Towards Evaluating the Impact of Anaphora Resolution on Text Summarisation from a Human Perspective

Mostafa Bayomi[1](✉), Killian Levacher[1], M. Rami Ghorab[3],
Peter Lavin[1], Alexander O'Connor[2], and Séamus Lawless[1]

[1] ADAPT Centre, Knowledge and Data Engineering Group, School of Computer Science and Statistics, Trinity College Dublin, Dublin, Ireland
{bayomim, killian.levacher, peter.lavin, seamus.lawless}@scss.tcd.ie

[2] ADAPT Centre, School of Computing, Dublin City University, Dublin, Ireland
alexander.oconnor@dcu.ie

[3] IBM Analytics, IBM Technology Campus, Dublin, Ireland
rami.ghorab@ie.ibm.com

Abstract. Automatic Text Summarisation (TS) is the process of abstracting key content from information sources. Previous research attempted to combine diverse NLP techniques to improve the quality of the produced summaries. The study reported in this paper seeks to establish whether Anaphora Resolution (AR) can improve the quality of generated summaries, and to assess whether AR has the same impact on text from different subject domains. Summarisation evaluation is critical to the development of automatic summarisation systems. Previous studies have evaluated their summaries using automatic techniques. However, automatic techniques lack the ability to evaluate certain factors which are better quantified by human beings. In this paper the summaries are evaluated via human judgment, where the following factors are taken into consideration: informativeness, readability and understandability, conciseness, and the overall quality of the summary. Overall, the results of this study depict a pattern of slight but not significant increases in the quality of summaries produced using AR. At a subject domain level, however, the results demonstrate that the contribution of AR towards TS is domain dependent and for some domains it has a statistically significant impact on TS.

Keywords: Text summarisation · Anaphora resolution · TextRank

1 Introduction

Natural Language Processing (NLP) has different tasks [1, 2]. One of these tasks is Automatic Text Summarisation (TS) that has been the subject of a lot of interest in the NLP community in recent years [2]. The goal of automatic summarisation is to process

M.R. Ghorab—Postdoctoral Researcher at Trinity College Dublin at the time of conducting this research.

E. Métais et al. (Eds.): NLDB 2016, LNCS 9612, pp. 187–199, 2016.
DOI: 10.1007/978-3-319-41754-7_16

the source text to produce a shorter version of the information contained in it then present this version in a way that suits the needs of a particular user or application. Text summaries attempt to provide concise overviews of content, and can allow a reader to make a quick and informed decision regarding whether a document contains the information they seek, and thus whether it would be worth the time and effort required to read the entire document. The rapid growth of the Web has resulted in a massive increase in the amount of information available online, which in turn has increased the importance of text summarisation.

Various techniques have been proposed in the literature for the automatic summarisation of text, some of which are supervised, while others are unsupervised. Supervised techniques involve the need for an existing dataset of example summaries [3]. In contrast, unsupervised techniques do not rely upon any external knowledge sources, models or on linguistic processing and interpretation to summarise text [4].

There are two primary approaches to automatic summarisation. Extractive methods work by selecting a subset of existing words, phrases, or sentences from the original text to form the summary. In contrast, abstractive methods build an internal semantic representation and then use natural language generation techniques to create a summary that is closer to what a human might generate. Such a summary may contain words that are not included in the original text. These approaches are outlined in more detail in Sect. 2, below.

Throughout the evolution of automatic text summarisation, attempts have continuously been made to improve the quality of the summaries produced. As a result, diverse NLP techniques have been combined together in an attempt to deliver advancements in the process, with mixed results. While some combinations of techniques demonstrated positive impact [5], some did not [6].

One approach that has been shown to deliver a positive impact on summarisation is Anaphora Resolution (AR). A number of studies have combined AR with the summarisation algorithm, and their results have shown a positive impact on summaries [6, 7].

Various systems were developed to carry out automatic evaluation, of the produced summaries, taking into consideration some measures and aspects in the evaluation process, such as comparing the summary to other (ideal) summaries created by humans [9]. An example of these measurements involves counting the number of overlapping units such as n-gram, word sequences, and word-pairs between the computer-generated summaries and the ideal summaries created by humans. The problems with these evaluations are: (1) the necessity for existing human summaries and the fact that (2) automatic systems lack evaluation measurements, such as conciseness and informativeness. Metrics such as these are often missing from evaluation systems, as they are difficult to quantify automatically and need to be assessed by human beings. Our research attempts to take these factors into consideration as part of the evaluation.

The contribution of this paper involves investigating the impact of AR on text summarisation, both in general, and with respect to different domains, by using human evaluation to judge the generated summaries. The evaluation includes factors that cannot be quantified by automatic systems, and therefore require a human to judge. These factors are: informativeness, readability & understandability, conciseness and the overall quality of the summary.

The summarisation system proposed in this paper builds upon TextRank [4] for the generation of summaries. In the evaluation carried out for this study, a comparison is conducted between two summarisation approaches. In the first approach, summarisation is performed directly on the original text. In the second approach, however, the anaphora of the original text is resolved before summarisation is performed.

2 Related Work

Many advances have been achieved in the field of text summarisation, which generate summaries of a variety of documents in different repositories, such as scientific repositories and meeting recordings [10]. The two main approaches to text summarisation consist of abstraction [11] and extraction [12] methods. Abstraction involves "understanding" the meaning of a text in order to build an internal semantic representation of it and subsequently use natural language generation techniques to create a summary that is close to what a human would generate. The challenge faced by abstraction techniques is that they rely upon the requirements for semantics and training data to automatically interpret and "understand" the meaning of the content in order to summarise the original text. As a result of this, summarisation techniques based upon abstraction have so far met with limited success [13]. On the other hand, extraction approaches require the identification of the most important sentences (or passages), in the text in order to extract the summary. Thus, the extraction method is deemed more feasible [4] and is selected as the approach to be used in this paper.

Various methods have been discussed in the literature to improve extractive summarisation systems. Early systems depended upon simple approaches, such as: *Sentence Location*, where the importance of the sentence is determined by: (1) its location in the text (being at the beginning or end of the text); (2) the emphasis of the text (e.g. being a heading); and (3) the formatting of the text (e.g. a sentence highlighted in bold would be considered more important than other sentences) [14]. The *Cue Phrase* approach focuses on words that are used to determine the appropriate sentences for the summary such as "this paper", "propose", and "concluding" [15]. The *most frequent words* approach extracts a summary by determining the most important concepts by exploring the term distribution of a document [16]. Another approach that extracts summaries by measuring the length of the sentences is the *Sentence Length* approach, which automatically ignores sentences that have a length that falls under a predefined threshold [15].

Robust graph-based methods for NLP tasks have also been gaining a lot of interest. Arguably, these methods can be singled out as key elements of the paradigm-shift triggered in the field of NLP. Graph-based ranking algorithms are essentially a way of deciding the importance of a vertex within a graph, based on global information recursively drawn from the entire graph. The basic idea implemented by a graph-based ranking model is that of "voting" or "recommendation". When one vertex links to another, it is basically casting a vote for that vertex. The higher the number of votes for a vertex, the higher the importance of that vertex. Moreover, the importance of the vote is determined by the importance of the vertex casting the vote. One of the most significant algorithms in this spectrum is the TextRank algorithm [4], which is based on the PageRank algorithm [17].

TextRank is an unsupervised graph-based ranking model that summarises text by extracting the most important sentences of the text after ranking them according to the number of overlapping words between them. TextRank can be applied to tasks ranging from automated extraction of keyphrases, to extractive summarisation and word-sense disambiguation, and it works as follows:

Formally, let *G = (V, E)* be a directed graph with the set of vertices *V* and set of edges *E*, where *E* is a subset of $V \times V$. In this model, the graphs are built from natural language texts where *V* represents a sentence from the text, and *E* is the edge that represents the connection between two sentences. Each edge acts as a weight that represents the similarity between two sentences. For a given vertex $V_{i.}$, let *In* (V_i) be the set of vertices that point to it (predecessors), and let *Out* (V_i). be the set of vertices that vertex V_i points to (successors). The score of a vertex V_i is defined as follows:

$$S(V_i) = (1 - d) + d * \sum\nolimits_{j \in In(V_i)} \frac{1}{Out(V_i)} S(V_i) \tag{1}$$

where *d* is a damping factor that can be set between 0 and 1, which has the role of integrating into the model the probability of jumping from a given vertex to another random vertex in the graph. The factor *d* is usually set to 0.85.

TextRank has been demonstrated to perform well as an algorithm for extracting summaries [4], however, combining it with other NLP approaches may enhance the quality of these summaries. These additional approaches can be employed in the text pre-processing step, or furthermore in the ranking process. In this research we combine TextRank with other approaches such as Sentence Location and Anaphora Resolution. Furthermore, we combine different approaches in the text pre-processing step such as stopword removal and stemming.

Information about anaphoric relations can be beneficial for text summarisation. Anaphora is a reference which points back to some previous item, with the 'pointing back' word or phrase called an anaphor, and the entity to which it refers, or for which it stands, its antecedent [6]. For example: "*The boy is clever, he is the first in the class*". Here, "*he"* is the reference (anaphor) that points back to its antecedent "*The boy"*.

Anaphoric information is used to enrich the latent semantic representation of a document, from which a summary is then extracted. Different algorithms for text summarisation are based on the similarity between sentences. AR can be used to identify which discourse entities are repeatedly mentioned, especially when different mentioning forms are used. Previous work involving the inclusion of AR in summarisation, reports some increase in performance. Vodolazova et al. [5] achieved an improvement of around 1.5 % over their summarisation system based on the lexical Latent Semantic Analysis (LSA) by incorporating anaphoric information into it. The performance was tested on the DUC 2002 data using the ROUGE evaluation toolkit [9]. The authors also mentioned two strategies for including anaphoric relations: (1) *addition*, when anaphoric chains are treated as another category of term for the input matrix construction; (2) *substitution*, when each representative of an anaphoric chain in the text is substituted by the chain's first representative. The evaluation results show that the substitution approach performs significantly worse than the addition approach and in some tests even worse than the same system without including AR [5]. Although the addition strategy is better, in some cases the output summary becomes incoherent. This happens when a pronoun appears in the summary and its antecedent does not. As a result, for coherency, in this research the addition strategy is used, and after that the anaphora chain is tracked. If a pronoun appears in the summary without its antecedent, the pronoun is substituted by its antecedent.

Various researchers have investigated how the quality of automatically generated summaries can be evaluated. Vodolazova et al. [12] attempted to analyse the impact of shallow linguistic properties of the original text on the quality of the generated summaries. Nenkova et al. [18] focused on how sentence structure can help to predict the linguistic quality of generated summaries.

Summarisation evaluation methods can be broadly classified into two categories [19]: *extrinsic* evaluation, where the summary quality is judged on the basis of how helpful summaries are for a given task, and *intrinsic* evaluation which is based on an analyses of the summary, and involves a comparison between the original text (the source document) and the generated summary to measure how many ideas of the source document are covered by the summary. Intrinsic evaluations can then be divided into: *content evaluation* and *text quality evaluation*. Content evaluation measures the ability to identify the key topics of the original document, while text quality evaluation judges the readability, grammar and coherence of automatic summaries [8].

Text Summarisation evaluation can be classified into two approaches: Automatic evaluation and Human evaluation. Many evaluation studies [4, 5, 8] used automatic evaluation systems such as ROUGE [9]. ROUGE is a tool which includes several automatic evaluation methods that measure the similarity between summaries. Other studies have asked human judges to evaluate the quality of generated summaries generated using their approach [20].

The problems with the automatic evaluation are that (1) they first of all necessitate the existence of human summaries for comparison with the generated summaries; additionally (2) automatic systems lack evaluation measurements such as conciseness and informativeness because these criteria are difficult to quantify automatically and need to be assessed by humans. We take these factors into consideration as part of the evaluation in this research where we adopted the intrinsic method and used human evaluation to judge the quality of summaries generated by two different approaches.

3 Design

In order to evaluate the impact of Anaphora Resolution from a human perspective a system has been developed. The system consists of the following modules:

Summarisation:

– *Summarisation Algorithm*: The task of summarisation is an unsupervised extractive task, so TextRank is used as the summarisation module.
– *Stemming*: TextRank measures the similarity between sentences by calculating the number of overlapping words between them. Therefore, to ensure that words that

share the same root (i.e. ones which are deemed very close to each other in meaning), words are stemmed before comparing them to each other. The Lancaster stemmer was selected to be used in this research. This follows on the successful approach of Augat and Ladlow [21], where they compared the performance of three stemmers: WordNet stemmer, Lancaster stemmer, and Porter stemmer; of the three, the Lancaster stemmer performed best.
- *Stopword Removal*: As stopwords are generally assumed to be of less, or no, informational value, the system performs stopword removal on the original text before stemming and summarisation.

Anaphora Resolution. The AR system used is the Stanford Deterministic Coreference Resolution System [22] which implements a multi-pass sieve coreference resolution (or anaphora resolution) system. The system relates pronouns to their nominal antecedents.

Addition Approach. As it has been shown by Vodolazova et al. [5], that the substitution approach performs significantly worse than the addition approach, the addition approach is used in this system.

Sentence Location. The sentence location approach is applied to extract the first sentence from the original document to be used as the first sentence in the summary, even if TextRank did not extract it. The selected dataset for this research is Wikipedia, and since the first sentence in any Wikipedia article states the definition of the article, the sentence location can provide valuable information to the summarisation process.

Chain Tracking. The Anaphora resolution system produces an anaphoric chain. This chain consists of an antecedent (first representative, a name for example) and the anaphor that refers to that antecedent (and may be more than one anaphor). This chain is used to check the coherence of the summary produced by the system, by checking that the anaphoric expressions contained in the summary still have the same interpretation that they had in the original text, and if it does not have the same interpretation, the anaphor is substituted by its antecedent.

The system implemented for this research carries out the text processing in the following order:

1. Anaphora in the original text is resolved.
2. Stopwords are removed.
3. Stemming is applied.
4. TextRank is executed on the new text.
5. Sentence ranks are produced.
6. The highest ranked sentences are selected and extracted from the original text.
7. The first sentence is selected to be used in the summary (if it was not already selected by TextRank).
8. If a pronoun is found in the summary without its antecedent, it is replaced by the antecedent.

4 Evaluation

4.1 Dataset

In order to investigate the impact of AR on summaries generated from different subject domains, 70 Wikipedia abstracts were selected from various subject domains. The abstract is the first section from the Wikipedia article. The abstracts have different length (from around 180 words to more than 560 words.) Wikipedia was chosen as an open source of content in multiple domains. The domains were selected at random and the abstracts were randomly chosen from within the following domains: Accidents, Natural Disasters, Politics, Famous People, Sports, and Animals. Two summaries were generated for each abstract, one summary was generated without applying Anaphora Resolution and the second summary was generated by applying Anaphora Resolution before summarisation[1].

When we tried to generate the two summaries for each article, we noticed in 45 articles that the two summaries were identical, which reflected the possibility that AR may have had no impact on the produced summaries.

We also noticed that in producing the two summaries, the identical summaries were often produced from the shorter articles. This means that, the length of text in the article proportional to the impact of AR on the produced summaries (i.e. the shorter the article, the less the impact of AR on the summary, and vice versa).

Hence, 45 abstracts were removed and the final selected set of abstracts, for the experiment, are 25 articles of the original document-set where each domain has 5 articles.

4.2 Experimental Setup

To conduct the comparative evaluation between the two summaries, a web application was built. The abstracts were divided into groups; each group had five abstracts that came randomly from different domains. When the first user logged in, they were randomly assigned a group. The next user was then randomly assigned a group from the remaining unassigned groups. This continued until all the groups were assigned to users. The process was then repeated for the next set of users who logged in to the system. This ensured an even spread of assessment. Each article in the group that is presented to the user was followed by the two generated summaries. The users were not aware of which summary was produced using which technique. Each participant in the experiment was asked to evaluate each summary according to the following characteristics:

1. **Readability and Understandability**: Whether the grammar and the spelling of the summary are correct and appropriate.
2. **Informativeness**: How much information from the source text is preserved in the summary.

[1] In our experiment, the first summary is marked as TR and the second summary is marked as TR + AR.

3. **Conciseness**: As a summary presents a short text, conciseness means to assess if this summary contains any unnecessary or redundant information.
4. **Overall**: The overall quality of the summary.

An open call for participation in the experiment was made on several mailing lists and through social media. The selected participants' ages were ranging from 24 to 50 years old and they were from different backgrounds.

The users were asked to evaluate each characteristic on a six-point Likert scale (ranging from one to six, where one is the lowest quality and six is the highest). For the sake of data completeness, each participant was asked to fill in answers for all the evaluation characteristics. A text box was provided in case the user had any comments regarding the quality of the summaries or regarding the difference between them. The users were allowed to leave this box empty in case they did not have any comments. After each summary was evaluated individually, the user was asked to indicate which summary they preferred. In most cases this characteristic was automatically set by the system based upon the user's individual evaluation of the two summaries, while still allowing for manual adjustment if the user so wished. However if the user evaluated the two summaries equally, then neither summary was preselected, and they had to manually make a selection.

5 Results and Discussion

Thirty-eight users participated in the experiment. Each user evaluated at least four abstracts in different domains and also added comments about the two summaries. The final results were analysed regarding:

(a) The general impact of Anaphora Resolution on Text Summarisation.
(b) The domain-specific impact of Anaphora Resolution on Text Summarisation.

5.1 General AR Impact

Table 1 reports the mean scores of the user evaluations, the variance, and p-value for each characteristic. The results show that the difference between the two approaches (TR and TR + AR) is not statistically significant, which coincides with results obtained by Mitkov et al. [6]. However, from the results obtained in our research, we can notice a pattern. Although the difference is not significant, the differences are constantly positive, which would suggest the combined approach does have a positive influence overall. Nevertheless, this pattern will need to be confirmed with future experiments and more detailed analyses.

As part of the analysis, the users' comments were also examined. In general, the comments show that the majority of the users preferred the second summary (TR + AR). The users' comments also showed that AR has no significant impact on the quality of the produced summaries. This is demonstrated by the following sample comment: "*I noticed that usually the difference between the two summaries is just one* *or two sentences. This is why it is hard to judge which one is better than the other. They are very close to each other with just a small difference.*"

Table 1. General Anaphora resolution impact

Criteria / Approach		TR	TR + AR
Readability& Understand-ability	Mean	4.8625	4.8688
	Variance	0.00625	
	P-Value	0.9291 (> 0.05)	
Informativeness	Mean	4.3875	4.4563
	Variance	0.0688	
	P-Value	0.4671 (> 0.05)	
Conciseness	Mean	4.2375	4.375
	Variance	0.1375	
	P-Value	0.1276 (> 0.05)	
Overall	Mean	4.35625	4.4
	Variance	0.04375	
	P-Value	0.6195 (> 0.05)	
Preferable	Mean	0.8375	1.09375
	Variance	0.2563	
	P-Value	0.1596 (> 0.05)	

5.2 Domain Specific AR Impact

Table 2 reports the mean scores of the user evaluations, the variance, and p-value for each characteristic in each of the different domains tested. In this experiment, abstracts were divided by domains to investigate the intuition that language characteristics differ by subject domain, and thus to assess if the impact of AR on summarisation also differed.

The impact is shown to be statistically significant in the "*Politics*" domain for Conciseness (p-value = **0.02**) and Readability (p-value = **0.05**). Also in the "*Animals*" domain, the difference between the two summarisation approaches is statistically significant with respect to two characteristics: Informativeness (p-value = **0.001**) and User Preference (p-value = **0.02**). In the "*Politics*" and "*Animals*" domains, AR has a positive impact on the generated summaries in all characteristics. In support of this argument, user comments confirmed this finding. The following is a sample of user testimonies: "*First summary did not summarise first part of the topic as efficient as the second summary.*" Where the first summary here refers to the TR summary and the second summary is the TR + AR summary.

In contrast, in the "*Accidents*" domain, the summaries generated with TR only are more preferable and are evaluated more positively with respect to all characteristics than summaries with AR.

Regarding the "*Famous People*" domain, although it was expected that AR would have a positive impact on its articles as they have more pronouns than others, the results actually show a negative impact of AR in all characteristics (variance ranges from −0.17 to −0.72). User comments on these domains show that the summary that was generated with TR only is more preferable than TR + AR. One of these comments is: "*The first summary* [the TR summary] *included an extra piece of information about*

Table 2. Domain specific Anaphora resolution impact

Domain / Criteria		**C1**[a]	**C2**[b]	**C3**[c]	**C4**[d]	**C5**[e]
Accidents	TR	5.21	4.50	4.71	4.50	0.93
	TR+AR	5.00	4.36	4.29	4.14	0.79
	Varience	-0.21	-0.14	-0.43	-0.36	-0.14
	P-Value	0.27	0.61	0.16	0.29	0.81
Natural Disasters	TR	5.00	4.60	4.33	4.56	0.82
	TR+AR	4.96	4.61	4.51	4.60	1.22
	Varience	-0.04	-0.09	0.18	0.04	0.40
	P-Value	0.75	0.64	0.27	0.79	0.28
Politics	TR	4.61	4.33	3.56	4.11	0.72
	TR+AR	5.00	4.33	4.44	4.28	1.11
	Varience	0.39	0.00	0.89	0.17	0.39
	P-Value	**0.05**	0.99	**0.02**	0.62	0.44
Famous People	TR	4.79	4.21	4.34	4.31	1.28
	TR+AR	4.62	4.17	4.21	4.21	0.55
	Varience	-0.17	-0.03	-0.14	-0.10	-0.72
	P-Value	0.38	0.89	0.53	0.63	0.07
Sports	TR	4.63	4.43	4.27	4.30	0.73
	TR+AR	4.80	4.37	4.30	4.37	1.10
	Varience	0.17	-0.07	0.03	0.07	0.37
	P-Value	0.28	0.71	0.84	0.71	0.38
Animals	TR	4.96	4.13	4.13	4.21	0.50
	TR+AR	4.92	4.96	4.42	4.54	1.67
	Varience	-0.04	0.83	0.29	0.33	1.17
	P-Value	0.80	**0.001**	0.18	0.12	**0.02**

[a] Readability & understand-ability, [b] Informativeness, [c] Conciseness, [d] Overall, [e] Preferable

travels which was omitted in the second [the TR + AR summary]." Which means that the TR summary is more informative than the TR + AR summary.

It was noted from the results in the "*Natural Disasters*" and "*Sports*" domains that the impact of AR varies from one characteristic to another. For example, the second summary (AR + TR) in both domains is more preferable than the first (positive variance, 0.40 and 0.37 respectively), however, the first summary is slightly more readable than the second one. A user comment shows that: "*These two are tight. The first is more readable.*"

6 Conclusion and Future Work

This paper presented an alternative approach to evaluate the impact of Anaphora Resolution on Text Summarisation. Various researchers have attempted to evaluate this impact automatically, however, automatic evaluation always lack evaluation characteristics that are difficult to effectively measure automatically and need to be evaluated manually. In this research we measured this impact by asking humans to evaluate two summaries, one of which was generated without Anaphora Resolution (AR), while the second was generated with Anaphora Resolution. The results showed that, in general, AR has a slight but not significant impact on the quality of the summaries produced. Further experimentation on the domain-specific impact of AR showed that AR is domain dependent and its impact varies from one domain to another. The results showed that AR has a statistically significant impact on TS for some criteria.

In the experimental setup stage, it was noticed that the length of the article and the density of the anaphoric references have an influence upon the summaries produced.

Future work will therefore be carried out in order to expand upon these findings. It will investigate the relation between text characteristics, such as the length, the style, the domain, and the number of anaphors in the original text, and their impact upon the generated summaries.

Acknowledgements. This research is supported by Science Foundation Ireland through the CNGL Programme (Grant 12/CE/I2267) in the ADAPT Centre (www.adaptcentre.ie) at Trinity College Dublin.

References

1. Bayomi, M., Levacher, K., Ghorab, M.R., Lawless, S.: OntoSeg: a novel approach to text segmentation using ontological similarity. In: Proceedings of 5th ICDM Workshop on Sentiment Elicitation from Natural Text for Information Retrieval and Extraction, ICDM SENTIRE. Held in Conjunction with the IEEE International Conference on Data Mining, ICDM 2015, Atlantic City, NJ, USA, 14 November 2015
2. Lawless, S., Lavin, P., Bayomi, M., Cabral, J.P., Ghorab, M.: Text summarization and speech synthesis for the automated generation of personalized audio presentations. In: Biemann, C., Handschuh, S., Freitas, A., Meziane, F., Métais, E. (eds.) NLDB 2015. LNCS, vol. 9103, pp. 307–320. Springer, Heidelberg (2015)

3. Cruz, F., Troyano, J.A., Enríquez, F.: Supervised TextRank. In: Salakoski, T., Ginter, F., Pyysalo, S., Pahikkala, T. (eds.) FinTAL 2006. LNCS (LNAI), vol. 4139, pp. 632–639. Springer, Heidelberg (2006)
4. Mihalcea, R., Tarau, P.: TextRank: bringing order into texts. In: Proceedings of EMNLP 2004, pp. 404–411. Association for Computational Linguistics, Barcelona, Spain (2004)
5. Vodolazova, T., Lloret, E., Muñoz, R., Palomar, M.: A comparative study of the impact of statistical and semantic features in the framework of extractive text summarization. In: Sojka, P., Horák, A., Kopeček, I., Pala, K. (eds.) TSD 2012. LNCS, vol. 7499, pp. 306–313. Springer, Heidelberg (2012)
6. Mitkov, R., Evans, R., Orăsan, C., Dornescu, I., Rios, M.: Coreference resolution: to what extent does it help NLP applications? In: Sojka, P., Horák, A., Kopeček, I., Pala, K. (eds.) TSD 2012. LNCS, vol. 7499, pp. 16–27. Springer, Heidelberg (2012)
7. Ježek, K., Poesio, M., Kabadjov, M.A., Steinberger, J.: Two uses of anaphora resolution in summarization. Inf. Process. Manag. **43**(6), 1663–1680 (2007)
8. Steinberger, J., Ježek, K.: Evaluation measures for text summarization. Comput. Inform. **28**(2), 251–275 (2012)
9. Lin, C., Rey, M.: ROUGE : a package for automatic evaluation of summaries. In: Text Summarization Branches Out: Proceedings of ACL-2004 Workshop, vol. 8 (2004)
10. Murray, G., Renals, S., Carletta, J.: Extractive summarization of meeting recordings. In: Proceedings of Interspeech 2005 - Eurospeech, 9th European Conference on Speech Communication and Technology, Lisbon, Portugal, 4–8 September 2005
11. Fiszman, M., Rindflesch, T.C.: Abstraction Summarization for Managing the Biomedical Research Literature (2003)
12. Vodolazova, T., Lloret, E., Muñoz, R., Palomar, M.: Extractive text summarization: can we use the same techniques for any text? In: Métais, E., Meziane, F., Saraee, M., Sugumaran, V., Vadera, S. (eds.) NLDB 2013. LNCS, vol. 7934, pp. 164–175. Springer, Heidelberg (2013)
13. Nenkova, A., Mckeown, K.R.: Automatic summarization. In: Proceedings of 49th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts of ACL 2011. Association for Computational Linguistics (2011)
14. Edmundson, H.P.: New methods in automatic extracting. J. ACM (JACM) **16**(2), 264–285 (1969)
15. Teufel, S., Moens, M.: Sentence extraction as a classification task. In: Proceedings of ACL, vol. 97 (1997)
16. Luhn, H.P.: The automatic creation of literature abstracts. IBM J. Res. Dev. **2**(2), 159–165 (1958)
17. Page, L., Brin, S., Motwani, R., Winograd, T.: The PageRank citation ranking: bringing order to the web. Stanford InfoLab (1999)
18. Nenkova, A., Chae, J., Louis, A., Pitler, E.: Structural features for predicting the linguistic quality of text. In: Krahmer, E., Theune, M. (eds.) Empirical Methods. LNCS, vol. 5790, pp. 222–241. Springer, Heidelberg (2010)
19. Sparck Jones, K., Galliers, J.R., Walter, S.M.: Evaluating Natural Language Processing Systems: An Analysis and Review. LNCS, vol. 1083. Springer, Heidelberg (1996)
20. Saggion, H., Lapalme, G.: Concept identification and presentation in the context of technical text summarization. In: Proceedings of 2000 NAACL-ANLP Workshop on Automatic Summarization, pp. 1–10. Association for Computational Linguistics, Stroudsburg, PA, USA (2000)

21. Augat, M., Ladlow, M.: An NLTK package for lexical-chain based word sense disambiguation (2009)
22. Lee, H., Peirsman, Y., Chang, A., Chambers, N., Surdeanu, M., Jurafsky, D.: Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 shared task. In: Proceedings of 15th Conference on Computational Natural Language Learning: Shared Task, pp. 28–34. Association for Computational Linguistics, Stroudsburg, PA, USA (2011)