Using a Keyness Metric for Single and Multi Document Summarisation

Mahmoud El-Haj
School of Computing and Communications
Lancaster University
United Kingdom
m.el-haj@lancaster.ac.uk

Paul Rayson
School of Computing and Communications
Lancaster University
United Kingdom
p.rayson@lancaster.ac.uk

Abstract

In this paper we show the results of our participation in the MultiLing 2013 summarisation tasks. We participated with single-document and multi-document corpus-based summarisers for both Arabic and English languages. The summarisers used word frequency lists and log likelihood calculations to generate single and multi document summaries. The single and multi summaries generated by our systems were evaluated by Arabic and English native speaker participants and by different automatic evaluation metrics, ROUGE, AutoSummENG, MeMoG and NPowER. We compare our results to other systems that participated in the same tracks on both Arabic and English languages. Our single-document summarisers performed particularly well in the automatic evaluation with our English single-document summariser performing better on average than the results of the other participants. Our Arabic multi-document summariser performed well in the human evaluation ranking second.

1 Introduction

Systems that can automatically summarise documents are becoming ever more desirable with the increasing volume of information available on the Web. Automatic text summarisation is the process of producing a shortened version of a text by the use of computers. For example, reducing a text document or a group of related documents into a shorter version of sentences or paragraphs using automated tools and techniques.

The summary should convey the key contributions of the text. In other words, only key sentences should appear in the summary and the process of defining those sentences is highly dependent on the summarisation method used. In automatic summarisation there are two main approaches that are broadly used, extractive and abstractive. The first method, the extractive summarisation, extracts, up to a certain limit, the key sentences or paragraphs from the text and orders them in a way that will produce a coherent summary. The extracted units differ from one summariser to another. Most summarisers use sentences rather than larger units such as paragraphs. Extractive summarisation methods are the focus method on automatic text summarisation. The other method, abstractive summarisation, involves more language dependent tools and Natural Language Generation (NLG) technology. In our work we used extractive single and multi-document Arabic and English summarisers.

A successful summarisation approach needs a good guide to find the most important sentences that are relevant to a certain criterion. Therefore, the proposed methods should work on extracting the most important sentences from a set of related articles.

In this paper we present the results of our participation to the MultiLing 2013 summarisation tasks. MultiLing 2013 was built upon the Text Analysis Conference (TAC) MultiLing Pilot task of 2011 (Giannakopoulos et al., 2011). MultiLing 2013 this year asked for participants to run their summarisers on different languages having a corpus and gold standard summaries in the same seven languages (Arabic, Czech, English, French, Greek, Hebrew or Hindi) of TAC 2011 with a 50% increase to the corpora size. It also introduced three new languages (Chinese, Romanian and Spanish). MultiLing 2013 this year introduced a new single-document summarisation pilot for 40 languages including the above mentioned languages (in our case Arabic and English).

In this paper we introduce the results of our

Proceedings of the MultiLing 2013 Workshop on Multilingual Multi-document Summarization, pages 64–71,
Sofia, Bulgaria, August 9 2013.

single-document and multi-document summarisers at the MultiLing 2013 summarisation tasks. We used a language independent corpus-based word frequency technique and the log-likelihood statistic to extract sentences with the maximum sum of log likelihood. The output summary is expected to be no more than 250 words.

2 Related Work

2.1 Automatic Summarisation

Work on automatic summarisation dates back more than 50 years, with a focus on the English language (Luhn, 1958). The work on Arabic automatic summarisation is more recent and still not on par with the research on English and other European languages. Early work on Arabic summarisation started less than 10 years ago (Conroy et al., 2006; Douzidia and Lapalme, 2004).

Over time, there have been various approaches to automatic text summarisation. These approaches include single-document and multi-document summarisation. Both single-document and multi-document summarisation use the summarisation methods mentioned earlier, i.e. extractive or abstractive. Summarising a text could be dependent on input information such as a user query or it could be generic where no user query is used.

The approach of single-document summarisation relies on the idea of producing a summary for a single document. The main factor in single-document summarisation is to identify the most important (informative) parts of a document. Early work on single-document summarisation was the work by Luhn (1958). In his work he looked for sentences containing keywords that are most frequent in a text. The sentences with highly weighted keywords were selected. The work by Luhn highlighted the need for features that reflect the importance of a certain sentence in a text. Baxendale (1958) showed the importance of sentence-position in a text, which is understood to be one of the earliest extracted features in automatic text summarisation. They took a sample of 200 paragraphs and found that in 80% of the paragraphs the most important sentence was the first one.

Multi-document summarisation produces a single summary of a set of documents. The documents are assumed to be about the same genre and topic. The analysis in this area is performed typically at either the sentence or document level.

2.2 Corpus-based and Word Frequency in Summarisation

Corpus-based techniques are mainly used to compare corpora for linguistic analysis (Rayson and Garside, 2000; Rayson et al., 2004). There are two main types of corpora comparisons, 1) comparing a sample corpus with a larger standard corpus (Scott, 2000). 2) comparing two corpora of equal size (Granger, 1998). In our work we adopted the first approach, where we used a much larger reference corpus. The first word list is the frequency list of all the words in the document (or group of documents) to be summarised which is compared to the word frequency list of a much larger standard corpus. We do that for both Arabic and English texts. Word frequency has been proven as an important feature when determining a sentence's importance (Li et al., 2006). Nenkova and Vanderwende (2005) studies the impact of frequency on summarisation. In their work they investigated the association between words that appear frequently in a document (group of related documents), and the likelihood that they will be selected by a human summariser to be included in a summary. Taking the top performing summarisers at the DUC 2003[1] they computed how many of the top frequency words from the input documents appeared in the system summaries. They found the following: 1) Words with high frequency in the input documents are very likely to appear in the human summaries. 2) The automatic summarisers include less of these high frequency words. These two findings by Nenkova and Vanderwende (2005) tell us two important facts. Firstly, it confirms that word frequency is an important factor that impacts humans' decisions on which content to include in the summary. Secondly, the overlap between human and system summaries can be improved by including more of the high frequency words in the generated system summaries. Based on Nenkova's study we expand the work on word frequency by comparing word frequency lists of different corpora in a way to select sentences with the maximum sum of log likelihood ratio. The log-likelihood calculation favours words whose frequencies are unexpectedly high in a document.

2.3 Statistical Summarisation

The use of statistical approaches (e.g. log-likelihood) in text summarisation is a common

[1] http://duc.nist.gov/duc2003/tasks.html

technique, especially when building a language independent text summariser.

Morita et al. (2011) introduced what they called "query-snowball", a method for query-oriented extractive multi-document summarisation. They worked on closing the gap between the query and the relevant sentences. They formulated the summarisation problem based on word pairs as a maximum cover problem with Knapsack Constraints (MCKP), which is an optimisation problem that maximises the total score of words covered by a summary within a certain length limit.

Knight and Marcu (2000) used the Expectation Maximisation (EM) algorithm to compress sentences for an abstractive text summarisation system. EM is an iterative method for finding Maximum Likelihood (ML) or Maximum A Posteriori (MAP) estimates of parameters in statistical models. In their summariser, EM was used in the sentences compression process to shorten many sentences into one by compressing a syntactic parse tree of a sentence in order to produce a shorter but maximally grammatical version. Similarly, Madnani et al. (2007) performed multi-document summarisation by generating compressed versions of source sentences as summary candidates and used weighted features of these candidates to construct summaries.

Hennig (2009) introduced a query-based latent Semantic Analysis (LSA) automatic text summariser. It finds statistical semantic relationships between the extracted sentences rather than word by word matching relations (Hofmann, 1999). The summariser selects sentences with the highest likelihood score.

In our work we used log-likelihood to select sentences with the maximum sum of log likelihood scores, unlike the traditional method of measuring cosine similarity overlap between articles or sentences to indicate importance (Luhn, 1958; Barzilay et al., 2001; Radev et al., 2004). The main advantage of our approach is that the automatic summariser does not need to compare sentences in a document with an initial one (e.g. first sentence or a query). Our approach works by calculating the keyness (or log-likelihood) score for each token (word) in a sentence, then picks, to a limit of 250 words, the sentences with the highest sum of the tokens' log-likelihood scores.

To the best of our knowledge the use of corpus-based frequency list to calculate the log-likelihood score for text summarisation has not been reported for the Arabic language.

3 Dataset and Evaluation Metrics

3.1 Test Collection

The test collection for the MultiLing 2013 is available in the previously mentioned languages.[2] The dataset is based on WikiNews texts.[3] The source documents contain no meta-data or tags and are represented as UTF8 plain text les. The multi-document dataset of each language contains (100-150) articles divided into 10 or 15 reference sets, each contains 10 related articles discussing the same topic. The original language of the dataset is English. The organisers of the tasks were responsible for translating the corpus into different languages by having native speaker participants for each of the 10 languages. In addition to the news articles the dataset also provides human-generated multi-document gold standard summaries. The single-document dataset contains single documents for 40 language (30 documents each) discussing various topics and collected from Wikipedia.[4]

3.2 Evaluation

Evaluating the quality and consistency of a generated summary has proven to be a difficult problem (Fiszman et al., 2009). This is mainly because there is no obvious ideal, objective summary. Two classes of metrics have been developed: form metrics and content metrics. Form metrics focus on grammaticality, overall text coherence, and organisation. They are usually measured on a point scale (Brandow et al., 1995). Content metrics are more difficult to measure. Typically, system output is compared sentence by sentence or unit by unit to one or more human-generated ideal summaries. As with information retrieval, the percentage of information presented in the system's summary (precision) and the percentage of important information omitted from the summary (recall) can be assessed. There are various models for system evaluation that may help in solving this problem. This include automatic evaluations (e.g. ROUGE and AutoSummENG), and human-performed evaluations. For the MultiLing 2013 task, the summaries generated by the participants

[2] http://multiling.iit.demokritos.gr/file/all

[3] http://www.wikinews.org/

[4] http://www.wikipedia.org/

were evaluated automatically based on human-generated model summaries provided by fluent speakers of each corresponding language (native speakers in the general case). The models used were, ROUGE variations (ROUGE1, ROUGE2, ROUGE-SU4) (Lin, 2004), the MeMoG variation (Giannakopoulos and Karkaletsis, 2011) of AutoSummENG (Giannakopoulos et al., 2008) and NPowER (Giannakopoulos and Karkaletsis, 2013). ROUGE was not used to evaluate the single-document summaries.

The summaries were also evaluated manually by human participants. For the manual evaluation the human evaluators were provided with the following guidelines: Each summary is to be assigned an integer grade from 1 to 5, related to the overall responsiveness of the summary. We consider a text to be worth a 5, if it appears to cover all the important aspects of the corresponding document set using fluent, readable language. A text should be assigned a 1, if it is either unreadable, nonsensical, or contains only trivial information from the document set. We consider the content and the quality of the language to be equally important in the grading.

Note, the human evaluation results for the English language are not included in this paper as by the time of writing the results were not yet published. We only report the human evaluation results of the Arabic multi-document summaries.

4 Corpus-based Summarisation

Our summarisation approach is a corpus-based where we use word frequency lists to compare corpora and calculate the log likelihood score for each word in the list. The compared corpora include standard Arabic and English corpora in addition to the Arabic and English summarisation datasets provided by MultiLing 2013 for the single and multi-document summarisation tasks. The subsections below describe the creation of the word lists and the standard corpora we used for the comparison process.

4.1 Word Frequencies

We used a simple methodology to generate the word frequency lists for the Arabic and English summarisation datasets provided by MultiLing 2013. The datasets used in our experiments were single-document and multi-document documents in English and Arabic. For the multi-document dataset we counted the word frequency for all the documents in a reference set (group of related articles), each set contains on average 10 related articles. The single-document dataset was straightforward, we calculated word frequencies for all the words in each document. Figure 1 shows a sample of random words and their frequencies for both Arabic and English languages. The sample was selected from the MultiLing dataset word frequency lists. As shown in the figure we did not eliminate the stop-words, we treat them as normal words.

word	frequency
المتحدة	33
ان	32
ايران	31
تم	29
قبل	25
المملكة	25
كان	24
البريطانية	24
ومئاة	19
البريطانيين	19

(a) Arabic Sample

word	frequency
government	21
personnel	21
release	20
Royal	17
would	17
this	16
into	16
United	15
Iraq	14
UK	14

(b) English Sample

Figure 1: Arabic and English Word Frequency List Sample

4.2 Standard Corpora

In our work we compared the word frequency list of the summarisation dataset against the larger Arabic and English standard corpora. For each of the standard corpora we had a list of word frequencies (up to $5,000$ words) for both Arabic and English using the frequency dictionary of Arabic (Buckwalter and Parkinson, 2011) and the Corpus of Contemporary American English (COCA) top 5,000 words (Davies, 2010).

The frequency dictionary of Arabic provides a list of the 5,000 most frequently used words in Modern Standard Arabic (MSA) in addition to several of the most widely spoken Arabic dialects. The list was created based on a 30-million-word corpus of Arabic including written and spoken material from all around the Arab world. The Arabic summarisation dataset provided by MultiLing 2013 was also written using MSA. The corpus of contemporary American English COCA is a freely searchable 450-million-word corpus containing text in American English of different number of genres. To be consistent with the Arabic

word frequency list, we used the top 5000 words from the 450 million word COCA corpus.

5 Summarisation Methodology

In our experiments we used generic single-document and multi-document extractive summarisers that have been implemented for both Arabic and English (using identical processing pipelines for both languages). Summaries were created by selecting sentences from a single document or set of related documents. The following subsections show the methods used in our experiments, the actual summarisation process and the experimental setup.

5.1 Calculating Log-Likelihood

We begin the summarisation process by calculating the log likelihood score for each word in the word frequency lists (see Section 4.1) using the same methodology described in (Rayson and Garside, 2000). This was performed by constructing a contingency table as in Table 1.

	Corpus One	Corpus Two	Total
Frequency of Word	a	b	a+b
Frequency of other words	c-a	d-b	c+d-a-b
Total	c	d	c+d

Table 1: Contingency Table

The values c and d correspond to the number of words in corpus one and corpus two respectively. Where a and b are the observed values (O). For each corpus we calculated the expected value E using the following formula:

$$E_i = \frac{N_i \sum_i O_i}{\sum_i N_i}$$

N_i is the total frequency in corpus i (i in our case takes the values 1 (c) and 2 (d) for the MultiLing Arabic Summaries dataset and the frequency dictionary of Arabic (or MultiLing English Summaries dataset and COCA corpus) respectively.

The log-likelihood can be calculated as follows:

$$LL = 2 * ((a * \ln(\frac{a}{E1})) + (b * \ln(\frac{b}{E2})))$$

5.2 Summarisation Process

We used the same processing pipeline for both the single-document and multi-document summarisers. For each word in the MultiLing summarisation dataset (Arabic and English) we calculated the log likelihood scores using the calculations described in Section 5.1. We summed up the log likelihood scores for each sentence in the dataset and we picked the sentences (up to 250 word limit) with the highest sum of log likelihood scores. The main difference between the single-document and multi-document summarisers is that we treat the set of related documents in the multiling dataset as one document.

6 Single-Document Summarisation Task

MultiLing 2013 this year introduced a new single-document summarisation pilot for 40 languages including (Arabic, Czech, English, French, Greek, Hebrew, Hindi, Spanish, Chinese, Romanian ...etc). In our case we participated in two languages only, English and Arabic.

The pilot aim was to measure the ability of automated systems to apply single document summarisation, in the context of Wikipedia texts. Given a single encyclopedic entry, with several sections/subsections, describing a specific subject, the pilot guidelines asked the participating systems to provide a summary covering the main points of the entry (similarly to the lead section of a Wikipedia page). The MultiLing 2013 single-document summaries dataset consisted of (non-parallel) documents in the above mentioned languages.

For the English language, there were 7 participants (peers) including a baseline system ($ID5$). The Arabic language had 6 participants including the same baseline system.

7 Multi-Document Summarisation Task

The Multi-document summarisation task required the participants to generate a single, fluent, representative summary from a set of documents describing an event sequence. The language of the document set was within a given range of languages and all documents in a set shared the same language. The task guidelines required the output summary to be of the same language as its source documents. The output summary should be 250 words at most.

The set of documents were available in 10 languages (Arabic, Czech, English, French, Greek, Hebrew, Hindi, Spanish, Chinese and Romanian). In our case we participated using the Arabic and English set of documents only.

For the English language, there were 10 participants (peers) including a baseline ($ID6$) and a topline ($ID61$) systems. The Arabic language had 10 participants as well, including the same baseline and topline systems.

The baseline summariser sorted sentences based on their cosine similarity to the centroid of a cluster. Then starts adding sentences to the summary, until it either reaches 250 words, or it hits the end of the document. In the second case, it continues with the next document in the sorted list.

The topline summariser used information from the model summaries (i.e. cheats). First, it split all source documents into sentences. Then it used a genetic algorithm to generate summaries that have a vector with maximal cosine similarity to the centroid vector of the model summary texts.

8 Results and Discussion

Our single-document summarisers, both English and Arabic, performed particularly well in the automatic evaluation. Ranking first and second respectively.

Tables 2 and 3 illustrate the AutoSummEng (AutoSumm), MeMoG and NPowER results and the ranking of our English and Arabic single-document summarisers (System **ID2**).

System	AutoSumm	MeMoG	NPowER
ID2	0.136	0.136	1.685
ID41	0.129	0.129	1.661
ID42	0.127	0.127	1.656
ID3	0.127	0.127	1.654
ID1	0.124	0.124	1.647
ID4	0.123	0.123	1.641
ID5	0.040	0.040	1.367

Table 2: English Automatic Evaluation Scores (single-document)

The evaluation scores of our single-document summarisers confirm with (Li et al., 2006) and (Nenkova and Vanderwende, 2005) findings, were they found that word frequency is an important feature when determining sentences importance and that words with high frequency in the input documents are very likely to appear in the human summaries, which explains the high correlation between our single-document and the human (model) summaries as illustrated in the evaluation scores (Tables 2 and 3). The single-document summaries were evaluated automatically only.

System	AutoSumm	MeMoG	NPowER
ID3	0.092	0.092	1.538
ID2	0.087	0.087	1.524
ID41	0.055	0.055	1.418
ID42	0.055	0.055	1.416
ID4	0.053	0.053	1.411
ID5	0.025	0.025	1.317

Table 3: Arabic Automatic Evaluation Scores (single-document)

System	Score
ID6	3.711
ID3	3.578
ID2	3.578
ID4	3.489
ID1	3.467
ID11	3.333
ID21	3.111
ID51	2.778
ID5	2.711
ID61	2.489

Table 4: Arabic Manual Evaluation Scores (multi-document)

Our Arabic multi-document summariser performed well in the human evaluation ranking second jointly with System ID2. Table 4 shows the average scores of the human evaluation process, our system is referred to as **ID3**. On the other hand, we did not perform well in the automatic evaluation of the multi-document summarisation task for both English and Arabic. Our systems did not perform better than the baseline. The automatic evaluation results placed our Arabic and English summariser further down in the ranked lists of systems compared to the human assessment. This is an area for future work as this seems to suggest that the automatic evaluation metrics are not necessarily in line with human judgements.

The low automatic evaluation scores are due to two main reasons. First, we treated the set of related documents (multi-documents) as a single big document (See Section 5.2), this penalised

our summaries as selecting the sentences with the maximum sum of log likelihood score lead to many important sentences being overlooked. This can be solved by running the summariser on each document to suggest candidate sentences and then selecting the top sentence(s) of each document to generate the final summary. Second, we did not work on eliminating redundancies. Finally, the log-likelihood score might be improved by the inclusion of a dispersion score or weighting to examine the evenness of the spread of each word across all the documents.

9 Conclusion

In this paper we presented the results of our participation in the MultiLing 2013 summarisation task. We submitted results for single-document and multi-document summarisation in two languages, English and Arabic. We applied a corpus-based summariser that used corpus-based word frequency lists. We used a list of the 5,000 most frequently used words in Modern Standard Arabic (MSA) and English. Using the frequency dictionary of Arabic and the corpus of contemporary American English (COCA).

Based on the automatic evaluation scores, we found that our approach appears to work very well for Arabic and English single-document summarisation. According to the human evaluation scores the approach could potentially work for Arabic multi-document summarisation as well. We believe that the approach could still work well for multi-document summarisation following the suggested solutions in Section 8.

References

R. Barzilay, N. Elhadad, and K. McKeown. 2001. Sentence Ordering in Multidocument Summarization. In *Proceedings of the First International Conference on Human Language Technology Research*, HLT'01, pages 1–7, Stroudsburg, PA, USA. Association for Computational Linguistics.

P. Baxendale. 1958. Machine-made index for technical literature: an experiment. *IBM Journal of Research and Development*, 2(4):354–361.

R. Brandow, K. Mitze, and Lisa F. Rau. 1995. Automatic Condensation of Electronic Publications by Sentence Selection. *Inf. Process. Manage.*, 31(5):675–685.

T. Buckwalter and D. Parkinson. 2011. *A Frequency Dictionary of Arabic: Core Vocabulary for Learners*. Routledge, London, United Kingdom.

J. Conroy, J. Schlesinger, D. O'Leary, and J. Goldstein. 2006. Back to Basics: CLASSY 2006. In *Proceedings of the 6th Document Understanding Conferences*. DUC.

M. Davies. 2010. The Corpus of Contemporary American English as the First Reliable Monitor Corpus of English. *Literary and Linguistic Computing*, 25:447–464.

F. Douzidia and G. Lapalme. 2004. Lakhas, an Arabic Summarising System. In *Proceedings of the 4th Document Understanding Conferences* , pages 128–135. DUC.

M. Fiszman, D. Demner-Fushman, H. Kilicoglu, and T. Rindflesch. 2009. Automatic Summarization of MEDLINE Citations for Evidence-based Medical Treatment: A Topic-oriented Evaluation. *Jouranl of Biomedical Informatics*, 42(5):801–813.

G. Giannakopoulos and V. Karkaletsis. 2011. AutoSummENG and MeMoG in Evaluating Guided Summaries. In *The Proceedings of the Text Analysis Conference*, MD, USA. TAC.

G. Giannakopoulos and V. Karkaletsis. 2013. Summary evaluation: Together we stand npower-ed. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volume 7817 of *Lecture Notes in Computer Science*, pages 436–450. Springer Berlin Heidelberg.

G. Giannakopoulos, V. Karkaletsis, G. Vouros, and P. Stamatopoulos. 2008. Summarization System Evaluation Revisited: N–Gram Graphs. *ACM Transactions on Speech and Language Processing (TSLP)*, 5(3):1–39.

G. Giannakopoulos, M. El-Haj, B. Favre, M. Litvak, J. Steinberger, and V. Varma. 2011. TAC 2011 MultiLing Pilot Overview. In *Text Analysis Conference (TAC) 2011, MultiLing Summarisation Pilot*, Maryland, USA. TAC.

S. Granger. 1998. The computer learner corpus: A versatile new source of data for SLA research. pages 3–18.

L. Hennig. 2009. Topic-based multi-document summarization with probabilistic latent semantic analysis. In *Proceedings of the International Conference RANLP-2009*, pages 144–149, Borovets, Bulgaria, September. Association for Computational Linguistics.

T. Hofmann. 1999. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '99, pages 50–57, New York, NY, USA. ACM.

K. Knight and D. Marcu. 2000. Statistics-Based Summarization – Step One: Sentence Compression. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference*

on Innovative Applications of Artificial Intelligence, pages 703–710, Menlo Park, CA. AAAI Press.

W. Li, B. Li, and M. Wu. 2006. Query Focus Guided Sentence Selection Strategy.

C. Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004)*, pages 25–26. WAS 2004).

H. Luhn. 1958. The Automatic Creation of Literature Abstracts. *IBM Journal of Research and Development*, 2(2):159–165.

N. Madnani, D. Zajic, B. Dorr, N. Ayan, and J. Lin. 2007. Multiple Alternative Sentence Compressions for Automatic Text Summarization. In *Proceedings of the 7th Document Understanding Conference at NLT/NAACL*, page 26. DUC.

H. Morita, T. Sakai, and M. Okumura. 2011. Query Snowball: A Co-occurrence-based Approach to Multi-document Summarization for Question Answering. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2*, HLT'11, pages 223–229, Stroudsburg, PA, USA. Association for Computational Linguistics.

A. Nenkova and L. Vanderwende. 2005. The impact of frequency on summarization. *Microsoft Research, Redmond, Washington, Tech. Rep. MSR-TR-2005-101*.

D. Radev, H. Jing, M. Sty, and D. Tam. 2004. Centroid-based Summarization of Multiple Documents. *Information Processing and Management*, 40:919–938.

P. Rayson and R. Garside. 2000. Comparing corpora using frequency profiling. In *Proceedings of the workshop on Comparing corpora - Volume 9*, WCC '00, pages 1–6, Stroudsburg, PA, USA.

P. Rayson, D. Berridge, and B. Francis. 2004. Extending the cochran rule for the comparison of word frequencies between corpora. In *Proceedings of the 7th International Conference on Statistical analysis of textual data (JADT 2004*, pages 926–936.

M. Scott. 2000. Focusing on the text and its key words. In *Burnard, L. and McEnery, T. (eds.) Rethinking language pedagogy from a corpus perspective: papers from the third international conference on teaching and language corpora*, pages 103–121.