Finding the Correct Interpretation of Swedish Compounds a Statistical Approach

Jonas Sjöbergh, Viggo Kann

KTH Nada
SE-100 44 Stockholm, Sweden
{jsh, viggo}@nada.kth.se

Abstract

This paper treats compound splitting for Swedish, where compounding is productive and very common. A method for splitting compounds and several methods for choosing the correct interpretation of ambiguous compounds are presented. 99% of all compounds are split, 97% of these are correctly interpreted.

1. Introduction

Swedish is a compounding language, where compounding is productive and very common. An introduction to Swedish word formation can be found for instance in (Thorell, 1981). An overview of Swedish written in English can be found in (Comrie, 1990).

Compounds are often ambiguous, as in ”bildrulle” (bad driver or a roll of film) which could be made from ”bil-drulle” (”car fool”) or ”bild-rulle” (”picture roll”). Many possible interpretations are unlikely and would be discarded by a human reader. Usually, humans have no problem deciding the correct interpretation of compounds in context, but taken out of context many compounds have several reasonable interpretations.

Finding the correct way to split a compound is in many ways similar to word sense disambiguation. Most of the ambiguity in the sense of a compound will be removed when the correct way to split the compound is given. Many natural language applications need or benefit from the ability to split compounds, including grammar checking, information retrieval, hyphenation, speech recognition, machine translation and text clustering.

Most common ways of splitting compounds use dictionary lookup. Other ways include splitting words on n-grams of characters that do not occur in non-compound words (Kokkinakis and Johansson Kokkinakis, 1999). A parallel corpus can also be useful for compound splitting (Brown, 2002). Research on compound splitting has been done for several different languages, for instance German (Koehn and Knight, 2003) and Norwegian (Johannessen and Hauglin, 1996). In (Dura, 1998) a linguistic approach to automatic analysis of Swedish compounds is studied.

In this paper a statistical approach is used, Swedish compounds were automatically split by a modified spell-checker. Then several different methods of choosing the correct interpretation were evaluated on manually annotated test data.

2. Evaluation Method

Evaluation was done by manually annotating the correct interpretation of all compounds found by the splitter in a test text. This test data was 50 000 words of written Swedish, taken from the Stockholm-Umeå Corpus (Ejerhed et al., 1992). There were 3 500 compounds, of which 1 300 were ambiguous (i.e. had more than one suggestion from the splitter). Unless otherwise stated, all accuracy figures in the evaluation are computed only on the ambiguous compounds.

Some (less than 1%) compounds were not found by the splitter. Most of these contained a proper noun and would have been unambiguous if the proper noun was recognized.

For 99% of the unambiguous compounds the suggestion from the splitter was correct. For 99% of the ambiguous compounds one of the suggestions was the correct interpretation. The remaining 1% usually contained a proper noun.

A lot of words were split by the splitter even though they were not compounds. These words were ignored in the tests. A simple lexicon lookup in the splitter to avoid splitting words that occur as a (non-compound) word in the lexicon would remove most of this overgeneration. Some types of words that were arguably compounds were also ignored. These were mostly Swedish family names, which often consist of two words, often from nature, i.e. “Sjö + berg” (“lake + mountain”). Such words are generally easy to disambiguate, but there is seldom a reason to do so.

3. Finding Possible Interpretations

We modified the spelling error detection program Stava (Domeij et al., 1994) to find all possible splittings of compounds. Stava uses three word lists:

1. the *individual word list*, containing words that cannot be part of a compound at all,
2. the *last part list*, containing words that can end a compound or be an independent word,
3. the *first part list*, containing altered word stems that can form the first or middle component of a compound.

When a word is checked, the algorithm consults the lists in the order illustrated in Figure 1. In the trivial case, the input word is found directly in the *individual word list* or the *last part list*. If the input word is a compound, only its last component is confirmed in the *last part list*. Then the *first part list* is looked up to acknowledge its first part. If the compound has more components than two, a recursive consultation is performed. The algorithm optionally inserts

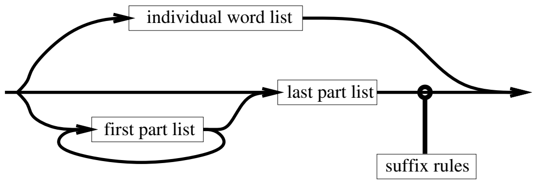

Figure 1: Look-up scheme for compound splitting.

an extra -*s*- between parts, to account for the fact that an extra -*s*- is generally inserted between the second and third components. As in "fotbollslag" ("fot-boll-s-lag", "football team").

Unlike Stava, where the algorithm stops the search when a possible splitting of the word is found, our algorithm instead proceeds to find all possible interpretations of the compound.

4. Choosing the Correct Interpretation

We have tried several methods of choosing the correct interpretation of ambiguous compounds. Below we describe the methods and their performance one by one and finally a hybrid method combining the methods.

4.1. Baseline, the Number of Components

A simple approach is to choose the suggestion with as few components as possible, selecting the suggestion with the longest last component in case of a tie (Kann et al., 2001). This approach works quite well, giving an accuracy of 90% on the ambiguous compounds in our test corpus.

4.2. Semantic Context

Human readers often base their disambiguation decisions on the semantic topic of the text. In a text on photography "bild-rulle" (roll of film) will be preferred over "bil-drulle" (bad driver).

A very simple method to capture this was tested. For each compound the number of occurrences of all suggested compound components in the context were counted. All occurrences of a component within a 100 word window centered on the compound were counted. The occurrences were weighted by the distance to the compound, with occurrences close to the compound given higher weight. The suggestions were then ranked by the mean count of their components and the suggestion with the highest mean was selected.

This method did not perform very well, only 72% accuracy. The problem was that compound components rarely occurred in the context. Using stemming on the components (though it is not always clear what the stem of a compound head is) and on the context helped a little, 74% accuracy. One way to improve this method would be to look for words related to a compound component instead of only the component itself, e.g. using LSA or similar techniques. The method also probably needs to be combined with some method to remove suggestions with words that are common words but rare in compounds. One example would be "för" which is a highly ambiguous word in Swedish and very common as a preposition. It sometimes, but not very often, occurs in compounds and is very often suggested as a compound component by the splitter.

Though this method did not perform well it did find the correct interpretation of some compounds that no other method handled correctly. Using the context is the only way in general to find the correct interpretation when there are several "real" interpretations (interpretations in actual use) of a compound (though such compounds are not very common in normal texts).

4.3. Component Frequencies

Another method used by human readers is to consider how common "bil" is as a compound head compared to "bild". Other things being equal, we prefer "bil-drulle" over "bild-rulle" since "bil" is more common as a compound head. To use this strategy statistics on compound head and tail frequencies were gathered. The suggestion consisting of the most common compound components was then selected. This was done by selecting the suggestion with the highest geometric mean frequency of its components.

This method did not perform as well as the baseline. Again, this is caused by data sparseness. Many compound components never occur in the training data. Since the statistics on component frequencies was too sparse different sources were tried to find the component frequencies. The following data was tried: frequencies of all the words in a 1 million words corpus (instead of frequencies of compound components); a lexicon of 84 000 compounds; the compound lexicon with frequencies for the compounds found in the corpus; a few hundred compounds from the same domain as the test data, hand annotated with their correct interpretation.

The best result was achieved using the lexicon with frequencies from the corpus and adding all the words found in the corpus to the compound tail statistics. Of the 84 000 compounds in the lexicon, only 10 000 occurred in the corpus (with each compound on average occurring four times). About 14% of the compounds occurring in the corpus were found in the lexicon. The best accuracy achieved was 86%.

A method similar to this has been used on German compounds (Koehn and Knight, 2003).

4.4. Syntactic Context

Often it is clear from the context of a compound which word class the compound should belong to. If different suggestions result in different word classes for the compound, such contextual information can be used.

This was tested by tagging the tail of each suggestion (the tail of a compound determines its part of speech in Swedish) in the context where the compound occurred. The tagging was done using TnT (Brants, 2000), a statistical part of speech tagger. Then the compound was replaced by a dummy word and the sentence tagged again. By assigning the dummy word equal lexical probability for all tags found in the first step, TnT will select the most likely tag (by TnT's measure) based solely on the context. Any suggestion with another tag was then discarded and the baseline method was used to break ties.

This method did not work as well as the baseline, 86% accuracy. This could probably be raised a little by tailoring the tagset to this specific application, but this was not tested. The method is rarely applicable, since most suggestions have the same part of speech. Also, it is only based on the tail of the compound, so suggestions with the same tail cannot be disambiguated.

4.5. Part of Speech of Components

Some combinations of word classes are more common than others in compounds. Noun noun combinations are the most common (more than 25% of all compounds) while for instance pronoun pronoun combinations are extremely rare.

To use this information two part of speech taggers were created, one for compound heads and one for tails. Both were very naive, just a dictionary lookup and a few simple morphology rules allowed for heads (which often change slightly in compounds). No disambiguation at all was performed, all possible tags were kept. No contextual information was used. The resulting taggers are not very accurate (especially the head tagger makes a lot of errors), but this is not really a problem, as explained below.

For each head component of a suggested interpretation of a compound the probability of this head PoS and tail PoS combination was calculated. If there were several possible PoS tags for the head or tail, the most favorable combination was used. The probability of a suggestion was then computed as the product of these probabilities for all the heads of this suggestion, and the suggestion with the highest probability was chosen.

The head PoS and tail PoS combination probabilities were computed from automatically tagged (with the above taggers) compounds (with the correct interpretation known). Since the taggers make the same types of errors on the training data as on the test data, it is not a great problem that they make a lot of errors. If the taggers were better, for instance by disambiguating the components more, the method would likely work even better, though.

The method works quite well, 91% accuracy.

4.6. Character n-grams

Some character combinations never occur in a non-compound word, but can occur in the head/tail border of a compound. This property has been used to split compounds (Kokkinakis and Johansson Kokkinakis, 1999). Most compounds do not contain such character combinations, though.

A method inspired by this was developed. Though not all head/tail borders contain character combinations not possible in non-compounds the character combinations common on such borders are often less common internally in compound components. The frequencies of all character 4-grams in compound heads and tails (not overlapping a head/tail border) were counted. This was counted in a lexicon of compounds, with frequencies added, by counting their occurrences in a corpus.

For each suggestion from the compound splitter all frequencies of the character 4-grams spanning the suggested splits were added. The suggestion with the lowest sum was then selected. This suggestion thus has the splits located at the positions most unlikely to not contain a split (low frequency of these 4-grams in compound components), i.e. the most likely positions.

Method	Accuracy (%)
Number of components	90
Semantic context	74
Component frequencies	86
Syntactic context	86
Part of speech of components	91
Character n-grams	91
Hybrid	94

Table 1: Accuracy on ambiguous compounds only.

This was the best method, with 91% accuracy. It is of course possible to use n-grams of different lengths than 4, or combinations of different lengths, though only 4-grams were tested.

4.7. Ad Hoc Rules

A few ad hoc rules were constructed to deal with some error sources common to most methods. One example is dealing with common inflectional suffixes that also happen to be a possible word. These words are very rare in compounds, but are often suggested by the splitter.

In Swedish three identical consonants in a row over a split in a compound are merged to two consonants. So "vinnyheter" could be made from "vin-nyheter" ("new wines") or "vinn-nyheter" ("winning news", unlikely interpretation). Many of the methods mentioned will always select the three consonant interpretation. For instance, it always has unlikely character 4-grams, since the frequency of three identical consonants is 0. Since most methods do not handle these suggestions in an intelligent way, an ad hoc rule was created to always select the most common interpretation (which is the two consonant interpretation). Some methods, for instance the semantic context method in section 4.2., do in fact handle these types of compounds intelligently, and thus does not need this ad hoc rule.

4.8. Hybrid Methods

Since the methods make errors on different compounds, combining the methods should give higher accuracy than the individual methods. Almost all compounds have the correct interpretation suggested by at least one method.

One simple hybrid system was tested. It combined the n-gram method in section 4.6. and the method using the PoS of the components in section 4.5. Two ad hoc rules briefly discussed in section 4.7. were also used to handle some common problems. The suggestion from the n-gram method was normally used, since this was the most accurate method. If the PoS method had a probability for its own suggestion that was more than five times higher than the probability of the suggestion from the n-gram method, the suggestion from the PoS method was used instead.

This method had an accuracy of 94% on the ambiguous compounds. This is much better than either method alone, see Table 1. It makes almost 40% less errors than the baseline method. This amounts to a total accuracy of 97% on all compounds.

5. Error Analysis

The errors made by the best method (the hybrid method in section 4.8.) can be divided into four categories:

The first category is splitting the compound in all the correct places, but splitting some components too much. 7% of the errors were of this type. Usually, this type of error occurs because the correct interpretation is not suggested by the splitter. Otherwise the correct interpretation is (usually) preferred, since it has fewer components (both the combined methods in the hybrid method have a strong bias towards fewer components).

The second category is splitting the compound only in correct places but failing to split some components that are actually compounds. This is the most common error type, with 70% of the errors. These errors could in large part be avoided by removing compound words from the lexicon of the splitter.

In Swedish three identical consonants in a row over a split in a compound is merged to two consonants (see section 4.7.). Making the wrong choice between the two or three consonant interpretation when splitting compounds causes 5% of the errors.

The final category is simply splitting the compound in the wrong position(s). 18% of the errors belong to this category.

6. Conclusions

Automatic methods for compound splitting can give good results: finding 99% of all compounds, and the correct interpretation for 97% of these. These methods require very few resources. In these experiments a lexicon of (correctly split) compounds, two manually constructed rules for problematic cases, a statistical part of speech tagger and unannotated text was all that was used.

If large amounts of manually annotated data was available, such as a corpus with the compound interpretations added, even better results could likely be achieved. Many of the methods use statistics based on compound frequencies, which was not really available. Using a lexicon is not a realistic substitute, since uncommon compounds are too common in a lexicon. Using free text is not good either, since then the correct interpretation of the compounds is not known. In our experiments adding frequencies from free text to the lexicon was usually used as a compromise, which worked well despite a lot of compounds in both the lexicon and the text being ignored this way.

All evaluations were done on Swedish, but all methods, except these specific ad hoc rules, should work on any language with similar properties, such as German or other Scandinavian languages.

Acknowledgments

This work has been funded by The Swedish Agency for Innovation Systems (VINNOVA), The Swedish Research Council (VR) and The Royal Institute of Technology (KTH).

7. References

Brants, Thorsten, 2000. TnT – a statistical part-of-speech tagger. In *Proceedings of the 6th Applied NLP Conference, ANLP-2000*. Seattle, USA.

Brown, Ralf, 2002. Corpus-driven splitting of compound words. In *Proceedings of the 9th International Conference on Theoretical and Methodological Issues in Machine Translation*. Keihanna, Japan.

Comrie, Bernard, 1990. *The World's Major Languages*. Oxford University Press.

Domeij, Rickard, Joachim Hollman, and Viggo Kann, 1994. Detection of spelling errors in Swedish not using a word list en clair. *J. Quantitative Linguistics*, 1:195–201.

Dura, Elżbieta, 1998. *Parsing Words*. Ph.D. thesis, Göteborg University, Göteborg, Sweden.

Ejerhed, Eva, Gunnel Källgren, Ola Wennstedt, and Magnus Åström, 1992. The linguistic annotation system of the Stockholm-Umeå Corpus project. Technical report, Department of General Linguistics, University of Umeå (DGL-UUM-R-33), Umeå, Sweden.

Johannessen, Janne Bondi and Helge Hauglin, 1996. An automatic analysis of Norwegian compounds. In *Papers from the 16th Scandinavian Conference of Linguistics*. Turku / Åbo, Finland.

Kann, Viggo, Rickard Domeij, Joachim Hollman, and Mikael Tillenius, 2001. Implementation aspects and applications of a spelling correction algorithm. In L. Uhlirova, G. Wimmer, G. Altmann, and R. Koehler (eds.), *Text as a Linguistic Paradigm: Levels, Constituents, Constructs. Festschrift in honour of Ludek Hrebicek*, volume 60 of *Quantitative Linguistics*. Trier, Germany: WVT, pages 108–123.

Koehn, Philipp and Kevin Knight, 2003. Empirical methods for compound splitting. In *Proceedings of EACL 2003*. Budapest, Hungary.

Kokkinakis, Dimitrios and Sofie Johansson Kokkinakis, 1999. Sense-tagging at the cycle-level using GLDB. Technical Report GU-ISS-99-4, Department of Swedish, Göteborg University.

Thorell, Olof, 1981. *Svensk Ordbildningslära*. Esselte Studium.