Nóbrega and Pardo *Journal of the Brazilian Computer Society* (2018) 24:11
https://doi.org/10.1186/s13173-018-0075-1

Journal of the
Brazilian Computer Society

RESEARCH Open Access

Update summarization: building from scratch for Portuguese and comparing to English

Fernando Antônio Asevedo Nóbrega* and Thiago Alexandre Salgueiro Pardo

Abstract

Update summarization aims at automatically producing a summary for a collection of texts for a reader that has already read some previous texts about the subject of interest. It is a challenging task, since it not only brings the demands from the summarization area (as producing informative, coherent, and cohesive summaries) but also includes the issue of finding relevant new/updated content. In this paper, we report a comprehensive investigation of update summarization methods for the Portuguese language, for which there are few initiatives. We also propose new methods that combine some summarization strategies and enrich a traditional method with linguistic knowledge (subtopics), producing better results and advancing the state of the art. More than this, we present a reference dataset for Portuguese, so far inexistent, and establish an experiment setup in the area in order to foster future research. To confirm some of our summarization results, we run experiments in a well-known benchmark dataset for English language and show that our methods still do well.

Keywords: Update summarization, Subtopics, Evaluation

Introduction

Text summarization aims at producing a summary from one or more related (source) texts/documents. In this research field, the more recent and specific task of update summarization (US) focuses on the production of a summary under the assumption that the reader has some previous knowledge about the subject of the texts. It is useful when the reader has already read some material and is looking for new and relevant information about some fact or event, being appropriate to handle the current web environment with the incredible amount of data and new content that is very quickly produced in many different sources.

As illustrations, Figs. 1 and 2 show a regular (generic) summary and an update summary with no more than 100 words, respectively (in Portuguese, which is the original language). The summaries were produced from the same text collection, and, in the case of the update summary, it was supposed that the reader had read another text collection on the same subject. The source material is about a change of positions in the National Agency of Civil Aviation in Brazil (ANAC). One may see that they are different, as they serve to different purposes.

There are several challenges that the US task must face. As it usually happens in the area, it must produce informative, coherent, and cohesive summaries, dealing with the multi-document phenomena (as the occurrence of redundant, contradictory, and complementary information in the texts) and temporally ordering the events and facts, among many others. The area also brings new challenges, as modeling the user previous knowledge (which is generally represented by the collection of texts that was previously read by the user) and finding relevant new information to compose the summary.

The US task was introduced in an evaluation track at the Document Understand Conference (DUC[1]) in 2007 and was present in some editions of the Text Analysis Conferences (TAC[2], in a new incarnation of DUC conferences). In DUC 2007, each test set had three text collections, named A, B, and C, which were sorted by their respective timestamps. A summary with no more than 100 tokens

*Correspondence: fasevedo@icmc.usp.br
University of São Paulo, Av. Trabalhador São-carlense, 400, 13566-590 São Carlos, Brazil

O ministro da Defesa, Nelson Jobim, deve encaminhar o nome da economista Solange Vieira para assumir uma das diretorias da Agência Nacional de Aviação Civil (Anac). Ainda não está definida a diretoria que a economista vai assumir. O ministro da Defesa, Nelson Jobim, informou no fim da noite desta terça-feira que a economista Solange Vieira, de 38 anos, será a nova presidente da Agência Nacional de Aviação Civil (Anac). - A Solange vai ser a nova presidente da Anac - disse Jobim, em jantar que celebrou os 50 anos da Rede RBS em Brasília. Mas, diante da dificuldade para encontrar

Fig. 1 An example of a regular (generic) summary

(whitespace tokenized) should be produced for each one of them, considering that, for a collection *i*, it was assumed that the reader knew the previous ones [1]. For instance, it was assumed that the reader had read the A and B collections (with the "old" texts) when automatically producing a summary for the collection C (with the "new" texts). The only exception was for the summarization of the collection A, in which the produced summary should not be an update summary (as the reader had no previous knowledge). In the more recent TAC conferences, the task was simplified and only two text collections were used in each test set.

Distinct approaches have been proposed in the area, as methods based on positional features, content ranking, graphs, and topic models. Most of them are for the English language. For Portuguese, there are many investigations on the traditional summarization tasks (see, e.g., [2–14]), but, to the best of our knowledge, there is only one previous effort in the US field, which is a preliminary work conducted by the authors of this paper [15].

In this paper, we report our efforts on building the US area from scratch for the Portuguese language. We start by detailing our comprehensive investigation of some of the main US methods (from distinct approaches) for Portuguese, adapting and evaluating them. We then propose two new methods. One is an alternative version of one of the best methods in the literature, introducing linguistic knowledge (subtopics), which outperforms the original method. The other one is a combined method that merges the summarization strategies of some other methods, achieving the best results and advancing the state of the art. To conclude so, we assemble a reference dataset (inexistent, so far) and establish an experimental environment in order to evaluate the methods, which we also report here. Finally, to confirm our results, we run experiments for English, using a reference dataset in the area.

This paper is organized in the following sections: we introduce the basic concepts in the summarization area and describe the main related work in the "Basic concepts in summarization" and "Related work" sections; our extended and new methods are presented in the "The new methods" section; we introduce the datasets that were used in this paper in the "Datasets" section; the experimental setup and the evaluation results are reported in the "Experiments and results" section; and we present some conclusions and final remarks in the "Final remarks" section.

Basic concepts in summarization

Overall, text summarization follows a generic three-step process, as synthesized in [16]: analysis, transformation, and synthesis. The analysis step is in charge of interpreting the source texts to summarize, producing a internal "computational" content representation. The transformation step is where the main summarization operations usually happen: it performs content selection to produce an internal representation of the summary, most of the times adopting some kind of sentence representation and some sentence ranking function to select the relevant sentences to compose the summary. Finally, in the synthesis step, the output summary is produced from the summary internal representation, observing the specified compression rate, which indicates the size of the summary (usually in number of words). Producing extractive or abstractive summaries presents different demands for each step. The summarization area has traditionally put more efforts on the content selection phase, focusing on producing extractive summaries.

O ministro da Defesa, Nelson Jobim, informou no fim da noite desta terça-feira que a economista Solange Vieira, de 38 anos, será a nova presidente da Agência Nacional de Aviação Civil (Anac). O relatório final da CPI do Apagão da Câmara, que começou a ser lido nesta terça-feira, recomenda o ingresso da iniciativa privada na administração da infra-estrutura dos aeroportos, hoje sob o comando de uma estatal, a Infraero. Ele disse que não está convencido da "participação objetiva" de Zuanazzi nas denúncias contra a agência: - Não podemos indiciar para agradar à oposição, ao governo ou a quem quer que seja

Fig. 2 An example of an update summary

The area started with the single document summarization task and evolved to more sophisticated initiatives, as multi-document summarization, including dealing with e-mails, scientific articles, dialogs, speech, and several other media, as discussed in [17]. The update summarization appeared as a type of the multi-document approach and gained a lot of attention due to its usefulness in the current information overload situation that we face nowadays.

As discussed in [16, 18], to do summarization, or, in more general terms, to produce any natural language processing (NLP) application, requires dealing with linguistic knowledge of varied levels, including, e.g., processing words and their morphology, syntax, semantics, and discourse. Discourse, in particular, has traditionally been investigated for Portuguese summarization (see, e.g., [13, 14, 19, 20]), including discourse relations (as predicted by the rhetorical structure theory—RST [21] and cross-document structure theory—CST [22]) and topics/subtopics. Subtopics are of special interest in this work and are introduced in more details in "The new methods" section when we present one of the methods that we test.

In what follows, we briefly describe the main related work in the US area.

Related work

Researchers have been proposing distinct approaches to produce update summaries. Below, we will present the most representative methods from the different approaches, from the simplest to the more complex ones, and their advantages and disadvantages.

Reeve and Han, and Varma et al. [23, 24] propose methods that rank the source sentences based on lexical features to identify clues of updated content. Reeve and Han [23] assume that a good summary must have a word distribution similar to its source texts, and it also shows that the frequencies of words in the old texts may be used to estimate how much outdated the sentences in the new texts are. Varma et al. [24] propose the Novelty-Factor method, which scores the sentences based on the vocabulary differences among old and new texts using the following equation:

$$\mathrm{NF}(s) = \frac{1}{|s|} \sum_{w \in s} \frac{|w \in D_{\mathrm{new}}|}{|w \in D_{\mathrm{old}}| + |D_{\mathrm{new}}|}$$

where s is a sentence, w is a word, and $D_\bullet$ is a collection of $\bullet$ (new or old) documents. As we may see, a sentence s receives a high score when its words occur more times in the texts of a new collection than in an old one.

Nóbrega et al., Katragadda et al., and Ouyang et al. [10, 25, 26] use positional features, and their results show that this kind of data is better to find salient information than updated information. Katragadda et al. [25] produce summaries based on the optimal position policy (OPP) rank, which estimates how much relevant a sentence is by its respective position in the text. The authors built the OPP rank by the analysis of the distribution of elementary discourse units (EDUs), as defined in the pyramid evaluation method [27], for each sentence position in the DUC 2007 dataset. Once the OPP rank is produced, it may be used as a scoring function for the sentences in order to produce the summaries. As it was expected, the selection of first sentences usually produces better results. The authors have referenced this method as a more robust baseline for update summarization. Nóbrega et al. [10] replicate the experiments with OPP rank for Portuguese language in the CSTNews corpus ([28, 29]). However, the authors use two different information instead of EDUs in order to build the positional rank: the frequencies of words and manually identified sentential alignments among sentences from summaries and their respective source texts in the corpus [30]. It is important to say that word frequency has been used in a lot of summarization researches and it is very useful to find salient information [31]. Nóbrega et al. [10] shows that the use of sentential alignments produces better results than frequencies of words because they result in a more sophisticated way to identify the content from source texts that was selected to the summary. Ouyang et al. [26] show experiments with many positional features of sentences and words, which are based on the idea that the most relevant content occurs first in texts. Thus, their ranking functions decrease the score of a sentence or a word according to their distance to the respective first instance (sentence or word). It is an interesting method because it assumes relevance for first occurrences of words, which may be in other parts of the texts, and it is not limited to the first sentences. They have presented four different positional functions, where n is the number of sentences in the document and i is the position of the sentence that will be ranked, as follows:

- Direct proportion: $f(i) = (n - i + 1)/n$;
- Inverse proportion: $f(i) = 1/i$;
- Geometric sequence: $f(i) = (1/2)^{i-1}$;
- Binary function[3]: $f(i) = 1$ *if* $i == 1$ *else* λ.

The methods above are simple and fast methods to rank sentences, but they adopt oversimplified text representations that do not identify the information flow among old and new texts in order to find updated content. Reeve and Han, and Varma et al. [23, 24] analyzed this information, but in a superficial way.

Steinberger and Ježek [32] propose a method based on the differences among LSA (latent semantic analysis) [33] topics from old and new texts. Each topic is scored by the subtraction of its weight in old and new texts. Thus, a topic gets a high score if it is more relevant in new texts than others. Iteratively, the best weighted sentence

from the topic with the highest score is selected to the summary, and the weights are recalculated.

Huang and He, and Li et al. [34, 35] associate labels (four and three labels, respectively) for LDA (latent Dirichlet allocation) topics based on their weights in the old and new texts. As an example, [34] defines the following topics: emergent (topics present only in new texts); active (topics present on both collections, but more relevant in new texts); not active (topics more relevant in old texts); and extinct (topics present only in old texts). These methods use different features in order to select the sentences for the summary. Huang and He [34] use word frequencies and [35] apply the maximal marginal relevance (MMR) [36] approach, which assumes that a good sentence must be similar to a target and dissimilar to another one, as the new and old texts, respectively. Both first select the sentences related to the topics with higher weights in the new texts.

Delort and Alfonseca [37] show a method based on probabilistic topic models, called DualSum. Each text in this approach is represented by a bag of words, and each word is associated with a latent topic similar to the LDA model. DualSum, which has a procedure similar to the TopicSum system [38], learns a distribution of topics that are organized into the following categories: general topic, which works as a language model in order to identify irrelevant information; topics for collections A and B, in which they represent the subjects that are more present in the old and new texts, respectively; and document specific topics. After this learning step, DualSum finds an output update summary with topics closest to a target distribution, which is based on the intuition that a good summary may be more similar to its respective texts in the collection B. At this point, it is also important to comment on how DualSum and also other summarization methods compare distributions. One of the most used metrics for comparing distributions is the Kullback-Leibler (KL) divergence (see., e.g., [7, 38–41]), which is usually referred by KLSum strategy when applied as a sentence ranking function in summarization. We introduce in more details such strategy in the next section, as we have extended it for our tests on update summarization.

Methods based on graph models have been widely investigated in automatic summarization (see, e.g., [8, 42–46]). To the best of our knowledge, in the context of US, the most expressive results were reached by the positive and negative reinforcement (PNR^2) system [47]. PNR^2 uses a graph for text modeling, in which each node indicates a sentence and each edge between two sentences is weighted by their Cosine similarity [48]. In PNR^2, given a graph that represents a text collection, its procedure runs an optimization algorithm in which the sentences share scores among themselves based on their similarities with positive and negative reinforcements. A positive reinforcement occurs only among sentences from the same text set, and it is represented by a positive β parameter in the algorithm. On the other hand, a negative relation occurs among sentences from different sets and it is indicated by a negative α parameter. This way, a sentence receives a more positive score if it is more similar to sentences from new texts. In the experiments reported in [47], PNR^2 outperforms the PageRank [49] algorithm, which the authors have also experimented for the US task.

Some other recent initiatives tried to use integer linear programming to combine relevant summarization features and to properly deal with redundancy treatment in the summarization process (see, e.g., [50, 51]), producing competitive results in the area. There are also some attempts to use US for specific situations, as to follow the news about human tragedies and disasters [52]. This kind of application seems a natural way to follow in the area.

All the previous efforts focused on the English language. To the best of our knowledge, the only previous work for Portuguese is our preliminary effort reported on [15], where we have tested some US methods. This paper builds upon this previous initiative by reporting new summarization strategies and their cross-lingual evaluation, which we start detailing in the next section.

The new methods

Besides the methods that we briefly described in the previous section, we have also tested two more methods, which we introduce in what follows.

An enriched version of KLSum: introducing subtopics

Hearst and Koch [53, 54] define a textual topic as the main subject or theme in a text, and this topic may be divided into minor portions, its subtopics, which contribute to the main topic. Therefore, the subtopics in a text are the components of its main subject[4].

A subtopic may be expressed by a coherent textual portion with one or more sentences in a row in a text. Thus, we may handle the identification of subtopics as a text segmentation task, in which each identified segment is a subtopic. For instance, Table 1 shows a text from the CSTNews corpus [29] with its subtopics separated by horizontal lines. In this example, we may see a text about an airplane crash segmented into three subtopics: sentences from 1 to 5; sentence 6; and sentence 7. The first subtopic is about the accident itself, while the others present more details about the airplane and its crew, respectively.

To automatically segment texts into their subtopics, several approaches were proposed in the literature (see, e.g., [53, 55, 56]). Of special interest to us is the TextTiling algorithm [53]. Basically, TextTiling analyzes each sentence pair (following the reading flow) in order to identify significant vocabulary changes that may indicate

Table 1 Example of text segmented into subtopics

[S1]	A plane crash in Bukavu, in the Eastern Democratic Republic of Congo, killed 17 people on Thursday afternoon, said the spokesman of the United Nations.
[S2]	The victims of the accident were 14 passengers and three crew members.
[S3]	Everyone died when the plane, hampered by the bad weather, failed to reach the runway and crashed in a forest that was 15 kilometers from the airport in Bukavu.
[S4]	The plane exploded and caught fire, said the UN spokesman in Kinshasa, Jean-Tobias Okala.
[S5]	"There were no survivors", said Okala.
[S6]	The spokesman said the plane, a Soviet Antonov-28, of Ukrainian manufacturing and under ownership of the Trasept Congo, a Congolese company, also took a mineral load.
[S7]	According to airport sources, the crew members were Russian.

subtopic boundaries. It has a good performance and is among the most used ones in the area. Such strategy was recently adapted for the Portuguese language, as reported by [57, 58], also performing well. Some other strategies for this language do exist, as the one that correlates discourse structure (following the RST model) with subtopic changes in a text [59], but they are more expensive and of less general application than the previous one.

Since subtopics have recently shown to be very useful in summarization (see, e.g., [13, 14]), producing better results, we have opted to explore this specific linguistic knowledge for US. We have included subtopics into the KLSum strategy, which is used by several summarization systems, as already commented in the previous section. For clear reference, our subtopic-enriched version of KLSum is named KLSum-Sub.

The systems based on the KLSum approach learn a target distribution T of content unities from the text collections, aiming to produce summaries with distributions S that are closer to T, using the KL divergence. The most common application of KLSum is based on n-gram distributions, in which, for each n-gram w in a vocabulary V, we may define $pT(w)$ and $pS(w)$ as the probabilities of the n-gram w to occur in the text collection ($f(w, \text{text collection})/|V|$) and in a set of sentences ($f(w,S)/|S|$), respectively, where $f(w,\bullet)$ is the frequency of w in $\bullet$; and $|\bullet|$ is the number of words in $\bullet$.

Once the distribution T is learned, the KL formulation may be used in order to select a subset of sentences that minimizes the divergence between the distributions S and T, as we may see in the equation below, where τ is a smoothing factor that is frequently used in order to avoid undefined values.

$$S* = \underset{S}{\text{argmin}}\, KL(T,S) = \sum_{w \in V} pT(w)\, log \frac{pT(w)}{pS(w) + \tau} \quad (1)$$

To avoid the computational time to produce the summary that minimize the KL divergence, the summarization methods based on KLSum frequently use a greedy algorithm, in which the systems interactively pick the sentence that produces the summary closest to the learned distribution of words (see, e.g., [7, 37, 38, 41]).

For the KLSum-Sub, we have assumed that the subtopics in a text collection show the proportion of different ideas that occur in it. Thus, the KLSum-Sub approach aims to produce summaries with the sentences that better represent the different ideas that also occur in the text collection. To do so, we have changed the KLSum formulation in order to analyze the distribution of words over the subtopics in a text collection. Thus, for each word $w \in V$, we have defined $pT(w)$ as the probability of w to occur in the subtopics of a text collection, as below, where c_j is a subtopic, and Sub is the set of all subtopics in the text collection:

$$pT(w) = \frac{1}{|Sub|} \sum_{c_j inSub} 1\ if\ w \in c_j\ else\ 0 \quad (2)$$

During our experiments, for subtopic segmentation, we have adopted the previously cited TextTiling algorithm [53], as it is available for both English and Portuguese languages and shows good results.

A combined method

As we may see in the related work section, there are many and distinct approaches for US, in which varied processes and information are used in order to identify the most relevant content that must be included in a summary. Under the assumption that these variations may contain different clues to the relevance of sentences, we may take advantage of a combined method that takes into consideration the answers of the corresponding methods to determine which sentences to select to the summary. For that, we simply sum up the resulting (normalized[5]) sentence scores produced by all the methods (excepting DualSum), and the highest scored sentences are selected to compose the summary.

We did not include DualSum in the combined method because it is a more complex and expensive method that already (indirectly) incorporates many of the relevance clues of the other methods. More than this, we are interested on testing the power of the combination of such clues for US.

Datasets

We performed experiments over two distinct datasets for the US task, the CSTNews-Update and the TAC 2009 datasets. The first one, the CSTNews-Update, is a distinct arrangement of the CSTNews corpus [28, 29], which has been used for many investigations on Automatic Summarization for Portuguese (see, e.g., [2–14]). We propose

such arrangement here to be a reference corpus to train and test US methods for Portuguese. The TAC 2009 dataset was used in one of the US tracks of the Text Analysis Conference, containing collections with news texts in English. It is widely used in the area, being considered a benchmark. We use it in this paper to confirm the main results that we achieve.

We detail each of the datasets in the following subsections.

The dataset for Portuguese: CSTNews-Update

CSTNews-Update is a different arrangement of the CSTNews corpus [28, 29], which has 140 news texts organized in 50 text clusters. Each cluster has two or three related texts that were collected from mainstream news agencies in Brazil, being labeled into one of the following categories: daily news, world news, sports, economy, politics, and science. The texts span a time period from August to September 2007. In general, CSTNews counts with 2088 sentences (15 sentences per text, in average) and 47,240 words (337 sentences per text, in average).

In a similar way to the datasets for US that were used in the TAC conferences, each cluster in CSTNews-Update has two collections, A (old) and B (new), that are also chronologically sorted. The idea is to produce an update summary from the second collection under the assumption that the reader has already read the texts in the first one.

We have defined 59 distinct clusters for the CSTNews-Update by using two different strategies: intra-cluster and inter-cluster approaches. For each one of them, we have followed the directions of those used in the datasets of DUC and TAC.

In the intra-cluster approach, we have picked all the clusters with three texts (in a total of 40) from CSTNews and then, for each cluster, we have labeled the oldest text as the collection A and the others as the collection B. In the inter-cluster procedure, we manually identified pairs of different clusters from CSTNews with similar subjects (in a total of 19), in which, for each resulting set, the cluster with the oldest texts was considered the collection A and the other one as the collection B. Here, it is important to notice that the same cluster in CSTNews may be used in the two approaches. For instance, a cluster with three texts (that is valid to be used in the intra-cluster approach) may also be paired with another cluster and labeled as collection A or B in the inter-cluster approach. This strategy for rearranging the clusters of CSTNews to build the CSTNews-Update corpus is schematically shown in Fig. 3.

All the resulting clusters have two or more texts in the B collections. This way, all update summaries produced using this dataset are multi-document, therefore. Overall, CSTNews-Update has 3320 sentences, 49,449 words, and 225 texts (95 that were labeled as A and 130 as B texts).

An interesting feature of CSTNews-Update is its different timestamp distances (from seconds to days) among the texts in the collections A and B for each cluster. This feature may model cases of real world, in which the users may read sequential texts that have low timestamp differences and also read others that have huge differences.

As expected, the timestamp differences among documents are low in the intra-cluster collections and huge in the inter-cluster ones. The maximal difference is approximately 216 h and the average difference is 175.51 h. Thus, CSTNews-Update enables investigations about the impact of the published time of documents to find updated and new information. However, it is expected that, in the clusters with higher timestamp distances among its collections A and B, the identification of the most relevant updated content is harder because there are probably more different information among the texts.

The dataset for English: TAC 2009

The TAC 2009 dataset has 44 clusters. Each one of them is also organized into A and B collections and sorted by timestamps.

For each collection, A and B, there are 10 related news texts. Furthermore, for each one of them, there are four respective human summaries. For the B collection, the summaries are update summaries, as it is assumed that the user has already read the texts in A.

As pointed in the TAC webpage, the texts in the corpus come from the AQUAINT-2 collection of news articles, which is a subset of the LDC English Gigaword Third Edition[6] and comprises approximately 2.5 GB of text (about 907K documents) spanning the time period from October 2004 to March 2006. The articles come from a variety of sources, including *Agence France Presse*, Central News Agency (Taiwan), Xinhua News Agency, Los Angeles Times-Washington Post News Service, New York Times, and the Associated Press.

Experiments and results

In order to evaluate the US methods, we have applied the ROUGE framework [60], which is the most used evaluation approach in automatic summarization. ROUGE computes the number of n-grams in common among automatic and reference texts. Usually, in NLP area, a reference text is the "ideal" output that a system should produce. In summarization, it is common to use manually produced summaries as reference texts, to which the automatic summaries are compared in order to be evaluated. Such comparisons result in Precision, Recall and F-measure figures, being indicative of the informativeness of the automatic summaries: the closer to 1 the results are, the more informative the summaries are. Precision indicates the proportion of relevant n-grams in the automatic summary; recall indicates the proportion of relevant

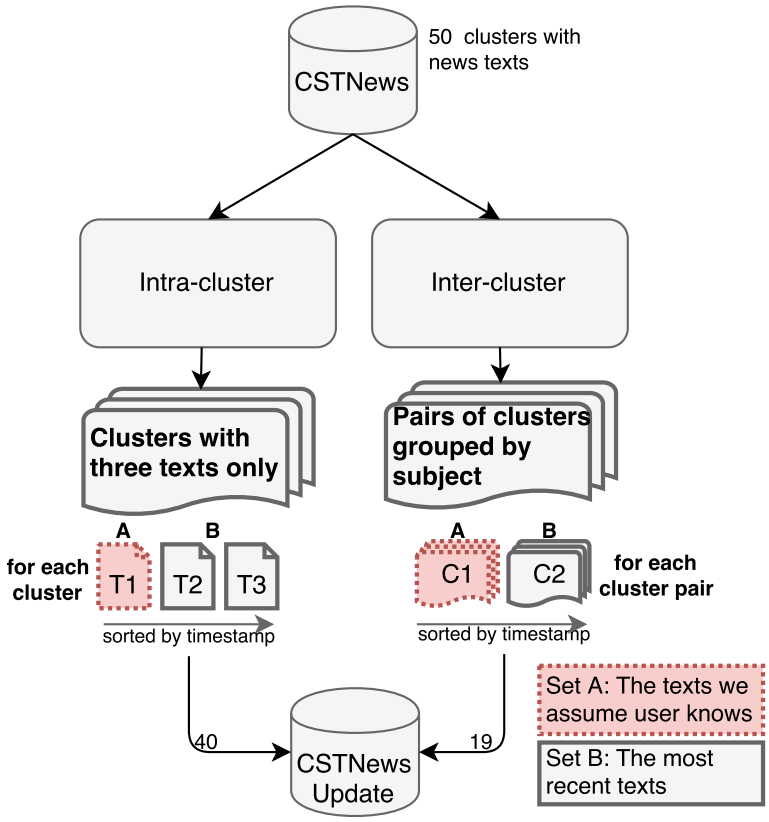

Fig. 3 Strategy for organizing CSTNews clusters (C) and their texts (T) for building the CSTNews-Update corpus

n-grams in the automatic summary in relation to the reference texts; and f-measure is a unique performance metric that combines the results of Precision and Recall. Relevant n-grams are those that occur in the reference texts. We show below the equations for these metrics. When two or more reference texts are used, a Jackknifing procedure is applied.

$$\text{Precision} = \frac{\text{number of relevant } n-\text{grams in the automatic summary}}{\text{number of } n-\text{grams in the automatic summary}} \tag{3}$$

$$\text{Recall} = \frac{\text{number of relevant } n-\text{grams in the automatic summary}}{\text{number of } n-\text{grams in the reference text}} \tag{4}$$

$$F-\text{measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5}$$

In addition to ROUGE, we have also employed the Nouveau-ROUGE method [61], which is a different application of ROUGE with focus on the US task. This metric assumes that a good update summary must be informative and updated. Thus, initially, Nouveau-ROUGE computes two ROUGE scores, $R^{(AB)}$ and $R^{(BB)}$, in which there are reference texts either from collection A or from collection B, respectively. After that, weights are used to combine $R^{(AB)}$ and $R^{(BB)}$ in order to approximate the results to those produced by the manual summarization evaluation approaches of Pyramid [27] and Responsiveness, resulting in correlated Precision, Recall and F-measure figures for both of them. In Pyramid evaluation, automatic summaries are scored by their content units (which usually are manually identified concepts represented by the occurring n-grams), which are weighted by the number of their occurrences in the reference texts. The more a content unit occurs, the better its associated weight is. Consequently, automatic summaries with better weighted content units are desired, since they probably are more informative summaries. Responsiveness, in turn, directly evaluates the "usefulness" of the automatic summaries, considering both their content and their linguistic quality. Pyramid and responsiveness evaluations are very relevant measures in the summarization area, but, as they require manual analysis of the texts, they usually are complementary performance indicators for the ROUGE values.

One final relevant issue about ROUGE and its related measures is the way that the achieved results must be

interpreted. As its author argues, ROUGE is good at comparing automatic summaries, being almost as good as humans in ranking different summaries according to their informativeness. Therefore, taken in isolation, a single ROUGE value is not a direct indication of the summary quality. ROUGE is a comparative measure and must be read and interpreted in this way. More than this, it is important to remember that the summarization task is usually cruel for evaluation, as there is not a unique good summary as reference. Instead, there are many possible summaries for a collection of texts, and this may influence ROUGE results.

For the English evaluation, we used the reference update summaries provided in the TAC 2009 dataset, as it is usually done in the area. For Portuguese, however, as there are not update summaries made by humans in the CSTNews dataset, we applied the automatic evaluation approach that was proposed in [62], in which the produced summaries are compared to their respective source texts. Louis and Nenkova [62] have shown that the ROUGE evaluation of summaries based on their source texts is a good approximation of evaluations based on human summaries. Therefore, for ROUGE, we compared the produced summaries to their respective source texts that are labeled as collection B.

We show the average scores for two of ROUGE running settings, ROUGE-1 and ROUGE-2, which calculate the scores based on unigrams and bigrams overlapping (that are the most used variations), respectively. Furthermore, we applied the same ROUGE parameters[7] that were applied in the TAC conferences.

Once Nouveau-ROUGE is focused on the US task and it requires two collections of reference texts (texts in collections A and B), we used the respective old and new texts for each produced summary. Here, we only report the *F*-measure score and its respective correlation with Responsiveness (Res) and Pyramid (Pyr) results [27].

Following what was observed at the DUC and TAC conferences, we produced update summaries based on the extractive approach, in which the systems pick some sentences from the source texts and put them in the output without content changes. Furthermore, all the produced summaries have no more than 100 words, and it was assumed that the reader had already read the old texts in each document set. As we focus in the US task, we present the average evaluation scores for update summaries only, and we do not consider the regular summaries created for collection A.

Besides our two new methods, we performed experiments with the most representative methods of the distinct summarization approaches that were investigated in the US task, as follows: DualSum [37], which uses a probabilistic topic model; the graph-based algorithms PNR^2 [47] and its variations with distinct setups of the PageRank algorithm [49]; the ranking functions based on positional features that were proposed by [26]; and the Novelty-Factor [24]. For Portuguese, we have also performed experiments with the RSumm system [8], which is among the best systems for Portuguese for general multi-document summarization, not being tailored for the US task. The purpose of this comparison was to show how (in)adequate such general systems are for the US task.

In our experiments with DualSum, we have used the same setups that were adopted in [37]. Thus, we applied the same preprocessing steps, but changed the resources and tools that are language depended. In the topic learning stage, we use the CSTNews-Update dataset in order to identify the general topics, once [37] has also applied the experimented dataset itself. Here, it is important to say that [37] has proposed that this kind of topic may be previously learned in order to reduce the required computational processing time.

We investigate the PageRank [49] algorithm in two different setups that were also used in [47], the PageRank (A + B) and PageRank (B). In the first one, we use the sentences from the A and B collections in order to build a sentential graph, in which each node is a sentence and each edge indicates the Cosine similarity [48] between two sentences. The second setup has a procedure identical to the first one, but we build the graph with sentences from the collection B only. It is important to notice that, independently of the approach, only sentences from Collection B are used to build the summary. We have also used these two setups in the RSumm method [8], once it is also based on graph algorithms. However, RSumm was not affected by this, producing the same evaluation scores.

For all the above methods, except for DualSum and KLSum-Sub that internally handle content redundancy, we have used the procedure for redundancy removal that was defined in [8]. For each produced summary, we firstly define as threshold t the average Cosine [48] among all the sentences in the collection B[8] and, after that, we ignore those sentences that have similarity above the threshold with any other sentence that has already been included in the output summary. This strategy of varying threshold (that depends on the test set) is interesting because it may better handle different sceneries. For instance, in collections with very similar texts, the used threshold is higher than in contexts in which there are very distinct texts.

Table 2 shows the obtained average evaluation scores for Portuguese. We sort the methods by their *F*-measure scores for ROUGE-2. We organize the methods in the rows and each evaluation score in the columns. For instance, one may see that our combined method shows 0.426 and 0.329 *F*-measure values for ROUGE-1 and ROUGE-2, respectively.

As we may see, the differences among the scores of some methods are not so high. It probably occurs because some

Table 2 ROUGE and Nouveau-ROUGE scores in the CSTNews-Update dataset based on the evaluation approach of [62]

	ROUGE-1			ROUGE-2			Nouveau-ROUGE	
Methods	Precision	Recall	*F*-measure	Precision	Recall	*F*-measure	Pyr	Res
Combined method	0.843	0.296	0.426	0.648	0.230	0.329	7.224	1.264
KLSum-Sub	0.825	0.297	0.421	0.640	0.233	0.329	6.744	2.271
DualSum	0.823	0.294	0.418	0.636	0.230	0.325	7.694	2.438
Novelty-Factor	0.837	0.285	0.414	0.645	0.220	0.319	6.942	2.424
Pos-Direct	0.831	0.289	0.418	0.632	0.222	0.319	6.639	2.240
PageRank(A+B)	0.817	0.288	0.413	0.620	0.221	0.317	6.410	2.118
PageRank(B)	0.822	0.284	0.411	0.627	0.218	0.314	6.642	2.260
Pos-Binary	0.797	0.277	0.400	0.603	0.211	0.304	6.370	2.136
Pos-Geometric	0.796	0.277	0.400	0.602	0.211	0.304	6.389	2.145
Pos-Inverse	0.795	0.277	0.400	0.601	0.211	0.304	6.383	2.141
PNR^2	0.785	0.271	0.392	0.595	0.207	0.299	6.239	2.087
RSumm	0.714	0.247	0.356	0.444	0.149	0.217	2.049	0.586

The methods are sorted by the *F*-measure scores of ROUGE-2

test cases in our dataset have short texts, and we have used a fixed summary length, which is the same used in the DUC and TAC conferences. In general, however, we see that some methods outperform others.

Usually, the methods that have more procedures to identify more recent and relevant content present better results, as DualSum, PNR^2, and Novelty-Factor approaches. It was not expected that the Position-Direct method would outperform many others. However, positional features have been used in many summarization investigations and have presented satisfactory results, being relevant features.

As expected, the RSumm system [8] method produced the worst results, as it is not focused on US, demonstrating that efforts are needed to tackle the summarization task specificities. It is interesting that the other graph methods have presented better results, as PageRank (A+B) and PageRank (B). This probably happens because they were tailored for the US task.

Overall, one may see that our new methods—the combined method and KLSum-Sub - were the methods that produced the most informative summaries, outperforming the literature methods that we investigated in this paper. Our combined approach was the best one, showing that simple features may be very useful to the task. It is interesting that KLSum-Sub was better than DualSum, showing that the use of subtopics does result in positive impact, as some previous tentatives have showed. DualSum was the third one in the evaluation, and this is not totally surprising, as this approach has constantly achieved good results in the area.

Although our dataset is small (which is a usual situation for several NLP tasks for Portuguese) and the results of statistical tests might not be reliable, we have run traditional paired t tests (over ROUGE-two f-measure values). The results showed that the performance differences of our combined method and KLSum-Sub are not statistically significant; however, both of them showed statistical difference to the next one in the rank—DualSum. Our statistical tests also showed that the difference in RSumm performance is also significant in relation to the update methods.

In our experiments, we have also observed that the average ROUGE and Nouveau-ROUGE scores for the investigated methods in the intra-cluster setup is a bit higher than in inter-cluster collections. As we said before, the timestamp differences among the texts that occur in the inter-cluster collections are higher than in the intra-cluster ones, and our results suggest that identifying the most relevant updated content is harder in these situations.

Table 3 shows the achieved results for the TAC 2009 dataset, for the English language. It is noticeable that there is great variation in relation to the results for the dataset in Portuguese. This is usual in the area, as the evaluation is very sensitive to the corpus. Nonetheless, our combined method is still the best method and KLSum-Sub performs better than DualSum, confirming our main results for Portuguese.

Directly comparing results across different experiments is unfair, as the evaluation results are very sensitive to the test data and the experiment setup. However, to give an idea to the interested reader of how the state of the art systems perform for the English language, we cite the results of two other well known papers in the area. In [37], the authors report that the best configuration of DualSum system achieved a ROUGE-2 recall value of 0.092 on a corpus with partial overlap with the one of TAC 2009. In [51], the

Table 3 Results for TAC 2009 dataset

	ROUGE-1			ROUGE-2			Nouveau-ROUGE	
Methods	Precision	Recall	*F*-measure	Precision	Recall	*F*-measure	Pyr	Res
Combined method	0.344	0.356	0.349	0.079	0.082	0.080	2.174	0.232
Novelty-Factor	0.325	0.336	0.330	0.077	0.080	0.078	2.248	0.248
Pos-Binary	0.341	0.353	0.347	0.076	0.079	0.077	2.117	0.221
Pos-Direct	0.341	0.353	0.347	0.076	0.079	0.077	2.117	0.221
Pos-Geometric	0.341	0.353	0.347	0.076	0.079	0.077	2.117	0.221
Pos-Inverse	0.341	0.353	0.347	0.076	0.079	0.077	2.117	0.221
KLSum-Sub	0.328	0.339	0.333	0.076	0.078	0.077	2.086	0.213
PageRank(B)	0.320	0.333	0.326	0.074	0.077	0.075	2.182	0.235
DualSum	0.312	0.323	0.317	0.073	0.076	0.074	2.078	0.210
PageRank(A+B)	0.315	0.328	0.322	0.068	0.071	0.070	2.058	0.209
PNR[2]	0.223	0.228	0.225	0.020	0.021	0.021	1.262	0.054

authors got a higher value of 0.106 in their best system configuration for the TAC 2009 dataset. This shows that, even considering the state of the art, there is still room for improvements in the US area.

Final remarks

We introduced an experimental setup and a reference dataset, the CSTNews-Update[9], that allow the investigation of Update Summarization methods for the Portuguese language. Based on them, we have evaluated the most representative methods of Update Summarization from different approaches, and have also introduced two new methods: a subtopic-enriched version of KLSum and a combined method that takes advantage of other methods. The evaluation scores show that some performance differences are small, but that some methods outperform others, indicating future research directions for US investigation for Portuguese. In particular, our combined method achieved the best results in the evaluation and our variation of KLSum was better than state of the art systems. Finally, we have also performed experiments for English, which confirmed our main results, but shows that there is still room for improvements.

Particularly, we believe that the US methods might significantly benefit from some more intelligent/linguistically motivated text representation for the collections of previously known/old/read and new texts, e.g., making use of some kind of semantic representation. More informed methods do exist for summarization, but this is practically an unexplored field for the specific US task.

Two recent advances in the NLP area might be useful for US. In one side, word embeddings might be tested, replacing words in the summarization related computations. Although this is more expensive, it might allow to the methods to incorporate semantics and produce more significant results. More interestingly, explicit semantics might be used, as the recently widely spread Abstract Meaning Representation (AMR) graphs [63], to represent the content of both previously read and new texts, allowing to produce update summaries for the content that is unique to the new texts. Currently, there are large word embedding repositories and AMR-based semantic parsers for both English and Portuguese languages. The existent semantic parsers produce results that are still far from the ideal, but they are fastly evolving and may achieve acceptable performance soon, being then ready for use in US.

Endnotes

[1] at http://duc.nist.gov/duc2007/tasks.html

[2] More details about the TAC conferences may be found at http://tac.nist.gov. TAC has been held annually since 2008, and the last edition with a summarization track was in 2014.

[3] [26] has suggested the use of a small positive real number for λ. We have used $\lambda = 0$.

[4] It is important to notice that the terms "topic" and "subtopic" that were used here are different of those used in topic models. To avoid confusion, whenever we use the term "subtopic", we will be referring to the definition above of [53, 54].

[5] Normalization is necessary because each method produces sentence scores in different scales. Inside each method, we have normalized its scores for the sentences in a standard way, i.e., dividing each score by the sum of all the scores, making the scores fall in the 0-1 interval.

[6] LDC catalog number LDC2007T07.

[7] -n 4 -w 1.2 -2 4 -u -c 95 -r 1000 -f A -p 0.5 -t 0 -a

[8] Remember that only sentences from this collection may be added in the summary.

[9] https://github.com/fernandoasevedo/CSTNews-Update

Abbreviations

DUC: Document understand conference; EDUs: Elementary discourse units; KL: Kullback-Leibler; LDA: Latent Dirichlet allocation; LSA: Latent semantic analysis; MMR: Maximal marginal relevance; NF: Novelty-Factor; OPP: Optimal position policy; PNR: Positive and negative reinforcement; TAC: Text analysis conference; US: Update summarization

Acknowledgements

The authors are grateful to CAPES and FAPESP for supporting this work.

Funding

This project was supported by CAPES and FAPESP.

Availability of data and materials

Most of the computational linguistic tools and resources that we have used in our experiments are available at http://www.icmc.usp.br/~taspardo/sucinto.

Authors' contributions

FN participated in the design, carried out the experiments, and drafted the manuscript. TP participated in the design and coordination and helped to draft the manuscript. Both authors read and approved the final manuscript.

Authors' information

FN is a PhD in Computer Science. He has worked with natural language processing during his Master's and PhD researches under the supervision of Prof. Thiago Pardo. TP is a PhD in Computer Science and a professor at ICMC-USP. He has worked with natural language processing on many tasks, as automatic summarization, opinion mining, discourse parsing, and others.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Received: 26 March 2018 Accepted: 6 August 2018
Published online: 21 September 2018

References

1. Witte R, Krestel R, Bergler S (2007) Generating update summaries for DUC 2007. In: Proceedings of the Document Understanding Conference (DUC). NIST, Rochester. p 5
2. Pardo TAS (2002) DMSumm: Um gerador automático de sumários, Master's thesis. Universidade Federal de São Carlos, São Carlos
3. Rino LHM, Pardo TAS, Jr CNS, Kaestner CAA, Pombo1 M (2004) A comparison of automatic summarization systems for Brazilian Portuguese texts. In: Proceedings of the 17th Brazilian Symposium on Artificial Intelligence (BRACIS). Springer, Sao Luis. pp 235–244
4. Muller E, Granatyr J, Lessing OR (2007) Comparativo entre o algoritmo de luhn e o algoritmo gistsumm para sumarização de documentos. Revista de Informática Teórica e Aplicada 22(1):584–599. (75–94)
5. Leite DS, Rino LHM, Pardo TAS, das Gracas V, Nunes M Extractive automatic summarization: Does more linguistic knowledge make a difference? In: Proceedings of the HLT/NAACL Workshop on TextGraphs-2: Graph-Based Algorithms for Natural Language Processing. ACL, Rochester. pp 17–24
6. Antiqueira L, Oliveira Jr ON, Costa LdF, Nunes MdGV (2009) A complex network approach to text summarization. Inf Sci 179(5):584–599
7. Castro Jorge ML, Pardo TAS (2011) A generative approach for multi-document summarization using the noisy channel model. In: Proceedings of the 3rd RST Brazilian Meeting. Sociedade Brasileira de Computação, Cuiabá/MT. pp 75–87
8. Ribaldo R, Akabane AT, Rino LHM, Pardo TAS (2012) Graph–based methods for multi-document summarization: exploring relationship maps, complex networks and discourse information. In: Proceedings of the 10th International Conference on Computational Processing of Portuguese (PROPOR). Springer, Coimbra. pp 260–271
9. Silveira S, Branco AH (2012) Enhancing multi-document summaries with sentence simplification. In: Proceedings of the 14th International Conference on Artificial Intelligence. IEEE, Las Vegas. pp 742–748
10. Nóbrega FAA, Agostini V, Camargo RT, Di Felippo A, Pardo TAS (2014) Alignment-based sentence position policy in a news corpus for multi-document summarization. In: Proceedings of the 11st International Conference on Computational Processing of Portuguese (PROPOR). Springer, São Carlos. pp 6–9
11. Cardoso P, Pardo TAS (2015) Joint semantic discourse models for automatic multi-document summarization. In: Proceedings of the 10th Brazilian Symposium in Information and Human Language Technology (STIL). Sociedade Brasileira de Computação, Natal. pp 81–90
12. Ângelo Abrantes Costa M, Martins B (2015) Uma comparação sistemática de diferentes abordagens para a sumarização automática extrativa de textos em português. Linguamática 7(1):23–40
13. Cardoso P, Pardo TAS (2016) Multi-document summarization using semantic discourse models. Processamiento de Lenguaje Nat 56: 57–64
14. Ribaldo R, Cardoso PF, Pardo TAS (2016) Exploring the subtopic-based relationship map strategy for multi-document summarization. J Theor Appl Comput – RITA 23(1):183–211
15. Nóbrega FAA, Pardo TAS (2017) Update summarization for portuguese. In: Proceedings of the 6th Brazilian Conference on Intelligent Systems (BRACIS). IEEE, Uberlândia. pp 348–353
16. Mani I (2001) Automatic Summarization vol. 3. John Benjamins Publishing Company, Amsterdam/Philadelphia
17. Nenkova A, McKeown K (2011) Automatic summarization. now Publishers Inc, Hanover/Massachusetts
18. Jurafsky D, Martin JH (2009) Speech and Language Processing: an Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition 2nd edn. Prentice Hall. Pearson, Upper Saddle River/New Jersey
19. Castro Jorge ML, Pardo TAS (2010) Experiments with cst-based multidocument summarization. In: Proc TextGraphs-5 - Workshop on Graph-based Methods for Natural Language Processing. Association for Computational Linguistics, Uppsala. pp 74–82
20. Castro Jorge MLR, Dias MS, Pardo TAS (2014) Building a language model for local coherence in multi-document summaries using a discourse-enriched entity-based model. In: Proceedings of the Brazilian Conference on Intelligent Systems – BRACIS. Springer, São Carlos. pp 44–49
21. Mann WC, Thompson SA (1987) Rhetorical structure theory: a theory of text organization, Technical report. Univerisity of Southern California, Information Science Institute, Los Angeles
22. Radev DR (2000) A common theory of information fusion from multiple text sources step one: cross-document structure. In: Proceedings of 1st ACL SIGDIAL Workshop on Discourse and Dialogue. ACL, Hong Kong. p 10
23. Reeve LH, Han H (2007) A term frequency distribution approach for the DUC-2007 update task. In: Proceedings of Document Understanding Conference (DUC). NIST, Rochester. p 7
24. Varma V, Bharat V, Kovelamudi S, Bysani P, GSK S, N KK, Kumar KRK, Maganti N (2009) IIIT hyderabad at TAC 2009. In: Proceedings of the Second Text Analysis Conference (TAC). NIST, Gaithersburg. pp 1–15
25. Katragadda R, Pingali P, Varma V (2009) Sentence position revisited: a robust light-weight update summarization baseline algorithm. In: Proceedings of the Third International Workshop on Cross Lingual Information Access: Addressing the Information Need of Multilingual Societies (CLIAWS). ACL, Boulder. pp 46–52
26. Ouyang Y, Li W, Lu Q, Zhang R (2010) A study on position information in document summarization. In: Proceedings of the 23rd International Conference on Computational Linguistics (Coling) – Posters. ACL, Beijing. pp 919–927
27. Nenkova A, Passonneau R (2004) Evaluating content selection in summarization: the pyramid method. In: Proceedings of the Human Language Technology and Conference of the North American Chapter of

the Association for Computational Linguistics (HLT-NAACL). ACL, Boston. pp 145–152
28. Aleixo P, Pardo TAS (2008) *CSTNews: Um Córpus de Textos Jornalísticos Anotados segundo a Teoria Discursiva Multidocumento CST (Cross-document Structure Theory)*. Technical Report, 326
29. Cardoso PCF, Maziero EG, Castro Jorge MLR, Seno EMR, Di Felippo A, Rino LHM, Nunes MdGV, Pardo TAS (2011) CSTNews – a discourse-annotated corpus for single and multi-document summarization of news texts in brazilian portuguese. In: Anais do III Workshop "A RST e Os Estudos do Texto". Sociedade Brasileira de Computação, Cuiabá. pp 88–105
30. Agostini V, López Condori RE, Pardo TAS (2014) Automatic alignment of news texts and their multi-document summaries: comparison among methods. In: Proceedings of the 11st International Conference on Computational Processing of Portuguese (PROPOR). Springer, São Carlos. pp 220–231
31. Nenkova A, Vanderwende L (2005) The impact of frequency on summarization. Technical report, Microsoft Research
32. Steinberger J, Ježek K (2009) Update summarization based on latent semantic analysis. In: Matoušek V, Mautner P (eds). Text, Speech and Dialogue. Springer, Pilsen Vol. 5729. pp 77–84
33. Landauer TK, Dutnais ST (1997) A solution to plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychol Rev 104:211–240
34. Huang L, He Y (2010) Lect Notes Comput Sci. In: Huang D-S, Zhang X, Reyes García C, Zhang L. (eds). Advanced Intelligent Computing Theories and Applications with Aspects of Artificial Intelligence. Springer, Changsha Vol. 6216. pp 641–648
35. Li J, Li S, Wang X, Tian Y, Chang B (2012) Update summarization using a multi-level hierarchical dirichlet process model. In: Proceedings of the 24th International Conference on Computational Linguistics (Coling). ACL, Mumbai. pp 1603–1618
36. Carbonell J, Goldstein J (1998) The use of MMR, diversity-based reranking for reordering documents and producing summaries. In: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, Melbourne. pp 335–336
37. Delort JY, Alfonseca E (2012) DualSum: a topic-model based approach for update summarization. In: Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL). ACL, Avignon. pp 214–223
38. Haghighi A, Vanderwende L (2009) Exploring content models for multi-document summarization. In: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL). ACL, Boulder. pp 362–370
39. Daumé III H, Marcu D (2006) Bayesian query-focused summarization. In: Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL. Association for Computational Linguistics, Sydney. pp 305–312
40. Lerman K, McDonald R (2009) Contrastive summarization: an experiment with consumer reviews. In: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL) – Short Papers. ACL, Boulder. pp 113–116
41. Wang D, Zhu S, Li T, Gong Y (2009) Multi-document summarization using sentence-based topic models. In: Proceedings of the ACL-IJCNLP 2009 Conference – Short Papers. ACL, Suntec. pp 297–300
42. Erkan G, Radev DR (2004) Lexrank: Graph-based lexical centrality as salience in text summarization. J Artif Intell Res 22(1):457–479
43. Leskovec J, Milic-Frayling N, Grobelnik M (2005) Extracting summary sentences based on the document semantic graph. Technical report, Microsoft Research
44. Lin Z, Kan MY (2007) Timestamped graphs: evolutionary models of text for multi-documentsummarization. In: Proceedings of the Second Workshop on TextGraphs: Graph-Based Algorithms for Natural Language Processing. ACL, Rochester. pp 25–32
45. Li X, Du L, Shen YD (2011) Graph-based marginal ranking for update summarization. In: Proceedings of the 2011 SIAM International Conference on Data Mining. SIAM, Mesa. pp 486–497
46. Li X, Du L, Shen YD (2013) Update summarization via graph-based sentence ranking. IEEE Trans Knowl Data Eng 25(5):1162–1174
47. Wenjie L, Furu W, Qin L, Yanxiang H (2008) PNR^2: ranking sentences with positive and negative reinforcement for query-oriented update summarization. In: Proceedings of the 22nd International Conference on Computational Linguistics (Coling). ACL, Manchester. pp 489–496
48. Salton G, Wong A, Yang CS (1975) A vector space model for automatic indexing. Commun ACM 18(11):613–620
49. Brin S, Page L (1998) The anatomy of a large-scale hypertextual web search engine. In: Proceedings of 17th International World-Wide Web Conference (WWW). ACM, Beijing. p 20
50. Li C, Liu Y, Zhao L (2015) Improving update summarization via supervised ilp and sentence reranking. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, Denver. pp 1317–1322
51. Mnasri M, de Chalendar G, Ferret O (2017) Taking into account inter-sentence similarity for update summarization. In: Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers). Asian Federation of Natural Language Processing, Taipei. pp 204–209
52. Kedzie C, McKeown K, Diaz F (2015) Predicting salient updates for disaster summarization. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics, Beijing. pp 1608–1617
53. Hearst MA (1997) Texttiling: segmenting text into multi-paragraph subtopic passages. Comput Linguist 23(1):33–64
54. Koch IGV (2009) Introdução à Linguística textual. Contexto, São Paulo
55. Choi FYY (2000) Advances in domain independent linear text segmentation. In: 1st Meeting of the North American Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, Seattle. pp 26–33
56. Riedl M, Biemann C (2012) TopicTiling: a text segmentation algorithm based on LDA. In: Proceedings of ACL Student Research Workshop. Association for Computational Linguistics, Jeju Island. pp 37–42
57. Cardoso PCF, Taboada M, Pardo TAS (2013) Subtopic annotation in a corpus of news texts – steps towards automatic subtopic segmentation. In: Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology (STIL), Cuiabá. pp 49–58
58. Cardoso PCF, Pardo TAS, Taboada M (2017) Subtopic annotation and automatic segmentation for news texts in brazilian portuguese. Corpora 12(1):23–54
59. Cardoso PCF, Taboada M, Pardo TAS (2013) On the contribution of discourse structure to topic segmentation. In: Proceedings of the 14th Annual SIGdial Meeting on Discourse and Dialogue. Association for Computational Linguistics, Metz. pp 92–96
60. Lin CY (2004) Rouge: A package for automatic evaluation of summaries. In: Text Summarization Branches Out: Proceedings of the ACL-04 Workshop. ACL, Barcelona. pp 74–81
61. Conroy JM, Schlesinger JD, O'Leary DP (2011) Squibs: Nouveau-ROUGE: A novelty metric for update summarization. Comput Linguis 37(1):1–9
62. Louis A, Nenkova A (2009) Automatically evaluating content selection in summarization without human models. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP). ACL, Singapore. pp 306–314
63. Banarescu L, Bonial C, Cai S, Georgescu M, Griffitt K, Hermjakob U, Knight K, Koehn P, Palmer M, Schneider N (2013) Abstract meaning representation for sembanking. In: Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse. Association for Computational Linguistics, Sofia. pp 178–186