COGNITIVE SCIENCE

MASTER THESIS

Automatic summarization and Readability

LIU-IDA/KOGVET-A–11/004–SE

Author:
Christian SMITH
chrsm588@student.liu.se

Supervisor:
Arne JÖNSSON
arne.jonsson@liu.se

List of Figures

List of Tables

Abstract

The enormous amount of information available today within different media gives rise to the notion of ways to reduce the inevitable complexity and to distribute text material to different channels or media. In an effort to investigate the possibilities of a tool to help eleviate the problem, an automatic summarizer called COGSUM has been developed and evaluated with regards to the informational quality of the summaries and with regards to the readability. COGSUM is based on word space methodology, including virtues such as problematic computational complexity and possibilities of inferring semantic relations. The results from the evaluations show how to set some parameters in order to get as good summary as possible and that the resulting summaries have higher readability score than the full text on different genres.

Contents

Chapter 1

Introduction

In todays' society, information is of essence and means of easily acquiring said information is an important area of research. The information should further be accessible to all humans regardless of handicaps or difficulties with language. Further, the web makes the information available everywhere at any time by anyone and published documents get visible instantly. The sheer amount of information gives rise to the notion of ways to reduce the inevitable complexity and together with the possibility of distributing text material to different channels or even media, some way of tailoring text encounters is desired.

To analyze the impact of a text on a reader, some analysis of the texts with regards to the perceived work load required by the reader is required. The notion of readability comes to mind, with research dating back to the 1920's. Research have through the years focused on investigating how readability can be measured, preferably automatically and how that correlates with the subjective experience within the reader. Properties within the text are often the focus, but the individual prerequisites should not be neglected.

The fact is, groups of readers are largely heterogeneous and individual adaptations might be required in order for text material be more easy to read. Further, types of texts might need to be treated differently as there is a difference in how texts are structured, the lengths of the texts, the media it is supposed to be read upon, etc.

One way of reducing the complexity of texts regardless of type or length is simply to reduce the amount of text to be read, thereby hopefull reducing the effort required by the reader. Due to the enormous amount of information available and the fact that it is an extremely time consuming task to do by hand, some way of automatically shortening the text without loosing too much information would be desired.

Automatic summarization is a challenging task however, especially if the process need to be quick, portable, independent of text type and genre and preferably language independent. If the summaries further can be tailored for different groups of people, big steps can be taken with regards to the accessability of information. Hopefully, by shorten texts automatically, it is possible to make texts easier to read as well as distributable to different medias, which is a start (or continuation) of a more accessible information society.

The work described in this thesis covers an automatic summarizer called COGSUM and some experiments regarding the readability of the resulting summaries on different lengths and text types, as well as some investigations on how the summarizer performs under different settings.

The purpose of the thesis is to review the word space technique behind the previously developed automatic summarizer called COGSUM(Jönsson et al., 2008) and to optimize its' performance, as well as to evaluate the readability of the resulting summaries. This is done through some experiments investigating the readability of Swedish texts of different genres, by using several automatic readability measures.

The thesis is laid out as follows. First, a theoretical background on the techniques used by the summarizer followed by a more detailed description of the developed summarizer. An overview of readability and automatic measures thereof is provided, with focus on the Swedish language. The summarizer is then explained in more detail, together with some evaluation of its' performance. The remainder contains various experiments on readability of the summaries created by the summarizer and their results.

Chapter 2

Background

In this chapter an overview of word spaces is presented, followed by some examples of implementations, with focus on Random Indexing which is one of the techniques used by COGSUM. Also, different parameter settings of word spaces that may have an impact on the quality of the summaries are reviewed. A brief overview of graph-based ranking is also provided. Background on readability is presented lastly.

2.1 The word space model

The word space model, or vector space model (Eldén, 2007), uses a spatial metaphor of a word's meaning (Sahlgren, 2006). The core idea is that semantic similarity can be measured by proximity in an n-dimensional space, where n can reach millions. More specifically, word spaces acts under the distributional hypothesis and the proximity hypothesis. In the distributional hypothesis, words with similar distributional characteristics, or words that occur in similar contexts have similar meanings. In essence, a *word* is the sum of it's contexts and the *context* is the sum of its' words. The *context* can be defined as some surrounding words or even an entire document or even corpus. The proximity hypothesis states that words close to each other in the word space have similar meaning while those far from each other have dissimilar meaning. Word spaces thus relies on distributional characteristics of words according to their contexts.

Practically, every word in a given context occupies a specific point in the space and has a vector associated to it that defines its meaning. One way of creating such a vector is to look at the co-occurence information of the words and plot it in a matrix. Each element in such a matrix tells of the co-occurance of the word on the specific row, the value will be 0 if the word doesn't occur with the word in the

column and 1 for each time the word occur with the word in the column. Each row can be seen as a vector with the dimensionality equal to the number of elements, or words in this case, thus effectively depicting a *context vector*, with the context being the immediate preceding and following word. Consider the sentence:

A red car is a vehicle too.

A corresponding co-occurence matrix can be seen in Figure 2.1.

Other types of matrices that have been used to build context vectors are words-by-documents. In the latter, each cell is occupied by a referenced document wherein a word can exist or not. Further, the word might be given a weight that specifies its' relative importance in that document, often calculated by its' frequency in the document and the overall document frequency of the word in total (Sahlgren, 2006).

	a	red	car	is	vehicle	too
a	0	1	0	1	1	0
red	1	0	1	0	0	0
car	0	1	0	1	0	0
is	1	0	1	0	0	0
vehicle	1	0	0	0	0	1
too	0	0	0	0	1	0

Table 2.1: Co-occurrence matrix for the sentence *A red car is a vehicle too*.

When the context vectors of a space has been constructed, it allows for some mathematical operations. This is done for instance by measuring the distance between the points in the space denoted by the vectors. The most common, however, is to calculate the cosine of the angle between the vectors to get a measurement of their similarity. A value of 1 means that the vectors are identical (the angle is 0). A value of 0 means that the vectors are orthogonal (the angle is 90 degrees) and a value of -1 means the vectors point in opposite directions (a degree of 180). Thus, the closer to 1 a comparison of two vectors (words or documents in this case) is, the more similar they can be considered to be.

A problem with vector spaces of this character in many dimensions is that they are very sparse and largely filled with zeroes, since words occurs zero times in the most contexts. The typical vector space of term-by-context also occupies tens of thousands or more dimensions and grows whenever data gets added. Many of the dimensions are therefore considered "wasted" and many words seem to have polysemous meanings by residing in contexts that is represented by many different lexical items. By the same virtue there is a problem with synonyms, words that appear in non-identical, but similar patterns in the space. The model in this untreated

way fails to generalize between terms that have non-identical patterns but could still be quite similar. Therefore it is practical to apply some form of dimension reduction or factor analysis to get a condensed representation of the space with co-occurrence patterns more elevated.

A word space can be implemented in several ways (Latent Semantic Analysis (LSA), Hyperspace Analogue to Language(HAL), Random Indexing (RI) to name a few) all using some kind of dimension reduction to increase the effectiveness of the space's semantic modelling and for the sake of computational efficiency.

In LSA, Singular Value Decomposition (SVD) is used to reduce the dimensionality. SVD is a matrix factorization technique that compresses the sparse space into a new space of much smaller dimensionality (a couple of hundred), while keeping the majority of the information intact. Many of the dimension lies latent when they are compressed due to the nature of SVD, so that vectors with similar contexts are grouped together. This makes for the possibiliy to find occurences with similar contexts that are non-identical, as well as a more managable space with fewer dimensions. Further, LSA uses a words-by-documents matrix with entropy-based weighting of the words (words containing much information is weighted higher). Latent Semantic Analysis can however be computationally demanding, much due to SVD. Gorrell (2006) proposes the use of Generalised Hebbian Learning for SVD to improve scalability.

Other methods to reduce the computational cost of LSA includes for instance Random Projection (Papadimitriou, Raghavan, Tamaki, & Vempala, 2000). This technique takes the LSA words-by-document matrix and projects it onto a random matrix of much lower dimensionality, where the distances between the points in the space are approximately preserved. SVD is then performed on the projected matrix, with a substantial gain in processing time as a consequence.

Hyperspace Analogue to Language (HAL) uses a different approach, with a directional word-by-words co-occurence matrix. This directional matrix accounts for the words appearing after the word in focus, with a weight associated with it based on the distance. Each row-column pair in the matrix is then concatenated to a vector with a dimensionality twice as large as the vocabulary. If the dimensionalitey gets too large to handle, a dimension reduction step is performed which discards the words that have the lowest variance, reaching 100-200 effective dimensions.

Random Indexing is another word space approach which presents an efficient, scalable and incremental alternative to standard word space methods.

2.1.1 Random Indexing

Random Indexing emerged from a family of dimension reduction techniques, including that of Random Projection (Sahlgren, 2005). Random Indexing is further

based on sparse distributed representations (Kanerva, 1988), a mathematical approach of representing human memory in an n-dimensional space. The basic idea of Random Indexing is to accumulate context vectors based on the occurrence of words in contexts. This technique can be used with any type of linguistic context, is inherently incremental, and does not require a separate dimension reduction phase as for instance Latent Semantic Analysis.

Random Indexing can be described as a two-step process:

Step 1 A unique d-dimensional *index vector* is assigned and randomly generated to each context (e.g. each document or each word). These index vectors are sparse and high-dimensional. They 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 on-the-fly. As scanning the text, each time a word occurs in a context (e.g. in a document, or within a sliding context window, w), that word's d-dimensional index vector is added to the context vector for the word in focus. 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.

Random Indexing poses several advantages over other word space techniques. No further dimension reduction technique is necessary, since the dimensionality is set beforehand to a much lower value than the number of contexts. Since the technique also is incremental, only a couple of examples need to be encountered before similarity computations are possible, when other techniques need to collect an entire dataset.

2.1.2 Dimensionality

Word spaces are often used on large data sets to get a good semantic representation of meaning. When dealing with large data sets, there is also a constant trade off between the quality of the word space and the complexity of it, affecting further operations within a reasonable processing time and storage space. Dimensionality is a parameter that has a large effect on time and storage, which is why several methods are proposed to reduce it. Much of the debate regarding word space is about what dimensionality to choose (Karlgren, Holst, & Sahlgren, 2008). The choice of dimensionality is often 100 for LSA and 1000 for Random Indexing (Karlgren et al., 2008), numbers that are acquired through trial and error on synonym tests. Karlgren and Sahlgren (2001) uses a dimensionality of 1800 to perform a TOEFL-test with good results. It is noted that what dimensionality to use is dependent

on the size of the text material and Chatterjee and Mohan (2007) uses a dimensionality of 100 for smaller documents of a couple of hundred words. In the task of automatic summarization however, Hassel (2007) performs experiments on different dimensionalities of Random Indexing and concludes that the dimensionality doesn't have a large impact on the quality of the extracts and that the random factor when creating the indexes is larger.

2.1.3 Stop words

The use of stop words has also been through some debate. By removing extremely high frequent words it is possible to drastically reduce the size of a text thus making it easier to handle computationally. The notion is that words that occurs in many different context doesn't add anything important to the meaning of the text and thus can be removed. Stop words often include words like prepositions, pronouns and participles. Not only does the text get computationally more managable, in some cases the quality of the word space may be increased by removing stop words (Hassel, 2007). In words-by-documents matrices or entire document queries stop words might be important to remove, because those words doesn't help to distinguish documents (Eldén, 2007). Conversely, other sources have used the most common words as important to denote the style in the text (Campbell, January 2003) rather than semantic content, indicating that the choice of what words should be included the space depends on what information one wants to provide.

2.1.4 Focus window

The size of the focus window is effectively the size of the context of each word. For instance, a window size of 5x5 denotes a context consisting of the five preceding words of the current focus word and the five following. The context can also be weighted so that words that are closer to the focus word gets a higher weight while those further away get lesser weight. One example of weighting is the inverse of the distance of the focus word to neighbouring words where the weights of a 3x3 window would look like $[0.25, 0.33, 0.5, 0, 0.5, 0.33, 0.25]$. Another example is according to the formula 2^{1-l} where l is the distance to the focus word, giving a weighting vector $[0.125, 0.25, 0.5, 1, 0, 1, 0.5, 0.25, 0.125]$ which gives a more aggressive falloff of the values where words closer to the focus words are considered more important. For determining syntactic category of a word, narrow context windows (1x1) with aggressive weighting is better, while a larger window (2x2, 3x3) gives better results when inferring semantic relations (Sahlgren, 2006).

2.2 Graph-based ranking

COGSUM uses a weighted PageRank algorithm in conjunction to the Random Index-space of a text to rank its' sentences, as previously have been shown by others to be successfull in the task of summarization (Chatterjee & Mohan, 2007). Graph-based ranking algorithms have been used in a variety of situations, most notably perhaps in the analysis of link structutre of web pages as in the case with Google's PageRank (Brin & Page, 1998). In general, such algorithms are used to decide the weight of vertices in a graph by taking into account the whole graph structure recursively as opposed to only local vertex specific characteristics (Mihalcea & Tarau, 2004). To use PageRank for summaries an undirected 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, see Figure 2.1.

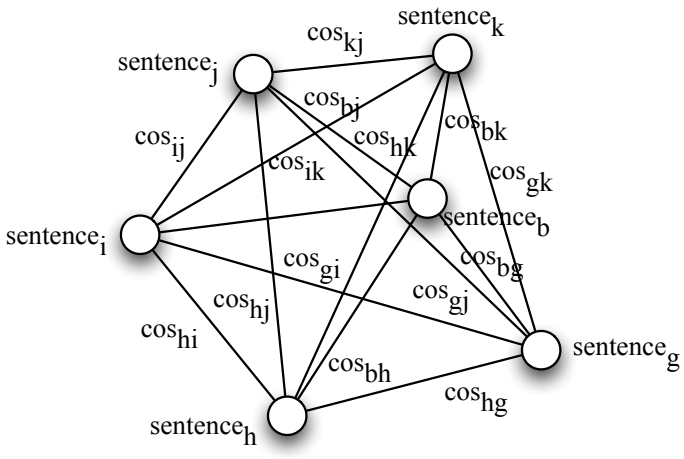


Figure 2.1: A simplified graph where sentences are linked and weighted according to the cosine values between them.

The algorithm rank ingoing and outgoing links to pages depending on the number of links as follows:

$$PR^W(s_i) = \frac{1-d}{N} + d \sum_{s_j \in In(s_i)} w_{ji} \frac{PR^W(s_j)}{\sum_{s_k \in Out(s_j)} w_{kj}} \tag{2.1}$$

where s_i is the sentence under consideration, $In(s_i)$ are the set of sentences that link to s_i, $Out(s_j)$ are the set of sentences that link from s_i and N is the total

number of sentences. df is the damping factor.

The damping factor is originally set to account for the possibility of a surfer clicking a "random web link when he gets bored" (Brin & Page, 1998). With regards to the ranking of sentences, we see the damping factor as the possibility of a sentence containing some implicit information that a certain reader might consider more important at the time. This is much in line with (Mihalcea & Tarau, 2004) that argues that the process can be seen as "text surfing" much alike "web surfing"; connected concepts exists through the discourse that can be followed via semantic or lexical links and that there are related facts spread out that ties the discourse together.

In COGSUM the vectors for whole sentences and the similarity between these and the average document vector are of interest. The average document vector is calculated by dividing the total document vector, which consists of the sum of all unique words' context vectors, with the number of unique words in the document following the formula 2.2

$$\vec{doc} = \frac{1}{N}\sum_{i=1}^{N}\vec{w_i} \tag{2.2}$$

where N denotes the number of unique words.

The sentence vectors are then calculated by subtraction of the average document vector from the context vectors of the words in the sentence which are summed together and divided by the number of words in the sentence as in the formula 2.3.

$$\vec{sent_j} = \frac{1}{S}\sum_{i=1}^{S}(\vec{w_i} - \vec{doc}) \tag{2.3}$$

where S denotes the number of words in sentence j.

The computation of the algorithm is carried out on all sentences iteratively until node weights converge. Sentences with similar content will then contribute with positive support to each other. This does not exclusively depend on the number of sentences supporting a sentence, but also on the rank of the linking sentences. This means that a few high-ranked sentences provide bigger support than a greater number of low-ranked sentences. This leads to a ranking of the sentences by their importance to the document at hand and thus to a summary of desired length only including the most important sentences.

2.3 Readability

In this section some notions of readability is presented, along with some ways of measuring them. The measurements that are examplified are all automatic and have

in commmon that they in one way or another correlates with human understanding of the readability of texts.

2.3.1 What is readability?

Generally, easy-to-read material is characterized by simple straightforward language without necessarily being simplistic or childish. Linguistically, this can be achieved by removing grammatical features while keeping close to the original meaning of the text, thus reducing the effort required by the reader to get the same information as in the original text. These features can operate on a syntactic, lexical or textual level to make text material more readable and comprehensible (Mühlenbock & Kokkinakis, 2009). Research on readability have been carried out since the 1920's, mainly in the US. The early work focused on vocabulary elements such as word length, percentage of multisyllabic words, subordinate clauses etc. Readability can also be seen from various angles. In part, it is the extent to which the reader can understand a written text, and the psychological processes involved within the reader. Here, focus lies on individual shortcomings with regards to perception and understanding of the written text and not on the text itself. Readability can also be seen as a measurable property of a given text where the individual prerequisites in terms of psychological abilities are often neglected. The latter have long been pursued as it has been useful to draw statistical conclusions on readaility in experiments. This have provided several formulas for different languages that have made automatic measures possible on larger texts.

2.3.2 Formulas

Formulas of readability for English is abundant, e.g. The Flesch Reading Ease Formula, Flesch-Kincaid Grade Level, Dale-Chall, the Coleman-Liau test, Gunning Fog and SMOG (DuBay, 2004).

The Flesch Reading Easy score can be computed as:

$$Score = 206.835 - (1.015 \times \frac{n(w)}{n(s)}) - (84.6 \times ASW) \tag{2.4}$$

where $n(w)$ denotes the number of words, $n(s)$ the number of sentences and ASW the number of syllables.

The measures correspond to how understandable a text is, e.g. a Flesch Reading Easy score between 70 and 80 is "Fairly Easy" which means that the text can easily be understood by a (U.S.) 7th grade student. The Flesch-Kincaid Grade level is a U.S. grade level version that normalizes (2.4) to correspond to readability for students in various grades.

Björnsson (1968) was the first to introduce a similiar formula for Swedish, namely LIX (läsbarhetsindex, readability index) which is the almost exclusively used readability measure for Swedish. LIX relies partly on the notion that Swedish is an inflecting and compounding language, by taking into account long words. LIX measures the number of words per sentence and also the number of long words (> 6 characters) in the text through the formula:

$$LIX = \frac{n(w)}{n(s)} + (\frac{n(words\ >\ 6\,chars)}{n(w)} \times 100) \tag{2.5}$$

where $n(s)$ denotes the number of sentences and $n(w)$ the number of words.

Contrary to Flesch's original formula (and many of its modifications) the LIX formula does not consider syllables but instead word length. As LIX only considers ratios, sentence length and proportion of long words, it does not depend on text length. A text with many long words and long sentences is considered more complex and therefore more difficult to read, indicated by a high LIX value, see Table 2.2, from Mühlenbock and Kokkinakis (2009).

Table 2.2: LIX-values for different genres.

LIX value	Text genre
-25	Children's books
25-30	Easy texts
30-40	Normal text/fiction
40-50	Informative text
50-60	Specialist literature
> 60	Research, dissertations

Other formulas for Swedish include NR and OVIX (Mühlenbock & Kokkinakis, 2009; Rybing, Smith, & Silvervarg, 2010). Lexical variation or OVIX (word variation index) measures the ratio of unique tokens in a text, calculated as:

$$OVIX = \frac{log(n(w))}{log(2 - \frac{log(n(uw))}{log(n(w))})} \tag{2.6}$$

where $n(w)$ denotes the number of words, $n(uw)$ the number of unique words, and $n(s)$ the number of sentences. OVIX does not depend on text length (Lundberg & Reichenberg, 2009).

NR is calculated by dividing the number of nouns, prepositions and participles with the number of pronouns, adverbs and verbs:

$$NR = \frac{n(noun) + n(prep) + n(part)}{n(pro) + n(adv) + n(v)} \tag{2.7}$$

where $n(noun)$ denotes the number of nouns, $n(prep)$ the number of prepositions, $n(part)$ the number of participles, $n(pro)$ the number of pronouns, $n(adv)$ the number of adverbs, and $n(v)$ the number of verbs.

A higher NR indicates a more professional and stylistically developed text, while a lower value indicate more simple and informal language. In some contexts a low NR can indicate a narrative style, such as in children's books.

2.3.3 A broader view

These measures can by themselves correlate with the perceived readability of a text, however, they should not be seen as the only way of specifying the readability of a text. One problem is that they cover only certain parts of what constitutes readability. Another difficulty in assessing methods for measures of readability is the hetergenuosity of the different groups of readers (Mühlenbock & Kokkinakis, 2009). LIX has therefore been considered insufficient for a complete readability assessment of a text as it is based only on the surface structureof the text. LIX was originally developed to assess the readability of school books and might not always be applicable and should be used with caution. Even though long words are considered as an indication on the readability of a text, short words can not be neglected simply as "easy". Bernhardt (1984) tested for the relationship between word length and difficulty in second language readers and found that words such as *the, them, they, their, these* are difficult despite their short length, much thanks to their graphemic similarity. Eye-movement studies have shown that prepositions and articles are processed for an excess time of $>$1000 milliseconds, three times longer than for longer words.

Chall (1958) concluded that there are generally four types of elements that seem to be significant for a readability criteria; vocabulary load, sentence structure, idea density and human interest and further states that vocabulary is of prime importance for readability, that is, the number of difficult words in a text is highly determinate of the readability. However, even if the number of difficult words in the text is stated as an important factor, it takes a large amount of difficult words for the readability to decrease. Therefore, it might be a mistake to interpret correlations between the results of vocabulary tests and reading profiency and to state vocabualry as such an important factor in text comprehension(Freebody & Anderson, 1983). Duffy and Kabance (1982) also found that a simplified vocabulary and sentences hade little to no effect on the performance on comprehension even though formulas such as Flesch-Kincaid stipulates this. They state that word and sentence

difficulty is of correlative but not causative factors in comprehension (Duffy & Kabance, 1982).

A way of getting a better view of a text's readability is to map the psychological criteria of the individual with several measurable readability properties of a text. This is done by mapping Chall's elements of readability to the automatic measures (Mühlenbock & Kokkinakis, 2009). LIX is for instance mapped to *vocabulary load*. Further, in an effort to enhance the LIX formula, Mühlenbock and Kokkinakis (2009) have included an additional parameter called Extra Long Words (XLW). Extra Long Words are words with more than 14 characters and indicates a larger proportion of compounds of usually three or more stems, relatively common in Swedish. OVIX can be used to indicate the *idea density*, in conjunction with the nominal ratio (NR). The degree of *human interest* is measured simply through the proportion of proper nouns (PN) and by measuring the length of sentences (ASL), *sentence structure* can broadly be gathered (Mühlenbock & Kokkinakis, 2009).

This way, a wider application of the measures is applied and by taking into account additional parameters, a better view of a text's readability can be achieved.

Chapter 3

The summarizer

This chapter reviews in more detail how the summarizer performs when summarizing texts.

3.1 Background

COGSUM is based on Random Indexing (RI) and PageRank to rank the sentences in a text according to importance. When the text has been processed using RI and PageRank, the most important sentences are extracted, for instance 30% of the original text, resulting in a condensed version of the original text with the most important information intact. Since all sentences are ranked, the length of the summary is easy to specify. It is important to note that the algorithm only takes the current document as total context and the information within the document, without any knowledge from an outside corpus. This makes it highly portable to different domains, genres and languages (Mihalcea & Tarau, 2004). It is further based on Java and uses the RI toolkit available at (Hassel, 2011). Initial evaluations of COGSUM with human users show that summaries produced by COGSUM are useful and considered informative enough (Jönsson et al., 2008). COGSUM has also been evaluated on gold standards for news texts and authority texts showing that it is better than another Swedish summarizer (SweSum, Dalianis et al. (2003)) on authority texts and almost as good on news texts, texts that the other summarizer was especially adapted to handle (Gustavsson & Jönsson, 2010).

Previous studies have recommended a certain setting on some parameters, mainly dimensionality, for texts of a particular length. The next section describes some trials to optimize the dimensionality for larger texts using the PageRank method, as research on the matter is sparse. It is suspected that different settings provide summaries of different qualities.

3.2 Evaluation

The traditional way of dealing with word spaces is to build a space on a very large data set to get a good statistically sound representation of the meaning of words. The method proposed by Chatterjee and Mohan (2007) that is used in COGSUM uses word spaces that are constructed only on a local much smaller context using only the document to be summarized rather than an outside training corpus. The sentences in a text are instead ranked by an iterative process of recommendation so that a summary can be extracted, thus how the settings of the word space affect the summaries is in this sense unclear. To investigate how the parameters affected the relatively small space with regards to the quality of the summaries, some experiments were performed.

3.2.1 Dimensionality

The impact of dimensionality to the quality of the summaries were evaluated, as research on the matter on smaller texts is sparse. As previous studies (Gustavsson & Jönsson, 2010; Chatterjee & Mohan, 2007) have shown, for shorter texts (200-300 words) a dimensionality of 100 is sufficient. It is unclear, however, how the dimensionality should be set when the text size increase to a size of about 500 words or more, such as often is the case with informative authority texts. Gustavsson and Jönsson (2010) used the dimensionality 250 to preserve a dimensionality reduction of 50% as in the case with smaller texts at 100 dimensions. Except for the dimensionality, the recommended context window size is 2x2 with a weighting of $[0.5, 1, 0, 1, 0.5]$.

One way to test what dimensionality that should be chosen is simply to summarize texts using different dimensionality and then test the resulting texts against a gold standard to check correspondence with humans. To evaluate the summaries, AutoSummENG (Giannakopoulos, Karkaletsis, Vouros, & Stamatopoulos, 2008) was used to compare the summaries on different parameter settings to a gold standard (manually created summaries which are considered optimal) created by Carlsson (2009). AutoSummENG is an evaluation system that compares graph n-gram similarities between different texts to get a value on how similar they are. It allows for a several texts (summaries) to be compared with several target texts (gold standard texts developed by humans) and get a value denoting the performance of the summarizer.

To evaluate the dimensionality, summaries (10% of the sentences in the original texts) on different dimensionalities were created on five different texts from Försäkringskassan, all around 1000 words in length. The different dimensionalities were 10, 20, 50, 100, 300, 500, 2000 and 10000. For each dimensionality, twenty

different summaries were created on different random seeds, each compared to the gold standard. The avarage score of the summaries on the different seeds are shown in figure 3.1. A dimensionality of 100 performed best (0.320) while worst at 10000 (.297). At 10 the results were .298. As a comparison, randomly selected sentences performs at (.236).

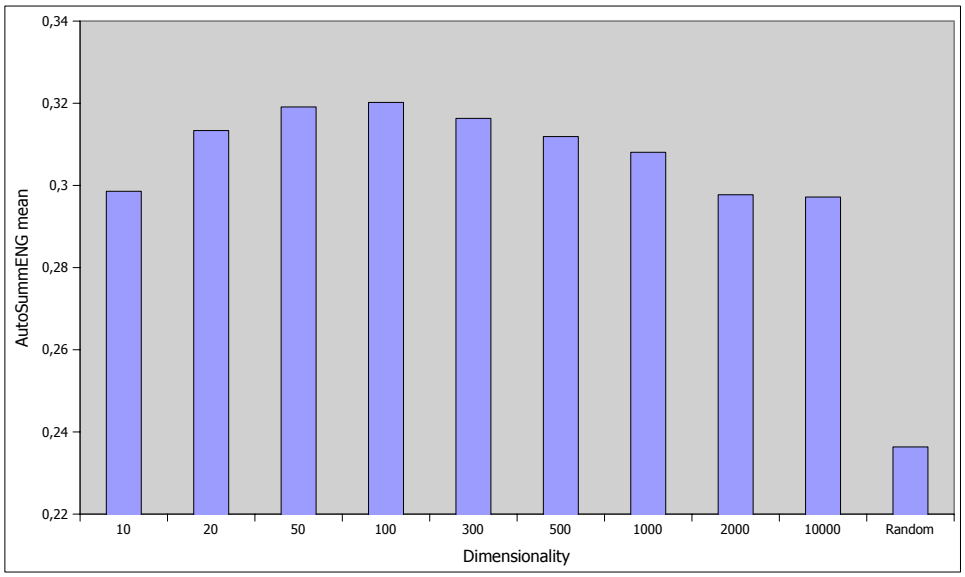


Figure 3.1: Evaluation of summaries on different dimensionalities. The X-axis denotes different dimensionalities, the Y-axis plots the mean value from evaluations on several seeds on each dimensionality.

The results indicate that dimensionality doesn't have a large impact on the quality of the summaries, see 3.1, as long as it is within a certain range. This is in agreement with Hassel (2007) who states that the random factor when building indexes is higher than the factor of dimensionality, as long as the dimensionality is large enough.

To further experiment with the behaviour and to find alternate means of investigating the effect of dimensionality, some additional experiments were performed. By visualizing each step in the ranking process, the behaviour of the PageRank algorithm when dealing with different parameters could be investigated. This was to hopefully get a better view of how some of the parameters are affecting the creation of the summary, and how one can evaluate the impact of parameters. The ranking of the sentences were plotted on each iteration of the algorithm for different settings of the parameters.

Each series in the in the graph depicted in Figure 3.2 represents one sentence.

The Y-axis includes the value of the sentences in the graph during the PageRank-algorithm; the higher the value, the more important a sentence is in the document. On the X-axis, each iteration is plotted. All sentences begin at zero before any iterations have been performed and the PageRank-algorithm is then iterated 50 times. Some sentences get a lower value each iteration while others get a higher value. By twenty or so iterations, the values stop changing (or the changes are really small) and have thus converged; no further iterations are necessary. Depending on the setting of for instance dimensionality, it should be possible to spot differences in how the graphs develop in terms of ranking of the sentences, if the parameters have an effect on the comparison process (cosine-measure between sentences).

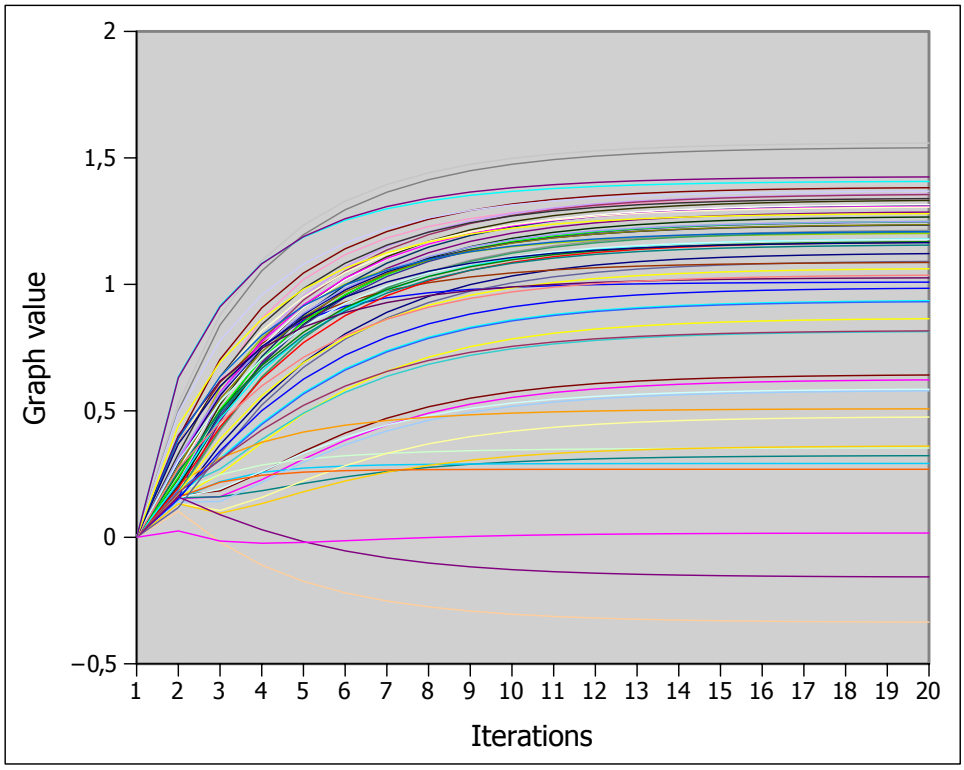


Figure 3.2: The iterations of PageRank. The figure depicts the ranks of the sentences plotted on the Y-axis and the iterations on the X-axis. Each series represents a sentence.

Figure 3.3 illustrates how the rank of each sentence is affected each iteration of PageRank in different dimensionalities. The graph to the furthest left uses a dimensionality of 10, the one next to it uses 100, the next 300, and the last uses a

dimensionality of 1000.

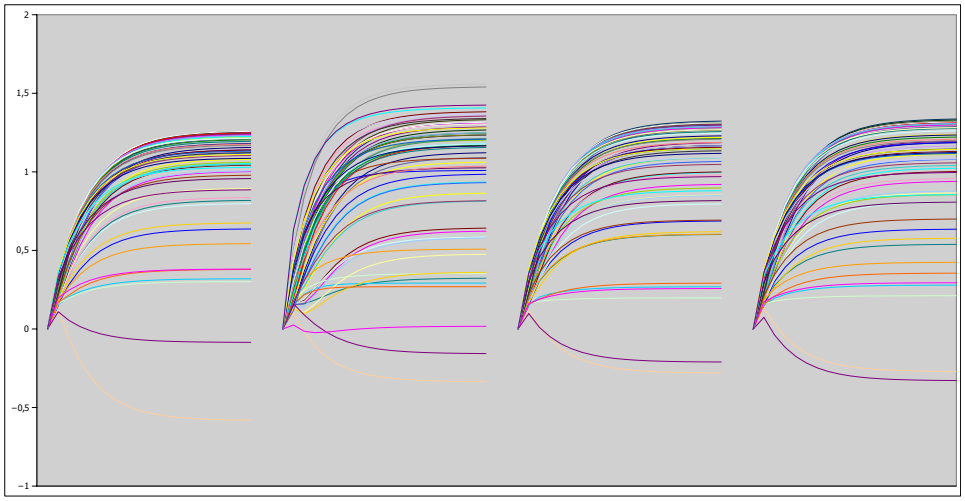


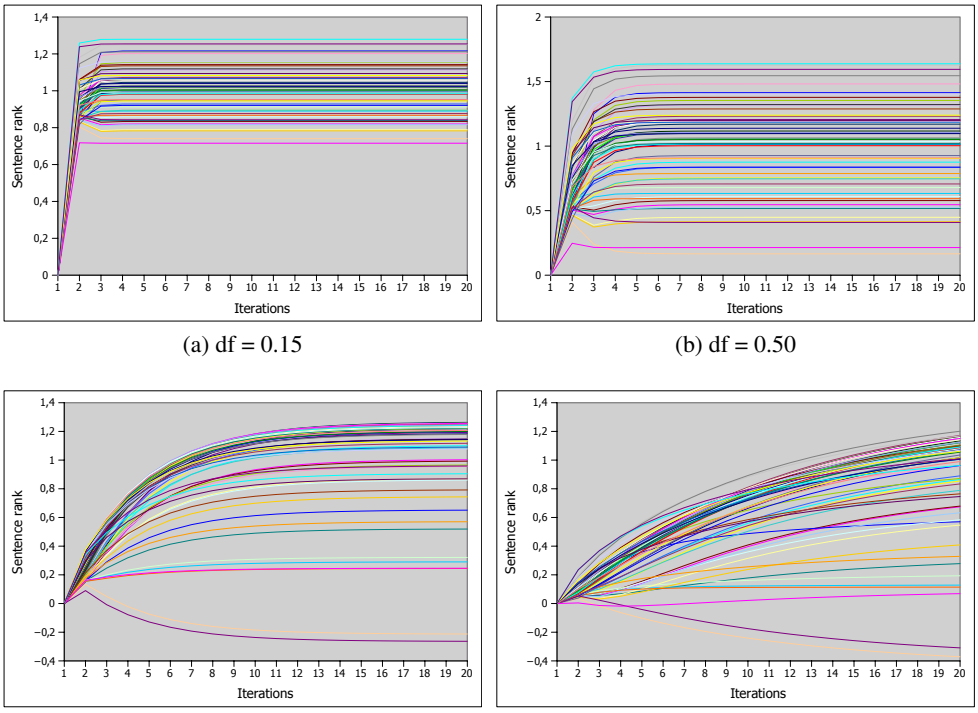
Figure 3.3: The iterations of PageRank in different dimensionalities of the RI-space. The figure depicts 4 different graphs, each representing the trial of a specific setting of the dimensionality. From the left the dimensionalities of 10, 100, 300 and 1000 were used. The ranks of the sentences is plotted on the Y-axis and the iterations on the X-axis. Each series represents a sentence.

There seems in this isolated case to be a difference in how the sentences are separated in a dimensionality of 10 compared to a dimensionality of 100 for instance. Between 100 and 300 there is also a difference, although seemingly smaller. Between 300 and 1000 there is an even smaller difference. Nonetheless, as the algorithm proceeds through the iterations the sentence ranks is separated and convergence happen after $\approx 15 - 20$ iterations.

In the Figure 3.4 the same thing is plotted, the sentence ranks on Y and iterations on X, in this case however the damping factor is altered in the PageRank algorithm on a static dimensionality of 100. The graphs use a damping factor (df) of $df = .15, df = .50, df = .85$ and $df = .95$ respecitvely. Note that at $.15$ the values converge already after 3 iterations while at $.95$, it takes much longer.

3.2.2 Randomness

In order to test the effect of the randomness in the space, the same settings on 10 different random seeds were used for d=10 up to d=10000 (Figure 3.5). A seed can be used to make sure that the same random numbers are generated each time. By changing the seed it can be specified when the random numbers should be different



(a) df = 0.15

(b) df = 0.50

(c) df = 0.85

(d) df = 0.95

Figure 3.4: Different usage of damping factor during the creation of summaries.

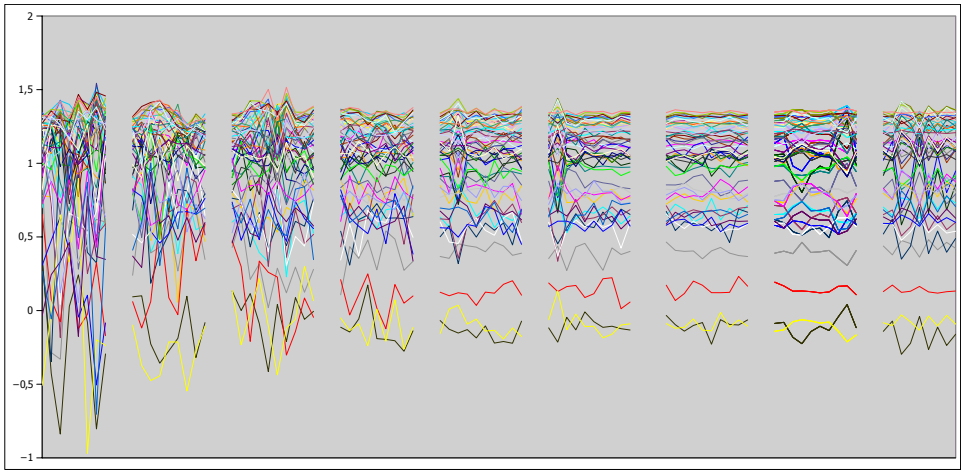


Figure 3.5: Effect of randomness, same text and settings on ten different random seeds. The final ranks of the sentences is plotted on the Y, with the different seeds on X. The graph depicts 9 trials at following dimensionalities, from left: 10, 20, 50, 100, 300, 500, 1000, 2000, 10000.

and not. This can be handy during evaluation.

In Figure 3.5 the final ranks of the sentences are plotted on the Y-axis, with the different seeds on the X-axis. There are 9 graphs in the diagram, each depicting a trial on a specific dimensionality. From the left the dimensionalities used are: 10, 20, 50, 100, 300, 500, 1000, 2000 and 10000. On each graph, the X-axis depicts trials on a different random seed. For lower dimensionality the lines are jagged and crosses everywhere, meaning the random effect is large, the values differ alot between the trials. For higher dimensionality however (d=500 & d=1000), the random effect seems to be smaller as the lines are more straight and it does not matter what seed that is used, the values are still the same. At 2000 and 10000, the lines are more jagged again, suggesting that a smaller dimensionality provides for smaller amount of randomness. This is in line with the results from the evaluation of summary quality in Figure 3.1; the quality is best where the effect of randomness is smaller.

3.2.3 A text that is too small

Sometimes, the ranks seem to get into a state where they don't converge (see Figure 3.6). This happens more frequently when the text is small, around 10 sentences. If the dimensionality is lowered however, the effect of convergence can be achieved. In Figure 3.6 the same text with a dimensionality of 100 is lowered to a dimension-

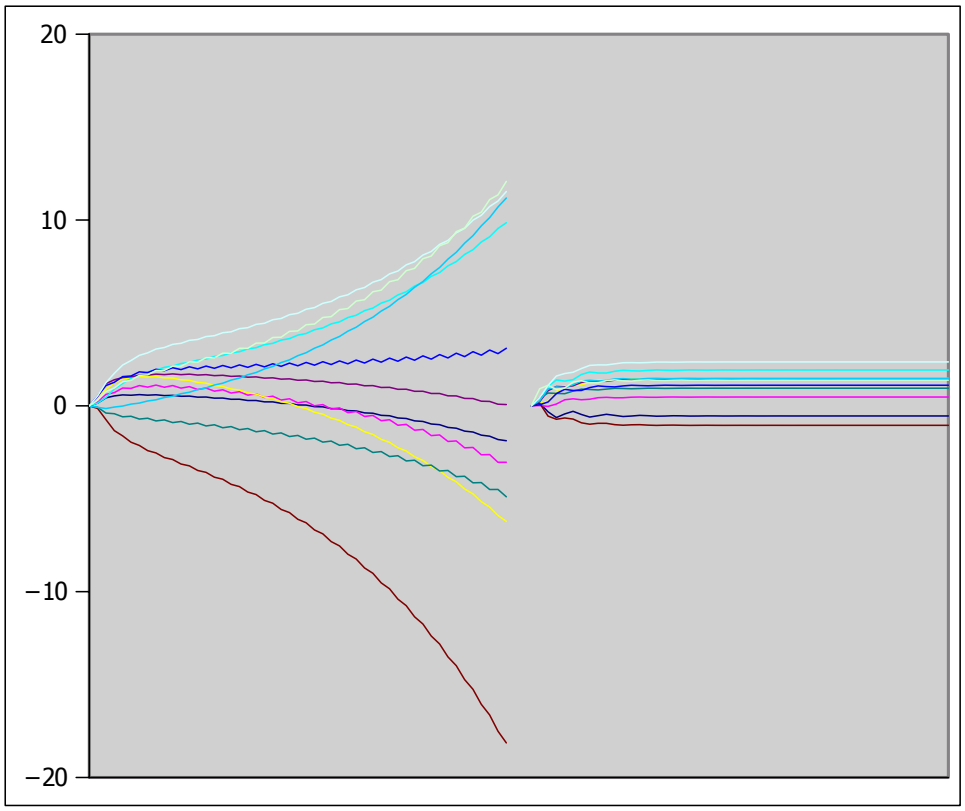


Figure 3.6: Values sometimes don't converge on smaller texts. The left graph depicts the text in d=100 and the right graph d=20.

ality of 20 to gain the effect of convergence as an example.

3.2.4 Stop words

By removing stop words from a large dataset, the quality of the semantic representation therein can be increased. Stop words are words that do not add anything to the meaning of the contexts, as they can appear in many different contexts. Chatterjee and Mohan (2007) use a different approach; by removing the avarage document vector from each sentence, the impact of words that appear in many different contexts is lessened in the sentences. To investigate the impact of stop words using both methods some trials were performed on a text consisting of approximately 2500 words.

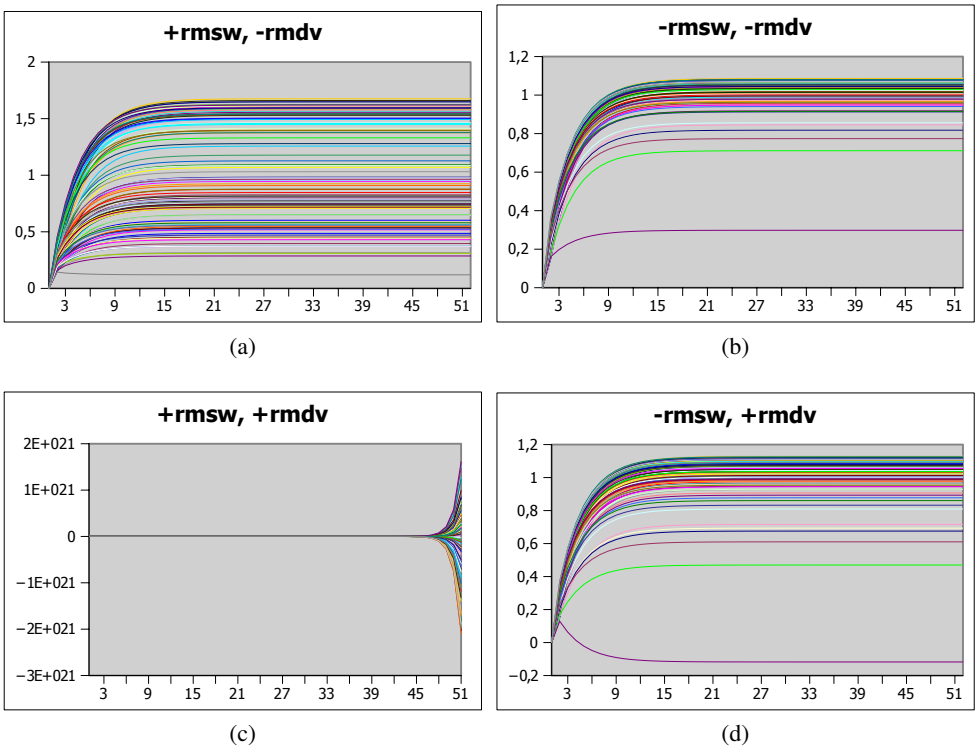


Figure 3.7: Different usage of stop words in the creation of summaries. (a) Only stop words are removed, (b) Nothing is removed, (c) Both stop words and the document vector are removed, (d) The document vector is removed.

The figures in 3.7 depict graphs at different settings regarding stop word usage.

On one occasion the stop words were removed instead of subtracting the document vector (+rmsw, -rmdv). In the second case neither the stop words or the document vector were removed (-rmsw, -rmdv). In the third case, both stop words and document vector were removed (+rmsw, +rmdv) and in the last case, only the document vector was removed (-rmsw, +rmdv).

The graphs all depicts all sentence ranks on Y and iterations on X, the same as before.

If nothing is done to account for stop words (Figure 3.7b), no sentence get negative weight and overall, a tighter cluster around the top can be observed. If the stop words are removed instead of the document vector (Figure 3.7a), the sentence ranks get spread out with no distinct top sentences, and also, no sentence with negative weight. If the stop words are removed in conjunction with the document vector (Figure 3.7c), too much information is removed and the graph fails to converge. If only the document vector is removed (Figure 3.7d), some sentences get lower or even negative weight, while there still is a cluster of top sentences.

With regards to the quality of the summaries, no significant difference could be observed.

3.2.5 Context window

Figure 3.8 depcits three graphs; the first on a 1x1 context window, the second a 2x2 and the third a 3x3. A narrower context window seems to create less distinct clusters of sentences, with some sentences getting a higher overall value. There were no significant difference in the quality of the summaries on different window sizes however.

3.3 Discussion

Instead of relying on a large semantic data sets, RI have been used in an iterative fashion where small local contexts goes through a recommendation process. In this sense it is not a universal semantic knowledge acquired from a large corpus that is valuable, but rather a measurement of shared local contexts in small documents. Thus it is not meaningful to talk about a general semantic knowledge in the space as in the case with large corpora, but rather local contextual awareness.

With regards to the quality of the summaries it can be concluded that the settings of the parameters doesn't have a large impact on the resulting summaries using the proposed method. However, since a relatively large amount of randomness is involved, it is necessary to see how the parameters can be tuned so that the results suffer the least from it. In the case with dimensionality, since it has an

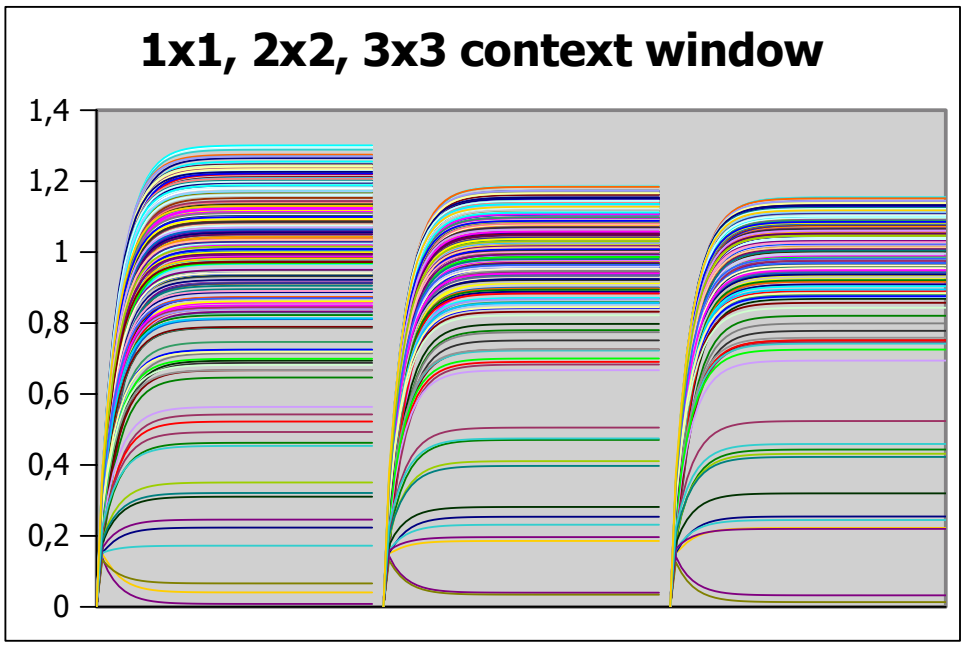


Figure 3.8: Three different trials on different context window sizes; 1x1 to the left, 3x3 to the right and 2x2 in the middle

impact on processing time, the lowest dimensionality possible would be desired. The results have indicated that a dimensionality of 100 is sufficient even for larger texts of around 1000 words. As observed in experiments, the random factor can be quite large for smaller texts, but as long as the dimensionality is sufficient the random factor should not be the prime factor.

It should also be noted that a certain care must be taken when trying to summarize smaller texts, as there is a risk of them not containing enough information for the recomendation process to be meaningful. The failure of convergence as seen in Figure 3.6 probably can be traced to the PageRank-algorithm and more precisely to where the document vector is subtracted from each sentence. If too many sentences are too close to the document vector, the effect of the reduction of the document vector can be that the sentences are left with a representation of a context that is dissimilar to every other sentence. In other words, if there isn't a 'central theme' in a text, then each sentence doesn't have anything to vote about or recommend. In the PageRank-algorithm, when summing up the 'outgoing' weights of a sentence, the value becomes close to zero making each iteration increase (or decrease) the ranks in an exponential fashion when the weight is divided with that value close to zero. Similarly, Eldén (2007) states that a web page that has no outlinks and only gets its' weight from inlinks may not converge. A summarized cosine measure for a particular sentence that is zero may thus translate to a web page that has zero outlinks. The effect of non-convergence can also happen if too much information is removed from the sentences, as seen in the case with the removal of stop words in Figure 3.7c.

By removing the document vector from each word when building the sentence vectors, the impact of sentences containing contexts that include non-function words can be lessened. In Figure 3.7d for instance, a particular sentence get negative weight in the text used in the trial, compared to when nothing was done to account for stop words as in Figure 3.7b. If stop words are removed as well as the document vector, the sentence ranks fail to converge (Figure 3.7c). This is probably due to that the sentences get to little information and are treated as having no out links. If stop words are removed instead of removing the document vector (Figure 3.7a) the sentence ranks are more spread out, with no distinct "top" sentences.

The damping factor has a large effect on the behaviour of the ranking algorithm as can be seen in Figure 3.4. The setting of the damping factor should be set so that enough recomendations between sentences can occur before convergence. At .15 only three iterations are performed before convergence and the resulting ranks differ alot from when more recomendations have been performed. At .50 a couple more is needed before convergence and at .85 around fifteen. When the damping factor increases, so does the number of iterations needed before convergence. Since the number of iterations is determinate of the processing time, a low enough

number should be set. .85 is recomended by other sources((Brin & Page, 1998; Chatterjee & Mohan, 2007)) and has also been used here. It would be interesting to investigate more deeply how the damping factor affects the actual quality of the summaries.

The effect of the context can be seen in Figure 3.8 where the size of the focus window is altered. A smaller context region makes for not as distinct separation of sentences, however, the differences in the quality of the summaries were not significant.

Chapter 4

Readability experiments

This chapter presents some experiments performed on automatically created summaries regarding readability and some conclusions thereof.

4.1 Readability of summaries

We have evaluated summarized texts from a readability perspective by creating summaries of texts from different genres and compared their readability to the original text.

We used three types of texts representing three different genres:

- **DN**. Newspaper texts from the Swedish newspaper ”Dagens Nyheter”; ca 25,000 words divided in 130 articles.
- **FOF**. Popular science texts from the Swedish Magasine ”Forskning och Framsteg”; ca 20,000 words divided in 31 articles.
- **FOKASS**. Authority texts from the Swedish Social Insurance Administration (Sw. Försäkringskassan); ca 25,000 words from 2 brochures. The brochures were divided so that each chapter was an article resulting in a total of 35 ”articles”

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 different lengths (30%, 50% and 70%) and compared with the originals (100%) with regards to the different readability measures.

The texts were summarized using COGSUM with a random index dimensionality, d, of 100, a focus window size, w, of 4 (2 left, 2 right) and $\rho = 4$, i.e. 2 positive 1:s and 2 negative 1:s, in line with (Chatterjee & Mohan, 2007).

The texts were also stemmed using the snowball algorithm, so that the same words with different endings get counted as the same. The PageRank damping factor was set to .85 (Brin & Page, 1998) and the number of iterations was set to 50.

The summaries were evaluated using 7 measures, see Table 4.1 for results on readability indices LIX, NR and OVIX, and Table 4.2 for additional parameters. The values of the measures of the summaries were also compared to the full length texts using a paired-samples T-test.

4.2 Results

Table 4.1 shows the mean values for the various readability measures used on the different texts. Roughly, low values on any readability measure means that the text is more readable. Table 4.2 also includes values on average word length (AWL).

The following significant differences were found:

DN For newspaper articles LIX got a lower score on all the summaries, ($p<.05$):

	Length	t(122)	Mean	SD
	30%	-8.092	43.60	10.14
LIX	50%	-9.147	45.21	8.70
	70%	-7.393	47.04	8.16
	100%		49.34	7.35

OVIX got a higher value for the 30% summary, ($p<.05$):

	Length	t(122)	Mean	SD
OVIX	30%	2.483	81.24	30.62
	100%		75.48	11.00

At 50% of the original text the words are also shorter on average (AWL) ($p<.05$):

	Length	t(122)	Mean	SD
AWL	50%	-3.4642	4.74	0.51
	100%		4.83	0.41

The sentences also became shorter (ASL) for all summarization lengths ($p<.05$):

	Length	t(122)	Mean	SD
	30%	-4.817	16.12	5.09
ASL	50%	-3.331	16.85	5.98
	70%	-5.115	17.04	4.27
	100%		18.00	3.91

FOF For popular science LIX was lower on all summarization lengths ($p<.05$):

	Length	t(30)	Mean	SD
	30%	-6.933	53.65	9.84
LIX	50%	-6.270	55.90	8.61
	70%	-5.327	57.57	8.36
	100%		59.92	7.81

At 50% and 70% OVIX got a lower score ($p<.05$):

	Length	t(30)	Mean	SD
OVIX	50%	-7.136	64.26	9.67
	70%	-6.017	66.26	8.89
	100%		69.24	7.94

At 30% and 50% we had lower average word length, ($p<.05$):

	Length	t(30)	Mean	SD
AWL	30%	-2.234	4.86	0.36
	50%	-2.465	4.89	0.27
	100%		4.94	0.23

We had a smaller proportion of extra long words for all summarization lengths ($p<.05$):

	Length	t(30)	Mean	SD
XLW	30%	-2.689	0.01	0.01
	50%	-2.458	0.01	0.01
	70%	-2.464	0.01	0.01
	100%		0.02	0.01

FOKASS Authority texts also displayed a lower LIX for all summarization lengths ($p<.05$):

	Length	t(34)	Mean	SD
LIX	30%	-8.497	46.28	12.76
	50%	-5.939	50.53	13.39
	70%	-4.642	52.92	13.09
	100%		55.46	13.00

OVIX was lower for 70% summarizations ($p<.05$):

	Length	t(34)	Mean	SD
OVIX	70%	-2.209	46.77	9.55
	100%		48.19	8.69

The sentences were longer at 50% and at 70% ($p < .05$):

	Length	t(34)	Mean	SD
ASL	50%	2.144	15.10	3.55
	70%	2.606	14.87	2.61
	100%		14.27	2.41

No significant differences could be observed in nominal ratio (NR) or proper nouns (PN) for any text genre or summarization length.

		LIX		**OVIX**		**NR**	
Length	TEXT	Mean	SD	Mean	SD	Mean	SD
0,30	DN	**43,60**	10,14	*81,24*	30,62	1,39	0,72
	FOF	**53,65**	9,84	65,29	16,46	1,56	0,33
	FOKASS	**46,28**	12,76	48,22	14,17	1,23	0,62
0,50	DN	**45,21**	8,70	74,41	16,45	1,37	0,61
	FOF	**55,90**	8,61	**64,26**	9,67	1,59	0,36
	FOKASS	**50,53**	13,39	47,96	14,61	1,29	0,71
0,70	DN	**47,04**	8,16	74,33	12,74	1,37	0,54
	FOF	**57,57**	8,36	**66,26**	8,89	1,56	0,28
	FOKASS	**52,92**	13,09	**46,77**	9,55	1,25	0,54
1,00	DN	49,34	7,35	75,48	11,00	1,35	0,44
	FOF	59,92	7,81	69,24	7,94	1,53	0,24
	FOKASS	55,46	13,00	48,19	8,69	1,25	0,63

Table 4.1: Results on LIX, OVIX and NR on three different texts at different summarization lengths. Significant differences with the values of the full texts are in bold.

A significantly lower LIX could be observed across the board of summaries, regardless of length and genre. This shows that the complexity of the text is reduced when the text is summarized by the summarizer. A lower LIX presents together with the amount of extra long words the vocabulary load required to read the text (Mühlenbock & Kokkinakis, 2009). For popular science, this seems the most prominent, as not only LIX but also the amount of extra long words decreased for all summarization lengths. Thus, for popular science, the vocabulary load decreased when articles were summarized.

OVIX also seems to be most effectively reduced for popular science texts when summarized, indicating that idea density is also reduced.

The average sentence length can be seen as a way of analyzing the structure of the sentence, without adopting syntactic parsing (Mühlenbock & Kokkinakis,

		AWL		ASL		XLW		PN	
Length	TEXT	Mean	SD	Mean	SD	Mean	SD	Mean	SD
0,30	DN	4,76	0,62	**16,12**	5,09	0,02	0,02	0,09	0,07
	FOF	**4,86**	0,36	17,46	4,40	**0,01**	0,01	0,03	0,03
	FOKASS	4,84	0,63	15,22	4,30	0,04	0,03	0,02	0,03
0,50	DN	**4,74**	0,51	**16,85**	5,98	0,02	0,02	0,09	0,07
	FOF	**4,89**	0,27	17,15	3,32	**0,01**	0,01	0,03	0,02
	FOKASS	4,89	0,58	*15,10*	3,55	0,04	0,03	0,02	0,03
0,70	DN	4,79	0,44	**17,04**	4,27	0,02	0,02	0,09	0,06
	FOF	4,90	0,25	17,01	3,27	**0,01**	0,01	0,03	0,02
	FOKASS	4,90	0,51	*14,87*	2,61	0,04	0,03	0,01	0,02
1,00	DN	4,83	0,41	18,00	3,91	0,02	0,01	0,09	0,06
	FOF	4,94	0,23	16,74	2,82	0,02	0,01	0,03	0,02
	FOKASS	4,89	0,47	14,27	2,41	0,04	0,05	0,01	0,02

Table 4.2: Results on Avarage Word Length (AWL), Avarage Sentence Length (ASL), amount of Extra Long Words (XLW) and the number of names (PN). Significant differences with the full text are in bold.

2009). This seems to be most prominent in newspaper articles. Newspaper texts had a high idea density from the start (by a high variation in words, OVIX) and a low LIX. They benefitted from summarization by getting a lower value on average sentence length, or sentence structure, for all summarization lengths.

Authority texts did not benefit as much from summarization, for all summarization lengths, as the other genres. OVIX was lower for 70% summaries but sentences got, for instance, longer for long summaries (50% and 70%).

No significant differences were found for any text on the amount of proper nouns (PN) and NR. A low NR might indiciate a stylistically simple text such as narrative children's books while a higher NR is more common in advanced texts and since the summarizer at this point does nothing to rewrite sentences, a change in NR is not to be expected.

Chapter 5

Discussion

In this thesis work on an automatic summarizer called COGSUM is presented. It can be concluded that automatic summarization may work as a first step in making a text easier to read. This is based on the results from some experiments regarding the readability of the resulting texts, acquired through the use of several automatic readability indices. The summaries are achieved by letting the sentences in the text recommend eachother so that sentences that share contexts that are important for the document is valued higher. By tuning different parameters it was hypothesized that the quality of the summaries would be different and therefore that it was possible to optimize the settings in order to get as high readability as possible. However, the parameters did not have a large effect on the quality of the summaries, other than that outside certain thresholds, the result were more or less random, meaning that an evaluation of the summaries should not be performed until some insight over how the parameters affects the ranking is gained. By plotting the different values of the sentences during the summarization process, it was possible to acquire some knowledge of what happens with the ranking with regards to different settings of the parameters, providing for more thorough investigation in future research.

5.1 Readability discussion

There seems to be a difference between different genres in how a summary is affecting the readability of the texts. For instance, with regards to LIX, it seems that the shorter the summary, the lower the score, regardless of text type. With ASL however, summarization on newspaper texts seem to keep sentences that are shorter overall, while summarization on the other genres tend to keep longer sentences. Thus for newspaper texts, the most important sentences are the short ones,

while the opposite is true for texts of the other genres. Popular science seems to benefit most by being summarized being the genre that gets a lower vocabulary load and a lower complexity of sentence structure, together with idea density.

However, these results are only indicative and should not be seen as a way of completely assess the readaility of the resulting summaries. One obvious flaw with extraction based summaries of the type used in these experiments is that they lack a way of providing anaphora resolution, meaning the summaries may very well contain sentences that refer to preceding sentences. This is currently not captured within the readability indices and a qualitative approach to evaluation would be favorable to assess how the lack of anaphora resoltion affects the subjective experience of readability.

5.2 Conclusions

By using an automatic extraction based summarizer, it is possible to make a text easier to read by shortening them and keep only the most important information. With regards to vocabulary load, the texts become easier to process for the reader. The sentences that are still present in newspaper texts after summarization are also of a simpler structure as they are shorther overall. This indicates that the important sentences in the text are also shorter. There is further a difference in how summarization affects the readability with regards to text type and text length; for instance, the length of the sentences are longer for authority texts and popular science when they are more important.

Different parameter settings of the summarizer was hypothesized to produce summaries of different qualities and thus an important aspect in creating more readable texts. This was also the case to a small extent; the value on the parameters does have a small impact on the quality of the summaries, especially dimensionality. The best dimensionality was spotted at 100 in the trials for texts around 1000 words. It is interesting since the same dimensionality generally is chosen for shorter texts of 200-300 words, indicating that a flat percentage with regards to dimension reduction may not be optimal.

5.3 Future work

For the future, some further investigation on how different text types are affected by summaries with regards to readability would be needed, in order to apply the appropriate tools to the appropriate texts. Summaries could however be seen as a first step of simplifying texts and achieve higher readability. Also, in depth evaluation of the quality of the summaries and how it affects the readability as well as how to

improve the quality of the summaries by tuning relevant parameters is an important task. By scaling up the visualization studies on several text types and lengths in future research, more conclusions on the matter would be possible. Further, some deeper qualitative research of the applicability of automated shallow readability measures when evaluating summaries would be desired. For instance, the dawning era of eye-tracking studies recently poses interesting capabilities of evaluation research and would provide interesting insights in how a reader interacts with a text, for instance, how important anaphora resolution is to readability.

References

Bernhardt, E. B. (1984). Toward an information processing perspective in foreign language reading. *The Modern Language Journal*, *68*(4), 322-331.

Björnsson, C. (1968). *Låsbarhet*. Stockholm: Liber.

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

Campbell, R. S. (January 2003). The secret life of pronouns: flexibility in writing style and physical health. *Psychological Science*, *14*, 60-65(6). Available from `http://www.ingentaconnect.com/content/bpl/psci/2003/00000014/00000001/art00010`

Carlsson, B. (2009). *Guldstandarder – dess skapande och utvärdering*. Unpublished master's thesis, Linköping University.

Chall, J. (1958). *Readability: An appraisal of research and application*. Columbus, OH: Ohio State University Press. Reprinted 1974. Epping, Essex, England: Bowker Publishing Company.

Chatterjee, N., & 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)* (p. 448-455).

Dalianis, H., Hassel, M., Wedekind, J., Haltrup, D., Smedt, K. de, & Lech, T. C. (2003). From SweSum to ScandSum: Automatic Text Summarization for the Scandinavian Languages. In H. Holmboe (Ed.), *Nordisk sprogteknologi 2002: Årbog for nordisk språkteknologisk forskningsprogram 2000-2004* (pp. 153–163). Museum Tusculanums Forlag. Available from `http://nlp.lacasahassel.net/publications/scandsum02.pdf`

DuBay, W. H. (2004). *Smart language: Readers, readability, and the grading of text*. Costa Mesa:Impact Information.

Duffy, T. M., & Kabance, P. (1982). Testing a readable writing approach to text revision. *Journal of Educational Psychology*, *74*(5), 733 - 748. Available from `http://www.sciencedirect.com/science/article/B6WYD-4NV6XF3-D/2/da9af4eccd0e69dbe160186c21ad99f4`

Eldén, L. (2007). *Matrix methods in data mining and pattern recognition*. Society for Industrial & Applied Mathematics (SIAM).

Försäkringskassan. (2011, January). *Försäkringskassans website.* Available from http://www.forsakringskassan.se (http://www.forsakringskassan.se)

Freebody, P., & Anderson, R. C. (1983). Effects of vocabulary difficulty, text cohesion, and schema availability on reading comprehension. *Reading Research Quarterly*, *18*(3), pp. 277-294. Available from http://www.jstor.org/stable/747389

Giannakopoulos, G., Karkaletsis, V., Vouros, G., & Stamatopoulos, P. (2008, October). Summarization system evaluation revisited: N-gram graphs. *ACM Trans. Speech Lang. Process.*, *5*, 5:1–5:39. Available from http://doi.acm.org/10.1145/1410358.1410359

Gorrell, G. (2006). *Generalized hebbian algorithm for dimensionality reduction in natural language processing*. Unpublished doctoral dissertation, Linköping University.

Gustavsson, P., & Jönsson, A. (2010). Text summarization using random indexing and pagerank. In *Proceedings of the third swedish language technology conference (sltc-2010), linköping, sweden.*

Hassel, M. (2007). *Resource lean and portable automatic text summarization*. Unpublished doctoral dissertation, ISRN-KTH/CSC/A–07/09-SE, KTH, Sweden.

Hassel, M. (2011). *Random indexing java-toolkit.* Available from http://www.nada.kth.se/~xmartin/

Jönsson, A., Axelsson, M., Bergenholm, E., Carlsson, B., Dahlbom, G., Gustavsson, P., et al. (2008). Skim reading of audio information. In *Proceedings of the the second swedish language technology conference (sltc-08), stockholm, sweden.*

Kanerva, P. (1988). *Sparse distributed memory*. Cambridge MA: The MIT Press.

Karlgren, J., Holst, A., & Sahlgren, M. (2008). Filaments of meaning in word space. In C. Macdonald, I. Ounis, V. Plachouras, I. Ruthven, & R. White (Eds.), *Advances in information retrieval* (Vol. 4956, p. 531-538). Springer Berlin / Heidelberg.

Karlgren, J., & Sahlgren, M. (2001). From words to understanding. In Y. Uesaka, P.Kanerva, & H. Asoh (Eds.), *Foundations of real-world intelligence* (p. 294-308). Stanford: CSLI Publications.

Lundberg, I., & Reichenberg, M. (2009). *Vad är lättläst?* Socialpedagogiska skolmyndigheten.

Mihalcea, R., & Tarau, P. (2004, July). TextRank: Bringing order into texts. In *Proceedings of EMNLP-04and the 2004 conference on empirical methods in*

natural language processing.

Mühlenbock, K., & Kokkinakis, S. J. (2009). Lix 68 revisited – an extended readability measure. In *Proceedings of corpus linguistics.*

Papadimitriou, C. H., Raghavan, P., Tamaki, H., & Vempala, S. (2000). Latent semantic indexing: A probabilistic analysis. *Journal of Computer and System Sciences*, *61*(2), 217 - 235. Available from http://www.sciencedirect.com/science/article/B6WJ0-45FC93J-W/2/1a6dfbe012f6fe2fcf927db62e2da5e2

Rybing, J., Smith, C., & Silvervarg, A. (2010). Towards a rule based system for automatic simplification of texts. In *Proceedings of the third swedish language technology conference (sltc-2010), linköping, sweden.*

Sahlgren, M. (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*.

Sahlgren, M. (2006). *The word-space model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces.* Unpublished doctoral dissertation, Stockholm University, Department of Linguistics.

Språkbanken. (2011, January). *Concordances of språkbanken.* Available from http://spraakbanken.gu.se/konk/ (http://spraakbanken.gu.se/konk/)