Decompounding query keywords from compounding languages

Enrique Alfonseca
Google Inc.
ealfonseca@google.com

Slaven Bilac
Google Inc.
slaven@google.com

Stefan Pharies
Google Inc.
stefanp@google.com

Abstract

Splitting compound words has proved to be useful in areas such as Machine Translation, Speech Recognition or Information Retrieval (IR). Furthermore, real-time IR systems (such as search engines) need to cope with noisy data, as user queries are sometimes written quickly and submitted without review. In this paper we apply a state-of-the-art procedure for German decompounding to other compounding languages, and we show that it is possible to have a single decompounding model that is applicable across languages.

1 Introduction

Compounding languages (Krott, 1999), such as German, Dutch, Danish, Norwegian, Swedish, Greek or Finnish, allow the generation of complex words by merging together simpler ones. So, for instance, the *flower bouquet* can be expressed in German as *Blumensträuße*, made up of *Blumen* (flower) and *sträuße* (bouquet), and in Finnish as *kukkakimppu*, from *kukka* (flower) and *kimppu* (bunch, collection). For many language processing tools that rely on lexicons or language models it is very useful to be able to decompose compounds to increase their coverage and reduce out-of-vocabulary terms. Decompounders have been used successfully in Information Retrieval (Braschler and Ripplinger, 2004), Machine Translation (Brown, 2002; Koehn and Knight, 2003) and Speech Recognition (Adda-Decker et al., 2000). The Cross Language Evaluation Forum (CLEF) competitions have shown that very simple approaches can produce big gains in Cross Language Information Retrieval (CLIR) for German and Dutch (Monz and de Rijke, 2001) and for Finnish (Adafre et al., 2004).

When working with web data, which has not necessarily been reviewed for correctness, many of the words are more difficult to analyze than when working with standard texts. There are more words with spelling mistakes, and many texts mix words from different languages. This problem exists to a larger degree when handling user queries: they are written quickly, not paying attention to mistakes. However, being able to identify that *achzigerjahre* should be decompounded as *achzig+jahre* (where *achzig* is a misspelled variation of *achtzig*) is still useful in obtaining some meaning from the user query and in helping the spelling correction system. This paper evaluates a state-of-the-art procedure for German splitting (Alfonseca et al., 2008), robust enough to handle query data, on different languages, and shows that it is possible to have a single decompounding model that can be applied to all the languages under study.

2 Problem definition and evaluation settings

Any set of query keywords contains a large amount of noisy data, such as words in foreign languages or misspelled words. In order to be robust enough to handle this kind of corpus, we require the following for a decompounder: first, obviously, compounds should be split, and non-compounds should be left untouched. This also applies if they are misspelled. Unknown words or words involving a part

Proceedings of ACL-08: HLT, Short Papers (Companion Volume), pages 253–256,
Columbus, Ohio, USA, June 2008.

in a foreign language are split if there is a plausible interpretation of them being a compound word. An example is *Turingmaschine* (Turing machine) in German, where Turing is an English word. Finally, words that are not really grammatical compounds, but due to the user forgetting to input the blankspace between the words (like *desktopcomputer*) are split.

For the evaluation, we have built and manually annotated gold standard sets for German, Dutch, Danish, Norwegian, Swedish and Finnish from fully anonymized search query logs. Because people do not use capitalization consistently when writing queries, all the query logs are lowercased. By randomly sampling keywords we would get few compounds (as their frequency is small compared to that of non-compounds), so we have proceeded in the following way to ensure that the gold-standards contain a substantial amount of compounds: we started by building a very naive decompounder that splits a word in several parts using a frequency-based compound splitting method (Koehn and Knight, 2003). Using this procedure, we obtain two random samples with possibly repeated words: one with words that are considered non-compounds, and the other with words that are considered compounds by this naive approach. Next, we removed all the duplicates from the previous list, and we had them annotated manually as compounds or non-compounds, including the correct splittings. The sizes of the final training sets vary between 2,000 and 3,600 words depending on the language. Each compound was annotated by two human judges who had received the previous instructions on when to consider that a keyword is a compound. For all the languages considered, exactly one of the two judges was a native speaker living in a country where it is the official language[1]. Table 1 shows the percentage of agreement in classifying words as compounds or non-compounds (Compound Classification Agreement, CCA) for each language and the Kappa score (Carletta, 1996) obtained from it, and the percentage of words for which also the decomposition provided was identical (Decompounding Agreement, DA). The most common source of disagreement were long words that could be split into two or more parts.

[1]This requisite is important because many queries contain novel or fashionable words.

Language	CCA	Kappa	DA
German	93%	0.86	88%
Dutch	96%	0.92	96%
Danish	89%	0.78	89%
Norwegian	93%	0.86	81%
Swedish	96%	0.92	95%
Finnish	92%	0.84	89%

Table 1: Inter-judge agreement metrics.

Language	Morphemes
German	∅,-e,+s,+e,+en,+nen,+ens,+es,+ns,+er
Dutch	∅,-e,+s,+e,+en
Danish	∅,+e,+s
Norwegian	∅,+e,+s
Swedish	∅,+o,+u,+e,+s
Finnish	∅

Table 2: Linking morphemes used in this work.

The evaluation is done using the metrics precision, recall and accuracy, defined in the following way (Koehn and Knight, 2003):

- Correct splits: no. of compounds that are split correctly.
- Correct non-splits: no. non-compounds that are not split.
- Wrong non-splits: no. of compounds and are not split.
- Wrong faulty splits: no. of compounds that are incorrectly split.
- Wrong splits: no. of non-compounds that are split.

$$Precision = \frac{\text{correct splits}}{\text{correct splits} + \text{wrong faulty splits} + \text{wrong splits}}$$

$$Recall = \frac{\text{correct splits}}{\text{correct splits} + \text{wrong faulty splits} + \text{wrong non-splits}}$$

$$Accuracy = \frac{\text{correct splits}}{\text{correct splits} + \text{wrong splits}}$$

3 Combining corpus-based features

Most approaches for decompounding can be considered as having this general structure: given a word w, calculate every possible way of splitting w in one or more parts, and score those parts according to some weighting function. If the highest scoring splitting contains just one part, it means that w is not a compound.

For the first step (calculating every possible splitting), it is common to take into account that modifiers inside compound words sometimes need *linking morphemes*. Table 2 lists the ones used in our system (Langer, 1998; Marek, 2006; Krott, 1999).

Method	Precision	Recall	Accuracy
Never split	-	0.00%	64.09%
Geometric mean of frequencies	39.77%	54.06%	65.58%
Compound probability	60.41%	**80.68%**	76.23%
Mutual Information	**82.00%**	48.29%	80.52%
Support-Vector Machine	**83.56%**	**79.48%**	**87.21%**

Table 3: Results of the several configurations.

Language	P	R	A
German	83.56%	79.48%	87.21%
Dutch	78.99%	76.18%	83.45%
Danish	81.97%	87.12%	85.36%
Norwegian	88.13%	93.05%	90.40%
Swedish	83.34%	92.98%	87.79%
Finnish	90.79%	91.21%	91.62%

Table 4: Results in all the different languages.

Concerning the second step, there is some work that uses, for scoring, additional information such as rules for cognate recognition (Brown, 2002) or sentence-aligned parallel corpora and a translation model, as in the full system described by Koehn and Knight (2003). When those resources are not available, the most common methods used for compound splitting are using features such as the geometric mean of the frequencies of compound parts in a corpus, as in Koehn and Knight (2003)'s back-off method, or learning a language model from a corpus and estimating the probability of each sequence of possible compound parts (Schiller, 2005; Marek, 2006). While these methods are useful for several applications, such as CLIR and MT, they have known weaknesses, such as preferring a decomposition if a compound part happens to be very frequent by chance, in the case of the frequency-based method, or the preference of decompositions with the least possible number of parts, in the case of the probability-based method.

Alfonseca et al. (2008) describe an integration of the previous methods, together with the Mutual Information and additional features obtained from web anchor texts to train a supervised German decompounder that outperforms the previous methods used as standalone. The geometric mean of the frequencies of compound parts and the probability estimated from the language model usually attain a high recall, given they are based on unigram features which are easy to collect, but they have some weaknesses, as mentioned above. On the other hand, while Mutual Information is a much more precise metric, it is less likely to have evidence about every single possible pair of compound parts from a corpus, so it suffers from low recall. A combination of all these metrics into a learning model is able to attain a high recall. An ablation study, reported in that paper, indicated that the contribution of the web anchor texts is minimal, so in this study we have just kept the other three metrics. Table 3 shows the results reported for German, training (i.e. counting frequencies and learning the language model) on the query keywords, and running a 10-fold cross validation of a SVM with a polynomial kernel using the German gold-standard. The supervised system improves over the single unsupervised metrics, attaining simultaneously good recall and precision metrics.

4 Experiments and evaluation

The first motivation of this work is to test whether the results reported for German are easy to reproduce in other languages. The results, shown in Table 4, are very similar across languages, having precision and recall values over 80% for most languages. A notable exception is Dutch, for which the inter-judge agreement was the highest, so we expected the set of words to be easier to classify. An analysis of the errors reported in the 10-fold cross-validation indicates that most errors in Dutch were wrong non-splits (in 147 cases) and wrong splits (in 139 cases), with wrong faulty splits happening only in 20 occasions. Many of the wrong splits are location names and trademarks, like *youtube*, *piratebay* or *smallville*.

While the supervised model gives much better results than the unsupervised ones, it still requires the construction of a goldstandard from which to train, which is usually costly. Therefore, we ran another experiment to check whether the models trained from some languages are applicable to other languages. Table 5 shows the results obtained in this case, the last column indicating the results when the model is trained from the training instances from all the other languages together. For each row, the highest value and those which are inside its 95% confidence interval are highlighted. Interestingly, apart from a few exceptions, the results are rather good for all the pairs of training and test language.

	Language for training						
	de	**nl**	**da**	**no**	**sv**	**fi**	**others**
de	P:83.56	P:78.69	P:74.96	P:**88.93**	P:82.72	P:**89.69**	P:80.89
	R:79.48	R:75.48	R:**92.77**	R:89.26	R:**90.79**	R:89.96	R:76.07
	A:87.21	A:82.76	A:83.53	A:**90.31**	A:86.53	A:**90.82**	A:88.15
nl	P:79.52	P:78.99	P:76.93	P:**92.81**	P:85.67	P:**90.98**	P:77.53
	R:75.74	R:76.18	R:**89.02**	R:55.08	R:**87.15**	R:86.73	R:76.54
	A:87.77	A:83.45	A:83.21	A:**91.00**	A:86.47	A:88.95	A:82.32
da	P:82.21	P:**90.86**	P:81.97	P:90.61	P:85.52	P:**92.65**	P:76.28
	R:45.01	R:42.94	R:87.12	R:80.25	R:81.41	R:82.46	R:**94.84**
	A:78.95	A:74.78	A:85.36	A:**89.30**	A:83.70	A:87.55	A:84.60
no	P:68.23	P:70.18	P:74.85	P:88.13	P:82.25	P:**90.08**	P:88.78
	R:83.33	R:87.18	R:**96.67**	R:93.05	R:94.21	R:91.84	R:90.88
	A:83.77	A:80.67	A:84.18	A:90.40	A:87.24	A:**91.41**	A:89.85
sv	P:76.57	P:77.33	P:76.31	P:89.00	P:83.34	P:**90.81**	P:83.89
	R:79.76	R:81.79	R:**94.66**	R:90.41	R:92.98	R:90.86	R:92.05
	A:87.18	A:83.38	A:84.57	A:89.67	A:87.79	A:**91.38**	A:87.69
fi	P:74.12	P:74.50	P:75.93	P:88.71	P:83.54	P:**90.79**	P:**90.70**
	R:80.12	R:81.67	R:**95.39**	R:91.46	R:92.70	R:91.21	R:90.62
	A:85.93	A:81.98	A:84.51	A:90.07	A:87.52	A:**91.62**	A:**91.18**

Table 5: Result training and testing in different languages.

Thus, the use of features like frequencies, probabilities or mutual information of compound parts is truly language-independent and the models learned from one language can safely be applied for decompounding a different language without the need of annotating a gold-standard for it.

Still, some trends in the results can be observed: training with the Danish corpus produced the best results in terms of recall for all the languages, but recall for Danish still improved when we trained on data from all languages. We believe that this indicates that the Danish dataset contains items with a more varied sets of feature combinations, so that the models trained from it have a good coverage on different kinds of compounds, but models trained in other languages are not able to identify many of the compounds in the Danish dataset. Concerning precision, training with either the Norwegian or the Finnish data produced very good results for most languages. This is consistent with the monolingual experiments (see Table 4) in which these languages had the best results. We believe these trends are probably due to the quality of the training data. Interestingly, the size of the training data is not so relevant, as most of the best results are not located at the last column in the table.

5 Conclusions

This paper shows that a combination of several corpus-based metrics for decompounding, previously applied to German, with big improvements with respect to other state-of-the-art systems, is also useful for other compounding languages. More interestingly, models learned from a goldstandard created for some language can be applied to other languages, sometimes producing better results than when a model is trained and tested in the same language. This should alleviate the fact that the proposed system is supervised, as there should just be the need of creating a goldstandard in one language in order to train a generic decompounder, thus facilitating the availability of decompounders for smaller languages like Faroese. For future work, we plan to investigate more deeply how the quality of the data affects the results, with a more detailed error analysis. Other open lines include exploring the addition of new features to the trained models.

References

S.F. Adafre, W.R. van Hage, J. Kamps, G.L. de Melo, and M. de Rijke. 2004. The University of Amsterdam at CLEF 2004. *CLEF 2004 Workshop*, pages 91–98.

M. Adda-Decker, G. Adda, and L. Lamel. 2000. Investigating text normalization and pronunciation variants for German broadcast transcription. In *ICSLP-2000*.

E. Alfonseca, S. Bilac, and S. Pharies. 2008. German decompounding in a difficult corpus. In *CICLING*.

M. Braschler and B. Ripplinger. 2004. How effective is stemming and decompounding for german text retrieval? *Information Retrieval*, 7:291–316.

R.D. Brown. 2002. Corpus-driven splitting of compound words. In *TMI-2002*.

J. Carletta. 1996. Assessing agreement on classification tasks: the Kappa statistics. *Computational Linguistics*, 22(2):249–254.

P. Koehn and K. Knight. 2003. Empirical methods for compound splitting. In *ACL-2003*.

A. Krott. 1999. Linking elements in compounds. LINGUIST, 7 Oct 1999. http://listserv.linguistlist.org/cgi-bin/wa?A2=ind9910a&L=linguist&P=6009.

S. Langer. 1998. Zur Morphologie und Semantik von Nominalkomposita. *Tagungsband der 4. Konferenz zur Verarbeitung naturlicher Sprache (KONVENS)*.

T. Marek. 2006. Analysis of german compounds using weighted finite state transducers. Technical report, BA Thesis, Universität Tbingen.

C. Monz and M. de Rijke. 2001. Shallow morphological analysis in monolingual information retrieval for Dutch, German and Italian. In *CLEF-2001*.

A. Schiller. 2005. German compound analysis with wfsc. In *Finite State Methods and NLP 2005*.