This also affects the context - Errors in extraction based summaries

Thomas Kaspersson, Christian Smith, Henrik Danielsson, Arne Jönsson

Santa Anna IT Research Institute AB & Linköping University
SE-581 83, Linköping, SWEDEN
thoka336@student.liu.se, christian.smith@liu.se, henrik.danielsson@liu.se, arnjo@ida.liu.se

Abstract

Although previous studies have shown that errors occur in texts summarized by extraction based summarizers, no study has investigated how common different types of errors are and how that changes with degree of summarization. We have conducted studies of errors in extraction based single document summaries using 30 texts, summarized to 5 different degrees and tagged for errors by human judges. The results show that the most common errors are absent cohesion or context and various types of broken or missing anaphoric references. The amount of errors is dependent on the degree of summarization where some error types have a linear relation to the degree of summarization and others have U-shaped or cut-off linear relations. These results show that the degree of summarization has to be taken into account to minimize the amount of errors by extraction based summarizers.

Keywords: summarization, text cohesion, evaluation

1. Introduction

An extraction based summary is created by extracting the most important sentences from the original text. Previous results have shown that broken or erroneous anaphoric references is a problem in extraction based summarizers (Hassel, 2000) breaking the cohesion of the summarized text and in some cases even altering the meaning of the text, making them hard for readers to understand. Thus, cohesion and discourse relations play a vital role in understanding summaries (Louis et al., 2010). None of these studies have investigated how the occurrence of errors is distributed over the summarized texts or how different levels of summarization are affected by the errors in terms of how the amounts of errors correlate with summary level.

In this paper we present results from investigations of the linguistic errors that occur in single document extract summaries. We focused mainly on discourse errors, such as referring expressions with missed antecedent and fragments; how well the text units in the summaries are linked. This can be seen as a type of cohesion, which is an important part of coherence, i.e is the reader able to get a coherent meaning conveyed when reading the texts? By measuring distinct error types having to do with cohesion, an objective measure of a part of the coherent structure of the texts can be calculated.

The investigation further focused on the impact different text summary levels had on the amount of error types, and if different genres had any impact on the amount of error types.

The results will show what type of errors in the summary that is the most pronounced, and at what summary level and in what genres.

2. The vector space model

Many extraction based summarizers utilize the vector space model. The vector space model (Eldén, 2007), is a spatial representation of a word's meaning where every word in a given context occupies a specific point in the space and has a vector associated to it that can be used to define its meaning.

The vector space can be constructed from a matrix where text units are columns and the words in all text units are rows. A certain entry in the matrix is nonzero iff the word corresponding to the row exists in the text unit represented by the column. The resulting matrix is very large and sparse, which makes for the usage of techniques for reducing dimensionality and get a more compact representation. Random Indexing (Sahlgren, 2005; Kanerva, 1988) is one such dimension reduction technique that can be described as a two-step operation:

Step 1 A unique d-dimensional, sparse and high-dimensional *index vector* is randomly generated and assigned to each context. Index-vectors consist of a small number, ρ, of randomly distributed +1s and -1s, with the rest of the elements of the vectors set to 0.

Step 2 *Context vectors* are produced by scanning the text. Each time a word occurs in a context, that context's index vector is added to the context vector for the word. A sliding window, w, defines a region of context around each word. Words are thus represented by d-dimensional context vectors that are 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 can be measured by calculating the cosine angle between their word vectors.

The summarizer used in our investigations is a Random indexing based summarizer called COGSUM (Smith and Jönsson, 2011). COGSUM also uses the Weighted PageRank algorithm in conjunction to its Random Indexing-space to rank the sentences (Chatterjee and Mohan, 2007). The results however are valid for other vector space based summarization approaches e.g. HolSum (Hassel and Sjöbergh, 2007), SummaryStreet (Franzke et al., 2005) and Gong (2001).

COGSUM is written in Java and utilizes a Random Indexing toolkit available at Hassel (2011). The summarizer is able to operate without any outside material, including an outside word space.

3. Linguistic quality of summarizations

A variety of investigations on errors in summarizations have been done, for instance the evaluation of linguistic quality for summarizations used in the DUC (Document Understanding Conference) summarization track. Over et al. (2007) describe the following five aspects:

1. Grammaticality; referring to the summary not having fragments or missing components
2. Non-redundancy; referring to the summary not having unnecessary repetitions
3. Referential clarity; meaning that pronouns and noun phrases should be properly referred to
4. Focus; meaning that the summary should have a focus
5. Structure and coherence; in that the summary should convey a coherent body of information.

Summarizers generally perform well in grammaticality and non-redundancy (Over et al., 2007). Grammatical errors may still arise in extraction based summaries, e.g. if lists or headings are not treated properly, but this more depends on how the summarizer converts documents to plain text.

Otterbacher et al. (2002) further identify five major categories on text cohesion related to multi-document summaries:

1. Discourse; relating to the relationships between sentences in a summary
2. Identification of entities; relating to resolution of referential expressions
3. Temporal relationships, i.e. establish the correct temporal relationship between events
4. Grammatical problems
5. Location problems; where an event takes place

Otterbacher et al. (2002) aim to revise multi-document summaries and find that the first three categories comprise the majority (82%) of revisions done. For single document summaries, the third category, temporal relationships, is less prominent, as the temporal order, as given in the text, often is retained, which is not necessarily the case if the text is assembled from multiple-documents.

There are other studies on linguistic quality, e.g. Lapata and Barzilay (2005), and on automatic vs. human judgements (Pitler et al., 2010) that also stress the importance of cohesion, but none of them investigate the distribution between e.g. genres and summary lengths.

4. Errors in extraction based summaries

We have conducted a pilot study to find error types in summarized texts that can have negative consequences on cohesion, coherence and readability, making the summarized text difficult to read, or even incomprehensible. The task was to read summarized texts from three different genres with five levels of summarization, tagging everything in the text that was considered an error with a description of the error.

We use three types of texts representing three different genres:

- **DN**. Newspaper texts from the Swedish newspaper ”Dagens Nyheter”; around 190 words per text
- **FOF**. Popular science texts from the Swedish Magazine ”Forskning och Framsteg”; around 650 words per article
- **FOKASS**. Authority texts from the Swedish Social Insurance Administration (Sw. Försäkringskassan); around 720 words per text

The texts were extracted from the concordances at Språkbanken (2011), except for the authority texts which were taken from the Swedish Social Insurance Administration’s web page (Försäkringskassan, 2011). They were summarized to five different lengths: 17%, 33%, 50% 67% and 83%.

Tagging was done by four independent analyzers and each were given four summarized texts. The errors found were then grouped into different categories, resulting in three categories and sub-categories:

1. Erroneous anaphoric reference (divided into three subtypes). When an anaphoric expression in the summarized text refers to an erroneous antecedent as the correct antecedent has not been extracted. For an example of this see Figure 1. The sub-types of erroneous anaphoric references are:
   (a) Noun-phrase
   (b) Proper names
   (c) Pronouns.
2. Absent cohesion or context. Sentences which in the summary lack any cohesion or context, necessary for understanding the extracted sentence.
3. Broken anaphoric reference, see Figure 2, (divided into three sub-types). When the summarized text contains one, or more, anaphoric expression(s) that has its antecedent in a sentence that has not been extracted. The sub-types of broken anaphoric references were:
   (a) Noun-phrase
   (b) Proper names
   (c) Pronouns.

Typical examples of cohesion errors in extraction based summaries occur when the antecedent to an anaphora is not included in the summary, Figure 1.

The pronoun ”such” in the summary in Figure 1 does not have an antecedent in any previously extracted text, creating a broken anaphoric reference. A slightly more difficult error type are erroneous anaphoric reference, when the correct antecedent has not been extracted and at the same time altering the meaning and understanding of the text, as in Figure 2.

”He” in the full text in Figure 2 refers to Fridtjof Nansen, but as this part was not extracted it refers to De Long in the summarized text.

Originally the return from Uppsala royal estate property should be enough for the kings support. What we nowadays call taxes was not in question - the free man could not be forced to pay any fees. The free man had, however, official duty.

<u>Such</u> official duty was the guesting, the obligation to receive and support the king and his escort when they travelled.

Figure 1: Example of broken anaphoric reference. Text in italics represents the non extracted sentences. Text in bold represents an extracted sentence, and the underlined words highlights the words making the sentence erroneous

But the trip towards the north pole became a disaster.

De Long and his crew sailed with the ship Jeannette through the Bearing sea 1879.

Soon they got stuck in ice north of the Wrangel island. In June 1881, Jeannette was crushed by the ice, and everyone onboard perished after a time of hardship. The theory about the open polar sea was declared dead. The disaster however, became of great importance for polar research. A few years after the foundering of the Jeannette wreck parts reached the east coast of Greenland - a revolutionary discovery. Fridtjof Nansen immediately got the idea to test the theory of an open sea filled with drift ice

<u>He</u> let build a powerful ship strong enough to drift unharmed with the thick pack ice for a long time.

Carried by the ice, the expedition would travel from Siberia to the North pole.

Figure 2: Example of erroneous anaphoric reference. Text in italics represents the non extracted sentences. Text in bold represents an extracted sentence, and the underlined words highlights the words making the sentence erroneous

As can be seen from the examples, the errors made by extraction summarizers can be quite severe. It is, however, unclear how common they are in a summary.

5. Evaluation

Based on the error types and categories, we developed guidelines for tagging summarized texts. The guidelines consisted of a document with the error types and one or more example(s) to illustrate how the errors could be identified in the actual summary. With the guidelines 30 texts, 10 from each of the three genres (news paper texts, authority texts and popular science texts), with five different summary levels, were tagged. The texts were presented line by line, with the extracted sentences presented in bold black text, and the non extracted sentences marked red. Three columns with separate columns for sub-categories, one for each error type followed each sentence. Each error found in the texts were tagged. If two or more errors occurred in the same sentence, all were tagged. Two peers were set to tag the 30 texts by reading the original text with the extracted sentences from the different summary levels in bold black text and the non-extracted sentences in red. By presenting the whole text, and not only the summaries, the peers were able to tag the specific error type with the correct subcategory, as they could easily read the non-extracted sentences to determine if the missing/erroneous antecedent belonged to subcategory nounphrase, proper names or pronouns. They were given 20 texts each, with five texts which were the same for both peers. This meant an overlap of 10 of the 30 texts and an inter judge reliability of 69.4% on these. When all the texts were tagged, the errors were summed up.

6. Results

The evaluations resulted in 30 texts tagged with errors of the different types presented in Section 4. These were summarized as to display the amount of errors per 100 sentences.

The percentages denote the amount of the original text that is retained, e.g, a 17% summary means a text consisting of 17% of the sentences from the original text.

We did not find any significant differences in the amount of errors between different genres.

For summarization length, however, we found several significant differences. Table 1 shows the number of errors and the standard deviation for the ten error types and five summary levels for all the different text genres together. In what follows we will present the significant differences found both between text summarization lengths. There were no significant differences between genres.

All results (except Table 1) are based on non-sentence normalized data, and mean values are the number of errors per sentence in the summarized text. All figures show 95% confidence interval on error bars

The errors were analyzed with analysis of variance, ANOVA, with the level of summarization and genre as within group variables. ANOVAs were made separately for all error types. Comparisons not showing significant effects and are not presented. The following significant differences were found:

Error type 1c: Erroneous anaphoric references, pronoun. There was a significant effect of summary level. $F(4, 108) = 2.87$ $p < .05$.

Figure 3 shows the mean values of erroneous anaphoric references, sub-type pronoun. The values show that the 50% and 67% level of summary contain the most errors, and significantly more than 17%, 33% and 83%.

Error type 2: Absent cohesion or context. There was a significant effect of summary level. $F(4, 108) = 14.01$ $p < .001$.

Figure 4 shows the mean values of absent cohesion or context. The values show that 17%, 33% and 50% have an even amount of errors. After 50%, the errors start to decrease and after reaching 67% the difference becomes significant.

Error type 3a Broken anaphoric references, nounphrase. There was a significant effect of summary level.

Table 1: Number of errors (and standard deviation (SD)) per one hundred sentences, based on sentence normalized data from all texts.

Error type	17%(SD)	33%(SD)	50%(SD)	67%(SD)	83%(SD)
1a	0,2(0,8)	0,3(0,9)	0,3(0,8)	0,3(0,8)	0,3(0,7)
1b	0,0(0,0)	0,0(0,0)	0,0(0,0)	0,0(0,0)	0,0(0,0)
1c	1,1(1,5)	1,3(1,8)	1,5(2,0)	1,5(2,1)	0,5(1,0)
2	9,9(6,0)	11,0(7,6)	9,6(8,2)	8,6(11,2)	2,6(3,7)
3a	3,8(3,9)	3,2(3,8)	2,5(3,1)	1,7(4,3)	0,7(1,4)
3b	1,1(3,2)	0,8(1,9)	0,5(1,6)	0,4(1,1)	0,2(0,6)
3c	3,1(3,4)	4,0(4,5)	3,8(4,3)	3,6(4,9)	0,7(1,4)

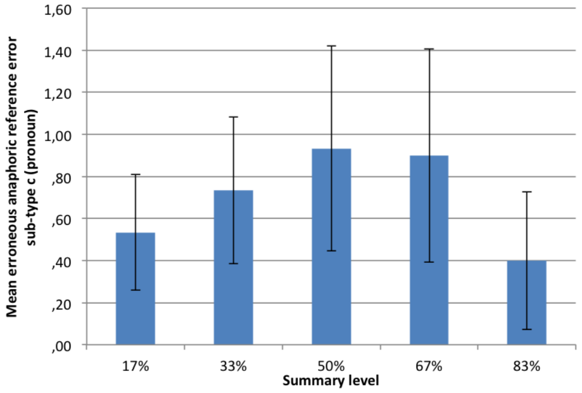

Figure 3: **Error type 1c.** Mean values of erroneous anaphoric references, sub-type pronoun. The effect on summary level is significant $F(4, 108) = 2.87$ $p < .05$.

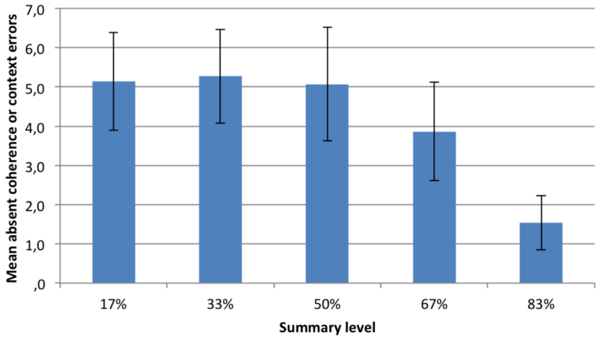

Figure 4: **Error type 2.** Mean values of absent cohesion or context. The effect on summary level is significant $F(4, 108) = 14.01$ $p < .001$.

$F(4, 108) = 7.35$ $p < .001$.

Figure 5 shows the mean values of broken anaphoric references, sub-type noun phrase. The values show a linear decrease in errors from 17% summary level to 83% summary level.

Error type 3c: Broken anaphoric references, pronoun. There was a significant effect of summary level. $F(4, 108) = 10.04$ $p < .001$.

Figure 6 shows the mean values of broken anaphoric references, sub-type pronoun. The values show that summary levels 17%, 33% 50% and 67% have a fairly even amount of errors and that summary level 83% has significantly less errors.

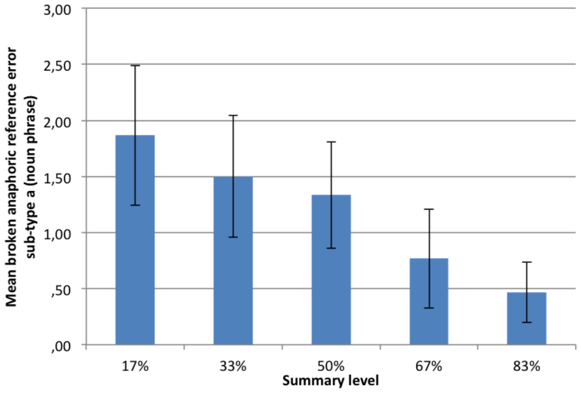

Figure 5: **Error type 3a.** Mean values of broken anaphoric references, noun-phrase. The effect on summary level is significant $F(4, 108) = 7.35$ $p < .001$.

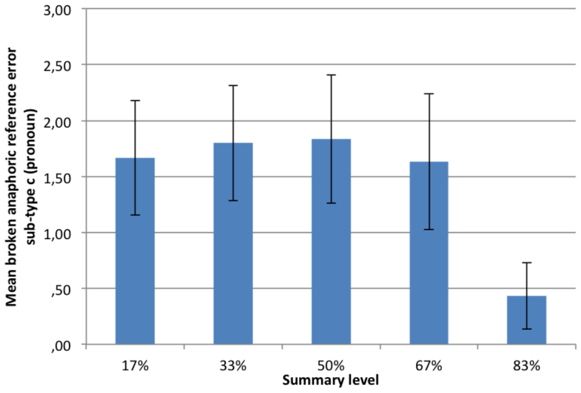

Figure 6: **Error type 3c.** Mean values of broken anaphoric references, pronoun. The effect on summary level is significant $F(4, 108) = 10.04$ $p < .001$.

Figure 7 shows how the different error types with significant differences are spread over the five summary levels.

As shown in Figure 7, Error type 2, absent cohesion or context, is the most dominant error type and occurs roughly once every tenth sentence depending on the summarization level.

Figure 7 also shows that different error types, though within the same family of errors (such as anaphoric references), show very different relations to the level of summarization. As can be seen for error types 1c and 3c there is no linear relation between how frequent different error types are,

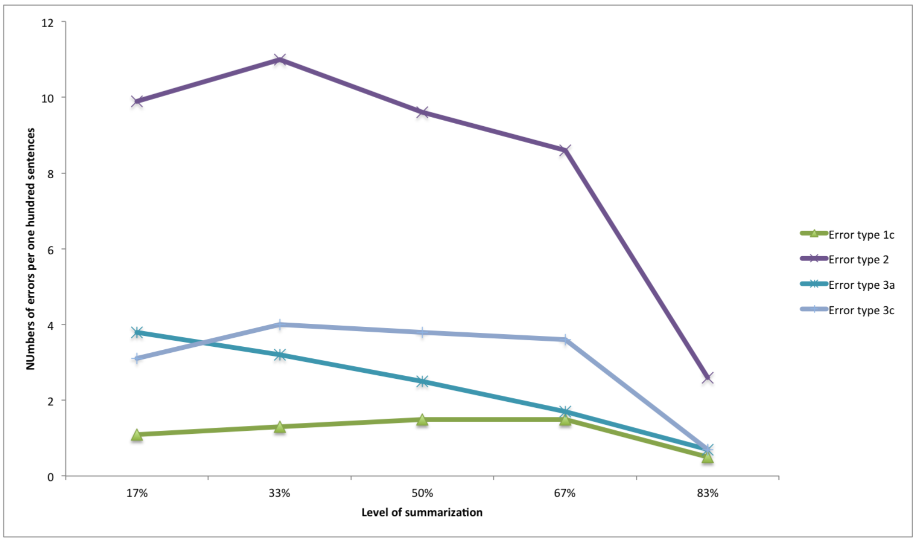

Figure 7: Error type relations to summary level

whereas error types 3a and 2 show such behaviour.

7. Discussion

Several differences could be observed across level of summarization. The following types of errors were found to be significant:

1c Erroneous anaphoric references, pronoun (Figure 3)

2 Absent cohesion or context (Figure 4)

3a Broken anaphoric references, noun phrase (Figure 5)

3c Broken anaphoric references, pronoun (Figure 6)

Furthermore, the error types with a significant effect also depend on summary level.

Error type 1c, erroneous pronoun reference, shows that the quantity of errors increase along with a decreased summary level, but only to a certain degree after which the quantity begins to decrease again. There is a significant difference between 17% and 33% compared to 50% and 67%, but also between 83% and 50%, 67%. A possible explanation for this is that at summary level 17% and 33% (the most summarized texts) the amount of extracted sentences are quite few, making the error decrease in amount as the sentences which make up the error are not extracted to the summary, and the extracted sentences at this level of summarization usually are adjacent. For summary level 83% however, the opposite happens as the amount of extracted sentences is high. The more extracted sentences, the lower the risk of erroneous anaphoric references as the risk of the correct antecedent not being extracted is much lower.

For Error type 2, absent cohesion or context, there is a significant difference in summary level between 17%, 33%, 50%, 67% and, 83%. This is interesting because the results show that 17%, 33% and 50% have an almost equal amount of errors, after which the amount of errors start decreasing almost linearly, and when passing 67% the difference becomes significant. This means that it is not a linear increase or decrease in errors, but that the amount of errors level out after a certain level of summarization. This suggests that in order to keep relevant cohesion or context, the level of summarization should be taken into consideration as a text summarized more than a certain given level will lose contextual information and lack cohesion. Error type 2 was also most dominant, Figure 7 which was expected as the cohesion of the text is expected to be affected by the extraction method, and is often the reason for errors like erroneous or broken references.

Error type 3a, broken anaphoric references sub-type noun-phrase, shows significant differences based on the level of summarization, and a trend of linear decrease when the summary level increased. This means that the fewer sentences extracted, the more noun phrases without an antecedent will occur, thus making it a broken anaphoric reference. This kind of linear decrease in errors is the kind of result we expected to see in most error types.

Error type 3c, broken anaphoric references sub-type pronoun, also shows a significant difference in level of summarization but the significance for this error is between summary level 83% and 17%, 33%, 50%, 67%. This indicates a cut-off in the amount of broken anaphoric references where a pronoun does not have an antecedent. This means that this error type is persistent throughout a 17% level of summary up to 67% after which it seems to rapidly decrease.

Thus, just like Error type 2, absent cohesion or context, Error type 3c, broken anaphoric references sub-type pronouns, follow the same pattern and show that the amount of errors is persistent until a certain cut-off point, after which the errors start to decrease linearly.

This suggests again, that the level of summarization must be taken into consideration, as it indicates that at a certain level, pronouns in extracted sentences will begin to loose their antecedent, thus making the summary incoherent.

The most interesting finding is that the different error types, though within the same family of errors (such as anaphoric references), show very different relations to the level of summary (as seen in Figure 7). Some errors show a linear decrease in errors along with a decrease in summary level, while some show a cut-off at a specific summarization percentage or an increase in errors parallel to higher level of summarization, only to decrease after reaching a specific summary level. Previous results show that broken or erroneous anaphoric references is a problem in extraction based summarizers (Hassel, 2000) and that cohesion play a vital role in summarized texts (Louis et al., 2010), though none of them have studied how different levels of summarization are affect by the errors in terms of how the amounts of errors correlate with summary level.

8. Conclusion

We have presented results on the distribution and frequency of linguistic errors in extract summaries. The results show that the most common errors are absent cohesion or context and various types of broken or missing anaphoric references. No significant difference between genres were found.

The results are based on only one vector space based summarizer, but we believe that they are relevant to any extraction based summarizer, regardless of technique.

The most interesting finding is that the different error types, though within the same family of errors (such as anaphoric references), show very different relations to the level of summary. Some errors present a linear decrease in errors along with a decrease in summary level, while some show a cut-off at a specific summarization percentage or an increase in errors parallel to higher level of summarization, only to decrease after reaching a specific summary level.

These results show that the degree of summarization has to be taken into account to minimize the amount of errors produced by extraction based summarizers. It is however not apparent that a shorter summary always is worse with regards to the relative amount of errors.

Errors like broken or erroneous anaphoric references and lack of cohesion or context are errors expected to be found in any extraction based summarizer that does not consider context. These kinds of errors are the ones that affect coherence and discourse and often make the text hard to read or incomprehensible.

The results also stress the importance of improving text generation for extraction based summarizers as the most dominant error types affect the coherence and discourse relations of the text, and also often alter its meaning.

9. References

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.

Lars Eldén. 2007. *Matrix Methods in Data Mining and Pattern Recognition*. Society for Industrial & Applied Mathematics (SIAM).

Försäkringskassan. 2011. Försäkringskassans website, January. http://www.forsakringskassan.se.

M Franzke, E Kintsch, D Caccamise, N Johnson, and S Dooley. 2005. Summary street®: Computer support for comprehension and writing. *Journal of Educational Computing Research*, 33(1):53–80.

Yihong Gong. 2001. Generic text summarization using relevance measure and latent semantic analysis. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Martin Hassel and Jonas Sjöbergh. 2007. Widening the holsum search scope. In *Proceedings of the 16th Nordic Conference of Computational Linguistics (Nodalida)*, Tartu, Estonia, May.

Martin Hassel. 2000. Pronominal resolution in automatic text summarisation. Master's thesis, Master thesis in Computer Science, Department of Computer and Systems Sciences (DSV), Stockholm University, Sweden.

Martin Hassel. 2011. Java random indexing toolkit, January 2011. http://www.csc.kth.se/~xmartin/java/.

Pentii Kanerva. 1988. *Sparse distributed memory*. Cambridge MA: The MIT Press.

Mirella Lapata and Regina Barzilay. 2005. Automatic evaluation of text coherence: Models and representations. In *Proceedings of the International Joint Conference On Artificial Intelligence (IJCAI)*.

Annie Louis, Aravind Joshi, and Ani Nenkova. 2010. Discourse indicators for content selection in summarization. In *Proceedings of SIGDIAL 2010: the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue, Tokyo, Japan*, pages 147–156.

Jahna C. Otterbacher, Dragomir R. Radev, and Airong Luo. 2002. Revisions that improve cohesion in multi-document summaries: A preliminary study. In *Proceedings of the Workshop on Automatic Summarization (including DUC 2002), Philadelphia*, pages 27–36.

Paul Over, Hoa Dang, and Donna Harman. 2007. Duc in context. *Information Processing & Management*, 43:1506–1520, Jan.

Emily Pitler, Annie Louis, and Ani Nenkova. 2010. Automatic evaluation of linguistic quality inmulti-document summarization. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Uppsala, Sweden*, pages 544–554.

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*.

Christian Smith and Arne Jönsson. 2011. 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*.

Språkbanken. 2011. Concordances of språkbanken, January. http://spraakbanken.gu.se/konk/.