A good space: Lexical predictors in vector space evaluation

Christian Smith, Henrik Danielsson, Arne Jönsson

Santa Anna IT Research Institute AB & Linköping University
SE-581 83, Linköping, SWEDEN
christian.smith@liu.se, henrik.danielsson@liu.se, arnjo@ida.liu.se

Abstract

Vector space models benefit from using an outside corpus to train the model. It is, however, unclear what constitutes a good training corpus. We have investigated the effect on summary quality when using various language resources to train a vector space based extraction summarizer. This is done by evaluating the performance of the summarizer utilizing vector spaces built from corpora from different genres, partitioned from the Swedish SUC-corpus. The corpora are also characterized using a variety of lexical measures commonly used in readability studies. The performance of the summarizer is measured by comparing automatically produced summaries to human created gold standard summaries using the ROUGE F-score. Our results show that the genre of the training corpus does not have a significant effect on summary quality. However, evaluating the variance in the F-score between the genres based on lexical measures as independent variables in a linear regression model, shows that vector spaces created from texts with high syntactic complexity, high word variation, short sentences and few long words produce better summaries.

Keywords: summarization, Random Indexing, corpus evaluation

1. Introduction

Extraction based summarizers extract the most important sentences from a text and present them as a compressed version of the original document. One way of representing the importance of information in a document is through the use of the vector space methodology.

The quality of the vector space used for representing the words and sentences is among other things dependent on what material that was used to create the vector space. Different spaces can be good at different things and the high parametrization often leads one to fine tune a vector space to a given task; being it text categorization, word sense disambiguation or automatic summarization (Sahlgren, 2006). The issue investigated in this paper is how a good vector space for the task of automatic summarization is to be constructed and what characteristics this space should have. Is, for instance, a vector space created from a corpus of newspaper texts best for summarization of newspaper texts?

Often genre is used to classify texts, and corpora are often built from text from several genres to get a heterogenity that represents the way a language is used in a small scale (Webber, 2009). Different text types have further been studied with regards to lexical features (Biber, 1986). Genres differ in surface characteristics such as word and sentence length and the purpose of this study is to see what characteristics of the genres that might be important in describing the potentiality of a certain genre for use in a word space for automatic summarization. The important characteristics of a text could then be used to build optimal word spaces for a given application, regardless of genre or source.

The performance of the summarizer using different vector spaces will be evaluated indirectly. There are two kinds of evaluations for vector spaces; direct and indirect evaluations (Sahlgren, 2006). Direct evaluations aim to investigate the actual geometry of the space to see whether it is capable of a sound semantic representation, whereas indirect evaluations are used to investigate the performance of a particular application utilizing the space. Thus, to compare two spaces indirectly, the performance of an application can be compared while utilizing different vector spaces.

The main goal of our research is to find the best corpus, or sub corpus, to be used for creating the vector space. To characterize a corpus we use genre and a variety of lexical measures.

In the paper, we present results from using corpora from various genres to train a vector space model based extraction summarizer. When comparing the summaries created by using the different spaces, we evaluate the spaces indirectly with regards to how well a particular application performs in terms of gold standard comparisons using the ROUGE F-score. Furthermore, a regression model is presented that predicts the performance of the summarizer based on the lexical measures.

2. The summarizer

In our experiments we use a summarizer called COGSUM (Smith and Jönsson, 2011b). COGSUM is an extraction based summarizer, using the vector space model Random Indexing (RI), c.f. Sahlgren (2005) and a modified version of PageRank (Brin and Page, 1998).

In Random Indexing (RI), contexts are built incrementally where every word consists of three parts; a string representation of the word itself, a random d-dimensional *index vector* consisting of a small number, ρ, of randomly distributed +1s and -1s, with the rest of the elements of the vectors set to 0, and a *context vector*.

Whenever a word occurs in a text, it gets assigned an index vector and has its context vector updated. A weighted sliding window, w, defines a region of context around each word. If the word has been encountered before, only the context vector is updated.

Words are thus represented by d-dimensional context vectors that are effectively the sum of the index vectors of all the contexts in which the word appears.

After the creation of word context vectors, the similarity between words could be measured by calculating the co-

sine angle between their word vectors, by taking the scalar product of the vectors and dividing by their norms.
Random Indexing is useful for acquiring the context vectors of terms, it is however not clear how a bigger context, such as a sentence, could be built from the word vectors. A crude way of creating sentence vectors from word vectors would be to simply summarize the vectors of the words in the sentence after they have been normalized to unit length. However, as the number of words in a sentence increase, so will the sentence similarity to the mean vector. Comparing sentences or documents in this way using cosine will make for larger similarity just by a larger number of words, regardless of relatedness. To alleviate this problem, the mean document vector is subtracted from each of the sentence's word vectors before summarizing the vectors (Higgins and Burstein, 2007), see Equation 1,

$$\vec{sent}_j = \frac{1}{S}\sum_{i=1}^{S}(\vec{w}_i - \vec{doc}) \quad (1)$$

where S denotes the number of words, w, in sentence j and $\vec{doc}$ is calculated as in Equation 2,

$$\vec{doc} = \frac{1}{N}\sum_{i=1}^{N}\vec{w}_i \quad (2)$$

where N denotes the number of unique words.
Words that are similar to the document vector will come closer to the zero vector, while those dissimilar to the document vector will increase in magnitude. When later summarizing the vectors, those of greater magnitude will have increased impact on the total sentence vector so that common, non-distinct, words do not contribute as much to the sentence vector. As this reduces the impact of common non-distinct words, there is essentially no need for a stop word list.
The vector space is created beforehand from a large corpora and is thus used to represent the meaning of words and sentences of a smaller document to be summarized (Smith and Jönsson, 2011b). Using an outside corpus, the summarizer processes each text by assigning each of the words in the document the corresponding semantic vector from a previously trained vector space. Sentence vectors are then created by calculating the mean vector of all the words contained within that sentence, subtracted by the mean space vector, as in Equation 1.
To extract the most important sentences, a variant of PageRank is used to find sentences in the vector space that share the most important information (Chatterjee and Mohan, 2007).
The method of using graph-based ranking algorithms for extracting sentences in summarization purposes was proposed by (Mihalcea, 2004), who introduce the TextRank model. In graph-based algorithms such as TextRank the text need to be represented as a graph, where each vertex depicts a unit of text and the edges between the units represent a connection between the corresponding text units. Graph-based ranking algorithms may be used to decide the importance of a vertex within a graph, by taking into account global information from the entire graph, rather than from only the local context of the vertices. The ranks are thus recursively computed so that the rank of a vertex depends on all the vertices' ranks.
For the task of sentence extraction, each sentence in a text is represented as a vertex and the relation between sentences are based on their overlap or "similarity", denoted by Equation 3.

$$Similarity(S_i, S_j) = \frac{|\{w_k | w_k \in S_i \& w_k \in S_j\}|}{log(|S_i|) + log(|S_j|)} \quad (3)$$

Thus, if a sentence addresses certain concepts, the other sentences that share content will get recommended by that sentence in the recursive fashion provided by PageRank.
To use PageRank and Random Indexing for summaries an undirected fully connected graph is created where a vertex depicts a sentence in the current text and an edge between two different vertices is assigned a weight that depicts how similar these are based on a cosine angle comparison of their meaning vectors.
When the text has been processed using RI and PageRank, the most important sentences are extracted using the final ranks on the sentences.

3. Text characteristics

Our goal is to characterize the best sub corpus to use when creating the vector space for a vector space based summarizer. Typical characteristics of a corpus include genre and various lexical measures.

3.1. Genres

Genres should differ in their lexical characteristics, the genres in the Brown Corpus (Francis and Kucera, 1979), for instance, are explained as primarily reflecting external purposes of the texts within and is divided into genres based on the communicative purposes of the texts making up the genres (Webber, 2009). The Swedish Stockholm-Umeå Corpus (SUC 2.0) (Ejerhed et al., 2006) is constructed similarly. A genre in this sense, should not be so broad as to not have any distinguishing features, nor so narrow not to have any general applicability; a genre should be variable in content. Thus, there will probably be differences in lexical characteristics between the different genres, possibly affecting the nature of a vector space built from them, but also some differences between texts within the same genre.
Some attempts have been made on automatically identifying genres in the Brown corpus. Karlgren and Cutting (1994) succeeded for instance well in classifying texts using discriminant analysis to recreate the partitioning in the Brown-corpus automatically using a set of linguistic parameters. In this way, the communicative purposes of the texts could be captured using various linguistic measures. By using these measures, it was possible to capture what, in communicative purpose, lies outside of the generality of the genre. Thus, there are features of texts that capture their nature, other than a binary decision placing a text in a particular genre. These features are of interest when trying to predict good qualities of a text collection making up a vector space.

Table 1: Genre characteristics of SUC, divided into partitions of similar size. Prose, general fiction was split into four equally sized chunks, whereas other genres was split according to sub type, i.e. Misc into Municipal publications and financial/company publications and Press into Editorials, Reportage and Reviews. The main goal was to keep the genres roughly the same size.

Genre	Description	Size
1	Biographies and essays	53723
2	Prose (General fiction)	43368
3	”	40228
4	”	40381
5	”	42250
6	Light reading	40306
7	Misc(Municipal publications)	42697
8	Misc (Financial and company publications)	40937
9	Prose (Mystery, sci-fi and humour fiction)	50528
10	Popular lore	98553
11	Press (editorials)	36751
12	Press (reportage)	94422
13	Press (reviews)	57768
14	Scientific (Technology, Mathematics and Medicine)	28844
15	Scientific (Social sciences)	44936
16	Scientific (Humanities)	56625
17	Scientific (Behavioural sciences, religion)	40849
18	Skills (Society press, religion)	27297
19	Skills (Hobbies)	51818
20	Skills (Union press)	41373
	SUC	1048325

3.2. Lexical Measures

To study the characteristics of the different genres in the corpus, each text collection that was used to create a vector space was analyzed using different lexical measures. The lexical measures comprise traditional measures, such as average sentence length, and a number of well known readability measures for Swedish. The measures are presented as they are used on one text. In the experiments presented in this paper the measures were averaged over all texts in a genre, see below, Equation 12. The measures are described below ($n(x)$ denotes the number of x):

OVIX OVIX (Hultman and Westman, 1977) is a measure on the ratio between the number of unique words and words in total. OVIX can be used to denote the idea density in a text (Mühlenbock and Kokkinakis, 2009).

$$OVIX = \frac{log(n(words\,in\,total))}{log(2 - \frac{log(n(unique\,words))}{log(n(words\,in\,total))})} \quad (4)$$

LIX LIX is a measure commonly used to describe the syntactic complexity of a text (Björnsson, 1968). The formula is as follows:

$$LIX = \frac{n(words)}{n(sentences)} + (\frac{n(words\ >\ 6\,chars)}{n(words)} \times 100) \quad (5)$$

NR NR or Nominal Ratio (Lundberg and Reichenberg, 2009) indicates the style of the text, a low NR is common in narrative texts while high NR are often seen in professional texts (Mühlenbock and Kokkinakis, 2009).

$$NR = \frac{n(nouns) + n(prepositions) + n(participles)}{n(pronouns) + n(adverbs) + n(verbs)} \quad (6)$$

ASL Average sentence length

$$ASL = \frac{n(words)}{n(sentences)} \quad (7)$$

AWL Average word length

$$AWL = \frac{n(characters)}{n(words)} \quad (8)$$

LWP Ratio of long words

$$LWP = \frac{n(words\ >\ 6\,chars)}{n(words)} \quad (9)$$

ANS Number of sentences

$$ANS = n(sentences) \quad (10)$$

PN Ratio of proper nouns

$$PN = \frac{n(proper\,nouns)}{n(words)} \quad (11)$$

Taken together, these measures provide a description of a text that can characterize different text collections. The characteristics can then be used as a basis for selecting the corpus that is used for creating a vector space. The measures also correlates with perceived readability, capturing different aspects of the texts, such as information load, sentence structure and syntactic complexity.

4. Evaluation

To evaluate how different corpora can be utilized to train the summarizer, 20 different text collections were used to build vector spaces. The text collections were taken from the Stockholm-Umeå Corpus (SUC 2.0) (Ejerhed et al., 2006), based on its genre distinctions. The genres of SUC comprise a number of novels, popular lore, publications etc., summarized in Table 1. The texts were partitioned to roughly the same size of $\approx 50,000$ (some were close to 100,000) words each, see Table 1.

A rule of thumb is that the larger the space, the larger the probability of containing the information necessary to specify a words meaning (Landauer et al., 2007). For this purpose, SUC in its entirety was also used as input to create the vector space, resulting in a vector space built from ≈ 1 million words.

Each of the 21 trained vector spaces were used to summarize 13 newspaper texts 10 times to account for possible randomness.

Each of the 20 text collections were analyzed with the lexical measures presented in Section 3.2. as a mean, $\bar{V}$, of all the texts contained within each genre, Equation 12,

$$\bar{V_j} = \frac{\sum_{i=0}^{N_g} V_j(doc_i^g)}{N_g} \tag{12}$$

where V_j is the result from a lexical measure applied to the texts, doc_i, in genre g. N_g denotes the total number of texts in each genre.

The vector spaces used a dimensionality, d, of 1800, a window size, $w = 2$ with a weighting of $[0.5, 1, 0, 1, 0.5]$, and $\rho = 8$, i.e. 8 non-zeroes in the index vectors, similar to Karlgren and Sahlgren (2001).

Table 2: F-scores and standard deviation of all genres across ten random seeds, including standard deviation between seeds on the entire corpus in the last row.

Genre	Decsription	F	St. dev.
1	Biographies and essays	0.582	0.0088
2	Prose (General fiction)	0.557	0.0092
3	”	0.622	0.0129
4	”	0.559	0.0065
5	”	0.603	0.0141
6	Light reading	0.570	0.0045
7	Misc(Mun. pub)	0.624	0.0051
8	Misc (Finance)	0.594	0.0057
9	Prose (Mystery etc)	0.612	0.0053
10	Popular lore	0.588	0.0106
11	Press (editorials)	0.561	0.0089
12	Press (reportage)	0.570	0.0069
13	Press (reviews)	0.612	0.0072
14	Scientific (Techn, Maths, Med)	0.593	0.0079
15	Scientific (Social sciences)	0.602	0.0121
16	Scientific (Humanities)	0.630	0.0034
17	Scientific (Behav. sc.)	0.599	0.0192
18	Skills (Society, religion)	0.567	0.0163
19	Skills (Hobbies)	0.584	0.0207
20	Skills (Union press)	0.570	0.0072
	SUC	0.582	0.0044

Table 3: Mean F and Standard deviation across all genres in Table 2.

	Mean F	St. dev.
All genres	0.582	0.0231

Previous studies have shown that a vector space built from a small corpus can show some randomness (Smith and Jönsson, 2011a) and that the vector space, thus, can be subject to random noise. Each genre was therefore used on ten different summarizations using different random seeds and the mean of the resulting ROUGE F-score after gold standard comparison was calculated. The results from the lexical measures, of course, stay the same. Table 2 depicts the standard deviation in the ROUGE F-score across the different genres. There is some difference in the variance between seeds, none is however larger than the variance between the genres. The standard deviation of the mean ROUGE F-scores between the genres is 0.0231, see Table 3. The standard deviation between seeds in SUC is the second lowest, probably benefiting from a larger text collection.

Using each of the 20 spaces, COGSUM was used to summarize 13 newspaper articles. Each summary was set to 30% of the original text. The resulting summaries were evaluated by comparing them to manually created gold standards (available through KTH eXtract Corpus(Hassel, 2011)) using the ROUGE-toolkit (Lin, 2004).

5. Results

The performance of the summarizer, using a vector space created based on the various genres', shows no significant difference between the genre being used. The third column in Table 2 shows each genre's ROUGE-1 F-score as a mean of the 13 summarized newspaper texts.

Partitioning SUC may result in parts that are not big enough. However, the performance of the summarizer was not affected by the various corpus sizes, as can be seen in Table 2.

Lexical measures, however, as depicted in Table 4, show some differences between genres. The scientific texts and municipal/communal publications (genres 7,8,14-18), display for instance a high LIX, as opposed to prose of various categories. The same applies for NR. For OVIX, press scored the highest whereas prose scored lowest.

To explain the variance in ROUGE-1 F-scores in Table 2, multiple linear regression with backward stepwise elimination was used on all lexical measures. This was done on the texts within the different genres to eliminate all non-significant variables affecting the ROUGE F-score.

These differences proved to be significant on predicting the ROUGE-1 F-score in a linear regression model containing the lexical measures OVIX, LIX, ANS and LWP as significant predictors, Equation 13 (all coefficients have a $p < .05$), $R^2 = .373$.

$$\begin{aligned} F = -0.12 + 0.013 * LIX + 0.007* \\ OVIX - 3.23 * LWP + 0.00078 * ANS \end{aligned} \tag{13}$$

Table 4: Results from using various lexical measures on the texts from SUC. For each genre, the mean of the measures from the different texts in that genre is presented. See Section 3.2. for explanation of the measures, and Table 1 for a description of the genres (omitted textual description here due to size). Genres 10 and 12 were of larger size. The last row displays the means on the entire corpus.

GENRE	ASL	AWL	ANS	PN	LWP	LIX	NR	OVIX
1	18.40	4.89	133.42	0.038	0.254	69.31	1.17	71.27
2	16.50	4.27	149.70	0.029	0.172	50.33	0.83	66.41
3	15.29	4.18	153.95	0.020	0.158	47.33	0.72	64.38
4	15.24	4.14	162.80	0.026	0.156	46.93	0.69	64.50
5	15.52	4.22	158.91	0.028	0.162	48.08	0.73	65.18
6	16.08	4.20	154.76	0.033	0.163	49.24	0.68	63.69
7	17.58	5.26	136.42	0.031	0.290	81.29	1.83	68.91
8	17.58	5.25	130.98	0.050	0.294	78.65	1.95	70.36
9	11.65	4.12	209.72	0.029	0.160	47.92	0.63	62.44
10	18.87	4.91	124.19	0.025	0.262	73.35	1.41	67.88
11	18.57	4.99	129.77	0.052	0.264	66.72	1.20	79.02
12	16.91	4.75	145.69	0.086	0.241	62.69	1.33	75.96
13	19.65	4.82	126.15	0.081	0.252	62.88	1.25	82.38
14	20.12	5.18	116.73	0.032	0.298	84.04	1.71	66.62
15	20.85	5.53	110.43	0.027	0.323	89.24	1.76	68.50
16	23.98	5.05	101.00	0.040	0.284	82.46	1.65	68.31
17	21.88	5.21	105.56	0.025	0.290	83.98	1.56	66.13
18	17.23	4.70	138.88	0.050	0.228	63.56	1.08	68.88
19	18.25	4.64	132.88	0.034	0.215	60.72	1.18	69.56
20	16.62	4.84	139.72	0.040	0.241	65.75	1.14	69.78
SUC	18.10	4.80	136.29	0.038	0.239	42.04	1.26	68.70

6. Discussion

An indirect evaluation of using various corpora to train vector spaces used for automatic summarization revealed that creating a vector space based on the genre of a text can not predict how well newspaper texts can be summarized using that space.

The model in Equation 13 shows the significant predictors, explaining some of the variance in performance. The coefficients does not reflect the strength of the predictors as the various measures are not normalized. What can be seen, though, is that LIX, OVIX and ANS are positive predictors whereas LWP is a negative predictor.

LIX explains variance in the F-score in terms of short sentences with short words, ANS explains the variance in terms of the average number of sentences per text, and as the texts are of roughly equal size this can be interpreted as short sentences, and OVIX explains variance in terms of ratio of unique words. LWP finally explains variance in terms of few long (> 6 characters) words.

This means that texts consisting of many short sentences with short words and high variability in words perform the best in a vector space and that those variables explain the variance in performance between genres.

Furthermore, our results somewhat contradicts the notion that a larger space contains better semantic information. However, with a large space, the meaning of the words are also more generalized and it might be that in some situations, it is more beneficial to have more specific or concrete meanings of words, rather than general ones. It is still somewhat unclear how the representation of the meaning of the words affect the summary results, as it is more to the process of creating a summary than only looking at the meaning of the words in the space, for instance the process of ranking, construction of sentence vectors, and how these steps are affected by the meaning of the words. It is not necessarily better for the words to mean the ”correct” most general thing, but to fit with a specific application and maximize the performance of it.

The model captures 37% of the variance, which is acceptable but it means that there presumably are other variables, not existing in our model, that can explain the variance. If other types of measures were included maybe that number could be increased to get an even more accurate predictive model. Additional variables may for instance include syntactic features on the phrase level, dependency features and additional part-of-speech features. It would also be interesting to look at tasks other than summarization and how the measures can predict performance for them.

The measures used in this work are tested for readability in Swedish, however, other languages, e.g. English, have similar measures that should make similar studies possible.

7. Conclusion

We have presented results on properties of a corpus that are important when creating the vector space for vector space based extraction summarizers.

Using a corpus built from newspaper texts to create the vector space is not necessarily best to use when summarizing

newspaper texts. On the contrary, lexical measures, previously most often used for readability studies, can be used to describe text features in a more fine grained way than for instance genre. Such lexical measures can then be used to identify corpora that provide better vector spaces for automatic extraction based summarizers, in a computational and comparable way. This allows, for instance, for predictive models of important features of texts, in this case to predict the performance of a summarizer using different vector spaces. More specifically, a collection of texts with high syntactic variation (LIX), a large number of unique words (OVIX), many sentences (ANS) and few long words (LWP) will produce vector spaces that, when used in the summarizer, provides better summaries.

Future work includes expanding the feature set to get a more accurate prediction of the summarizer's performance and scaling up the experiments to include more corpora and summary evaluation material. Furthermore, using the coefficients in for instance a genetic algorithm might make it possible to breed a nearly optimized vector space for use in summarization.

Acknowledgements

The research was funded by Santa Anna IT Research Institute AB and the Swedish Post and Telecom Agency, PTS.

8. References

Douglas Biber. 1986. Spoken and written textual dimensions in english: Resolving the contradictory findings. *Language*, 62(2):pp. 384–414.

C.H. Björnsson. 1968. *Läsbarhet*. Stockholm: Liber.

Sergey Brin and Lawrence Page. 1998. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1-7):107–117.

Nilhadri Chatterjee and Shiwali Mohan. 2007. Extraction-based single-document summarization using random indexing. In *Proceedings of the 19th IEEE international Conference on Tools with Artificial intelligence – (ICTAI 2007)*, pages 448–455.

Eva Ejerhed, Gunnel Källgren, and Benny Brodda. 2006. Stockholm umeå corpus version 2.0.

W. N. Francis and H. Kucera. 1979. Brown corpus manual. `http://icame.uib.no/brown/bcm.html`.

Martin Hassel. 2011. Kth extract corpus (kthxc), January 2011. `http://www.nada.kth.se/~xmartin/`.

Derrick Higgins and Jill Burstein. 2007. Sentence similarity measures for essay coherence. In *Proceedings of the 7th International Workshop on Computational Semantics (IWCS), Tilburg, The Netherlands*.

Tor G. Hultman and Margareta Westman. 1977. *Gymnasistsvenska*. LiberLäromedel.

Jussi Karlgren and Douglass Cutting. 1994. Recognizing text genres with simple metrics using discriminant analysis. In *Proceedings of the 15th conference on Computational linguistics - Volume 2*, volume 2 of *COLING '94*, pages 1071–1075, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jussi Karlgren and Magnus Sahlgren. 2001. From words to understanding. In Y. Uesaka, P.Kanerva, and H. Asoh, editors, *Foundations of Real-World Intelligence*, chapter 26, pages 294–308. Stanford: CSLI Publications.

T. K. Landauer, D.S. McNamara, S. Dennis, and Kintsch W., editors. 2007. *Handbook of Latent Semantic Analysis*. Mahwah NJ: Lawrence Erlbaum Associates.

Chin-yew Lin. 2004. Rouge: a package for automatic evaluation of summaries. In *ACL Text Summarization Workshop*, pages 25–26.

Ingvar Lundberg and Monica Reichenberg. 2009. *Vad är lättläst?* Socialpedagogiska skolmyndigheten.

Rada Mihalcea. 2004. Graph-based ranking algorithms for sentence extraction, applied to text summarization. In *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*, ACLdemo '04, Morristown, NJ, USA. Association for Computational Linguistics.

Katarina Mühlenbock and Sofie Johansson Kokkinakis. 2009. Lix 68 revisited – an extended readability measure. In *Proceedings of Corpus Linguistics*.

Magnus Sahlgren. 2005. An Introduction to Random Indexing. *Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering, TKE 2005*.

Magnus Sahlgren. 2006. Towards pertinent evaluation methodologies for word-space models. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*.

Christian Smith and Arne Jönsson. 2011a. Automatic summarization as means of simplifying texts, an evaluation for swedish. In *Proceedings of the 18th Nordic Conference of Computational Linguistics (NoDaLiDa-2010), Riga, Latvia*.

Christian Smith and Arne Jönsson. 2011b. Enhancing extraction based summarization with outside word space. In *Proceedings of the 5th International Joint Conference on Natural Language Processing, Chiang Mai, Thailand*.

Bonnie Webber. 2009. Genre distinctions for discourse in the penn treebank. *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, (August):674–682.