The impact of phrases in document clustering for Swedish

Magnus Rosell and **Sumithra Velupillai**
KTH Nada
100 44 Stockholm
Sweden
{rosell, sumithra}@nada.kth.se

Abstract

We have investigated the impact of using phrases in the vector space model for clustering documents in Swedish in different ways. The investigation is carried out on two text sets from different domains: one set of newspaper articles and one set of medical papers.

The use of phrases do not improve results relative the ordinary use of words. The results differ significantly between the text types. This indicates that one could benefit from different text representations for different domains although a fundamentally different approach probably would be needed.

1 Introduction

For document clustering one normally uses the vector space model to represent texts. It is based on the distribution of single words over the texts in a set. We have investigated the impact of introducing phrases in this representation for Swedish in different ways and in different domains. Our hypothesis was that phrases would improve results and that the improvement would be greater for the medical papers than for the newspaper articles as we believe that phrases carry more significance in the medical domain.

To calculate similarity between documents with respect to their phrases we use a word trie (in one set of experiments). This approach has a lot in common with the method presented in (Hammouda and Kamel, 2004). They show improvements in clustering results on web pages using phrases combined with single words, using other algorithms than we. Another related method is the Phrase-Intersection Clustering method which has been proven efficient on web pages (Zamir and Etzioni, 1998). It is based on word-n-grams rather than phrases.

2 Text sets

We have used a set of 2500 newspaper articles from KTH News Corpus (AB) (Hassel, 2001) and a set of 2422 medical papers from Läkartidningen[1] (Med). In Table 1 some statistics for the sets are given.

We need categorizations of the text sets for the evaluation. The newspaper articles have been categorized by the paper into five sections such as Economy and Sports etc.

The medical papers are categorized with The Medical Subject Headings (MeSH) thesaurus[2]. This thesaurus is (poly)hierarchical with a term and a unique code at each place in it. The terms are not unique and may occur at several places in the hierarchy. There are 15 *broad headings* at the top level.

[1] http://www.lakartidningen.se/

[2] http://www.nlm.nih.gov/mesh/meshhome.html

S. Werner (ed.), *Proceedings of the 15th NODALIDA conference, Joensuu 2005*, Ling@JoY 1, 2006, pp. 173–179
ISBN 952-458-771-8, ISSN 1796-1114. http://ling.joensuu.fi/lingjoy/

Text Set	Categories	Documents	Words	Unique Words
AB	5	2500	119401	5896
Med	15, 814	2422	4383169	26102

Table 1: Text Sets

Each paper has one or more terms from the thesaurus assigned to it. This categorization is very extensive, but also very hard to handle for clustering evaluation. Hence we have made four attempts to flatten and disambiguate it so that each paper belongs to only one of a set of non overlapping categories.

We have made three categorizations where we try to put each document into one of 15 categories corresponding to the 15 broad headings. The first, which we call General, is constructed by choosing the broad heading to which most of the MeSH-terms assigned to the paper belongs.

By choosing the broad heading under which the most specific term (the term deepest into the hierarchy) is found for each paper we have constructed the second categorization, which we call Specific.

Many of the papers have as one of the terms assigned to it one or several broad headings. In the third categorization we have chosen this (always one) as the categorization of those papers. The other papers are categorized using the same system as for our categorization Specific. We call this categorization Combined.

We have made a fourth categorization which we call Term. In this every paper is assigned the MeSH-term that has the highest frequency among the terms assigned to it. This leads to a categorization with 817 categories.

The categorizations General and Combined are those that seem most trustworthy. A paper may probably have a very specific term assigned without having its broad heading as the general focus (see Specific). Terms at different levels of the MeSH-hierarchy probably make up an unequal categorization (see Term).

3 Linguistics

We used the grammar checking program Granska[3] to extract nominal phrases from the texts and a stemmer (Carlberger et al., 2001) to stem all words. To prevent very similar but not identical phrases to be deemed unsimilar we removed stopwords within the phrases as well as from the single words.

Swedish solid compounds often correspond to phrases (or compounds) in other languages. We use the spell checking program Stava (Kann et al., 2001) to split them. An earlier study (Rosell, 2003) has proven this to improve clustering results for newspaper articles. We also try to represent the split compounds as phrases and try to split compounds within phrases (see Section 5).

4 Similarity

When calculating the similarity between two documents using phrases two natural alternatives are at hand. Either one chooses to deem phrases similar only if they are identical or one looks at the overlap of words between them. We have tried both. In the first case we have calculated the weight for each phrase in a document as the frequency of its appearance in that document multiplied by the sum of the idf-weight for the single words in it.

To find the overlaps of phrases in documents we have built a trie based on words for each document from the phrases appearing in them. Each phrase is put into

[3] http://www.nada.kth.se/theory/projects/granska/

the trie in its entire and with all but the first word, with all but the first two words, etc. In each node of the trie we save the number of times it has been reached. To calculate the overlap of phrases between two documents we follow all common paths in the tries and multiply relative appearances in each node weighted by the sum of the idf-weights for the words along the path.[4]

5 Representations

From the phrases and single words we built several different representations. Refer to Table 2 through this section.

Combining all the described possibilities (full phrases or overlap, using split compounds as phrases or not, and split compounds within phrases or not) we get eight different representations based on phrases. By combining[5] these with the ordinary single word representation with split compounds we get eight more. This gives 16 representations (representations 3 through 18 in Table 2). We also made the reference representation (only words, 1) and the representation where solid compounds have been split (2), giving in total 18 different representations.

Finally, for comparison we also try a random “clustering” (Rand) and in the evaluation we present the theoretical worst (Worst) and best (Best) possible results (see Sections 7 and 8).

6 Clustering algorithm

The clusterings have been made using the divisive algorithm Bisecting K-Means (Steinbach et al., 2000) which splits the worst cluster (i.e. largest) in two, using K-Means, until the desired number of clusters are reached. We have let the K-Means algorithm iterate ten times and for each split we ran it five times and picked the best split (evaluated using the similarity measure). Average results are calculated over ten runs to ten clusters for each representation.

7 Evaluation

As we compare different representations we use extrinsic evaluation measures that requires a categorization of the the same text set to compare with. Among the extrinsic evaluation measures that have been used for text clustering are *the purity* and *the entropy*. These measures are well suited for evaluation of single clusters, but for evaluation of whole clusterings *the mutual information* is better. (Strehl et al., 2000)

Consider a text set N with n texts. Let C be a clustering with γ clusters, c_1 through c_γ. By n_i we mean the number of texts in cluster c_i ($\sum_{i=1}^{\gamma} n_i = n$). Similarly, let K be a categorization with κ categories, $k^{(1)}$ through $k^{(\kappa)}$ and let $n^{(j)}$ denote the number of texts in category $k^{(j)}$.

The γ by κ matrix M describes the distribution of the texts over both C and K; that is $m_i^{(j)}$ is the number of texts that belong to c_i and $k^{(j)}$.

The mutual information of clustering C and categorization K is:

$$MI(C,K) = \sum_{i=1}^{\gamma}\sum_{j=1}^{\kappa}\frac{m_i^{(j)}}{n}\log(\frac{m_i^{(j)}n}{n_i n^{(j)}}) \qquad (1)$$

A theoretical tight upper bound is $MI_{max}(C,K) = \log(\kappa\gamma)/2$, the mean of the theoretical maximal entropy of the clustering and the categorization. By dividing the mutual information by this we get a normalized measure. (Strehl, 2002)

This normalization is theoretical and particular for each clustering-categorization-setting. We want to compare results on different such settings, with different text

[4] Compare with Phrase-Intersection Clustering in (Zamir and Etzioni, 1998).

[5] We use equal weight on the two different representations. In (Hammouda and Kamel, 2004) they try different weightings.

Repr.	Description			
Worst	The worst possible result			
Rand	Random partiton of the set – average for ten iterations			
Best	The best possible result			
1	Only words, stemming			
2	Only words, stemming and splitting of compounds			
3	P	PM	NSP	NSC
4	P	PM	NSP	SC
5	P	PM	SP	NSC
6	P	PM	SP	SC
7	P	POM	NSP	NSC
8	P	POM	NSP	SC
9	P	POM	SP	NSC
10	P	POM	SP	SC
11	P&W	PM	NSP	NSC
12	P&W	PM	NSP	SC
13	P&W	PM	SP	NSC
14	P&W	PM	SP	SC
15	P&W	POM	NSP	NSC
16	P&W	POM	NSP	SC
17	P&W	POM	SP	NSC
18	P&W	POM	SP	SC

Abbr.	Explanation
P	Similarity only between phrases
P&W	Similarity using both phrases and words
PM	Phrase-match
POM	Phrase-overlap-match
SP	Use splitted compounds as phrases
NSP	Do not use splitted compounds as phrases
SC	Split compounds within phrases
NSC	Do not split compounds within phrases

Table 2: Representations

sets, having varying clustering complexity/difficulty. Therefore we need to normalize with regard to something else.

Since we want to know how much introducing phrases improve results we use the result from a clustering using only words as a reference. By comparing the results with this reference we take the complexity of the different text sets into account.

There are two simple and reasonable ways of normalizing the result using the word clustering result, $MI(C_{word}, K)$. We can divide the result by it:

$$MI_{word}(C, K) = \frac{MI(C, K)}{MI(C_{word}, K)}, \qquad (2)$$

or we can divide the improvement by the maximum possible improvement from the word clustering result:

$$MI_{imp}(C, K) = \frac{MI(C, K) - MI(C_{word}, K)}{MI_{max}(C, K) - MI(C_{word}, K)} \qquad (3)$$

The first normalization is suitable when we have a decrease in performance. It puts the decrease in relation to the greatest possible decrease. The second normalization is suitable when we have an increase in performance.

8 Results

We present the results of our investigation in Tables 3 and 4. All values are average results over ten clusterings with standard deviation within parenthesis.

The first row of each part of these tables gives the results for the newspaper articles and the following the results on the medical papers compared to the different categorizations. In Table 4 we only give results for representations Term and General as the results for Combined, General and Specific are very similar.

The columns represent the different representations which were described in Section 2 and summarized in Table 2. In Table 3 we present the result for a random “clustering” (the average of 10 random partitions of the text set) and the theoretical worst and best possible results.

	Measures	Worst	Rand		Best	1		2	
AB	MI	0.000	0.009	(0.003)	2.822	0.947	(0.043)	1.093	(0.084)
	MI_{word}	−100.0%	−99.0%	(0.3%)	198.0%	0.0%	(4.6%)	15.4%	(8.9%)
	MI_{imp}	−50.5%	−50.0%	(0.2%)	100.0%	0.0%	(2.3%)	7.8%	(4.5%)
Combined	MI	0.000	0.038	(0.006)	3.614	0.407	(0.016)	0.415	(0.010)
	MI_{word}	−100.0%	−90.6%	(1.4%)	787.9%	0.0%	(4.0%)	2.0%	(2.4%)
	MI_{imp}	−12.7%	−11.5%	(0.2%)	100.0%	0.0%	(0.5%)	0.3%	(0.3%)
General	MI	0.000	0.041	(0.005)	3.614	0.478	(0.013)	0.486	(0.016)
	MI_{word}	−100.0%	−91.5%	(1.1%)	656.0%	0.0%	(2.7%)	1.7%	(3.4%)
	MI_{imp}	−15.2%	−13.9%	(0.2%)	100.0%	0.0%	(0.4%)	0.3%	(0.5%)
Specific	MI	0.000	0.038	(0.005)	3.614	0.396	(0.010)	0.397	(0.017)
	MI_{word}	−100.0%	−90.4%	(1.2%)	812.6%	0.0%	(2.6%)	0.1%	(4.2%)
	MI_{imp}	−12.3%	−11.1%	(0.1%)	100.0%	0.0%	(0.3%)	0.0%	(0.5%)
Term	MI	0.000	1.450	(0.008)	6.498	1.850	(0.023)	1.868	(0.018)
	MI_{word}	−100.0%	−21.6%	(0.5%)	251.2%	0.0%	(1.2%)	1.0%	(0.9%)
	MI_{imp}	−39.8%	−8.6%	(0.2%)	100.0%	0.0%	(0.5%)	0.4%	(0.4%)

Table 3: Text Clustering Results (stdv)

	Measures	3		4		5		6	
AB	MI	0.067	(0.020)	0.071	(0.017)	0.086	(0.024)	0.080	(0.032)
	MI_{word}	−93.0%	(2.1%)	−92.5%	(1.8%)	−91.0%	(2.6%)	−91.5%	(3.4%)
General	MI	0.112	(0.008)	0.117	(0.012)	0.028	(0.005)	0.030	(0.002)
	MI_{word}	−76.6%	(1.7%)	−75.4%	(2.5%)	−94.2%	(1.1%)	−93.7%	(0.4%)
Term	MI	1.547	(0.020)	1.547	(0.013)	0.574	(0.096)	0.585	(0.022)
	MI_{word}	−16.4%	(1.1%)	−16.4%	(0.7%)	−69.0%	(5.2%)	−68.4%	(1.2%)

	Measures	7		8		9		10	
AB	MI	0.095	(0.020)	0.150	(0.024)	0.071	(0.021)	0.058	(0.010)
	MI_{word}	−90.0%	(2.1%)	−84.1%	(2.5%)	−92.5%	(2.2%)	−93.9%	(1.0%)
General	MI	0.148	(0.011)	0.178	(0.015)	0.031	(0.005)	0.037	(0.025)
	MI_{word}	−69.0%	(2.4%)	−62.7%	(3.1%)	−93.5%	(1.0%)	−92.2%	(5.2%)
Term	MI	1.565	(0.033)	1.607	(0.027)	0.506	(0.045)	0.694	(0.269)
	MI_{word}	−15.4%	(1.8%)	−13.2%	(1.4%)	−72.6%	(2.5%)	−62.5%	(14.6%)

	Measures	11		12		13		14	
AB	MI	0.820	(0.051)	0.809	(0.057)	0.946	(0.078)	0.919	(0.100)
	MI_{word}	−13.4%	(5.4%)	−14.6%	(6.0%)	−0.1%	(8.2%)	−3.0%	(10.6%)
General	MI	0.148	(0.016)	0.168	(0.018)	0.210	(0.013)	0.216	(0.013)
	MI_{word}	−69.0%	(3.4%)	−64.8%	(3.8%)	−56.0%	(2.7%)	−54.9%	(2.8%)
Term	MI	1.562	(0.022)	1.566	(0.021)	1.314	(0.052)	1.336	(0.064)
	MI_{word}	−15.6%	(1.2%)	−15.4%	(1.1%)	−29.0%	(2.8%)	−27.8%	(3.5%)

	Measures	15		16		17		18	
AB	MI	0.746	(0.090)	0.734	(0.063)	0.954	(0.063)	0.940	(0.061)
	MI_{word}	−21.3%	(9.5%)	−22.5%	(6.7%)	0.8%	(6.7%)	−0.8%	(6.4%)
General	MI	0.226	(0.022)	0.230	(0.007)	0.217	(0.029)	0.247	(0.020)
	MI_{word}	−52.8%	(4.5%)	−52.0%	(1.5%)	−54.7%	(6.1%)	−48.3%	(4.3%)
Term	MI	1.642	(0.026)	1.649	(0.033)	1.460	(0.054)	1.486	(0.048)
	MI_{word}	−11.2%	(1.4%)	−10.9%	(1.8%)	−21.1%	(2.9%)	−19.7%	(2.6%)

Table 4: Results for Text Clustering with Phrases (stdv)

9 Discussion

When, in the following discussion, we refer to the results on the medical papers we consider the results on the categorization General (which is very similar to results on Combined and Specific). The results with respect to the categorization Term of the medical papers are a bit different than for the others. As we believe the other categorizations to be better we do not discuss this further.

To split *compounds* in the representation based only on words (representation 2 compared to 1) improve results when clustering the newspaper articles but not when clustering the medical papers. This may be because compounds in the medical papers would need a different analysis. We have also used a stoplist for certain words that should not be split based on other newspaper articles as described in (Rosell, 2003). An optimization for medical compounds here would perhaps improve results.

All variations of clustering using *phrases* performs worse than clustering using only words. Clustering performs worse when using only phrases (representations 3-10) than when using the combination of words and phrases (representations 11-18). Since clustering using words is superior the impact of phrases is diminished in the combined representations (11-18).

Looking at the *representations based only on phrases* (3-10) we see that results on newspaper articles are almost as bad as random clustering for all of them. The performance on the medical papers, on the other hand, is better than random clustering as long as we do not use split compounds as phrases. It is also better here to use the word trie representation (POM) rather than the simple phrase match (PM). In all this is an indication that phrases contain more information in the medical papers than in the newspaper articles.

For the *combined representations* (11-18) the results are much harder to analyze as the word representation is so much better than the phrase representation. The results on the newspaper articles are much better than on the medical papers here. This could be since the phrase representations do not contain as much information for the newspaper articles as for the medical papers and they thereby obscure the clustering to a lesser extent. Concerning the medical papers, all what is stated for the representations using only phrases holds, except that here it is not negative to use the split compounds as phrases (SP). For the newspaper articles there is even a great increase in performance when using the split compounds as phrases. This could be explained if the phrase representations using split compounds gives no information, which the results for representations 3-10 indicates. There is no reliable difference between the use of simple phrase match and the word trie representations for the newspaper articles as the standard deviation is very high.

No cases show any change in performance when splitting compounds within phrases (SC) or not. The reason for this could be beacuse the amount of compounds within phrases is small.

It is important to bear the great *differences of the two text sets* in mind. The differences in results between them show that clustering works differently on corpora with different contents (i.e. medical text vs. newspaper text). However, this difference might as well to a great extent be explained by other things, such as the structure and size of the texts and the difference of the categorizations. The medical papers are much longer than the newspaper articles. This could in fact explain all of the differences between them regarding information found in the phrases and the compounds. The categorization of the newspaper articles is probably much better than our categorizations of the medical articles.

10 Conclusions and further work

Phrases do not improve clustering in Swedish. At least with the representations tried here. The impact of phrases is more obvious on the medical papers. Overlap match between phrases performs better than simple match. It seems to be bad to consider split compounds as phrases and it does not matter whether one splits compounds within phrases or not.

Splitting solid compounds for the ordinary word representation improves results for the newspaper articles and does not make results worse for the medical papers.

The results are very different for the two text types, the newspaper articles and the medical papers. Maybe one would need to develop different representations for different text types. The information found in the phrases of the medical papers could perhaps be exploited using some other representation. But the same information might also be found in the ordinary representation using only words.

Our results are different from those presented in (Hammouda and Kamel, 2004). This is presumably, at least partially, because of differences between Swedish and English. Swedish solid compounds often correspond to phrases in English.

It could be interesting to try other variations of the representations using phrases presented here, but to really use the information that phrases contain relative to mere words a fundamentally different approach is probably needed. One interesting obvious extension of the present work is, however, to look at word-n-grams instead of phrases as these has proven useful in other works.

Acknowledgements

Our thanks go to professor Viggo Kann for supervision, to VR (Vetenskapsrådet – the Swedish Research Council) for funding and to all participants in the project “Infomat” from KTH Nada (department of Computer Science and Numerical Analysis at the Royal Institute of Technology) and the department of Medical Epidemiology and Biostatistics (MEB) at Karolinska Institutet, both in Stockholm, Sweden.

References

J. Carlberger, H. Dalianis, M. Hassel, and O. Knutsson. 2001. Improving precision in information retrieval for Swedish using stemming. In *Proc. 13th Nordic Conf. on Comp. Ling. – NODALIDA ’01*.

K. M. Hammouda and M. S. Kamel. 2004. Efficient phrase-based document indexing for web document clustering. *IEEE Transactions on Knowledge and Data Engineering*, 16(10):1279–1296. Student Member-Khaled M. Hammouda and Senior Member-Mohamed S. Kamel.

M. Hassel. 2001. Automatic construction of a Swedish news corpus. In *Proc. 13th Nordic Conf. on Comp. Ling. – NODALIDA ’01*.

V. Kann, R. Domeij, J. Hollman, and M. Tillenius, 2001. *Text as a Linguistic Paradigm: Levels, Constituents, Constructs. Festschrift in honour of Ludek Hrebicek*, volume 60, chapter Implementation aspects and applications of a spelling correction algorithm.

M. Rosell. 2003. Improving clustering of swedish newspaper articles using stemming and compound splitting. In *Proc. 14th Nordic Conf. on Comp. Ling. – NODALIDA ’03*.

M. Steinbach, G. Karypis, and V. Kumar. 2000. A comparison of document clustering techniques. In *Proc. Workshop on Text Mining, 6th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*.

A. Strehl, J. Ghosh, and R. Mooney. 2000. Impact of similarity measures on web-page clustering. In *Proc. AAAI Workshop on AI for Web Search (AAAI 2000), Austin*, pages 58–64. AAAI/MIT Press, July.

A. Strehl. 2002. *Relationship-based Clustering and Cluster Ensembles for High-dimensional Data Mining*. Ph.D. thesis, The University of Texas at Austin.

O. Zamir and O. Etzioni. 1998. Web document clustering: A feasibility demonstration. In *Research and Development in Information Retrieval*, pages 46–54.