A More Cohesive Summarizer

Christian Smith[1], *Henrik Danielsson*[1], *Arne Jönsson*[1]
(1) Santa Anna IT Research Institute AB, Linköping, Sweden
christian.smith@liu.se, henrik.danielsson@liu.se, arnjo@ida.liu.se

Abstract

We have developed a cohesive extraction based single document summarizer (COHSUM) based on coreference links in a document. The sentences providing the most references to other sentences and that other sentences are referring to, are considered the most important and are therefore extracted. Additionally, before evaluations of summary quality, a corpus analysis was performed on the original documents in the dataset in order to investigate the distribution of coreferences. The quality of the summaries is evaluated in terms of content coverage and cohesion. Content coverage is measured by comparing the summaries to manually created gold standards and cohesion is measured by calculating the amount of broken and intact coreferences in the summary compared to the original texts. The summarizer is compared to the summarizers from DUC 2002 and a baseline consisting of the first 100 words. The results show that COHSUM, aimed only at maintaining a cohesive text, performed better regarding text cohesion compared to the other summarizers and on par with the other summarizers and the baseline regarding content coverage.

Keywords: Summarization, Coreference resolution, Cohesion.

Proceedings of COLING 2012: Posters, pages 1161–1170,
COLING 2012, Mumbai, December 2012.

1 Introduction

Extraction based summarizers are often prone to create texts that are fragmented, where sentences are extracted without considering the context, resulting in for instance broken anaphoric references. As pointed out by Nenkova (2006), the linguistic quality of automatically generated summaries can be improved a lot. For current popular measures summarizers often score relatively well on measures regarding content coverage that incorporates a comparison to humanly created gold standard summaries. The cohesiveness of the summaries is often left out in the evaluations since the measures favor inclusion of certain information, disregarding how well the text fits together. Brandow et al. (1995) revealed that summaries of news articles consisting only of the lead sentences are difficult to beat; when it comes to newspaper articles this type of summary fits well since the structure of the text is built around first presenting the gist in the lead sentences and then focusing the rest of the article on elaborating the information. These texts will of course also be cohesive since the sentences are extracted in the order they were written.

Barzilay and Elhadad (1999) proposed to improve cohesion in summaries by using lexical chains to decide which sentences to extract and Bergler et al. (2003) used coreference chains. Other attempts propose the use of a variety of revisions to the text based on cohesion and discourse relations (Mani et al., 1998; Otterbacher et al., 2002) or using both revisions and lexical chains (Alonso i Alemany and Fuentes Fort, 2003). Such approaches require a thesaurus, e.g. WordNet (Barzilay and Elhadad, 1999). Boguraev and Neff (2000) show that cohesion can be improved by utilizing lexical repetition. Coreference information has also been used, for instance, for creating summaries with a focus on answering queries on a text (Baldwin and Morton, 1998).

Pitler et al. (2010) attempted to develop and validate methods for automatic evaluation of linguistic quality in text summarization. They concluded that the topics of Referential clarity and Structure/Coherence seems to be most important when dealing with extraction based single document summarization. Furthermore, anaphoric expressions are important for a text's cohesion (Mani et al., 1998). Errors regarding broken anaphoric references are, however, common in extraction based summaries, especially (not surprisingly) in short summaries (Kaspersson et al., 2012), and in particular for summarizers that focus on content coverage and disregard how sentences are related to each other.

In this paper, we focus on cohesion and referential clarity, creating summaries that hopefully are more readable in that they maintain text cohesion. This can be contrasted to summarizers that are focused only on extracting the most important information in the text, without taking into account cohesion e.g. (DUC, 2002; Smith and Jönsson, 2011b; Chatterjee and Mohan, 2007; Hassel and Sjöbergh, 2007; Gong, 2001; Mihalcea and Tarau, 2004). Such summarizers have performed well when compared to gold standards, the studies lack however results on how cohesive the summaries are. The hypothesis is that a summarizer focused on creating a cohesive text without regarding content coverage will score well on cohesive measures while scoring worse at measures aimed at summary content, and vice versa.

2 Coreferences in Newspaper Texts

Coreferences are commonly used as a feature when evaluating cohesion and coherence (Graesser et al., 2004; Pitler et al., 2010) and we therefore conducted experiments on the distribution of coreferences in summaries. We analyzed the 533 news paper texts used for single text summarization at the 2002 Document Understanding Conference (DUC, 2002). The original documents were

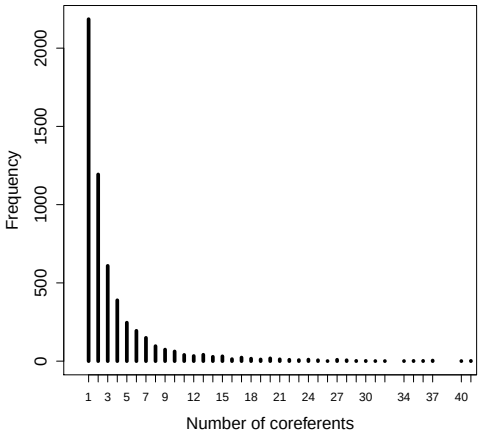

Figure 1: Frequency of the number of coreferents. The X-axis depicts the number of coreferents. The Y-axis shows the frequency for each number of coreferents for all 533 news paper texts. For example, most of the coreferents are between two sentences (one representative mention and one refererent).

tagged for coreference using the Stanford CoreNLP package (Lee et al., 2011)[1]. The coreference resolution system first extracts mentions together with relevant information such as gender and number. These mentions are processed in multiple steps (sieves), which are sorted from highest to lowest precision. For example, the first sieve (i.e., highest precision) requires an exact string match between a mention and its antecedent, whereas the last one (i.e., lowest precision) implements pronominal coreference resolution. At this stage, noun phrases, possessive pronouns and named entities have been considered for reference. In the last step, after coreference resolution has been done, a post-processing step is performed where singletons are removed. The results from the experiments are summarized in Figures 1, 2, and 3.

Figure 1 shows the length of the coreference chains (the number of sentences in the chains) that are most frequent. Note, however, that most sentences, 7281, does not have a reference at all (not included in the figure). Most references, 2185, are between two sentences; one with the representative mention and one additional mention. Approximately 1200 reference chains include three sentences and so on. There are very few reference chains of length 10 or more.

Figure 2 shows the average distance of the coreferences, that is, how many sentence indices are between a current sentence and the sentence it references. The X-axis shows the sentence index and the Y-axis shows the distance or number of sentences between the referents. In the beginning of the document the distance between referring sentences is around 4, increasing until index (sentence) 20 where the distance is approximately 8, probably because they refer to the first sentences. Then the

[1] nlp.stanford.edu/software/index.shtml

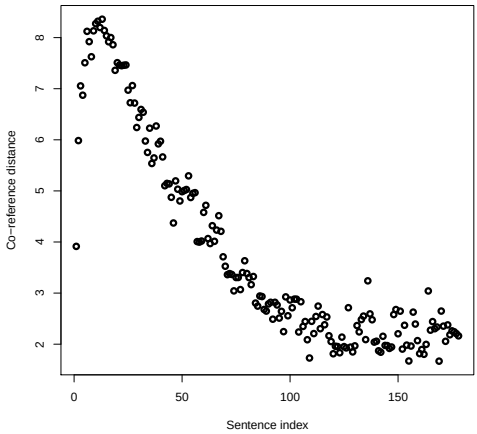

Figure 2: Coref distances, text. The earliest sentences have short distances, quickly followed by long distances in middle sentences and a shorter distance again concerning sentences further into the document.

distance decreases rapidly.

Figure 3 shows a plot where the sentence indices are on the X- and Y-axis and the size of the circle depicts the number of times a coreference exists in a given sentence pair. This figure shows that sentences early in the document have the most coreferences and that they corefer to each other. In the later parts of the document the sentences are mostly referring to the sentence before. Looking back at Figure 2 we see that long distance references occur mostly in the middle (around sentence number 20) of the document and is referring to the beginning of the document.

To summarize, the results reveal that, for news texts, the beginning of the document is terse with sentences coreferring each other. In the middle of the document, sentences are most often coreferring to the sentence before. Also, in the middle of the document, there is a longer average coreference distance, meaning that the sentences in the middle probably refers to the beginning of the document. This means that many of the coreferences can not be captured by, for instance, picking the previous sentence in an effort to glue together a summary to increase its cohesion.

3 The Summarizer

Based on the results from our investigations of coreferences in news paper texts presented above, we have developed a summarizer (COHSUM) that takes into account the distribution of coreferences indirectly, by calculating a rank for the sentences based on how many out-links (how many other sentences are a representative sentence referring to) and in-links (how many sentences are referring to a current sentence). To calculate the ranks, a variant of PageRank (Brin and Page, 1998) is used, similar to TextRank (Mihalcea, 2004). Mihalcea (2004) further notes that the nature of PageRank is probably enough for summaries to exhibit some kind of coherence, since sentences that contain

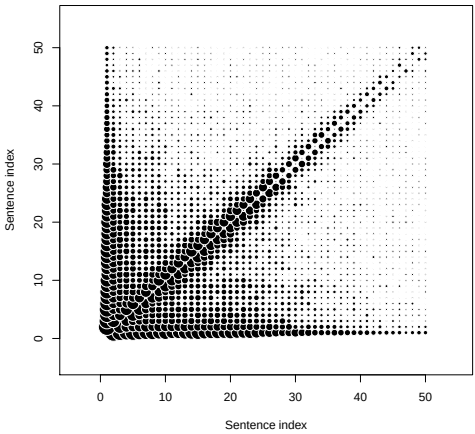

Figure 3: The figure shows a sentence by sentence matrix, where the radii of the circles depicts the number of times in average over 370 texts that a sentence corefers to another sentence. Early sentences and adjacent sentences corefer the most. The first 50 sentences are plotted. Coreferences within sentences are omitted.

similar information will be extracted. In COHSUM we take this one step further and extract coreferring sentences only. Coreferences have, as previously discussed, been used when creating summaries, however, using coreference chains in graph based ranking algorithms for summarization has not been done to our knowledge.

Each document that was to be summarized was first parsed and tagged using the CoreNLP-toolkit for coreference resolution. The coreference chains provided by the parser were used to create a graph, were each sentence is a node and all sentences having a referential relation to the sentence being in-/out links in the graph. In more detail; for each sentence, check if it exists in any coreference chain. For every coreference chain it exists in, count the number of sentences it refers to (with regards to noun phrases, possessive pronouns and named entities as mention earlier). Let these be the number of links. A reference consists, in the simplest case, of a two-way link, that is, if sentence A is referencing sentence B, then sentence A is also referenced by sentence B. A sentence can exist in multiple coreference chains but possible references within a sentence are not considered.

It is also possible for the parser to select the *representative mention*. In COHSUM mentions headed by proper nouns are preferred to mentions headed by common nouns, and nominal mentions are preferred to pronominal ones. In case of ties, the longer string is selected. The representative mention in the coreference chain can be considered as the preferred mention, or the most elaborate. The edges in the graph are weighted to prefer sentences with representative mentions; only sentences with representative mentions can be considered to have out links. Thus, sentences in a coreference chain that does not have the representative mention, will have 0 out links and X in-links, where X is the number of sentences in the chain minus one.

The sentences were ranked according to the number of links provided by the coreference chains using Equation 1, c.f. PageRank, which recursively calculates the number of links for a number of iterations (50 in our experiments, with d set to .85, c.f. Smith and Jönsson (2011a)). For our purposes, sentences containing the representative mention contain out-links while sentences lacking representative mentions only have in-links. Sentences with representative mentions referencing a high number of other sentences that are also referenced by a high number of sentences will thus receive a high rank. This means that sentences existing in multiple coreference chains will receive a higher rank, especially if that sentence has the representative mention for several chains.

$$PR^W(s_i) = \frac{1-d}{N} + d \sum_{s_j \in In(s_i)} w_{ji} \frac{PR^W(s_j)}{\sum_{s_k \in Out(s_k)} w_{kj}} \quad (1)$$

From the graph, weighted using Equation 1, COHSUM extracted the highest ranked sentences one sentence at a time until the summary consisted of roughly 100 words, to match the output from other systems and models. Focus was thus not for the summaries to retain the highest amount of coreference chains, but to be in comparable size to the resource data.

4 Evaluation

The summaries were evaluated using two measures; content coverage and cohesion. Content coverage is used to compare our summarizer with the systems from DUC (2002) as well as a baseline consisting of the first part of the documents. For evaluation of content coverage, ROUGE 1-gram F-measure (Lin, 2004) was used to compare summaries created by COHSUM to the summarization systems from DUC 2002. Other ROUGE measures are possible, but for this part of the DUC 2002 dataset (single document, 100 word summaries), ROUGE-1 has been shown not to differ significantly from other ROUGE measures. In total 533 texts were used[2], summarized by all 13 systems from DUC (concerned with producing single document 100 word abstracts), COHSUM and the baseline, FIRST, consisting of the first 100 words.

Cohesion is meant to be contrasted to content coverage; if content coverage is up to par, how cohesive are the texts? Looking at summaries as cohesive units and measuring the cohesion in them based on first parsing them with current parsers may be erroneous (Pitler et al., 2010). Current metrics may work for texts that are produced the way they are supposed to be read; in its entirety. Measures utilizing parsers (a common way of measuring cohesion, e.g through coreferences) used directly on summaries might not provide expected results, since the parsers expect the input texts to be correct. Thus, we have chosen to compare the summaries to the original documents. To calculate text cohesion when summarizing them, the coreferences in the summaries were logged in terms of what sentences coreferenced each other in the original documents. Depending on what sentences were retained in the summary, a coreference in the original document could be intact or broken:

Intact The amount of intact coreferences, that is, the amount of sentences that were retained in the summary that are coreferencing in the original document.

Broken Broken coreferences, sentences not extracted that contain the representative mention in the coreference chains. This case often leaves dangling anaphoric expressions without antecedent, leading to less cohesion.

[2]Duplicate texts from the corpus were removed.

Using the Stanford CoreNLP-toolkit, the original documents were parsed, followed by the DUC summaries, the 100 word summary, and the summaries created by COHSUM. The parser was used for the summaries even though coreference information from the summaries were not. This was to ensure that comparable outputs from the original documents and the summaries were created. By calculating the number of sentences in the summaries compared to the coreference chains in the original texts, we achieve a measure on how much of the cohesion that has been retained, given our measures of cohesion.

Table 1: Results on content coverage and cohesion. Results significantly worse than COHSUM in boldface.

System	Content	Intact	Broken
15	**0.442**	**3.318**	**3.775**
16	**0.425**	**2.991**	**3.294**
17	**0.158**	**1.959**	1.986
18	**0.432**	**2.973**	3.218
19	0.459	**3.531**	**3.878**
21	0.459	**3.805**	**4.292**
23	**0.410**	**4.409**	**4.939**
25	**0.443**	-	-
27	0.446	**3.806**	**4.282**
28	0.465	**4.231**	**4.692**
29	0.45	**4.217**	**4.817**
30	**0.114**	-	-
31	**0.443**	**2.505**	2.796
COHSUM	0.458	5.276	2.528
FIRST	0.459	10.587	0.417

5 Results

Table 1 shows the results from running the DUC summarizers, FIRST, and COHSUM on the 533 DUC 2002 news paper texts. The table shows the systems and their performance on gold standard comparison (Content) and cohesion (Intact and Broken). The blanks in the table are due to the systems 25 and 30 altering the summaries[3], making a coreference comparison fruitless.

We see that COHSUM is the fifth best system compared to FIRST and the DUC summarizers with regards to content coverage. The systems 28, 19, COHSUM 21, 29, 27 and FIRST perform best and compared to COHSUM no significant difference is obtained. COHSUM however, performs significantly better ($p < .05$) than the rest of the systems, 15, 16, 17, 18, 23, 25, 30, and 31, with regards to content coverage.

Comparing COHSUM to the DUC-systems with regards to coreference chains, reveals that one system's summary has fewer coreference chain breaks than COHSUM, no. 17. Compared to COHSUM there is a significant difference to all systems except systems 31, 18 and 17 with regards to broken coreferences ($p < .05$). Again, systems that perform significantly worse than COHSUM are marked as bold in Table 1. COHSUM has the most intact coreferences compared to the DUC systems. The number of intact coreferences is significantly higher in COHSUM than in all other summarisers ($p < .05$). FIRST is significantly better than all summarizers on both number of broken coreferences and intact coreferences.

[3]System 25 was a multi-document summarizer that was also tried on the single document summarization task, while system 30 focused on producing informative headlines.

6 Discussion

The performance on content coverage for COHSUM is surprisingly on par to the systems from the DUC 2002 competition (Table 1). Actually, most systems perform well on content coverage, the differences between the top systems were not significant. While performing on par with the DUC systems, the COHSUM summaries also have the highest amount of intact coreferences, and the second fewest breaks of coreference chains. The system with least broken coreferences, number 17, scores, however, low on content coverage. This indicates that by only taking into account the coreferences in a newspaper text, a summary that contains a high degree of important information can be created that also have a more cohesive structure in that they have fewer breaks in the coreference chains compared to other systems.

The baseline, FIRST, is still the clear winner. The nature in which newspaper articles are produced (where the gist of the story is presented first, with the rest of the article containing more detailed explanations, quotes and general development of the text) makes this kind of summary function well. COHSUM, making use of coreferences, will also often extract the beginning of the document, since this is where most of the coreferences are (c.f. Figure 3). The sentences in a document is often referring to the sentence before, however, most of the content seems to be introduced in the beginning which later parts of the document refer to. Thus, only doing a flat pick of the sentence before when trying to improve cohesion on a summary is not feasible (Smith et al., 2012).

The coreferences used in COHSUM are not weighted in any way, all sentences with coreferences are possible candidate sentences for inclusion in the summary. An informed decision on the type of coreference that should be allowed/weighted might affect the results. Our simplistic approach does not make this distinction since we were interested in sentences "being about" other sentences regardless of type. Currently all coreferences are considered as both in- and out-links if they contain a representative mention. The type of coreference could be further used to decide whether a link should be in one direction or another.

Using news texts has its limitations, as also pointed out by Over et al. (2007), but this is where most current research is conducted, and is, thus, important for benchmarking. It is, however, time for a new single text summarization competition where other text types are considered, texts that are important for the public to read and understand but where e.g. persons with reading disabilities have difficulties, such as authority texts and information texts, but also academic texts. Summarizing such texts (in Swedish) is in our focus of research, c.f. Smith and Jönsson (2011a) and our next step is to use COHSUM on these texts.

When it comes to other text types, the beginning of the document might not be as important. We have carried out some initial experiments on a variety of other text types. Looking at plots of the distribution of coreferences, similar to Figure 3, for other genres we find that scientific texts and financial publications seem even more terse with coreferences across the entire document, even though the first couple of sentences seem to contain a lot of coreferences in all the genres. This indicates that for these genres, the distribution of coreferences is different and taking for instance the lead sentences will break more coreferences and thus cohesion of the texts.

To summarize, COHSUM performs comparatively well with regards to content coverage, not significantly beaten by any system or the baseline but it has significantly fewer broken coreference chains and more intact coreferences compared to the other summarizers. It, thus, seems that coreferences are an important factor that can be tied to important sentences when summarizing news texts.

References

Alonso i Alemany, L. and Fuentes Fort, M. (2003). Integrating cohesion and coherence for automatic summarization. In *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics - Volume 2*, EACL '03, pages 1–8, Stroudsburg, PA, USA. Association for Computational Linguistics.

Baldwin, B. and Morton, T. S. (1998). Dynamic coreference-based summarization. In *In Proceedings of the Third Conference on Empirical Methods in Natural Language Processing (EMNLP-3*.

Barzilay, R. and Elhadad, M. (1999). Using lexical chains for text summarization. In Mani, I. and Maybury, M. T., editors, *Advances in Automatic Text Summarization*. The MIT Press.

Bergler, S., Witte, R., Khalife, M., Li, Z., and Rudzicz, F. (2003). Using knowledge-poor coreference resolution for text summarization. In *in DUC, Workshop on Text Summarization, May-June*, pages 85–92.

Boguraev, B. and Neff, M. S. (2000). The effects of analysing cohesion on document summarisation. In *COLING*, pages 76–82. Morgan Kaufmann.

Brandow, R., Mitze, K., and Rau, L. F. (1995). Automatic condensation of electronic publications by sentence selection. *Information Processing and Management*, 31(5):675 – 685.

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

Chatterjee, N. and Mohan, S. (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.

DUC (2002). Document understanding conference. `http://duc.nist.gov/pubs.html#2002`.

Gong, Y. (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*.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., and Cai, Z. (2004). Coh-metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, 36(2):193–202.

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

Kaspersson, T., Smith, C., Danielsson, H., and Jönsson, A. (2012). This also affects the context - errors in extraction based summaries. In *Proceedings of the eighth international conference on Language Resources and Evaluation (LREC), Istanbul, Turkey*.

Lee, H., Peirsman, Y., Chang, A., Chambers, N., Surdeanu, M., and Jurafsky, D. (2011). Stanford's multi-pass sieve coreference resolution system at the conll-2011 shared task. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, CONLL Shared Task '11, pages 28–34, Stroudsburg, PA, USA. Association for Computational Linguistics.

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

Mani, I., Bloedorn, E., and Gates, B. (1998). Using cohesion and coherence models for text summarization. In *AAAI Technical Report SS-98-06*.

Mihalcea, R. (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.

Mihalcea, R. and Tarau, P. (2004). TextRank: Bringing order into texts. In *Conference on Empirical Methods in Natural Language Processing*, Barcelona, Spain.

Nenkova, A. (2006). *Understanding the process of multi-document summarization: Content selection, rewriting and evaluation*. PhD thesis, Columbia University.

Otterbacher, J. C., Radev, D. R., and Luo, A. (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.

Over, P., Dang, H., and Harman, D. (2007). Duc in context. *Information Processing & Management*, 43:1506–1520.

Pitler, E., Louis, A., and Nenkova, A. (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.

Smith, C., Danielsson, H., and Jönsson, A. (2012). Cohesion in automatically created summaries. In *Proceedings of the Fourth Swedish Language Technology Conference, Lund, Sweden*.

Smith, C. and Jönsson, A. (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*.

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