Available online at www.sciencedirect.com

ScienceDirect

Computer Speech & Language 52 (2018) 141–164

www.elsevier.com/locate/csl

Automatic cohesive summarization with pronominal anaphora resolution

Jamilson Antunes[a,*], Rafael Dueire Lins[a,b], Rinaldo Lima[a,b], Hilário Oliveira[a], Marcelo Riss[c], Steven J. Simske[d]

[a] *Informatics Center, Federal University of Pernambuco, Recife, Pernambuco, Brazil*
[b] *Department of Statistics and Informatics, Federal Rural University of Pernambuco, Recife, Pernambuco, Brazil*
[c] *Hewlett-Packard Brazil, Porto Alegre, Rio Grande do Sul, Brazil*
[d] *Hewlett-Packard, Fort Collins, CO 80528, USA*

Received 20 August 2016; received in revised form 29 July 2017; accepted 29 May 2018
Available online 7 June 2018

Abstract

Automatic Text Summarization is the process of creating a compressed representation of one or more related documents, keeping only the most valuable information. The extractive approach for summarization is the most studied and aims to generate a compressed version of a document by identifying, ranking, and selecting the most relevant sentences or phrases from a text. The selected sentences go verbatim into the summary. However, this strategy may yield incoherent summaries, as pronominal coreferences may appear unbound. To alleviate this problem, this paper proposes a method that solves unbound pronominal anaphoric expressions, automatically enabling the cohesiveness of the extractive summaries. The proposed method can be applied to two distinct scenarios. The first one aims to find and fix unbound anaphoric expressions present in the generated summaries at a post-processing stage; whereas the second one is performed at the preprocessing stage of the proposed pipeline and generates an intermediate version of the input document that resolves the unbound pronominal coreferences. The proposed solution was evaluated on the CNN news corpus using the seventeen summarization techniques most widely acknowledged in the literature and four state-of-the-art summarization systems. Moreover, it also provides a comparative evaluation concerning two distinct assessment scenarios which are compared to a baseline. The experiments performed achieved very encouraging quantitative and qualitative results.

Keywords: Automatic summarization; Cohesive summarization; Extractive summarization; Anaphoric expressions

1. Introduction

The task of selecting relevant and concise information on the Internet is an increasingly difficult challenge due to the growing amount of available data. Automatic means to assist users in sieving and managing such a huge amount

* Corresponding author.
E-mail address: jba@cin.ufpe.br (J. Antunes), rdl@cin.ufpe.br (R.D. Lins), htao@cin.ufpe.br (H. Oliveira), marcelo.riss@hp.com (M. Riss), steven.simske@hp.com (S.J. Simske).

https://doi.org/10.1016/j.csl.2018.05.004
0885-2308/

of data is therefore of rising importance. In particular, Automatic Text Summarization (ATS) systems have been pointed out as offering a possibility to identify and present information in a clear and concise way. Thus, it has received considerable attention by researchers in the last few years. Essentially, ATS is the process of mechanically creating a compressed version of one or more related documents, extracting their central information.

According to Lloret and Palomar (2012), ATS approaches can be divided into two major branches: (i) *extractive summarization* which aims at selecting the most important sentences from a set of documents, copying them verbatim into the final summary; and (ii) *abstractive summarization* which attempts to summarize documents by paraphrasing or modifying the higher-ranked sentences in order to improve text cohesion eliminating redundancies.

There is a growing interest in using ATS in several Natural Language Processing (NLP) applications, such as Text Indexing and Classification (Manning et al., 2008). ATS can also be used as a preliminary step in selecting the core information to be viewed on mobile devices such as cell phones (Cabral et al., 2015).

Extractive techniques are the most studied ATS strategies in the literature (Das and Martins, 2007; Lloret and Palomar, 2012), but it is prone to create fragmented summaries with sentences whose meaning depend on their context, i.e., other sentences. Such fragment summaries usually contain *broken chains of anaphoric references* which are known to introduce cohesion problems in Automatic Text Summarization (Mani et al., 1998). Coreference chains are built by coreference resolution (CR) systems that are able to match all references to a single entity in a document, regardless of their possible syntactical forms (Nenkova et al., 2011). CR systems usually match nouns, noun phrases, or pronouns in a document. Errors regarding broken chains of anaphoric references are, however, very common in extractive summaries, especially (not surprisingly) in short ones (Kaspersson et al., 2012); and in particular for summarizers that focus on the content coverage and disregard how sentences relate to each other. Extractive summarizers, as pointed out by Nenkova (2006), often score relatively well because they are evaluated against gold standard summaries created by humans using measures that prioritize content coverage, such as ROUGE (Lin, 2004). However, summary cohesiveness is usually not considered in such evaluations, as the assessment measures favor the coverage of the key information content, disregarding how well the text fits together (Smith et al., 2012). Rennes and Jonsson (2014) conducted an eye tracking investigation of the summary cohesiveness generated by ATS systems. Their study aimed at assessing how different types of cohesion errors affect the reading of a text summarized by an extractive automatic summarizer. More precisely, by using an eye-tracking camera, they focused on the nature of three different types of cohesion errors occurring in extractive summaries, namely: the *erroneous anaphoric reference*, the *absent cohesion or context*, and the *broken anaphoric reference*. Statistical analysis of the data revealed that absent cohesion or context and broken anaphoric reference (pronouns) caused some disturbance in reading the summary.

The main goal of this work is to address such cohesion problems in ATS by proposing, implementing, and evaluating a method for the analysis and correction of anaphoric expressions (nouns and pronouns). The method presented can be regarded as a classical extractive summarization approach, but with the bonus that it improves the overall cohesion of the generated summaries. In other words, by analyzing the coreference chains produced by natural language parsers, the method presented here can both identify and filter out the spurious coreference chains, thereby reducing the cohesion errors that are common in the extractive summarization approaches. Furthermore, the proposed solution is independent of the extractive summarization system or method implemented.

The proposed rule-based solution can be applied to analyze and solve broken coreference expressions in two distinct scenarios: (a) improving the cohesion of the extractive generated summaries; (b) pre-processing the input text in such a way as to eliminate coreferences in sentences before the extractive selection.

In order to evaluate the proposed solution, two experiments – a quantitative and a qualitative one – were conducted in two application scenarios. The quantitative experiment aims to evaluate the impact of the proposed method regarding the content selection issue by adopting the traditional ROUGE (Lin, 2004) scores and the cosine similarity method (Donaway et al., 2000). Meanwhile, the qualitative evaluation performed manually by humans evaluated the cohesion of the generated summaries after using the proposed method. Seventeen sentence scoring techniques and four extractive summarization systems were considered in the experiments performed. Such evaluations were conducted in the news article domain adopting the CNN corpus (Lins et al., 2012). The tests made use of the current version of the CNN-corpus, which is composed of 3000 news articles in English, all of which have an extractive summary (the *gold standard*), produced by three experts following a computer-assisted sentence selection methodology.

The contributions of this paper are twofold:

- Proposing and implementing a new method for handling anaphoric expressions in extractive summarization tasks in order to improve the cohesion of the final summaries;
- Providing a comparative assessment of the proposed method against several extractive summarization techniques and systems.

The remainder of this paper is organized as follows: Section 2 presents a brief account of the literature on coreference resolution in summarization systems. The methodological steps in the proposed approach aiming a more cohesive summarization technique are presented in Section 3. The evaluation methodology and description of the corpus used are detailed in Section 4. In particular, several state-of-the-art extractive summarization strategies and systems are evaluated and discussed in three different scenarios in Section 5. The conclusions of this paper and perspectives for further work are presented in Section 6.

2. Related work on coreference resolution in summarization

One of the shortcomings of the current extractive summarization systems is that they usually consider words and sentences in isolation, ignoring their relationship. As a result, the final summaries produced by such systems tend to contain sentences with dangling or broken anaphoric references which hamper comprehension of the summary as a whole. To mitigate this problem, several automatic summarization systems that take coreference resolution into account have been proposed.

Steinberger et al. (2007) proposed two methods for exploiting coreference resolution in automatic summarization. The first solution is based on *Latent Semantic Analysis* (LSA) (Landauer and Dumais, 1997), which exploits the anaphoric information extracted by their coreference resolution system (GUITAR). The second approach, similarly to the present work, scans summaries looking for anaphoric expressions. Their strategy is to change anaphoric nouns for the first element in their corresponding coreference chain. Both systems were evaluated using the DUC 2002 corpus[1] and achieved significantly better performance than the versions of the system not processing anaphoric information.

Gonçalves et al. (2008) introduce a summarization system (CorrefSum) that improves the referential cohesion of the extractive summaries by using knowledge about coreference chains. Their system improves Steinberger's work by enabling a more flexible choice of the most representative entity in a coreference chain instead of always taking its first occurrence, as done by Steinberger. CorrefSum evaluation was based on the Summ-it (Collovini et al., 2007) corpus containing 50 newspaper texts (in Brazilian Portuguese) from the Folha de S̃ao Paulo newspaper.

Gonçalves et al. (2008) and the second method proposed in Steinberger et al. (2007) both check the generated summaries with the goal of handling broken anaphoric expressions. However, their methods do not perform a preliminary analysis of the coreference chains to filter out the spurious coreference chains, as the method proposed here does. In fact, the preliminary experimental results obtained here, using the Stanford CR system, show that almost 50.31% of the coreference chains found by these systems are not suitable for the summarization task (Section 5.1).

Orăsan (2009) used anaphora resolution to improve a term-based summarizer rooted on the simple Term Frequency–Inverse Document Frequency (TF–IDF). The authors argue that the most important sentences in a text can be determined on the basis of the importance of the words it contains. The summarizer was evaluated on several versions of the CAST corpus (Hasler et al., 2003) modified by six automatic coreference resolution systems and one human coreference annotator. The experimental results, evaluated using the cosine similarity measure (Donaway et al., 2000), suggest that pronominal coreference resolution was beneficial in improving the legibility of the produced summaries. Moreover, if the human-annotated corpus version is used, the term-based summarizer yields the best results for several compression rates.

Smith et al. (2012) proposed the COHSUM summarizer which is indirectly based on the distribution of coreferences in the source texts. COHSUM calculates a rank for each sentence, computing their in-/out coreference links to other sentences. The importance of the sentences was computed using a variant of PageRank (Brin and Page, 1998). COHSUM underlying idea is that the sentences providing the most in/out references to other sentences are considered the most important and, therefore, they should be selected. The summaries produced by COHSUM were

[1] http://duc.nist.gov/pubs.html#2002.

evaluated on the DUC 2002 corpus using two measures: ROUGE (for content coverage) and cohesion (intact and broken coreference chains) compared to the original document. The results revealed that COHSUM performed comparatively well in terms of content coverage, and it produced significantly fewer broken coreference chains and more intact coreferences compared to other summarizers.

This work differs from the ones by Orăsan (2009), Smith et al. (2012), and the first method proposed by Steinberger et al. (2007) in the sense that all those studies integrate coreference resolution either as a weighting factor for ranking sentences or as additional heuristics during the summarization process; whereas the proposed method for anaphora resolution can be applied in the original text either as a preprocessing or as a post-processing step over the extractive summaries independently of the technique or summarization system employed in their generation.

Christensen et al. (2013) proposed a system for producing cohesive summaries from multiple documents. The proposed system, so-called G-FLOW, attempts to balance coherence and salience among sentences, estimating the level of cohesiveness of a candidate summary. The G-FLOW model is essentially a graph representing the discourse relations across sentences based on several cohesion clues present in the text, including discourse phrases, deverbal nouns, and coreference referents. The author uses coreference mentions as features for both weighting (ranking) sentences and connecting G-FLOW nodes (or sentences). Differently, this work employs coreference resolution to analyze anaphoric expressions and replace them by the most representative referent.

Silveira (2015) investigated the impact of post-processing procedures on the extractive summaries aiming at obtaining coherent summaries. She combined several tasks that modify and relate sentences to each other, such as sentence simplification, paragraph creation, and insertion of discourse connector phrases, putting it all together as an attempt to improve the quality of the final summary. Her method is applicable only in the post-processing step, while the one proposed here can also be applied in the preprocessing step. Furthermore, Silveira's work does not employ any coreference resolution.

The rule-based method presented here either performs a corpus pre-processing or a summary post-processing step, replacing coreference instances by their most important entity in a coreference chain. The proposed method is independent of extractive summarization systems while related studies are tightly bound to a particular extractive summarization system. Moreover, the method presented here introduces specific criteria for such substitutions, preventing many repetitions of anaphoric expressions in text, while the related work always replaces the pronominal mentions by their referring entities. Finally, all the previous studies have not conducted an evaluation as extensive as the one reported here which employed an assessment methodology involving several extractive summarization techniques and systems, adopting a much larger summarization corpus.

3. Automatic cohesive summarization

As already stated, the summaries generated by extractive summarization systems usually contain broken references (Smith et al., 2012; Nenkova et al., 2011). To address this problem, a flexible software architecture (Fig. 1) that integrates a state-of-the-art coreference resolution system and several rules was proposed.

Fig. 1. Functional architecture of the Anaphoric Expressions Solver employed in two distinct scenarios.

Fig. 1 shows the functional architecture of the *Anaphoric Expressions Solver* (AES), the proposed method for analyzing, filtering, and resolving anaphoric coreferences chains in texts. The AES module can be applied in two distinct, but related contexts (Fig. 1(a) and (b)).

In Fig. 1(a), a set of documents is given as input to the Text Pre-processing component which, besides the traditional natural language processing subtasks, including tokenization and POS tagging, it also performs coreference resolution (CR). CR is still a very challenging task as the performance of the state-of-the-art systems is around 60% in terms of F1-measure according to the CoNLL-2011 shared task (Lee et al., 2011). Indeed, a significant number of false positives and false negatives can still be found even using the current state-of-the-art CR systems. To mitigate this problem, the AES component in Fig. 1(a) aims at improving the output analysis of the Text Pre-processing step by applying a set of rules that:

- filter the most relevant coreference chains;
- find the most representative entities and their corresponding referents as well;
- correct many kinds of errors made by CR systems.

As a result, the original source documents are enriched with solved coreferences, originating new or intermediate documents. Such documents, containing more precise coreference information, can be processed by an extractive summarization system that will produce more cohesive summaries.

Fig. 1(b) schematizes the task of combining the generated extractive summaries with the source documents in order to produce more cohesive summaries. The cohesive summaries are built after applying the following rules:

- search for broken references in the extractive summary;
- identify and extract the entity instances whose referents are broken in the extractive summary and find out the most representative entity for such mentions;
- generate a new version of the extractive summary by replacing the pronouns or anaphoric expressions with the most representative entity, but controlling the number of such repetitions in the final version of the summary.

The motivation guiding this research work is to investigate to what extent the current extractive summarization systems and techniques can profit from the heuristics introduced by the AES method in the application scenarios presented above. Another goal of the proposed solution is investigating the influence of anaphoric resolution on the performance of extractive summarizers.

The components of the functional architecture of the AES (Fig. 1) are detailed next.

3.1. Text Pre-processing

The Text Pre-processing step provides the morphosyntactic analysis of the input documents. The current implementation of the AES method relies on the Stanford CoreNLP toolkit, a state-of-the-art NLP system able to perform a myriad of natural languages subtasks including sentence splitting, tokenization, POS tagging, among others.[2]

The following NLP subtasks of the CoreNLP were chosen:

- *Sentence splitting:* delimits the sentences boundaries in the text;
- *Tokenization:* identifies the individual words or symbols (tokens) within sentences;
- *POS tagging:* provides part-of-speech categories to the tokens;
- *Lemmatization:* removes inflectional endings, such as the plural form of nouns, returning the base or dictionary form of a word;
- *Named entity recognition:* identifies and classifies a word or a group of consecutive words in a sentence into preselected categories such as Person, Organizations, Locations, among others;
- *Coreference resolution:* discovers all the relevant entities and their referents (nominal and pronominal) in a text.

[2] Stanford Coreference Resolution System. http://nlp.stanford.edu/software/dcoref.shtml.

Due to its paramount importance in the proposed solution, the Stanford Coreference Resolution system is further described next.

3.1.1. Stanford Coreference Resolution System

The Stanford Coreference Resolution System (SCRS) extends the multi-pass sieve system of Lee et al. (2013) which consists of a collection of deterministic coreference resolution models that incorporate lexical, syntactic, semantic, and discourse information. The system propagates global information by sharing attributes (e.g., gender and number) across mentions in the same cluster.

The SCRS was selected because it provides state-of-the-art performance on coreference resolution as reported in Lee et al. (2013), being ranked first at the CoNLL-2011 shared task (Pradhan et al., 2011), yielding a score of 57.8 in the closed[3] track and 58.3 in the open[4] track.

Although SCRS achieved good results, there is still room for improvements. Thus, the present work attempts to improve the performance of SCRS by integrating a set of heuristics as one of the contributions of the Anaphoric Expressions Solver, described next.

3.2. Anaphoric Expressions Solver

The Anaphoric Expressions Solver (AES), shown in Fig. 1, is one of the main contributions of this paper. It uses the output of the Text Pre-processing step, mainly the output of the CR system.

The AES implementation extends the previous work by applying a set of rule-based heuristics for the following tasks:

- *(Task 1)* Filtering out the spurious coreference chains before identifying the most representative entities (MRE) and referents, and;
- *(Task 2)* Improving the text quality of the generated summaries. In other words, reducing redundancy of the most representative entities in the final summaries. Such redundancy is usually due to the simple strategy of replacing every referent with the corresponding entity, as done by Steinberger et al. (2007).

The rule-based heuristics in Task 1 are rooted in the notion of the MRE in a coreference chain, defined as the entity represented by its full name followed by the shortest of its appositions present in the text. Another acceptable form of the most representative entity can have up to five tokens not separated by commas, e.g., multi-word entities.

The heuristics in Task 2 address the problem of avoiding the resolution of all anaphoric references which would lead to redundant information and repetitive entity references in the document. In order to improve the cohesiveness of summaries, Task 2 checks the distance, in terms of the number of sentences, between the entity and its referring pronoun. More precisely, if the entity information is found in the nearest preceding sentence, then there is no need to resolve the reference given by the pronoun found in the next sentence. A straightforward application of this idea is related to the substitution of anaphoric expressions that have their contexts not present in the generated summary.

In the following, an example of representative entity, intermediate document and cohesive summary as defined by AES are provided.

(Example 1) News article at http://www.edition.cnn.com/2013/01/24/business/davos-uk-cameron/ : **Cameron: We must focus on trade, taxes, transparency**
S1: Free trade, transparency and a crackdown on tax cheats will be at the heart of Britain's G8 presidency, [**Prime Minister David Cameron**]$_1$ told the World Economic Forum in Davos on Thursday as [*he*]$_1$ set out [*his*]$_1$ vision for a more competitive Europe.
S2: The speech comes a day after [**Cameron**]$_1$ made headlines by promising the British people a vote on European Union membership if [*he*]$_1$ wins the next general election in 2015.
An ideal coreference solver would find the following coreference chain:

[3] Only the provided data can be used, i.e., WordNet and gender gazetteer.
[4] Any external knowledge source can be used. They used additional animacy, gender, demonym, and country and states gazetteers.

Chain 1:
he in sentence S1 (task 1),
his in sentence S1 (task 1),
Cameron in sentence S2 (task 2),
he in sentence S2 (task 1)
MRE: *Prime Minister David Cameron* in S1 (task 2)
Output of the Example (*Replace_Document_MRE*):
S1: Free trade, transparency and a crackdown on tax cheats will be at the heart of Britain's G8 presidency, **Prime Minister David Cameron** told the World Economic Forum in Davos on Thursday as *he* set out *his* vision for a more competitive Europe.
S2: The speech comes a day after **Prime Minister David Cameron** made headlines by promising the British people a vote on European Union membership if *he* wins the next general election in 2015.
Output of the Example (*Replace_Summary_MRE*):
S1: Free trade, transparency and a crackdown on tax cheats will be at the heart of Britain's G8 presidency, **Prime Minister David Cameron** told the World Economic Forum in Davos on Thursday as *he* set out *his* vision for a more competitive Europe.
S2: The speech comes a day after *Cameron* made headlines by promising the British people a vote on European Union membership if *he* wins the next general election in 2015.

In all the examples above, one can observe that the proposed heuristics prefer shorter, but more informative, representative entities. The two previous tasks are performed by the following AES algorithm:

The *Find_Most_Representative_Entity* function is in charge of finding the most representative entity in a given coreference chain, according to the criteria mentioned earlier on, whereas the *Replace_Document_MRE* routine replaces the anaphoric expressions with the most representative entity of the chain in a document, but controlling the

number of the referent repetitions. To find the MRE the maximum number of 10 tokens to an entity was empirically defined in order to avoid unnecessary information.

The *Replace_Summary_MRE* routine solves the broken coreferences in the summaries generated by extractive summarizers. It verifies whether or not the representative entity is contained in a previous sentence to avoid redundancy in the final summary.

3.2.1. Discussion on the substitution algorithms in related work

The proposed summary post-edition algorithm (AES) differs from the one proposed by Steinberger et al. (2007) in the sense that the method here imposes further restrictions both over the definition of a representative entity in a coreference chain, and the way substitutions are performed. More precisely, Steinberger's work always selects the first nominal expression in the coreference chain as its representative entity. However, such an approach is likely to introduce repetitive entity expressions that can strongly influence the final length of the resulting sentences (cf. Example 1). Therefore, as the experimental results performed in this work revealed, many of the extractive summarization systems favor the selection of longer sentences.

Another closely related problem is that the longer versions of the representative entities invariably alter the frequency distribution of their constituent terms which can mislead the summarizers based on term frequency techniques. The substitution method proposed here chooses the best representative entity for a given coreference chain giving preference to, among all of the representative entity candidates, the shortest and the most informative one, using the rules in Task 1 (Section 3.2). Finally, the substitution method proposed by Steinberger et al. (2007) may introduce redundancy in the final summaries, contrarily to the proposed solution that further applies heuristics to control the number of repetitions of the representative entity in the same sentence or others phrases in the same summary.

4. Experimental setup

This section describes the experimental setup which comprises the corpus (Section 4.1), the measures used for evaluating the proposed solution for anaphoric expression solving, the extractive summarization systems and techniques used (Section 4.2), and the evaluation scenarios (Section 4.3). In Section 4.3.1, a baseline summarization scenario which evaluates all the summarization systems and techniques on the CNN corpus is introduced. Sections 4.3.2 and 4.3.3 describe the two last assessment scenarios corresponding to the two ways of employing the AES method: either in the preprocessing stage or the post-processing stage of the proposed summarization pipeline. In addition, the quantitative and qualitative measures used for evaluating the proposed solution for anaphoric expression solving are presented in Sections 4.4–5.2.1.

4.1. Evaluation dataset: the CNN corpus

All the experiments carried out in this work are in the news articles domain, using the CNN corpus composed by Lins et al. (2012). The current version of the CNN corpus encompasses 3000 news articles written in English collected from the CNN website[5] containing news from all over the world. The documents address general themes such as sports, politics, business, and health, among others. Each CNN news article has its corresponding *highlights*, a high quality and concise abstractive summary composed of up to four sentences written by the original author of the news article. The highlights cover the main topics of the news headline in a very succinct form. Thus, the highlights can be considered as a single-document abstractive summary that (expertly) reflects the gist of the news article.

The highlights were used to guide a semi-automatic process to create an extractive summary model for each document of the CNN corpus. This process was performed by human annotators that mapped each sentence of the article highlight onto one or more sentences of the original article content. A group of six proficient, but non-native, English human annotators was designated to perform this mapping task. To ensure the quality of the extractive summaries generated, two annotators mapped each document, and in the case of divergence, a third annotator conducted the divergence resolution process. The set of sentences mapped by the annotators constitutes the gold standard model,

[5] http://www.cnn.com.

Table 1
Example a highlight, extractive gold summary and semi-extractive summary of a news article the CNN corpus.

Models	Summaries
Highlight	Coca-Cola – the world's ubiquitous brown fizzy drink – is staying afloat as the soda market shrinks. Many point to a marketing strategy around the so-called "secret recipe" as key to its resilience. It is never been patented, to keep the formula secret, but many say they have discovered the recipe.
Extractive gold standard	Coca-Cola – the world's ubiquitous brown fizzy drink – is staying afloat as the soda market shrinks, and many point to a marketing strategy around the so-called "secret recipe" as key to its resilience in a struggling industry. "They kept the formulas secret, partly in order to increase sales with a sense of special mystery and to prevent competition, but also to keep people from knowing how cheap the ingredients were and how large the profits", **he** says. The company has never patented the formula, saying to do so would require its disclosure.
Semi-extractive gold standard	Coca-Cola – the world's ubiquitous brown fizzy drink – is staying afloat as the soda market shrinks, and many point to a marketing strategy around the so-called "secret recipe" as key to its resilience in a struggling industry. "They kept the formulas secret, partly in order to increase sales with a sense of special mystery and to prevent competition, but also to keep people from knowing how cheap the ingredients were and how large the profits", **pharmacist John Pemberton** says. The company has never patented the formula, saying to do so would require its disclosure.

an extractive summary of the article that can be used as a reference summary to evaluate extractive summarization systems. The gold standard models of the entire CNN corpus contain 10,755 sentences, which represents approximately 10% of the total number of sentences of the documents. Table 1 shows an example of a highlight, extractive summary, and its semi-extractive version of a CNN news article.

As the mapped sentences were extracted without any modification, problems related to the cohesion of extractive summaries can still be found. Seeking to alleviate this problem, a semi-extractive summary version of the gold standard was created for each highlight by performing the coreference resolution process followed by human validation.

In this work, the semi-extractive reference models of the CNN documents are used in all of the conducted experiments presented in Section 5. In the following, both basic statistics and the distribution of the pronouns found in the CNN corpus are provided.

CNN corpus basic statistics. Table 2 summarizes some basic statistics about the original corpus, the set of highlights, and the golden standards.

Pronoun distribution. Fig. 2 presents the distribution of pronouns in the CNN corpus.

The morphological class of a given token in the CNN corpus is determined by its POS tag. Fig. 2 shows that the subject pronouns (*Subj*) (I, you, he, she, it, we, you, and they) are the most frequent, followed by the possessive adjective (*Poss Adj*) pronouns. Object pronouns (*Obj*) are fairly frequent. In fact, they outnumber the possessive (*Poss*) and the reflexive pronouns (Reflex) taken altogether. Among all of the pronouns found in the CNN corpus, the most frequent are: *it* (18,496), *he* (15,697), *I* (10,487), *his* (10,413), *they* (9221), and *we* (8233). The total number of pronouns (130,689) as shown in Fig. 2 corresponds to 5.59% of all tokens in the entire corpus. Evans (2001) found approximately the same percentage of pronouns (5.7%) in the British National Corpus (BNC) (Burnard, 1995). Such

Table 2
Basic statistics of the CNN-corpus.

News	
Total number of news articles	3000
Total number of sentences	115,396
Average number of sentences per article	38.47
Total number of tokens	2,296,693
Average number of tokens per sentence	19.90
Highlights	
Total number of sentences	10,674
Average number of sentences per article	3.56
Total number of tokens in the highlights	130,844
Average number of tokens per highlight	12.26
Gold standard	
Total number of sentences	10,755
Average number of sentences	3.59
Total number of tokens	269,434
Average number of tokens/sentence	25.05

Fig. 2. Frequency of pronouns by type found in the CNN corpus.

results justify the choice of the news broadcast domain for evaluating the proposed solution for anaphoric expression resolution as pronouns are found more frequently in the news domain than in others (Biber et al., 1998; Orăsan, 2009).

4.2. Extractive summarization systems and techniques

For the three assessment scenarios of the proposed solution for anaphoric expression solving (Section 3), four extractive summarization systems were selected: AutoS, C4J, HP-UFPE FS, and Aylien. These systems were chosen because of their good performance, as reported by Batista et al. (2015). In addition, the following techniques were evaluated: Aggregate Similarity (AS), Word Co-Occurrence (N-GRAM), Sentence Centrality 2 (SC), Bushy Path (BP), Sentence Length (SL), TextRank Score (TS), Cue-phrase (CP), Sentence Centrality 1 (BLEU), Sentence Position in Paragraph (SPP), Lexical Similarity (LS), Term Frequencies (TF/IDF), Word Frequency (WF), Upper Case (UC), Resemblance to the Title (RT), Inclusion of Numerical Data (ND), Proper Noun (PN), Sentence Position in Text (SPT). More details about all the aforementioned sentence scoring techniques can be found elsewhere (Ferreira et al., 2013).

The summarization techniques above were evaluated with the goal of estimating which one produces the highest number of broken references. The results of such an evaluation would shed light on their impact on the two assessment scenarios compared to the baseline summarization one.

Summarization systems differ from the sentence scoring techniques in the sense that the former can integrate a combination of several scoring techniques, i.e., it is a specific solution with particular settings and design decisions. The systems considered in this work are briefly described in the remainder of this section.

4.2.1. Classifier4J

Classifier4J (C4J) (Nick, 2003) is a Java library that essentially performs text classification. It also provides additional services such as text summarization by means of a single-document extractive summarization method based on word frequency (Luhn, 1958). C4J selects sentences that contain the most frequent words in a document to compose a summary.

4.2.2. HP-UFPE Functional Summarization

The HP-UFPE Functional Summarization (HP-UFPE FS) (Ferreira et al., 2013; 2014b) system is based on 17 different techniques found in the recent literature. These features were largely evaluated using news, blogs and scientific articles documents. The best combination of the summarization techniques for each document type was selected to compose a hybrid system that first classifies the documents using predefined categories, and then summarizes the document using the methods that are more suitable to the specific document category, thus dealing with the generalization problem.

In this work, the HP-UFPE FS used the best combination of summarization techniques found in Ferreira et al. (2014b) for news articles. The selected features were: Term Frequency–Inverse Document Frequency (TF–IDF), Lexical Similarity, Sentence Position, and Resemblance to the title.

- *TF–IDF*: this algorithm computes the importance of a sentence based on the following steps: (i) removal of all stop words; (ii) calculates the TF–IDF of all remaining words using the formula presented in Ferreira et al. (2013); (iii) for each sentence, it sums up the TF–IDF score of each word in a sentence; and (iv) finally, all scores are normalized.
- *Lexical similarity*: it is based on the assumption that important sentences are identified by strong chains.
- *Sentence position*: this algorithm assumes that the first sentences of a paragraph are the most important ones. The sentences are ranked as follows: the first sentence in a paragraph has a score value of 5/5, the second sentence has a score 4/5, and so on. In short, the last sentences in a paragraph are less significant.
- *Sentence resemblance to the title*: it represents the vocabulary overlap between each sentence and a given document title.

The above features were combined using the arithmetic mean of the individual scores. Each method returns a score between 0 and 1, which are added up and divided by the number of tested features. The *N* sentences with the highest scores are selected to compose the summary, where *N* depends on the selected compression rate.

HP-UFPE FS also has two extensions providing language independent summaries (Cabral et al., 2014) and multi-document summarization (Ferreira et al., 2014a).

4.2.3. Autosummarizer (AutoS)

Autosummarizer (Autosummarizer, 2016) is a summarization system that produces an extractive single-document summary by splitting and ranking the most important sentences in a document. Its drawback resides in the fact that the number of selected sentences to be included into the summary cannot be set by the user. A four-sentence long summary is provided regardless of the size of the input text.

4.2.4. Aylien Text Analysis API (Aylien)

The Aylien Text Analysis API (Aylien, 2016) is a commercial tool that offers different natural language processing services, including text summarization, text classification, and sentiment analysis, among others. Its summarization service provides an extractive single document summary for texts in English, along with experimental methods for German, French, Italian, Spanish, and Portuguese. There is also a set of Software Development Kits (SDK) for different programming languages such as Java, Python, Ruby, Node.js, PHP, and .NET.

4.3. Evaluation scenarios

4.3.1. Baseline: Standard Summarization Flow (SSF)

Fig. 3 shows the Standard Summarization Flow (SSF) that denotes the classical extractive summarization process (Radev et al., 2002; Lloret and Palomar, 2012). The SSF will be used as the baseline performance for the evaluation methodology adopted in this work.

4.3.2. Post-processing: Anaphoric Expressions Solver in Summary (AESS)

Fig. 4 displays the application scenario of the AES method (see Section 3) after either applying the summarization techniques or the systems on the CNN corpus. As already presented, the main goal of the Anaphoric Expressions

Fig. 3. Standard summarization flow.

Fig. 4. AESS flow.

Solver is to address the problem of broken pronominal anaphoric expressions produced by the extractive summarization methods. Therefore, by applying the AES over the generated summaries from the Baseline Scenario, one is able to eliminate all the broken pronominal coreferences in the summaries. This directly impacts the cohesion of the summaries since many pronouns are resolved (*step post-processing*).

4.3.3. Preprocessing: the Anaphoric Expressions Solver on Corpus (AESC)

Fig. 5 depicts the last AES method application scenario, which was already described in Section 3. The basic difference between this scenario and the one described in the previous section is that either the techniques or the summarization systems take a new version of the CNN corpus as input. The main goal in this AES application scenario is to correct the mentions that can cause broken coreferences in the extractive summaries. In other words, the input corpus is preprocessed before being analyzed by the AES component (Fig. 5) that finally generates documents in a specific intermediate format (*step preprocessing*).

4.4. Quantitative evaluation

Two measures were used for assessing the effectiveness of the summarization systems and techniques evaluated in this section, as follows.

- *ROUGE* (Recall-Oriented Understudy for Gisting Evaluation) (Lin, 2004) measures the content similarity between system-developed summaries and the golden standard. The precision, recall, and *F*-measure (Baeza-Yates and Ribeiro-Neto, 1999) provided by ROUGE were used to perform the quantitative assessment in this paper.
  The ROUGE-N score is computed as presented in Eq. (1):

$$c_n = \frac{\sum_{c \in RSS} \sum_{gram_n \in C} Count_{match}(gram_n)}{\sum_{c \in RSS} \sum_{gram_n \in C} Count(gram_n)} \tag{1}$$

  where *Count_match (gram$_n$)* is the maximum number of n-grams co-occurring in a candidate summary and a reference summary (Golden Standard), while *Count*(*gram$_n$*) is the number of n-grams in the reference summary. One should notice that the average n-gram coverage score, C_n, is in fact a measure of recall. The present work adopted ROUGE-1 and ROUGE-2.
- *Cosine similarity* (Donaway et al., 2000): the degree of similarity between the sentences of the generated summary and the golden standard can be calculated using cosine similarity. The terms (*T*) in the sentences are weighted using Term Frequency−Inverse Document Frequency (TF−IDF) as in Eq. (4). The term frequency in a given document is defined as the number of times a given term appears in it, as shown in Eq. (2):

$$TF_i = \frac{T_i}{\sum_{k=1}^{n} T_k} \tag{2}$$

  where T_i is the number of occurrences of the term and T_k is the sum of the occurrences of all the terms in the

Fig. 5. AESC flow.

document. The inverse document frequency is a measure of the importance of the term:

$$IDF = \log\left(\frac{N}{n_i}\right) \tag{3}$$

where N is the number of sentences in the document, and n is the number of sentences containing the significant term. The corresponding weight is, therefore, computed as,

$$W_t = TF_i * IDF_i \tag{4}$$

The degree of similarity between sentences can be measured using the cosine similarity as in Eq. (5):

$$\cos(X_i, Y_i) = \frac{x_i \cdot y_i}{\left|\,|x_i| \cdot |y_i|\,\right|} \tag{5}$$

<?show -aptara_TEMP_aptara-?>

$$= \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} (x_i)^2}\sqrt{\sum_{i=1}^{n} (y_i)^2}} \tag{6}$$

where X and Y are representations of a summary and its golden standard based on the vector space model (Steinberger et al., 2007).

All the measures mentioned above focus on the informativeness of the extractive summary. Such features enable the direct comparison among the selected systems and the techniques presented in Section 4.2.

Finally, as the golden standard summaries of the CNN corpus comprise about 10% of the size of the original text corpus, the same compression rate was used in all the experiments reported on in this section.

5. Experimental results and discussion

The proposed method was employed in two assessment scenarios involving all the extractive summarization systems and techniques presented in Section 4.2.

The evaluation methodology adopted here first performs a quantitative analysis of the ratio of both spurious coreference chains and the invalid anaphoric expressions produced by the pre-processing step using the CNN corpus with 3000 news articles (Section 5.1). The main goal here is to filter out only coreference chains containing valid anaphoric expressions in the summaries.

Section 5.3.1 presents the automatic evaluation with a comparative analysis involving all the evaluation scenarios described in Sections 4.3.1–4.3.3. Finally, Section 5.2 presents the human evaluation using the proposed method.

5.1. Preliminary results on the identification of anaphoric expressions

This section describes the two preliminary experiments made, aiming at motivating the treatment of anaphoric expressions for achieving cohesiveness in extractive summarization.

5.1.1. Pronoun distribution

This preliminary experiment employs the Stanford CoreNLP toolkit to find the distribution of anaphoric pronouns in the CNN corpus. Such a distribution can also be regarded as a rough estimate of the possible number of sentences with broken coreferences in the generated summaries.

The CNN corpus contains 130,689 pronouns, which corresponds to approximately 5.69% of words in the corpus, as shown in Fig. 2. The coreference resolution system in CoreNLP tool identified 115,791 pronouns making part of some coreference chain. Such a number corresponds to 88.6% of the total number of pronouns in the CNN Corpus.

To obtain more cohesive extractive summaries, a growing interest has been focused on pronominal coreference resolution (Orasan, 2009; Steinberger et al., 2007; Smith et al., 2012). More precisely, on those pronouns which their

Table 3
Pronoun distribution in the CNN corpus.

	Subject (Subj)	Object (Obj)	Possessive (Poss)	Possessive Adjectives (Poss Adj)	Reflexive (Reflex)
Count	42,600	10,773	7701	20,489	1112

Table 4
Final evaluation of valid chains and mentions.

	Source chains	Replaced chains	Chain elimination rate (%)	Source mentions	Replaced mentions	Mention elimination rate (%)
Total	49,329	24,511	50.31	115,791	45,149	61.01
Average/Doc	16.44	8.17	–	38.60	15.05	–

referent entity appear in a different sentence. Table 3 shows the distribution of the pronoun references categorized by the main types of pronouns.

According to the results shown in Table 3, subject pronouns correspond to 63.26% of the total number of pronouns in the CNN corpus. Interestingly, almost the same subject pronoun ratio was reported in Orăsan (2004) using a corpus comprising 76 scientific articles. Orăsan also pointed out that only a third of the pronouns had their referencing entities in the same sentence, suggesting that if a sentence containing a pronoun is extracted, special care needs to be taken to avoid dangling references. In another work, Evans (2001) reported that 67.94% of the pronouns in a tailored corpus composed of documents extracted from the SUSANNE (Sampson, 2002) and the BNC corpora (Burnard, 1995) were anaphoric.

On the one hand, such preliminary results suggest the need for pronominal resolution in automatic extractive summarization as a means of avoiding broken coreferences in summaries. In order to keep control of the number of entity repetitions, some chains and mentions should not be resolved or replaced by their referencing entity to avoid a recurrence of the representative entity in the same sentence in both the corpus or in the summaries.

5.1.2. Filtering out valid coreference chains

The main goal of this experiment is to assess the task of filtering out the valid coreference chains from which the representative entities were extracted and used by the proposed AES algorithm. The results of this experiment are summarized in Table 4.

The analysis of the texts suggests that many chains and mentions should not be considered for the summarization task. For example, Fig. 6 shows a sentence processed by the Stanford CR system,[6] in which the pronoun *her* should not be replaced by its representative entity *(Angela Merkel)*, for the sake of avoiding the repetition of the representative entity in the same sentence according to the proposed AES method.

Another example can be found in Fig. 7, in which a coreference between two pronouns is ignored by the summarization task in the AES method.

Fig. 6. Mentions valid by AES method.

[6] http://nlp.stanford.edu:8080/corenlp/process.

Fig. 7. Mentions discarded by AES method.

It is worth mentioning that only the set of all valid coreference chains were used in the experimental assessment scenarios described in the remainder of this section.

5.1.3. Analysis of broken coreferences

The results of this experiment on the CNN corpus (3000 news articles) are summarized by Tables 5 (systems) and 6 (techniques) showing:

- #Summaries: the number of summaries with broken coreferences;
- #SentSummaries: the number of sentences in #Summaries;
- #SentGolden: the number of sentences in the corresponding golden standard summaries;
- #Percentage(%): the ratio between #Summaries and the total number of summaries in the CNN corpus (3000);
- #SentBrokenCoref: the number of sentences with broken coreferences in #Summaries;
- #Average: the average number of sentences with broken coreference per summary.

An important aspect to be evaluated is the broken coreferences considered in the summaries generated in the SSF scenario. For that, the same assessment methodology in Smith et al. (2012) was adopted here in order to evaluate the cohesion level of the summaries by counting the number of broken coreferences found in the summaries.

Table 5
Distribution of the summaries with broken coreferences by systems.

Systems	#Summaries	#SentSummaries	#SentGolden	%	#SentBrokenCoref	#Mean
AutoS	1686	7979	6148	56.20	3095	2.32
Aylien	1490	7161	5482	49.67	2761	1.85
HP-FS	1219	6120	4701	40.63	2152	1.77
C4J	1185	6929	4377	39.50	2136	1.80
Mean	*1395*	*7047*	*5177*	*46.50*	*2536*	*1.82*

Table 6
Distribution of the summaries with broken coreferences by techniques.

Techniques	#Summaries	#SentSummaries	#SentGolden	%	#SentBrokenCoref	#Mean
AS	954	3894	3484	31.80	2074	2.17
N-GRAM	935	3655	3401	31.17	2087	2.23
SC	931	3774	3369	31.03	2003	2.15
BP	930	3823	3390	31.00	1984	2.13
SL	923	3842	3378	30.77	2029	2.20
TS	811	3382	2946	27.03	1529	1.89
CP	809	3474	2927	26.97	1561	1.93
BLEU	784	3211	2855	26.13	1433	1.83
SPP	777	3367	2825	25.90	1513	1.95
LS	768	3299	2800	25.60	1427	1.86
TF/IDF	750	3234	2746	25.00	1373	1.83
WF	737	3236	2697	24.57	1348	1.83
UC	694	3067	2555	23.13	1195	1.72
RT	672	3153	2498	22.40	1241	1.85
ND	656	2459	2083	21.87	1207	1.84
PN	650	2837	2410	21.67	1080	1.66
SPT	564	2882	2377	18.80	972	1.72
Mean	*785*	*3329*	*2867*	*26.17*	*1532.71*	*1.93*

Furthermore, the distribution of the broken coreferences generated by the summarization systems also provides a baseline performance concerning the cohesion level of the summaries generated by the evaluated systems and techniques. This distribution is summarized by Tables 5 and 6, which provide the number of broken pronominal coreferences by summary. These tables reflect the fact that the higher the number of broken coreferences in the summaries, the lower is their cohesion level. According to this criterion, the C4J system achieved the best performance among the evaluated systems because it selected fewer sentences with broken coreferences (1185 in total). The AutoS system comes at the last position as it had the worst results for all the pronominal coreference categories evaluated in this experiment. The results presented in Table 5 show that more than half (56%) of the summaries generated by AutoS had broken coreferences, while C4J was the more resilient among the evaluated systems. Indeed, all the summaries have, in the mean, 46.50% of broken coreferences, a very high rate.

Table 6 shows that the summarization techniques yield much less cohesion problems (26% in average). The AS technique, in particular, generated the highest number of less coherent summaries, approximately 32% of the total number of summaries, while SPT was the less affected.

The results presented in Table 6 corroborate the results presented in the previous works (Ouyang et al., 2010; Ferreira et al., 2013) that pointed out the importance of sentence position as a feature for extractive summarization, mainly in the news articles domain. A possible explanation resides in the fact that usually, the authors introduce the most representative entities at the beginning of the article. Then, the following references to such entities are replaced by pronouns. Thus, the simple heuristic of choosing the first sentences of the text is less likely to generate summaries with broken coreferences in such a domain. The PN technique achieved the second best result, possibly because it assigns higher scores to the sentences with proper nouns.

Considering the results in Tables 4 and 5, one may conclude that:

- The summaries created by all of the systems and techniques (Table 5) suffer from cohesion problems. The possible reason is that the best summarization techniques are based on word frequency. Thus, the individual technique or the combination applied does not matter; the same cohesion issues will arise.
- If there is neither further handling of the pronominal coreferences at the time of the sentence selection during summary construction or some way of dealing with the broken pronominal coreferences, one can always expect cohesion problems in the generated summaries. This fact fully justifies the proposed solution introduced in this paper aiming at eliminating the broken pronominal coreferences in the final extractive summaries.

5.2. Human evaluation

5.2.1. Qualitative evaluation

The difficulty in resorting to human evaluation has been a considerable hurdle in the development of automatic summarization systems (Silveira, 2015).

There are several problems involved, such as cost, time, evaluator training, etc. On the one hand, the selection of the linguistic and textual aspects of the summaries to evaluate is not straightforward. Properties such as cohesion, fluency or readability are hard to objectively define, and their assessment differs from person to person and also depends on several aspects such as background knowledge or even linguistic skills. In addition, the size of the input data makes the evaluation even more complex and arduous. When evaluating the automatic correction of broken coreferences and the quality of a summary, the original text must be read to verify if the automatic replacement of the pronoun by its representative entity is correct.

Being aware of these challenges, 374 texts of the CNN corpus with problems of broken coreferences were randomly selected. The proposed AES method suggested for each pronoun an entity for automatic substitution. Adopting the Amazon Mechanical Turk[7] platform, humans evaluators were responsible for assessing the quality of each coreference resolution performed through surveys.

Each survey is represented by a Human Intelligence Task (HIT), in which it has a summary highlighting one or more pronouns and entities indicated by AES, a link to the original news and two questions related to the (i) Readability/Fluency − Reference clarity ; (ii) Textual quality; and (iii) Cohesion. Reference clarity addresses anaphoric disfluencies related to whether nouns, pronouns and other referential expressions are appropriately used. These

[7] https://requester.mturk.com/.

aspects are based on the linguistic features proposed by Over et al. (2007) and Steinberger and Ježek (2009). The following is an example of a HIT used by the evaluators.

She (Arianna Huffington) describes herself as a "sleep evangelist," has nap rooms in her offices at the AOL headquarters in New York and tries to start every day with meditation. Huffington, 62, founded Huffington Post in 2005, and two years ago sold it to AOL for $315 million.

Original text: `http://edition.cnn.com/2013/03/07/business/arianna-huffington-leading-women/`

1. Does the mention in parentheses correspond to their respective pronoun? (Yes or No)
2. After the replacement(s), has the summary become easier to read and understand? (Yes or No)

The summaries were evaluated by human evaluators holding university degrees with postgraduate level. The Amazon Mechanical Turk platform can determine the profile of the evaluators, such as educational attainment, confidence level in the platform, locality, mother language, etc. Thus, only evaluators who professed have knowledge in natural language processing and have English as their first language were selected. The results of this evaluation are reported in the following section.

5.2.2. Results

The human evaluation was conducted to assess the quality of the generated summaries and empirically providing support to confirm the working hypothesis formulated: The automatic anaphoric resolution of an extractive summary improves its textual quality (cohesion, fluency, readability, etc.). The results are shown in Fig. 8.

The analysis of Fig. 8 allows observing that the AES method was able to perform the correct coreference substitution improving the cohesion of 302 summaries, equivalent to 81% of the total evaluated. This statistics means that human evaluators gave positive answers to the two questions shown in Section 5.2.1, i.e., the AES method improved 81% of the summaries both in reference clarity and text quality.

The second evaluation analyzed only the results of the first question "The subject in parentheses corresponding to their respective pronoun particularly?". This question aims to assess if the reference clarity of the summaries was improved by the AES method, i.e., the mention suggested in parentheses by AES method is the correct subject referenced by the pronoun. The AES method improved the reference clarity in 86% (322 summaries) of the summaries evaluated. Although the state-of-the-art coreference resolution systems did not present high performance figures, the AES method was able to extract the most assertive coreference chains from the SCRS due to the filters employed to remove spurious coreference chains.

In 20 summaries (5%), although the AES method indicated the correct binding, the human evaluators did not consider the substitutions to have improved the cohesion of the summaries. Each of these five summaries was further analyzed to identify the possible problems that happened during the coreference resolution that decreased the quality of the summaries. Table 7 shows fragments of summaries with such cohesion problems.

Fig. 8. Results of the human evaluation.

Table 7
Fragments of summaries with text quality problems.

[1] Canning alleged in her lawsuit that her parents forced her out of their home and that **she (Rachel Canning)** was unable to support herself financially.
[2] **She (a meat lover than a vegetable lover)** says Indians are developing new tastes because their incomes have grown – India's economy is slowly heading in the right direction again; people are earning more, they are traveling more and are being exposed to new, international cuisines.
[3] "I regret any hurt or anguish such comments may have caused any party and I look forward to greater understanding for peace and cooperation in future," **he (China's ambassador to Australia Monday)** wrote.
[4] In 1990, **he (a future Dr. Harold Freeman first envisioned in the 1980s)** pioneered the first-ever patient navigation program, training people from the community to listen and answer questions after a diagnosis.
[5] While laws in the West protect against discrimination, "it is kosher here in Asia to push youth and beauty," **he (Nok Air's Sarasin himself hedges on)** says.
[6] "He should stay on board that vessel until **He (The captain)** knows everybody is safely evacuated.
[7] "But we didn't know what was the maximum speed, so I thought it was normal," **he (One victim)** said.

Fig. 9. Performance of the AES method.

As shown in Table 7, the entity (*Rachel Canning*) indicated in fragment [1] was repeated in the sentence. This problem occurs because the SCRS was not able to identify that the word "Canning" and the entity "Rachel Canning" are in the same coreference chain. Thus, the AES method pointed out that the pronoun *she* was disconnected (unbound). One possible improvement for this problem is to treat nominal coreferences, making the AES method able to identify the mention "Canning" referred to the entity "Rachel Canning".

It was possible to observe analyzing fragments [2]–[5], that the SCRS included several unnecessary words in the MER content, which led to a decrease in the quality of those summaries. In fragment [6], the AES method performed the substitution only in the second occurrence of the pronoun, since the SCRS did not identify the first pronoun in the same coreference chain. Finally, in fragment [7], the AES method correctly identified the MER and replaced the pronoun, but according to the human evaluation, this did not improve the quality of the summary. Analyzing the complete news article,[8] it was seen that the author did not describe the victim's name. Although the substitution has generated a nominal coreference problem, it is not possible to identify who is the victim.

The evaluators point out that in 52 summaries (14%), the AES method erroneously identified MERs and thereby reduced the cohesion of the summaries. That happened due to errors in the coreference chains produced by the SCRS. Fig. 9 summarizes the results discussed above.

5.3. Automatic evaluation

5.3.1. Comparative evaluation of the summarization scenarios

This section aims to conduct a comparative analysis on the three previous evaluation scenarios. For that, the statistical-based testing methodology adopted in this work is introduced, then the comparative results are provided as a way to verify the working hypothesis raised in this work.

[8] http://edition.cnn.com/2013/07/26/world/europe/spain-train-crash/.

Statistical significance tests – The evaluation methodology employed is based on the statistical significance tests structured in three steps as follows:

1. The Shapiro and Wilk (1965) test is used to verify whether the scores (R1–F1, R2–F1 and CS) follow a normal distribution;
2. If the above test has a positive outcome, then the paired T-test (Gibbons and Chakraborti, 2003) is applied; otherwise the Wilcoxon signed-rank test (Gibbons and Chakraborti, 2003) is used;
3. Finally, the statistical test selected in the previous step is performed twice at 5% significance level (p-value < 0.05): in the first time, the null hypothesis denoting *equal averages*) is assumed; at the second time, the test is carried out assuming the null hypothesis as *higher* or *equal averages* at the same significance level.

This testing methodology is used for determining whether there is a significant difference in performance among all of the evaluated scenarios.

Detailed comparison among all of the evaluated scenarios – Tables 8 and 9 summarize the comparative assessment results among all of the three evaluation scenarios (SSF, AESS and AESC) presented in Sections 4.3.1–4.3.3. The scores (R1–F1 and CS) are provided for the same systems (Table 8) and techniques (Table 9) discussed earlier on. Already the scores of R2–F1 are presented in Table 10 for the systems and Table 11 for the techniques.

A closer look at the results in Table 8 show that:

1. AESS achieved the best results in terms of CS measure among all the evaluated systems.
2. Between AESS and AESC, the first assessment scenario obtained a significant improvement regarding both R1–F1 and SC compared to the baseline. In particular, the AESS was able to improve the baseline performance of the AutoS, HP-UFPE FS and Aylien systems concerning R1–F1.
3. For the HP-UFPE FS summarization system, AESS and AESC obtained statistically similar performance figures.
4. Finally, for the C4J system, no significant improvement was achieved in either of the AES methods tested.

A broader comparative analysis of the performance of the three assessment scenarios above, using statistical significance tests, is provided in the next section.

According to the results in Table 9, the following conclusions may be drawn:

- The summaries treated by the AESS method achieved the highest F1 scores on eleven of the seventeen techniques tested.
- The AESC method obtained the second best overall result with no significant difference in performance for the twelve techniques concerning the AESS results.
- According to the results of the systems shown in Table 8, none of the results of the baseline techniques (SFF) was better than the AES results, regarding both R1–F1 and CS scores;
- In terms of CS measure, the results are very encouraging for the AESS scenario, as it had a significant performance difference compared to the baseline. In addition, for fourteen techniques, the AESC scenario had a significant improvement compared to the baseline one, presenting results similar with AESS.
- Regarding R1–F1 scores, N-GRAM and ND techniques are the only cases where there is no significant performance difference between the AES methods and the baseline.

Table 8
ROUGE-1 and CS. Comparative performance evaluation (%) and standard deviation in parentheses of the summarization systems. The overall highest performance is marked in bold and the group of scenarios statistically similar to best performance is indicated by a †.

Systems	Baseline				AES method							
	Standard Summarization Flow (SSF)				Anaphoric Exp. Solver in Summary (AESS)				Anaphoric Exp. Solver on Corpus (AESC)			
	R	P	F1	CS	R	P	F1	CS	R	P	F1	CS
AutoS	49.87	47.02	48.40 (16.71)	64.62 (16.00)	59.26	40.59	**48.18** (14.70)	**66.68** (15.53)	55.62	38.75	45.68 (14.56)	64.00 (16.50)
Aylien	56.08	39.51	46.36 (15.61)	62.70 (16.78)	55.31	41.14	**47.18** (14.96)	**64.43** (15.97)	53.58	39.82	45.72 (14.73)	63.23 (16.28)
HP-UFPE FS	57.47	38.14	45.85 (15.67)	62.69 (16.70)	56.41	39.52	**46.48** (15.07)	**64.04** (16.07)	56.10	39.29	46.21† (15.15)	63.92† (16.47)
C4J	51.10	45.64	48.22† (16.10)	63.13 (16.26)	56.01	43.47	**48.95** (14.38)	**48.95** (15.43)	53.24	44.12	48.25† (15.29)	64.06 (16.26)

Table 9
ROUGE-1 and CS. Comparative performance evaluation (%) and standard deviation in parentheses of the techniques. The overall highest performance is marked in bold and the group of scenarios statistically similar to highest performance is indicated by a †.

Systems	Baseline				AES Method							
	Standard Summarization Flow (SSF)				Anaphoric Exp. Solver in Summary (AESS)				Anaphoric Exp. Solver on Corpus (AESC)			
	R	P	F1	CS	R	P	F1	CS	R	P	F1	CS
AS	35.44	36.75	36.08 (12.44)	47.16(16.77)	35.81	37.91	36.83† (12.24)	49.67† (16.38)	35.76	38.74	**37.19** (12.67)	**50.16** (16.43)
N-GRAM	41.08	35.92	38.33† (15.39)	50.08 (19.61)	41.34	37.33	39.23† (14.74)	52.67† (18.42)	41.39	37.98	**39.61** (15.44)	**52.79** (18.88)
SC	16.87	30.17	21.64 (11.23)	28.70 (18.65)	18.65	33.66	**24.00** (11.01)	**34.82** (17.47)	17.16	33.15	22.61 (10.99)	33.67 (17.70)
BP	35.04	36.36	35.69 (12.53)	47.26 (16.74)	35.49	37.61	36.52† (12.32)	49.98† (16.23)	35.88	38.85	**37.31** (13.15)	**50.28** (16.83)
SL	46.58	33.52	38.99 (14.64)	54.11 (18.46)	46.62	35.00	39.98† (14.12)	55.93† (17.51)	46.72	35.48	**40.33** (14.10)	**56.19** (17.18)
TS	44.28	37.04	40.34 (15.27)	53.81 (18.72)	44.11	38.55	**41.14** (14.60)	**55.90** (17.98)	42.96	38.51	40.61† (14.79)	55.51† (18.34)
CP	30.99	35.11	32.93 (12.72)	44.34 (18.22)	31.91	36.88	**34.22** (12.33)	47.36† (17.18)	31,68	36.77	34.04† (12.40)	**47.56** (16.97)
BLEU	22.77	36.15	27.94 (16.36)	34.34 (22.31)	24.13	38.82	**29.76** (15.53)	**40.46** (20.57)	22.85	38.40	28.65† (15.70)	38.84 (20.90)
SPP	30.78	35.38	32.92 (13.27)	44.65 (18.16)	31.54	37.11	**34.10** (13.01)	**47.46** (17.41)	31.18	36.85	33.78† (13.01)	47.07† (17.15)
LS	50.38	37.14	42.76 (15.71)	58.28 (18.21)	50.07	38.37	**43.45** (15.30)	**59.73** (17.82)	48.50	37.94	42.57† (15.20)	58.88† (17.92)
TF/IDF	53.08	36.84	43.49 (16.87)	59.04 (18.59)	52.59	38.15	**44.22** (16.21)	**60.57** (18.00)	51.86	37.61	43.60 (15.72)	60.33† (17.66)
WF	52.57	37.35	43.68 (16.07)	59.71 (17.84)	52.06	38.68	**44.39** (15.27)	**61.35** (17.16)	49.79	38.07	43.15 (15.21)	60.07 (17.75)
UC	44.30	35.72	39.55 (14.62)	53.58 (18.34)	44.45	37.06	**40.12** (14.14)	**55.72** (17.54)	43.85	36.13	39.61† (13.95)	54.66† (17.60)
RT	47.42	40.03	43.41 (14.10)	57.41 (17.47)	47.22	41.19	**44.00** (13.71)	**59.11** (16.83)	46.73	40.77	43.55† (14.01)	58.71† (17.08)
SPT	35.13	39.47	37.17 (15.18)	48.25 (20.10)	34.97	40.46	37.52† (14.59)	50.56† (18.96)	35.28	40.92	**37.89** (14.61)	**50.87** (19.08)
ND	38.11	37.18	37.64† (14.44)	49.68 (18.81)	38.06	38.59	**38.32** (13.75)	**52.15** (18.09)	37.76	38.20	37.98† (13.61)	51.86† (17.81)
PN	43.26	35.65	39.09 (14.38)	52.66 (17.86)	43.33	37.16	**40.01** (13.76)	**55.05** (17.02)	43.25	36.50	39.59† (13.86)	54.68† (17.36)

In summary, one may conclude that the overall performance of both application scenarios (AESS and AESS) of the proposed solution for cohesive extractive summarization outperforms the baseline performance in almost all the summarization techniques tested regarding the traditional ROUGE-1 and cosine similarity measures.

For the results of the R2−F1 (Tables 10 and11) scores it can be concluded that:

1. The systems did not present statistically relevant results for the scenarios evaluated.
2. The AESC scenario presented lower results than the baseline for the AutoS and Aylien systems. The AESS scenario remains statistically equal to the baseline for all systems.
3. The AESS and AESC scenarios surpassed the baseline for 8 techniques. For the scenario AESS, those surpassing baseline were the techniques SC, CP, SPP and UC, and for the scenario AESC those surpassing baseline were the techniques AS, N-GRAM, BP and SL. The other techniques did not present statistically relevant results in relation to the baseline.
4. In general, the AESS and AESC scenarios present statistically similar or better results than the baseline.

Overall comparison among the evaluated scenarios − Fig. 10 summarizes the overall performance assessments of all summarization systems and techniques reported by Tables 8−11. In Fig. 10, the range between 0.0 and 0.2

Table 10
ROUGE-2. Comparative performance evaluation (%) and standard deviation in parentheses of the summarization systems. The overall highest performance is marked in bold and the group of scenarios statistically similar to best performance is indicated by a †.

Systems	Baseline			AES Method					
	Standard Summarization Flow (SSF)			Anaph. Exp. Solver in Summary (AESS)			Anaph. Exp. Solver on Corpus (AESC)		
	R	P	F1	R	P	F1	R	P	F1
AutoS	32.41	32.34	**32.38** (22.24)	38.31	26.88	31.59† (19.69)	34.06	24.20	28.30 (18.97)
Aylien	35.49	25.70	29.82† (20.57)	34.45	26.45	**29.92** (19.49)	32.29	24.90	28.11 (18.96)
HP-UFPE FS	36.60	25.33	**29.94** (20.42)	35.27	25.93	29.89† (19.36)	35.14	25.85	29.79† (19.36)
C4J	34.41	31.27	32.76† (21.52)	36.55	28.89	**32.27** (19.14)	34.73	29.31	31.79† (20.33)

Table 11
ROUGE-2. Comparative performance evaluation (%) and standard deviation in parentheses of the techniques. The overall highest performance is marked in bold and the group of scenarios statistically similar to highest performance is indicated by a †.

Systems	Baseline			AES method					
	Standard Summarization Flow (SSF)			Anaph. Exp. Solver in Summary (AESS)			Anaph. Exp. Solver on Corpus (AESC)		
	R	P	F1	R	P	F1	R	P	F1
AS	15.34	16.77	16.02 (15.53)	15.59	17.49	16.48† (15.08)	15.76	18.44	**16.99** (15.70)
N-GRAM	20.32	18.90	19.58 (19.37)	20.17	19.70	19.93† (18.46)	20.64	20.68	**20.66** (19.29)
SC	7.82	12.38	9.58 (9.73)	7.65	13.91	**9.87** (9.98)	7.14	13.40	9.32 (9.90)
BP	15.11	16.12	15.60 (15.54)	15.42	16.92	16.13† (15.05)	16.25	19.03	**17.53** (16.09)
SL	24.66	18.11	20.88 (18.57)	24.16	18.65	21.05† (17.74)	24.26	19.40	**21.56** (18.01)
TS	24.53	20.96	**22.60** (19.72)	23.79	21.52	22.59† (18.68)	22.79	21.54	22.15† (18.67)
CP	12.72	15.34	13.90 (15.14)	13.12	16.22	**14.51** (14.67)	12.68	16.06	14.17† (14.90)
BLEU	12.36	19.36	**15.09** (17.00)	11.82	19.89	14.83† (16.39)	10.91	18.74	13.79 (15.97)
SPP	13.34	15.78	14.46 (15.38)	13.47	16.44	**14.81** (15.06)	13.05	16.26	14.48† (15.14)
LS	29.68	22.91	25.86† (20.61)	29.16	23.38	**25.95** (19.92)	27.47	22.67	24.84† (19.69)
TF/IDF	31.95	23.65	**27.18** (22.15)	31.01	24.06	27.10† (21.10)	29.73	23.21	26.07† (20.25)
WF	31.80	23.69	**27.15** (21.16)	30.82	24.12	27.06† (20.02)	28.14	23.17	25.41† (19.77)
UC	24.06	20.91	22.37 (18.79)	23.90	21.46	**22.61** (18.06)	22.51	20.14	21.26 (17.82)
RT	28.49	24.64	**26.43** (18.58)	27.62	24.98	26.24† (17.92)	27.13	24.81	25.92† (18.20)
SPT	19.72	22.48	**21.01** (18.71)	18.82	22.37	20.44† (17.66)	19.05	23.04	20.86† (17.83)
ND	19.88	19.90	19.89† (17.84)	19.45	20.56	**19.99** (16.96)	19.08	20.38	19.71† (16.76)
PN	22.56	19.92	21.16† (18.70)	22.57	20.78	**21.63** (17.74)	22.05	20.28	21.13† (17.98)

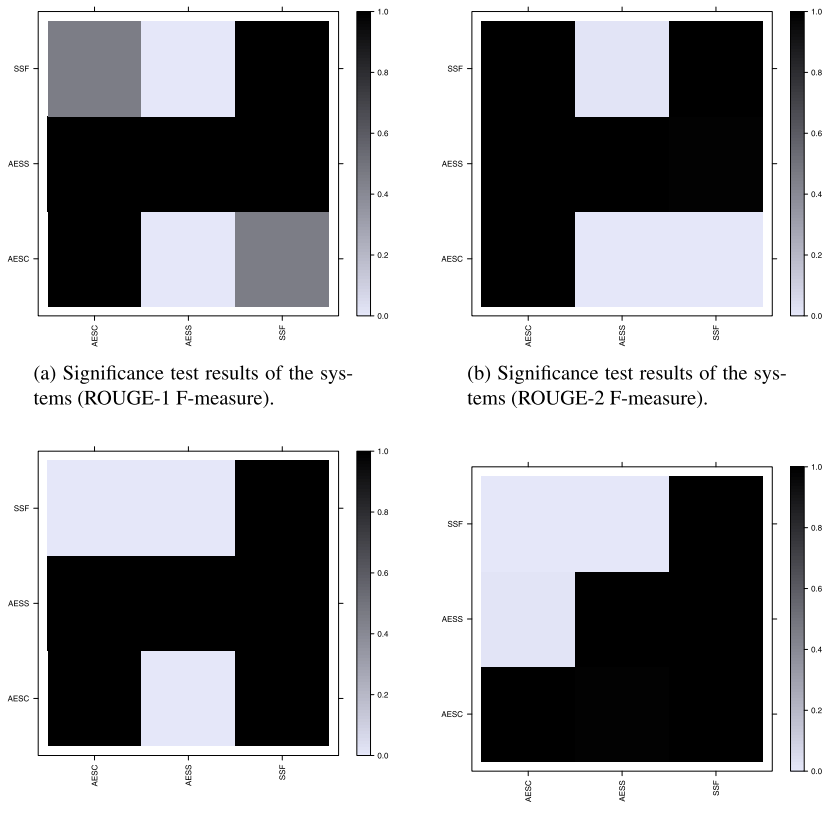
denotes a p-value < 0.01 and p-value < 0.05, respectively, i.e., such a range indicates that the results are statistically different at 95% and 99% of confidence level, respectively, and they are depicted in Fig. 10 in a lighter hue of gray. On the other hand, the p-value ranging from 0.2 to 1.0 means that there is no statistical difference between the two given scenarios, which is represented using a dark hue of gray. In all of the figures, the scenario in the X axis is compared to the one in Y axis.

According to Fig. 10, AESS and AESC outperform the baseline (SSF) regarding R1−F1, R2−F1 and CS for the majority of systems and techniques. Indeed, the AESS scenario obtained the highest overall R1−F1 and R2−F1 scores, whereas the AESC reached the second best overall result, as it can be seen in Fig. 10(a) and (c). The AESC scenario lost to the baseline only in the R2−F1 scores for the systems, see Fig. 10(d). In terms of CS, the AESC scenario (Fig. 10(b)) yielded a significant improvement compared to the other scenarios for the techniques; whereas, for the systems (Fig. 10(c)), the AESS scenario was the clear winner.

The bottom line is that the overall results confirmed the working hypothesis formulated in this paper that a deep analysis and correction of anaphoric expressions involving nouns and pronouns, either in the pre-processing or in the post-processing of the generated summaries, improves summary cohesiveness and also boosts the performance in terms of the traditional evaluation quality measures used by the extractive summarization community.

6. Conclusions and lines for further work

This paper presented a new method for extractive text summarization that attempts to produce more coherent summaries by solving pronominal anaphoric expressions. The rule-based implementation of the proposed method is able to both identify and filter out the spurious coreference chains from the input corpus (leaving intact the most relevant entities), and to avoid dangling references that could hamper the legibility of the text in the extractive summaries generated. The method proposed extends the related work by improving the cohesion of the summaries generated by the classical extractive summarization approaches. The method was extensively evaluated under two distinct application scenarios using several systems and techniques for extractive summarization. The AES method achieved satisfactory results in both quantitative and quantitative assessments. The overall results obtained using the CNN corpus demonstrated its effectiveness when compared to the baseline summarization



(a) Significance test results of the systems (ROUGE-1 F-measure).

(b) Significance test results of the systems (ROUGE-2 F-measure).

(c) Significance Test Results of the systems (cosine similarity) and techniques (ROUGE-1/ROUGE-2 F-measure).

(d) Significance Test Results of the techniques (cosine similarity).

Fig. 10. Overall comparison of the summarization systems and techniques on the three assessment scenarios.

scenario. Moreover, the proposed architecture for extractive summarization is independent of the summarization system employed.

The exhaustive number of evaluations for 17 techniques and 4 systems was necessary to analyze the impact of the AES method on each technique and system in the summarization process for the two evaluated scenarios. Besides, it shows which system has the highest and lowest number of summaries with broken references. For example, the AES method was able to correctly determine that the PN and SPT techniques have the smallest number of broken transfers and, consequently, less legibility and cohesion problems. As expected, these two techniques should have a smaller number of broken references, since the first one is based on the selection of sentences with the highest number of proper nouns and the second one selects the first or last sentences of the text.

Despite the encouraging results obtained, there is still room for many improvements. Current work considers alternatives for generating pronouns to avoid repeating entities within a summary and the shortening of repeated entities, i.e., exchanging the second occurrence of an entity by a shorter reference. The authors believe that a sentence simplification algorithm would allow the insertion of new contents in summaries, increasing its information coverage. The feasibility of exploring several features integrated with regression models will be analyzed to estimate the cohesion of the automatically generated summary. Such a mechanism may enable the selection of the most cohesive summary from a set of several summary candidates.

Acknowledgments

This research work has been partially funded by a R&D project between Brazilian HP and UFPE originated from tax exemption (IPI – Law number 8.248, of 1991 and later updates).

References

Autosummarizer, 2016. Automatic Text Summarizer. Retrieved from http://autosummarizer.com/. Last accessed 16 June 2016.

Aylien, 2016. Aylien Text Analysis API. Retrieved from http://aylien.com/text-api. Last accessed 16 June 2016.

Baeza-Yates, R., Ribeiro-Neto, B., 1999. Modern Information Retrieval. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

Batista, J., Ferreira, R., Oliveira, H., Ferreira, R., Lins, R.D., Pereira e Silva, G., Simske, S.J., Riss, M., 2015. A quantitative and qualitative assessment of automatic text summarization systems. In: Proceedings of the 2015 ACM Symposium on Document Engineering. ACM, New York, NY, USA, pp. 65–68. doi: 10.1145/2682571.2797081.

Biber, D., Conrad, S., Reppen, R., 1998. Corpus Linguistics: Investigating Language Structure and Use. Cambridge University Press, New York, NY, USA.

Brin, S., Page, L., 1998. The anatomy of a large-scale hypertextual web search engine. Comput. Netw. ISDN Syst. 30 (1–7), 107–117. doi: 10.1016/S0169-7552(98)00110-X.

Burnard, L., 1995. Users Reference Guide British National Corpus Version 1.0.

Cabral, L., Lima, R., Lins, R., Neto, M., Ferreira, R., Simske, S., Riss, M., 2015. Automatic summarization of news articles in mobile devices. In: Proceedings of the Fourteenth Mexican International Conference on Artificial Intelligence (MICAI), pp. 8–13. doi: 10.1109/MICAI.2015.8.

Cabral, L.S., Lins, R.D., Mello, R.F., Freitas, F., Ávila, B., Simske, S., Riss, M., 2014. A platform for language independent summarization. In: Proceedings of the 2014 ACM Symposium on Document Engineering. ACM, New York, NY, USA, pp. 203–206. doi: 10.1145/2644866.2644890.

Christensen, J., Soderl, S., Etzioni, O., 2013. Towards coherent multi-document summarization. In: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, Atlanta, Georgia, pp. 1163–1173.

Collovini, S., Carbonel, T., Fuchs, J.T., Coelho, J.C., Rino, L.H.M., Vieira, R., 2007. Summit: Um corpus anotado com informações discursivas visando à sumarização automática. In: Proceedings of the SBC, Workshop em Tecnologia da Informação e da Linguagem Humana (TIL'2007). Rio de Janeiro, RJ, p. 1.

Das, D., Martins, A.F.T., 2007. A Survey on Automatic Text Summarization. Technical Report. Literature Survey for the Language and Statistics II Course at Carnegie Mellon University.

Donaway, R.L., Drummey, K.W., Mather, L.A., 2000. A comparison of rankings produced by summarization evaluation measures. In: Proceedings of the 2000 NAACL-ANLPWorkshop on Automatic Summarization, 4. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 69–78. doi: 10.3115/1117575.1117583.

Evans, R., 2001. Applying machine learning toward an automatic classification of *It*. LLC 16 (1), 45–57. doi: 10.1093/llc/16.1.45.

Ferreira, R., Cabral, L.S., Freitas, F., Lins, R.D., Silva, G.F., Simske, S.J., Favaro, L., 2014a. A multi-document summarization system based on statistics and linguistic treatment. Expert Syst. Appl. 41 (13), 5780–5787. doi: 10.1016/j.eswa.2014.03.023.

Ferreira, R., Freitas, F.L.G., Cabral, L.S., Lins, R.D., Lima, R., Pereira e Silva, G.F., Simske, S.J., Favaro, L., 2014b. A context based text summarization system. In: Proceedings of the Eleventh IAPR International Workshop on Document Analysis Systems. Tours, France, pp. 66–70. April 7–10, 2014.

Ferreira, R., Souza Cabral, L., Lins, R.D., Pereira e Silva, G., Freitas, F., Cavalcanti, G.D.C., Lima, R., Simske, S.J., Favaro, L., 2013. Assessing sentence scoring techniques for extractive text summarization. Expert Syst. Appl. 40 (14), 5755–5764. doi: 10.1016/j.eswa.2013.04.023.

Gibbons, J.D., Chakraborti, S., 2003. Nonparametric Statistical Inference (Statistics: A Series of Textbooks and Monogrphs). fourth CRC.

Gonçalves, P.N., Rino, L.H.M., Vieira, R., 2008. Summarizing and referring: towards cohesive extracts. In: Proceedings of the 2008 ACM Symposium on Document Engineering. Sao Paulo, Brazil, September 16–19, 2008, pp. 253–256. doi: 10.1145/1410140.1410193.

Hasler, L., Orăsan, C., Mitkov, R., 2003. Building better corpora for summarisation. In: Proceedings of the 2003 Corpus Linguistics. Lancaster, UK, pp. 309–319.

Kaspersson, T., Smith, C., Danielsson, H., Jönsson, A., 2012. This also affects the context - Errors in extraction based summaries. In: Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12). Istanbul, Turkey, pp. 23–25.

Landauer, T.K., Dumais, S.T., 1997. A solution to Plato's problem: the latent semantic analysis theory of the acquisition, induction, and representation of knowledge. Psychol. Rev. 104, 211–240.

Lee, H., Chang, A., Peirsman, Y., Chambers, N., Surdeanu, M., Jurafsky, D., 2013. Deterministic coreference resolution based on entity-centric, precision-ranked rules. Comput. Linguist. 39 (4), 885–916. doi: 10.1162/COLI_a_00152.

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

Lin, C., 2004. Rouge: a package for automatic evaluation of summaries. In: Marie-Francine Moens, S.S. (Ed.), Proceedings of the ACL-04 Workshop on Text Summarization Branches Out. Association for Computational Linguistics, Barcelona, Spain, pp. 74–81.

Lins, R.D., Simske, S.J., Cabral, L.S., Silva, G.F., Lima, R., Mello, R.F., Favaro, L., 2012. A multi-tool scheme for summarizing textual documents. In: Proceedings of the Eleventh IADIS International Conference, WWW/INTERNET 2012, pp. 1–8.

Lloret, E., Palomar, M., 2012. Text summarisation in progress: a literature review. Artif. Intell. Rev. 37 (1), 1–41. doi: 10.1007/s10462-011-9216-z.

Luhn, H.P., 1958. The automatic creation of literature abstracts. IBM J. Res. Dev. 2 (2), 159–165. doi: 10.1147/rd.22.0159.

Mani, I., Bloedorn, E., Gates, B., 1998. Using cohesion and coherence models for text summarization. In: Proceedings of the 1998 Intelligent Text Summarization Symposium, pp. 69–76.

Manning, C.D., Raghavan, P., Schütze, H., 2008. Introduction to Information Retrieval. Cambridge University Press, New York, NY, USA.

Nenkova, A., 2006. Understanding the Process of Multi-document Summarization: Content Selection, Rewriting and Evaluation. Columbia University, New York, NY, USA Ph.D. thesis. AAI3203761

Nenkova, A., Maskey, S., Liu, Y., 2011. Automatic summarization. In: Proceedings of the Forty-Ninth Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts of ACL 2011. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 3:1–3:86.

Nick, L., 2003. Classifier4j. http://classifier4j.sourceforge.net/. Last accessed 16 June 2016.

Orăsan, C., 2004. The influence of personal pronouns for automatic summarisation of scientific articles. In: Proceedings of the Fifth Discourse Anaphora and Anaphor Resolution Colloquium. Furnas, Portugal, pp. 127–132.

Orăsan, C., 2009. The influence of pronominal anaphora resolution on term-based summarisation. In: Nicolov, N., Angelova, G., Mitkov, R. (Eds.), Recent Advances in Natural Language Processing V. Current Issues in Linguistic Theory. 309, John Benjamins, Amsterdam & Philadelphia, pp. 291–300.

Ouyang, Y., Li, W., Lu, Q., Zhang, R., 2010. A study on position information in document summarization. In: Proceedings of the Twenty-Third International Conference on Computational Linguistics: Posters. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 919–927.

Over, P., Dang, H., Harman, D., 2007. DUC in context. Inf. Process. Manag. 43 (6), 1506–1520. doi: 10.1016/j.ipm.2007.01.019.

Pradhan, S., Ramshaw, L., Marcus, M., Palmer, M., Weischedel, R., Xue, N., 2011. CoNLL-2011 shared task: Modeling unrestricted coreference in ontonotes. In: Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 1–27.

Radev, D.R., Hovy, E., McKeown, K., 2002. Introduction to the special issue on summarization. Comput. Linguist. 28 (4), 399–408.

Rennes, E., Jonsson, A., 2014. The impact of cohesion errors in extraction based summaries. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). European Language Resources Association (ELRA), Reykjavik, Iceland, p. 1.

Sampson, G., 2002. Briefly noted — english for the computer: the SUSANNE corpus and analytic scheme. Comput. Linguist. 28 (1), 102–103. doi: 10.1162/089120102317341800.

Shapiro, S.S., Wilk, M.B., 1965. An analysis of variance test for normality (complete samples). Biometrika 52 (3/4), 591–611.

Silveira, S.M.S.B., 2015. Enhancing Extractive Summarization with Automatic Post-processing, Ph.D. thesis. University of Lisboa, Alameda da Universidade, 1649–004 Lisboa, Portugal.

Smith, C., Henrik, D., Arne, J., 2012. A more cohesive summarizer.. In: Proceedings of the 2012 COLING: Posters, pp. 1161–1170.

Steinberger, J., Ježek, K., 2009. Text Summarization: An Old Challenge and New Approaches. Springer, Berlin, Heidelberg, pp. 127–149. doi: 10.1007/978-3-642-01091-0_6.

Steinberger, J., Poesio, M., Kabadjov, M.A., Ježek, K., 2007. Two uses of anaphora resolution in summarization. Inf. Process. Manag. 43 (6), 1663–1680, Text Summarization. doi: 10.1016/j.ipm.2007.01.010.