follow-up project of the Danish PAROLE lexicon (see LE-PAROLE 1998) with the aim of creating a large size Danish lexicon for natural language processing. The lexicon is planned to contain 50,000 lemmas divided between 35,000 from general language and 15,000 from different domains of language for special purposes.

The lemmas are provided with detailed morphological and syntactic information while the semantic information in this first version for a large part of the vocabulary is reduced to domain information.

The first version of lexicon should be complete by the end of 2003, but extension of the linguistic information especially at the semantic level, addition of pronunciation information as well as an extension of the number of domains will be possible at a later date, depending on funding. For further information on the STO project, see Braasch (2002).

2. The lemma selection process

2.1. The domains and the corresponding corpora

The Danish PAROLE lexicon contained 20,000 entries and the vocabulary consisted of frequent words of different word classes, belonging to the general language. Since the STO project is planned to consist of vocabulary from various different domains as well as from general language, the selection of specific domains and the lemma selection from these was one of the first tasks to initiate.

The selection of the specific domains is based on the potential future applications and at present the following domains have been selected, while – at least – one is still to come: IT, public administration, environment, commerce and health.

Existing lemma lists or dictionaries from the different domains are not directly suitable for our purpose, for various reasons.

Firstly the delimitation of a domain varies a lot, e.g. it seems that the domain IT in some lemma lists includes lemmas from domains like commerce and marketing, while other lemma lists have a more narrow definition merely including technical terms.

Secondly lemma lists and domain specific dictionaries are made for different purposes and address different user profiles. Mostly they are highly specialised term lists, while some cover a broader vocabulary like the one appearing in newspapers or words from general language having another meaning in the language of the domain in question. The STO database is not supposed to cover the most specialised terms of the different domains but rather the vocabulary that laymen might encounter in various contexts. Specialised termlists can later be added by future users.

Thirdly Danish is a less widely spoken language and up-to-date lemma lists or dictionaries are not available for all domains.

For these reasons, the lemma selection for the domain specific vocabulary of STO is primarily based on text corpora. As a matter of fact this goes for the STO project as such. STO is supposed to be corpus based, which means that not only the lemma selection for both general language and domain specific language but also the morphological, syntactic and semantic encoding of each lemma mainly depends on what is found in the relevant corpora.

For each of the domains we build a text corpus of at least 1 million tokens consisting of texts from textbooks, user manuals, informative articles from magazines, newspapers, the web etc. We do not include texts written by experts for experts since the language of such texts is too specialised leading to the specialised termlists mentioned above.

We are aware that the vocabulary of the domains changes very fast and that the vocabulary of the STO database will be out of date after a short time, lacking the new central terms. But since the process of collecting texts on the web for the corpora and the first editing of the lemma candidate lists is automatic, it will be rather uncomplicated to extend the vocabulary of each domain in the future and thus keep it up-to-date.

At this moment we have built two corpora, i.e. IT and environment. From the IT-domain we have selected and encoded – morphologically and syntactically - about 2000 words while the lemma selection and the morphological and syntactic encoding of the environment domain is the next task.

2.2. The selection of domain specific lemmas

The overall method for the lemma selection is sketched in the following table:

Step 1	List of all word forms found in the corpus includng frequency
Step 2	Comparison of this list with a list of general language lemmas already encoded
Step 3	Normalisation and truncation of the word forms -> lemma candidates
Step 4	Manual examination and part of speech marking of the lemma list
Step 5	Comparison with other dictionaries of the domain

Step 1: From each domain specific corpus we make up a list of the entire amount of words with their frequency.

Step 2: We automatically compare the word forms on this list with a list of encoded word forms of general language, thus sorting out lemmas already encoded in the database.

Step 3: The last letters are truncated in a specific order and afterwards the words are normalised, i.e. capitalisation and other special symbols like the special Danish characters are substituted by small letters or other letters respectively. The resulting forms are grouped together as candidate lemmas, keeping the word forms found in parenthesis, and the frequency of each word form is put together for the lemma frequency, e.g.

Frequency: 58	**Lemma candidate:** *område*
normalised word forms	omraade: 14, omraader: 27, omraaderne: 13, omraadernes: 2, omraaders: 2,
word forms found	Område: 14, Områder: 27, Områderne: 13, områdernes:2, områders: 2,

This truncation and grouping of word forms is a temporary step since at present we do not have access to a lemmatiser for Danish, capable of treating unknown words. The STO project has however initiated the development of a Danish lemmatiser which we hope will be ready for use for the lemma selection of the coming domains.

Step 4: The resulting list of lemmas is sorted by frequency and the lemmas with a frequency of 1 are removed from the list. The list then contains the candidates for the vocabulary of the actual domain. This list has to be examined manually since correction of the normalised forms and the correct lemmatisation of the truncated word forms cannot be made automatically. The lemmas of the first domain which we have encoded so far, were not part-of-speech tagged which means that the lemma candidate list had to be part-of speech tagged too during the manual examination. For the next domains we might try tagging the corpora with part of speech before the lemma selection. But since the lemma lists have to be examined manually anyway, it is not clear yet if there is any time saved by doing so.

Step 5: Finally, the resulting vocabulary is compared to existing dictionaries or lemma lists of the domain. Though we do not find such dictionaries suitable as a primary source for our vocabulary, we use them to check whether some central words of the domain accidentally did not occur in the corpus and therefore do not figure in the selected vocabulary. Again possible lemma candidates have to be evaluated manually.

3. Problems with the lemma selection method

Of course the implementation of the method outlined above reveals a variety of problems:

- Words that appear in the general language of STO will not be included in a domain specific lemma list,
- words that only appear once are sorted out,
- the candidate lemma list for a specific domain will contain words that have nothing to do with the domain in question but are low-frequent words from the general language.
- the need for another method to select collocations and multiword units.

3.1. Lemmas occurring in both general language and domain specific language

The first problem to be discussed here is the case where words appear in the general language vocabulary and in one or more domain specific corpora too.

Words that appear in the general language vocabulary of STO will automatically be excluded from a candidate list of domain specific lemmas according to step 2 of the lemma selection described above. But a word might belong to the general language and at the same time be part of the language of a specific domain, so we need a method to detect words that have already been encoded as general language words but that are also found with another meaning and perhaps another syntactic and/or morphological behaviour in a specific domain, e.g.

Semantic difference:
bus **general language**: a passenger vehicle
IT domain: a data channel

port **general language**: a large door or gate
IT domain: an external computer connection

Morphological difference:
indeks (eng. index)

general language plural:	*indeks,* *indekser*
IT and Mathematics plural:	*indeks,* *indekser,* *indices*

In order not to lose track of these lemmas, in the STO database we mark all entry words with source reference indicating in which corpus a lemma appears. Thus, the lemma 'bus' will be source-marked both for 'general language' and 'IT' since it appears in both corpora.

This means that a single word can have source reference to general language as well as to all the specific domains. This will be the case for the most common general language words. Words from the general language with low frequency will (hopefully) only appear in the general language corpus and will not be object to further treatment but words that are marked with source reference from a general language corpus and from one or two domain specific corpora have to be picked out for special treatment.

These lemmas will be object to a special encoding process. For each lemma it has to be decided whether the linguistic behaviour of this lemma in each domain in which it occurs, differs from the existing encoding of the general language lemma at all the three linguistic levels. Any linguistic behaviour – morphological, syntactic or semantic – that differs from the general language encoding demands an encoding reflecting this behaviour and in which it appears that the encoding is specific for a certain domain. Thus the 'indeks' example above will have two morphological units containing the inflectional patterns connected to it, one of which will be marked as valid for the IT-domain only.

Since we have only encoded one domain so far, we have not yet started the process of encoding the lemmas with a particular linguistic behaviour when appearing in a domain specific context. It is not possible to see whether a lemma with both a general language source reference and a source reference for the IT-domain is a common general language lemma or a domain specific lemma that only appears in one domain. So this process will not be started until a couple of other domains have been encoded.

3.2. Other problems connected to the lemma selection method

Some other problems turn up in connection with the lemma selection method we have chosen.

The lemmas that only appear once in a domain specific corpus are sorted out and will not be candidates on the lemma lists. Since the lemma lists have to be examined manually, we have to find a way to reduce the number of lemmas, making this task less time-consuming. We assume that the majority of the lemmas appearing only once in a corpus are not central words of that vocabulary. We are well aware that the size of our domain specific corpora and the fact that these are more or less biased because of the difficulty we have getting texts for this purpose, mean that some central lemmas only occurring once are sorted out. We try to reduce the number of such lemmas by comparing our final vocabulary with existing lemma lists and dictionaries. If we do not encounter them there, it might be because we are dealing with rather new lemmas of that domain, and hopefully these will become part of the vocabulary when this is extended at a later stage.

Some of the words on the candidate lemma lists have nothing to do with the domain in question but are low-frequent words from the general language that have not been encoded already. Since the lemma lists have to be examined manually anyway, this fact is not a real problem. If the examination of the list reveals such words, these are put aside on another list and afterwards they are evaluated (examining the frequency of the lemma in the different corpora) with the purpose of seeing whether they should be part of the general language vocabulary despite of their low frequency in the general language corpora.

The lemma selection described so far only covers single words. While we have been working with the issue of how to encode collocations in the STO description model (see Braasch and Olsen, 2000), we have not, so far, dealt with the selection of collocations and multiword units. We assume that we can make some rather good candidate lemma lists by making lists of tri-grams and bi-grams from the corpora, sorted by frequency. But of course we will have to study the subject further and see what has been done in this field before we establish a selection procedure for collocations.

4. Loanwords – the problem of recognition and normalisation

One of the biggest problems that we have come across in the lemma selection process is the recognition of loanwords, and afterwards the normalisation of these. This problem has less to do with our specific lemma selection method but constitutes a problem for all automatic and semi-automatic lemma selection and the morphological encoding process.

All languages extend their vocabulary borrowing words from other languages. This is particularly frequent in domain specific language. In Denmark, the language advisory committee, the Danish Language Council, is in charge of establishing official spelling rules and morphological behaviour, published in the official Spelling Dictionary for Danish (see Dansk Sprognævn 1996). This dictionary is normative but does of course not cover all kinds of spelling problems and does not include recently borrowed words.

The lemma selection process has to deal with the problem of detecting/recognising loanwords spelt in two or more different ways – a frequent phenomenon in domain specific language – classifying them as the same lemma and not as two (or more) unrelated words.

Though not part of the established vocabulary, new loanwords including acronyms and abbreviations of a kind follow a couple of rules in the process of being integrated in Danish. Foreign abbreviations and acronyms tend to be written with capital letters when they are new and unknown. The more established such a lemma is, the bigger the chance that it is spelled without capital letters. This is evident when we compare corpus occurrences in older and newer texts. E.g. compounds starting with the abbreviation 'EDB' (eng. EDP) used to be spelled with capital letters years ago,

EDB-udstyr (eng. computer gear)
EDB-firma (eng. computer company)

In the oldest corpus to which we have access, one third of the occurrences are spelt with capitals. In newer texts only a sixth of the occurrences with 'EDB' are spelt with capitals.

The word 'WEB' has undergone the same changes just much faster. In 10 years old texts the word does not occur at all. In some texts from the mid-nineties the word occurs with capital letters both as a part of a compound and as an independent word, but in all newer texts there are only a few occurrences of 'web' as an independent word with capital letters. All other occurrences - independent words as well as parts of a compound - are spelt with small letters.

web-baseret (eng. web based).

In Danish, abbreviations or foreign words that are part of a compound can be followed by a hyphen according to the official spelling rules like the examples above,

edb-udstyr, *web-baseret*

However it seems to be the case that frequent words are often spelled without the hyphen. Again not all the compounds with foreign words or abbreviations lose the hyphen. As to the two examples above, 'web' without the hyphen is much more frequent than 'edb'.

In step 3 of the lemma selection process word forms found in the corpora are normalised, i.e. capitalised parts and special symbols are removed. Thus some of the different word forms can be grouped as being forms of the same lemma. This normalisation fits very well together with the rules described above. Removing capital letters and hyphens from the word forms found in the corpora will often group together all the forms of a lemma. So the problem of recognition of words with a non-established spelling form can be reduced a lot.

But the recognition of loanwords without an established spelling form is not the only problem. After a lemma has been recognised, it has to be defined which of the spelling variants found shall be included in the lexicon. The STO-lexicon will not only contain the approved spelling forms (if any form has been approved for the lemma in question) but also frequent forms that are not approved by the Danish Language Council since future applications might want to parse and analyse texts where other spelling conventions are used. This means that a lemma might enter the database with three different spelling variants, e.g.

EDB-firma (eng. computer company)
edb-firma
edbfirma

of which only one should be used for generation. The non-approved forms will be marked as such to keep them from being used in text generation.

The decision of which spelling variants of the lemma should be included in the lexicon depends entirely on corpus occurrences. Old forms might be obsolete but if they are still found in many texts it seems relevant to include them. Very new forms – appearing a few times in the newest texts – might soon disappear again or might be the established norm in the future. For lemmas that have no approved spelling form, the most common form in up-to-date texts should be used for generation.

The problem of deciding which form should be the established form of the lemma is closely related to another problem at the encoding stage, namely how to decide the inflectional pattern of a newly borrowed word with no established morphology. The highly frequent words often show one or two typical morphological patterns and some times other less frequent inflectional variants too. The task of the encoding team is to decide which inflectional alternatives should be included in the lexicon.

In cases where it is hard to decide which form should be chosen as the 'approved' form of a lemma and which inflectional pattern should be the established one, we cooperate with the Danish Language Council. They are given a lemma list of the dubious cases and will then decide on which word form(s) and inflectional behaviour should be the established ones.

5. Concluding remarks

The issues dealt with above show that the automatic lemma selection for a computational lexicon for domain specific language encounters some problems that are not easily solved without human interference. Though the lemma selection process revealed some problems of how to keep track of words that appear both as general language words and domain specific words, it also became apparent that the normalisation which is part of the lemma selection was very suitable for the problem of loanwords that have no established lemma or inflection.

6. References

Braasch, A. (2002). ' Current Developments of STO – the Danish Lexicon Project for NLP and HLT applications' in: *Proceedings from the Third International Conference on Language Resources and Evaluation,* Las Palmas.

Braasch, A. & S. Olsen (2000). 'Towards a Strategy for a Representation of Collocations – Extending the Danish PAROLE-lexicon' in: *Proceedings form the Second International Conference on Language Resources and Evaluation,* Athens.

Braasch, A., A. B. Christensen, S. Olsen & B.S. Pedersen (1998). 'A Large-Scale Lexicon for Danish in the Information Society', in*: Proceedings from First International Conference on Language Resources & Evaluation*, Granada.

Dansk Sprognævn (1996). *Retskrivningsordbogen*, Copenhagen.

LE-PAROLE (1998) *Danish Lexicon Documentation*. *Internal report*. Center for sprogteknologi, Copenhagen.

Ruus, H. (1995). *Danske kerneord*, Tusculanum, Copenhagen.