Exploring Events and Distributed Representations of Text in Multi-Document Summarization

Luís Marujo[a,c,d], Wang Ling[a,c,d], Ricardo Ribeiro[a,b], Anatole Gershman[d], Jaime Carbonell[d], David Martins de Matos[a,c], João P. Neto[a,c]

[a] *INESC-ID Lisboa, Rua Alves Redol, 9, 1000-029 Lisboa, Portugal*
[b] *ISCTE-Instituto Universitário de Lisboa, Av. das Forças Armadas, 1649-026 Lisboa, Portugal*
[c] *Instituto Superior Técnico, Universidade de Lisboa, Av. Rovisco Pais, 1, 1049-001 Lisboa Portugal*
[d] *Language Technologies Institute, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA, 15213-3891, USA*

Abstract

In this article, we explore an event detection framework to improve multi-document summarization. Our approach is based on a two-stage single-document method that extracts a collection of key phrases, which are then used in a centrality-as-relevance passage retrieval model. We explore how to adapt this single-document method for multi-document summarization methods that are able to use event information. The event detection method is based on Fuzzy Fingerprint, which is a supervised method trained on documents with annotated event tags. To cope with the possible usage of different terms to describe the same event, we explore distributed representations of text in the form of word embeddings, which contributed to improve the summarization results. The proposed summarization methods are based on the hierarchical combination of single-document summaries. The automatic evaluation and human study performed show that these methods improve upon current state-of-the-art multi-document summarization systems on two mainstream evaluation datasets, DUC 2007 and TAC 2009. We show a relative improvement in ROUGE-1 scores of 16% for TAC 2009 and of 17% for DUC 2007.

Keywords: Multi-document summarization, Extractive summarization, Event detection, Distributed representations of text

1. Introduction

Many automatic summarization systems have been proposed in order to cope with the growing number of news stories published online. The main goal of these systems is to convey the important ideas in these stories, by eliminating less crucial and redundant pieces of information. In particular, most of the work in summarization has been focused on the news domain, which is strongly tied to events, as each news article generally describes an event or a series of events. However, few attempts have focused on the use of automatic techniques for event classification for summarization systems for the news domain [1]. In fact, most of the work on multi-document summarization are either based on Centrality-based [2, 3, 4, 5], Maximal Marginal Relevance (MMR) [6, 7, 8, 9], and Coverage-base methods [10, 11, 12, 13, 1, 14, 15]. Generally, centrality-based models are used to generate generic summaries, the MMR family generates query-oriented ones, and coverage-based models produce summaries driven by topics or events.

The use of event information in multi-document summarization can be arranged in the following categories: initial **hand-based experiments** [16]; **pattern-based approaches** based on enriched representations of sentences, such as the cases of the work presented by Zhang et al. [15] and by Wenjie Li et al. [13], which define events using an event key term and a set of related entities, or centrality-based approaches working over an event-driven representation of the input [1], where events are also pattern-based defined; and, **clustering-based** event definition [17].

The major problem of these approaches is that is difficult to relate different descriptions of the same event due to different lexical realizations. In our work, we address this problem by using an event classification-based approach and including event information supported by two different distributed representations of text—the skip-ngram and continuous bag-of-words models [18]. Our event detection and classification framework is based on vector-valued fuzzy sets [19, 20]. We evaluate our work using the standard summarization evaluation metric, ROUGE [21]. Moreover, to better understand the impact of

using event information, we also perform a human evaluation using the Amazon Mechanical Turk[1].

Our main goal in this work was to produce event-based multi-document summaries that are informative and could be useful for humans. The human evaluation shows that our summaries are on average more useful for humans than the reference summaries. While we conducted our experiments in the news domain, our methods are also applicable to other domains, such as opinion and meta-review summarization in consumer reviews [22].

In this document, the next section describes the related work to contextualize the findings obtained in the experimental results. Section 3.2 introduces the Event Detection framework; which is enhanced by the Continuous Skip-gram Model presented in Section 3.3; both are included in a Event-based Multi-Document Summarization framework (Section 3). The experimental results are included and discussed in Section 4. Section 5 details the conclusions and discusses future research directions.

2. Related Work

An early attempt at event-based multi-document summarization, proposed by [16], manually annotated events and showed that events are an useful cue for summarization systems. However, manually extracting events is undesirable as if hampers the automation of summarization systems.

Most of the work in automatic summarization concentrates on extractive summarization. In fact, extracting the important content is the first step of a generic summarization system. The extracted information can subsequently be further processed if the goal is to generate *abstracts*. For this case, the important content is generally devised as a set of concepts that are synthesized to form a smaller set and then used to generate a new, concise, and informative text. The alternative goal can also be to generate *extracts* where the identified content

[1] https://www.mturk.com/

consists of sentences that are concatenated to form a summary.

The most popular multi-document summarization baselines follow into one of the following general models: Centrality-based [2, 3, 4], Maximal Marginal Relevance (MMR) [6, 7, 8, 9], and Coverage-base methods [10, 12, 13, 14, 15, 23, 11, 24, 1].

Traditionally, *Centrality-based* models are used to produce generic summaries, the *MMR* family generates query-oriented ones, and *Coverage-base* models produce summaries driven by topics or events.

The most popular centrality-based method is the centroid [2] for multi-document summarization distributed in the MEAD framework. Expected n-call@k [7, 8, 9] adapted and extended MMR with new similarity and ranking methods.

Concerning the idea of using event information to improve summarization, previous work [12, 13, 14, 15, 1] defines events as triplets composed by a named entity, a verb or action noun, and another named entity, where the verb/action noun defines a relation between the two named entities. This information is then included in a generic unit selection model, often trying to minimize redundancy while maximizing the score of the important content. Others have tried to use time information and word overload to summarize the same events [25, 26]

In our work, we use, not only event information, but also their classification according to ACE [27]; we additionally explore the possibility of using events to filter out unimportant content; and, to our best of our knowledge, we present the first analysis of the impact of using this type of information on multi-document summarization.

Over the past years, the research community has been exploring event detection. The bulk of the event detection work started in the end of 1990s with the Topic Detection and Tracking (TDT) effort [28, 29, 30, 31]. The TDT project had two primary tasks: First Story Detection or New Event Detection (NED), and Event Tracking. The objective of the NED task was to discover documents that discuss breaking news articles from a news stream. In the other task, Event Tracking, the focus was on the tracking of articles describing the

same event or topic over a period of time. More recent work using the TDT datasets [32, 33, 34] on Event Threading tried to organize news articles about armed clashes into a sequence of events, but still assumed that each article described a single event. Passage Threading [33] extends the event threading work by relaxing the one-event-per-news-article assumption. For this purpose, it uses a binary classifier to identify "violent" events in paragraphs.

Even though the TDT project ended in 2004, new event detection research continued. The most well-known example is Automatic Content Extraction (ACE. The goal of ACE research is to detect and recognize events in text. Beyond the identification of events, the ACE 2005 [27] task identifies participants, relations, and attributes of each event. This extraction is an important step towards the overarching goal of building a knowledge base of events [35]. More recent research [36] explores bootstrapping techniques and cross-document techniques augmenting the ACE 2005 with other corpora, including MUC-6 (Message Understanding Conference).

The idea of augmenting the ACE 2005 corpus stems from the low occurrence of some event types in the sentences of the dataset. Most sentences do not contain any event or describe an event that does not exist in the list of event types, which makes the identification of events a complex task. Additional features combined with supervised classifier [37], such as SVM, improved the identification of events. But a more simple and efficient approach based on Fuzzy Logic outperformed the best results. For this reason, we are using it in this work.

As discussed above, events are hard to detect. However, the identification of anomalous events makes the task simpler [38]. Still, determining if two events are the same or are related is, as noted by Hovy et al. [39], an unsolved problem. Even event co-reference evaluation is not a trivial problem [40].

While word embeddings have been used in many NLP tasks [41, 42], they have not been used in event detection or summarization to the best of our knowledge. The closest work found is a summarization work that trains a neural network to learn the weights for a small set of features.

Even considering that clustering-based event definition approaches could handle this type of problem, the work of Li et al. [17] models events in a similar way of topics.

3. Event-based Multi-Document Summarization

Our multi-document summarization approach is based on a single-document centrality summarization method, KP-Centrality [43] (Figure 1). This method is easily adaptable [44] and has been shown to be robust in the presence of noisy input. This is an important feature, since the multiple documents given as input in multi-document summarization are more likely to contain unimportant information compared to single-document summarization.

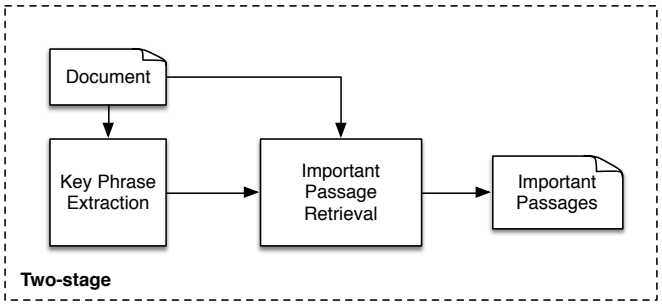

Figure 1: Two-stage single-document architecture.

3.1. From Single-Document to Multi-Document Summarization

Our goal is to extend the KP-Centrality method for multi-document summarization. The simplest method would be to concatenate all documents and use the single-document method to produce the summary. We shall use this approach as a baseline. This baseline works quite well for a small number of documents, but the performance decreases as the number of documents increases. This means that KP-Centrality has limitations identifying redundant content, such as events, when it is written with different words. Another

limitation of the baseline method is to ignore temporal information as more recent news documents tend to contain more relevant information and sometimes include brief references to the past events to provide some context.

To overcome the first limitation, we consider two simple but effective alternative approaches for improving the baseline method. The first approach is a two-step method where we summarize each document individually in such a way that each of the summaries have the size of the final multi-document summary. This is followed by the concatenation of all the resulting summaries, which is then summarized again into the final summary. In both steps, we use the KP-Centrality method to generate the summaries. The advantage of this approach is to reduce the redundancy of information at document level (intra-document). This means that we also need to reduce the redundancy of information between document (inter-documents). The second method we propose is similar reduces the redundancy inter-documents. Rather than considering all summaries simultaneously, we take one summary s_1, concatenate with another summary s_2, summarize the result to obtain a summary of documents s_1 and s_2, which we denote as $s_{1...2}$. Next, we take $s_{1...2}$ and perform the same operation with s_3, obtaining $s_{1...3}$. This is done recursively for all the N documents in the from the input, and the final summary is the one obtained in $s_{1...N}$.

We will denote these methods as hierarchical single-layer and waterfall. These are illustrated in Figures 2 and 3, respectively.

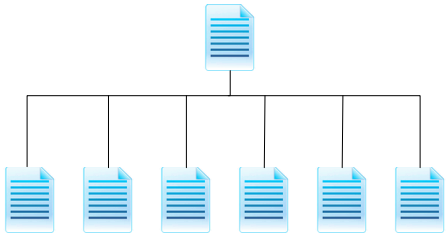

Figure 2: Single-layer architecture.

The waterfall method is sensitive to the order of the input documents. Since

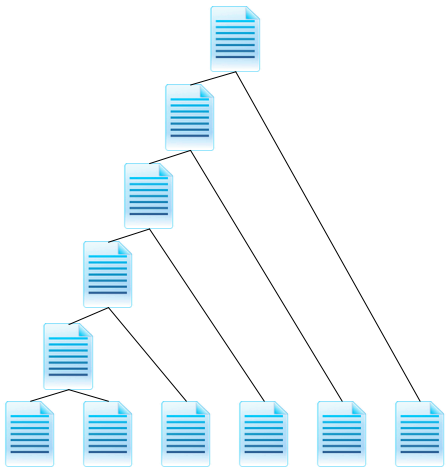

Figure 3: Waterfall architecture.

at each iteration the summaries of the documents are merged with the summary of the previous documents, the content of the initial documents is more likely to be removed than the content in the last documents. Thus, it is important to consider the order of the documents. We chose to organize the documents chronologically where the older documents are summarized and merged in the first iteration of the waterfall method. The waterfall method has two drawbacks. One limitation is the size of the intermediate summaries. Once we decided the size of the final summary, we obtain the intermediate summaries with the size of the final summary. In practice, this work well, but in some cases the size of the intermediate summary is not enough to contain all necessary information for the summarization process. From this limitation also emerges the second, which is the identification of redundant content between documents when written with different words.

Our solution to the first limitation of the waterfall method is as we merge more documents recursively, the intermediate summaries that contains the information of the documents so far, will grow in size to avoid losing important information. For that reason, we increased the number of sentences in the intermediate summary as a function of the number of documents that have been

covered. More formally, the size of the summary at a given time or document t is defined as:

$$L = \delta \times K \times log(t + \phi) \quad (1)$$

where K is the maximum number of words in the final summary, ϕ is a constant to avoid zeros ($\phi = 2$). δ is a scale factor that is 1 for the generation of the initial documents summaries and 200 for the remaining cases. Since the more recent documents contain more important content, we also increased the size of initial documents summaries created by the hierarchical single-layer based on Eq. 1 to not give an unfair advantage to the waterfall method.

The identification of redundant sentences written in different ways is not an easy task. For instance, the sentence "The Starbucks coffee co. plan to acquire Pasqua coffee is leaving a bitter aftertaste in the mouths of some patrons of the San Francisco-based coffeehouse." and "Starbucks , the nation 's largest coffee retailer , announced Tuesday that it would buy Pasqua for an undisclosed amount." have essentially the same meaning: a company plans to buy another. Nevertheless, the only common content between the two sentences are the company names. For this purpose, we propose two alternatives that complement each other. On the one hand, news documents describe events (e.g., Company acquisitions), thus sentences that cover the same event are good candidates to contain redundant information. On the other hand, different lexical realizations with the same meaning can be addressed using distributed word representations.

From this point, we present the two extensions to our multi-document summarization framework.

3.2. Supervised Event Classification

Our event detection method is based on the Fuzzy Fingerprints classification method [20], which is based on the work by Homem and Carvalho's [19]. This work approaches the problem of authorship identification by using the crime scene fingerprint analogy that leverages the fact that different authors have different writing styles. The algorithm is computed as follows: (1) Gather the top-k word frequencies in all known texts of each known author; (2) Build the

fingerprint by applying a fuzzifying function to the top-k list. The fuzzified fingerprint is based on the word order and not on the frequency value; (3) For each document, perform the same computations to obtain a fingerprint and assign the author with the most similar fingerprint.

Our motivation for the use of event information is the existence of secondary events that are not relevant to the main event of the documents, which need to be excluded from the summary. To do this, we use the event fingerprint method to identify sentences that describe events. Since we needed training data to build the event fingerprint of each event type, we used the ACE 2005 Multilingual Corpus [27]. These event fingerprints are used to generate each sentence fingerprint. For example, the fingerprint of the sentence "ETA, whose name stands for Basque Homeland Freedom, has killed nearly 800 people since 1968 in its campaign for Basque independence" considering, for example, only four event types would be the following vector: [Die = 0.1061, Attack = 0.0078, Divorce = 0.0, Null or No-event = 0.01907]. All sentences that the event fingerprint method classified as not containing any event are removed (F.E. - filtering events). The exception to this simple rule occurs when the method is not confident in the classification result (confidence less than 0.0001, obtained when we compute the fingerprint of the sentence). This event filtering is an optional pre-processing step of the multi-document summarization.

After filtering out the sentences that do not describe events, we also need to identify similar events. This is accomplished by using the sentences event fingerprints as features in the summarization process. This means that each sentence has 27 new features, each corresponding to one of the 27 different event types: Appeal, Arrest-Jail, Attack, Be-Born, Charge-Indict, Convict, Declare-Bankrupcy, Demonstrate, Die, Divorce, Elect, End-Org, End-Position, Fine, Injure, Marry, Meet, N (Null/No Event), Phone-Write, Release-Parole, Sentence, Start-Org, Start-Position, Sue, Transfer-Money, Transfer-Ownership, Transport, Trial-Hearing.

Our approach to the extraction of event information does not fall in any of the previously known categories (exploratory hand-based experiments; pattern-

based approaches; and, clustering-based), since it is a supervised classification method.

3.3. Unsupervised Word Vectors

Although the event detection method described above is supervised, where features are extracted from annotated data, we also need to leverage the large amount of raw text (without annotation) in an unsupervised setup. The small size of the annotated data is insufficient to cover also possible ways of describing events. Large amounts of raw text without event annotations are easy to obtain and contain different descriptions about the same event. Thus, we need a method to relate the event descriptions. For this purpose, we use the method recently introduced by Mikolov et al. [18], which uses raw text to build a representation for each word, consisting of a d-dimensional vector. Two models were proposed in this work, the skip-ngram model and the continuous bag-of-words model, which we shall denote as SKIP and CBOW, respectively. While both models optimize their parameters by predicting contextual words, the models differ in terms of architecture and objective function. SKIP iterates through each word w_i at index i, and predicts each of the neighbouring words up to a distance c. More formally, given a document of T words, the model optimizes its parameters by maximizing the log likelihood function:

$$\mathcal{L} = \frac{1}{T} \sum_{t=1}^{T} \sum_{\substack{-c \leq j \leq c, \\ j \neq 0}} \log p(w_{t+j} \mid w_t) \tag{2}$$

where the probability $p(w_{t+j} \mid w_t)$ is the output probability given by the network. The log likelihood function is optimized using gradient descend.

CBOW is similar to SKIP, in the sense that it uses word vectors to predict surrounding words, but predicts each word w_i conditioned on all surrounding words up to a distance of c. That is, we estimate the parameters that maximize the probability $p(w_t \mid w_{t-c}, ..., w_{t+c})$.

To use this information as features in our summarization model, we added to the representation of each sentence a vector consisting in the average of the

vectors representing each word in that sentence. Each word is described by 50-features vector.

We have also experimented using a distributed representation of sentences [45], but the results were worse than averaging word vectors due to overfitting.

4. Experiments

We evaluate our work in two distinct ways: through the automatic estimation of the informativeness, using ROUGE; and through a human study, designed according to two previous reference studies [46, 47], using the Amazon Mechanical Turk.

4.1. Datasets

To empirically analyse the performance of our event-based multi-document summarization methods, we use two standard evaluation datasets: DUC 2007[2] and TAC 2009[3]. However, the set of events types occurring in evaluation datasets only partially overlaps with the events types detected by our event detector. Hence, we created a subset for each of the evaluation datasets. Tables 1 and 2 identify the selected topics.

4.1.1. DUC 2007

The main summarization task in DUC 2007 is the generation of 250-word summaries of 45 clusters of 25 newswire documents and 4 human reference summaries. Each document set has 25 news documents obtained from the AQUAINT corpus [48].

4.1.2. TAC 2009

The TAC 2009 Summarization task has 44 topic clusters. Each topic has 2 sets of 10 news documents obtained from the AQUAINT 2 corpus [49]. There

[2] http://www-nlpir.nist.gov/projects/duc/duc2007/tasks.html
[3] http://www.nist.gov/tac/2009/Summarization/

Table 1: Subset of DUC 2007 topics containing several event types in the ACE 2005 list.

Topic	Description
D0705A	Basque separatism.
D0706B	Burma government change 1988.
D0712C	"Death sentence" on Salman Rushdie.
D0718D	Starbucks Coffee attempted to expand and diversify through joint ventures, acquisitions or subsidiaries.
D0721E	Mathew Sheppard's death.
D0741I	Day trader killing spree.
D0742J	John Kennedy Jr. Dies in plane crash.

are 4 human 100-word reference summaries for each set, where the reference summaries for the first set are query-oriented multi-document summaries, and for the second set are update summaries. In this work, we used the first set of reference summaries.

4.2. Evaluation Setup

To assess the performance of our methods, we compare them against other representative models: namely MEAD, MMR, Expected n-call@k [9], the Portfolio Theory [50], Filatova's event-based summarizer [12] (our implementation), TopicSumm [51], and LexRank [3]. MEAD is a centroid-based method and one of the most popular centrality-based methods. The MMR family is represented by the original MMR, Expected n-call@k [9], and the Portfolio Theory [50]. Expected n-call@k adapts and extends MMR as a probabilistic model (Probabilistic Latent MMR). The Portfolio Theory also extends MMR under the idea of ranking under uncertainty. Filatova's event-based summarizer is a summarization method that also explores event information in a pattern-based way. TopicSum models topics in documents and uses them for content selection, making it close to event-based summarization. LexRank is well-known PageRank-based summarization method often used as baseline. As our baseline method, we used the straightforward idea of combining all input documents

Table 2: Subset of TAC2009 topics containing several event types in the ACE 2005 list.

Topic	Description
D0904A	Widespread activities of white supremacists and the efforts of those opposed to them to prevent violence.
D0910B	Struggle between Tamil rebels and the government of Sri Lanka.
D0912C	Anti-war protest efforts of Cindy Sheehan.
D0914C	Attacks on Egypt's Sinai Penninsula resorts targetting Israeli tourists.
D0915C	Attacks on Iraqi voting stations.
D0922D	US Patriot Act, passed shortly after the September 11, 2001 terrorist attacks.
D0934G	Death of Yassar Arafat.
D0938G	Preparations and planning for World Trade Center Memorial
D0939H	Glendale train crash.
D0943H	Trial for two suspects in Air India bombings.

into a single one and then submit the resulting document to the single-document summarization method.

To evaluate informativeness, we used ROUGE [21], namely ROUGE-1, ROUGE-2, and ROUGE-SU4, a commonly used evaluation measure for this scenario. The ROUGE metrics measure summary quality by counting overlapping units, such as n-gram word sequences, between the candidate summary and the reference summary. ROUGE-N is the n-gram recall measure defined in Equation 3, where N is the length of the n-gram (we use $N = 1$, and $N = 2$), $\text{Count}_{\text{match}}(n\text{-}gram)$ is the maximum number of n-grams co-occurring in a candidate summary and a set of reference summaries, and $\text{Count}(n\text{-}gram)$ is the number of n-grams in the reference summaries.

$$\text{ROUGE-}N = \frac{\sum_{S \in \{\text{RefSums}\}} \sum_{n\text{-}gram \in S} \text{Count}_{\text{match}}(n\text{-}gram)}{\sum_{S \in \{\text{RefSums}\}} \sum_{n\text{-}gram \in S} \text{Count}(n\text{-}gram)} \tag{3}$$

ROUGE-SU4 is similar to ROUGE-N, but allows gaps of at most 4 words apart in matching bigrams.

For the human evaluation, we used the Amazon Mechanical Turk.

We assess the performance of the various models by generating summaries with 250 words.

4.3. Results

The default features of the summarizer models include the bag-of-words model representation of sentences (TF-IDF), the key phrases (80) and the query. The query is obtained from the descriptions of the topics.

Regarding the event-based features, they are obtained from the Event Fuzzy Fingerprint method and consist of scores associated with event fingerprints as described in Sections 3.2 and 3.

To create the CBOW and SKIP models we used New York Times articles covering a 16-year period from January of 1994 to December of 2010, included in the English Gigaword Fifth Edition [52]. Since the results obtained with both models were very similar, we opted to present only the results with the SKIP model.

Internally, the KP-Centrality method uses a distance metric to compute semantic similarity between the sentences. In these experiments, we explored the several metrics presented by Ribeiro and de Matos [5], but only present the results using the Euclidean distance, as it was best-performing one in this context.

In the next sections, we analyze the results of the automatic informativeness evaluation and of the human study. Although we have experimented both the single-layer and waterfall architectures in both datasets, we only present the best performing model for each dataset.

4.3.1. Informativeness Evaluation

Table 3 provides the results on the DUC 2007 dataset using the waterfall summarization model. Our first observation is that our proposed approach, even

without using any event information, filtering or the temporal dilation of the size of the initial and intermediate summaries, achieves better results than the baseline. Note that, although the presented results are for the waterfall architecture, the single-layer approach using all features (event information and filtering in addition to average word embeddings of sentences and temporal dilation) also achieved better results than the baseline (0.3522 ROUGE-1 score). The same does not happen for other summarization models: MEAD and Portfolio achieved better results than the baseline, but Filatova's event-based summarizer, MMR ($\lambda = 0.3$ was the best performing configuration), Expected n-call@k, TopicSum, and LexRank did not.

Another important aspect is that, in the DUC 2007 except the use of event information without event filtering, word embeddings, and temporal dilatation, all our variants improve over not using event information or temporal dilation. After we observed the summaries, we find out that the intermediate summaries were not large enough to keep all important events till the generation of the final summary. At the same time, the sentences describing the same event types were not exactly the same events, but follow up events (which are semantic similar), such as a new strike, or another company acquisition.

The best performing baseline was MEAD and only achieved a performance similar to the default model without event information or the temporal dilation. The best results in the DUC 2007 were obtained when using the average word embeddings of the sentences (SKIP model) combined with the event distribution scores and using event filtering and temporal dilation.

Figure 4 shows an example of a summary produced by our best method on the DUC 2007 dataset and the corresponding reference summary.

Table 3 also presents the obtained results on the TAC 2009 dataset. Note that, in this dataset, our best results were achieved using the single-layer architecture instead of the waterfall architecture. Nonetheless, the best result achieved by the waterfall approach (using all features) was better than our baseline (0.5163 ROUGE-1 score). On the other hand, all other approaches, achieved worse results than the baseline. The results in the TAC 2009 results

Table 3: ROUGE results.

			DUC 2007 (waterfall)			TAC 2009 (single-layer)		
Features	F.E.	T.D.	R1	R2	RSU4	R1	R2	RSU4
default + AWE + events info.	yes	yes	**0.381**	**0.092**	**0.160**	0.523	0.142	0.138
default + AWE + events info.	yes	no	0.353	0.067	0.139	0.530	0.154	0,134
default + AWE + events info.	no	yes	0.361	0.087	0.147	**0.550**	**0.163**	0.140
default + AWE + events info.	no	no	0.352	0.067	0.123	0.508	0.148	0.128
default + events info.	yes	yes	0.372	0.091	0.154	0.533	0.154	0.139
default + events info.	yes	no	0.353	0.075	0.126	0.528	0.149	0.134
default + events info.	no	yes	0.364	0.091	0.155	0.533	0.149	0.138
default + events info.	no	no	0.349	0.072	0.121	0.513	0.155	0.131
default + AWE	yes	yes	0.379	0.090	0.151	0.526	0.144	0.138
default + AWE	yes	no	0.353	0.080	0.130	0.538	0.162	0.134
default + AWE	no	yes	0.367	0.088	0.145	0.540	0.154	**0.143**
default + AWE	no	no	0.351	0.81	0.127	0.522	0.157	0.133
default	yes	yes	0.368	0.090	0.151	0.515	0.138	0.135
default	yes	no	0.352	0.080	0.130	0.523	0.152	0.136
default	no	yes	0.361	0.088	0.144	0.525	0.141	0.135
default	no	no	0.352	0.081	0.127	0.520	0.132	0.129
baseline			0.326	0.051	0.106	0.475	0.128	0.124
MEAD			0.352	0.089	0.150	0.469	0.128	0.128
Portfolio			0.349	0.088	0.142	0.422	0.086	0.095
Filatova's event-based summarizer			0.301	0.046	0.096	0.379	0.049	0.067
MMR			0.299	0.075	0.147	0.370	0.080	0.108
E.n-call@k			0.280	0.065	0.116	0.364	0.066	0.085
TopicSum			0.171	0.009	0.031	0.271	0.007	0.010
LexRank			0.170	0.009	0.031	0.262	0.017	0.030

exhibit the same behavior in term of features and temporal dilation observed in the DUC 2007 dataset: the best results use all features and temporal dilation of the size of the initial and intermediate summaries.

The event filtering consistently lower the results in the TAC 2009. The smaller number of documents to summarize 10 vs. 25 suggest that there is less redundant content in the TAC 2009 than in the DUC 2007. Some of the topics in the TAC 2009 are more complex, in the sense, that there are more relevant events, but with distributed lower relevance of those events making the distinction between primary and secondary events hard even for humans as topic

D0910B exemplifies. Under this conditions, an event classification error have more impact in the final outcome and should be avoided. Our event filtering results were also inline with Filatova's event-based summarizer, which had worse performance than Expected n-call@k and MMR on the TAC 2009.

We have also observed that when the connection between news documents covering a topic is weak, the cascade method performs worse than the singe-layer. This fact also helps to explain the performance differences between the hierarchical methods and datasets.

In order to give a better perspective over the results shown in Table 3, we need to know the ROUGE-1 of the perfect summary. This results corresponds to the optimal selection of important sentences achievable in the evaluation datasets (oracle) and it is shown in Table 4. We also included the results obtained using our best summarizer configuration. These values are obtained by testing all summaries that can be generated and extracting the one with the highest score. The precise calculation of this exponential combination problem is, in the most cases, unfeasible. As result, we restricted the size of the oracle to 3 sentences. The comparison of results of the oracle and our summarizer's show that our best methods are in the 70-80% range of the oracle summaries.

Table 4: Results of maximum ROUGE-1 scores and of our best performing methods.

#Sent.	Corpus	Oracle	Summarizer
1		0.242	0.193
2	TAC 2009	0.410	0.310
3		0.528	0.387
1		0.118	0.090
2	DUC 2007	0.215	0.167
3		0.396	0.229

Another interesting aspect that we observed is related to the representation of dates and numbers when using word embeddings. Since the frequency of this information is low in the used training data, it is not well captured by these

Event-based Summary

Iranian Foreign Minister Kamal Kharrazi, who made the announcement in New York, and his British counterpart, Robin Cook, had portrayed the move as a way to improve ties that have remained strained over the issue and agreed to exchange ambassadors. LONDON _ The British government said Wednesday that it would continue to press Iran to lift the death sentence against the author Salman Rushdie when its foreign secretary, Robin Cook, meets the Iranian foreign minister in New York on Thursday. VIENNA, Austria (AP) – The European Union on Monday welcomed a move by the Iranian government to distance itself from an Islamic edict calling for British author Salman Rushdie's death even as two senior Iranian clerics said the ruling was irrevocable. The move follows the Iranian government's distancing itself last month from bounties offered for the death of Rushdie and a strong reaction by hard-liners who support the killing of the Booker Prize-winning author. He said that Iran will ask the United Nations to effectively put a ban on insulting religious sanctities in a bid to prevent disputes such as the Rushdie affair. On February 14, 1989, late Iranian leader Ayatollah Khomeini issued a religious edict, pronouncing a death sentence on the Indian-born British author Salman Rushdie and his publishers in protest against the publication of Rushdie's novel "The Satanic Verses", which was believed by Moslems as defaming Islam, and exhorting all Moslems to carry out the sentence.

Reference

In 1989, Ayatollah Khomeini of Iran issued a death sentence on British author Salman Rushdie because his book "Satanic Verses" insulted Islamic sanctities. Rushdie was born in India, but his book was banned and his application for a visit was denied. British Airways would not permit Rushdie to fly on its airplanes. Reacting to diplomatic pressures by Britian and other European Nations, Iran announced in 1996 that the death sentence was dropped. President Rafsanjani said there was a difference between a fatwa (ruling) and a hokm (command) and that Khomeini did not mean the sentence to be a command. Despite official retraction of the death sentence, Iranian Islamic fundamentalists continue to demand Rushdie's death. The Khordad Foundation raised the reward for Rushdie's death to 2.5 million dollars and announced, "There is nothing more important to the foundation than seeing Imam Khomeini's decree executed." In 1998, Grand Ayatollah Lankarani and Grand Ayatolla Hamedani said the fatwa must be enforced and no one can reverse it. More than half of Iran's parliament signed a letter saying the death sentence against Rushdie still stands. A hard-line student group offered $333K to anyone who kills Salman Rushdie; residents of a village in northern Iran offered land and carpets to anyone who kills him and thousands of Iranian clerics and students pledged a month's salary toward a bounty. In February 2000, the Islamic Revolutionary Guard said in a radio report that the death sentence was still in force and nothing will change it.

Figure 4: Example of summary produced by our summarizer and the reference summary from the Topic D0712C DUC 2007 - "Death sentence" on Salman Rushdie.

models. The result is that this type of information is not well represented in the summaries generated by our methods, when using word embeddings. For example, in Figure 4, the reference summary contains four date entities and two money entities and in the automatic summary only one date entity appears.

4.3.2. User Study

The initial informativeness evaluation of our multi-document summarization framework was performed using the ROUGE evaluation metric.

The ROUGE metric does not measure how pragmatical the summaries are for humans. To evaluate usefulness, we needed a set of summaries from our event-based summarizer with the corresponding evaluation scores. We also needed a similar set for the baseline system to establish a proper comparison. Obtaining such sets presents both conceptual and practical difficulties. Defining usefulness or relevance of summaries are subjective decisions of each reader that can be influenced by their background.

Our solution was to use multiple judges for the same news story and provide a Likert scale to assign a score to each question. We used a five-level Likert scale, ranging from strongly disagree (1) to strongly agree (5).

We used the Amazon's Mechanical Turk service to recruit and manage our judges. To the best of our knowledge, this has not been done before for this purpose. Each assignment (called HIT) consisted of answering 9 evaluation questions. Evaluating one summary was a HIT and it paid $0.05 if accepted. We selected the reference summaries from each topic of the subsets of the TAC 2009 and DUC 2007 datasets.

We obtained 8 summaries for each topic: one using our event-based summarizer, another using the reference summary, and 7 using the baseline systems. Then, we created 5 HITs, one per judge, for each of the 17 topics. An individual judge could only do one HIT per summary of a topic and summarizer.

The use of the Mechanical Turk created the practical problem of the uneven quality of the judges: some of the judges used bad shortcuts to accomplish a HIT, producing meaningless results. We used several heuristics to weed out bad

HITs. For example, very fast work completion (less that 30 seconds), or missing answers to one or several questions usually indicated a bad HIT. As a result, we were able to keep 99% of HITs.

We created a "Gold Standard" set of 680 annotated summaries. For each summary, we used the 5 questions' quality description developed by Nenkova [46] to assess the linguistic quality of the summaries. In addition, we developed an additional set of questions to evaluate the usefulness of the summaries based on the work of McKeown et al. [47] and we included a question to measure the overall quality of the summary.

To be more precise, each HIT had a description of the task. It indicated that we were conducting a survey about computer-generated summaries. The evaluation was performed without reference to the original texts. We did not distinguish the reference summaries from the automatically generated summaries.

Each HIT contains the following questions:

1. To which degree do you agree with the following information:
   (a) *Background* - Familiarity with the main topic before reading it, that is: "I was familiar with the main topic of the summary before reading it".
2. Please indicate to which degree do you agree that the summary possessed the following qualities:
   (a) *Usefulness* - The summary informs you about the ¡TopicDescription¿ (variable replaced by the description of the topic included in Table 1 and 2)
   (b) *Coherence* - The summary is well-structured and organized. The summary should not just be a heap of related information, but should build from sentence to sentence to a coherent body of information about a topic.
   (c) *Referential clarity* - It should be easy to identify in the summary to whom or what the pronouns and noun phrases are referring to. If a person or other entity is mentioned, it should be clear what their

role in the story is. So, a reference would be unclear if an entity is referenced but its identity or relation to the story remains unclear.

(d) *Non-redundancy* - There should be no unnecessary repetition in the summary. Unnecessary repetition might take the form of whole sentences that are repeated, or repeated facts, or the repeated use of a noun or noun phrase (e.g., "Barack Obama") when a pronoun ("he") would suffice.

(e) *Focus* - The summary should not have extraneous information.

(f) *Context Coverage* - The summary should cover all main events of a story and give a brief context about them.

(g) *Grammaticality* - The summary should have no datelines, system-internal formatting, capitalization errors or obviously ungrammatical sentences (e.g., fragments and missing components) that make the text difficult to read.

(h) *Overall* - What is the overall quality of the summary?

Table 5: DUC 2007 human results.

Question	*Reference*	*MEAD*	*MMR*	*E.ncall*	*Portf.*	*EventSum*	*Filatova et al.*	*TopicSum*	*LexRank*
Background	3.000	2.742	2.926	2.682	3.125	**3.143**	2.765	2.727	3.088
Usefulness	3.966	3.419	3.556	3.500	3.750	**4.000**	3.471	2.970	3.206
Coherence	3.759	2.903	3.519	3.364	3.375	**3.857**	3.618	3.242	2.706
Referential Clarity	**3.966**	3.419	3.482	3.364	3.583	3.821	3.647	2.909	3.118
Non-redundancy	3.655	2.903	3.482	3.136	3.458	**3.857**	3.471	2.970	3.059
Focus	3.828	3.774	3.741	3.682	3.750	**3.929**	3.471	2.849	2.824
Context Coverage	4.034	3.452	3.667	3.455	3.708	**4.107**	3.588	2.879	3.088
Grammaticality	**4.138**	3.710	3.889	3.773	4.000	3.893	3.529	2.909	3.324
Overall	**4.000**	3.226	3.667	3.409	3.583	3.893	3.618	2.879	2.882

Tables 5 and 6 show the average scores obtained in the user study. As we can observe in both tables, the judges rated our event-based multi-document summaries as more useful than reference summaries and the baseline systems.

Table 6: TAC 2009 human results.

Question	*Reference*	*MEAD*	*MMR*	*E.ncall*	*Portf.*	*EventSum*	*Filatova et al.*	*TopicSum*	*LexRank*
Background	2.737	2.925	2.849	2.919	3.000	**3.063**	2.723	2.660	2.646
Usefulness	3.684	3.975	3.697	3.595	3.737	**4.031**	3.660	3.064	3.542
Coherence	**3.790**	3.650	3.667	3.487	3.500	3.781	3.638	3.489	2.938
Referential Clarity	**3.974**	3.875	3.667	3.595	3.395	3.969	3.596	3.149	3.333
Non-redundancy	**4.105**	3.550	3.788	3.324	3.421	3.719	3.809	3.277	3.625
Focus	3.816	**4.075**	3.667	3.838	3.868	4.000	3.660	2.851	3.250
Context Coverage	3.474	3.850	3.636	3.595	3.737	**3.969**	3.809	3.170	3.479
Grammaticality	**4.079**	3.975	3.849	3.865	3.868	4.031	3.830	3.106	3.583
Overall	3.684	3.775	3.697	3.649	3.711	**3.813**	3.809	3.192	3.417

They also reported that they better recognize the topic of the summaries using our summarization method.

In terms of coherence of the summaries, event-based summaries were perceived as more coherent than the references for DUC 2007. While on TAC 2009, the judges judged the coherence of our event-based summaries to be nearly the same. We empirically observed that the waterfall method produces more coherent summaries than the single-layer method, which is explained in part by the fact that most of the extracted sentences belong to few documents (in general, the most recent ones).

The reference summaries clearly outperformed our summaries in the Referential Clarity and Grammaticality categories. These are expected results because the reference summaries do not contain news source names (possibly motivated by the presence in the generated summaries of extracts like "VIENNA, Austria (AP)") and because all pronoun references can be resolved.

The evaluation scores for the Focus category highlight an important difference in the topics of the datasets. While in TAC 2009 most topics describe several equal-importance sub-topics/events spread in time, there is a single main topic center on a date in several topics of DUC 2007. One implication is that

our event-based multi-document summaries does not discard the sub-topics, which penalizes the Focus score in the TAC 2009 dataset when compared to the centroid-based method (MEAD) that selected the sentences for the summary using a single topic (centroid). Another implication is that increasing the focus in a single sub-topic can reduce the Context Coverage. However the results are not conclusive.

Even though the overall results are higher for our event-based multi-document summaries in TAC 2009, we cannot conclude that our method is better than the reference. The reason lies in the smaller size of reference summaries when compared to the remaining summaries (100 vs. 250 words).

Among the event-based and topic-based baselines, the human evaluation clearly shows that the Filatova et al. event-based method performed better than the topic based summarizer (TopicSum). More interesting is the fact that the overall human score of the Filatova et al. event-based were either the best or second best baseline.

In summary, the automatic evaluation of the informativeness results show that the proposed framework achieves better results than previous models. To this contributed, not only the single-document summarization method on which our multi-document approach is based, but also the use of event information and the better representation of text. Note that a simple baseline that combines all input documents and summarizes the resulting meta-document achieves better results than all other approaches in the TAC 2009 dataset and also achieves better results than five of the reference methods in the DUC 2007 dataset. Nevertheless, our best performing configurations relative improvement in ROUGE-1 scores of 16% for TAC 2009 and of 17% for DUC 2007 (8% for TAC 2009 and DUC 2007 over the performing of the reference systems).

In what concerns the human study, the judges preferred our event-based summaries over all automatically generated summaries, which included other event-based summaries produced by our own implementation of Filatova et al. [12] method. Moreover, in the TAC 2009 dataset, the summaries generated by the proposed methods were even preferred over the reference summaries.

In terms of usefulness, our event-based summaries were again preferred over all other summaries, including the reference summaries in both datasets. This is related to the scores obtained for context coverage, where our event-based summaries obtained the highest scores. It is also interesting to observe that, although being extractive summaries, as it happens in all other approaches, our summaries obtained high scores on readability aspects such as grammaticality, referential clarity, and coherence. In fact, they were better than all other automatically generated summaries (except for Portfolio, on grammaticality, in DUC 2007). The best coherence score achieved in DUC 2007 might be related to the use of the waterfall architecture, that boosted the number of sentences selected from the last documents (the most recent ones). Concerning grammaticality, we believe that our event-based method could be improved by the inclusion of a pre-filtering step to remove news sources and datelines.

5. Conclusions

In this work, we explore a multi-document summarization framework based on event information and word embeddings that achieves performance above the state-of-the-art.

The multi-document summarization framework was developed by extending a single-document summarization method, KP-Centrality, in two hierarchical ways: single-layer and waterfall. The single-layer approach combines the summaries of each input document to produce the final summary. The waterfall approach combines the summaries of the input documents in a cascade fashion, in accordance with the temporal sequence of the documents. Event information is used in two different ways: in a filtering stage and to improve sentence representation as features of the summarization model. Related to event information, we also explored the temporal sequence of the input documents by increasing the size of the initial and intermediate summaries, used by our framework. To better capture content/event information expressed using different terms, we use two distributed representations of text: the skip-ngram

model, the continuous bag-of-words model, and the distributed representation of sentences. Event detection is based on the Fuzzy Fingerprint method and trained on the ACE 2005 Corpus.

To evaluate this multi-document summarization framework, we used two different setups: an automatic evaluation of the informativeness of the summaries using ROUGE-1, and a user study.

Our experiments showed that the use of event information combined with a distributed text representation (the SKIP model) further improved a generic multi-document summarization approach above state-of-the-art. Although we propose two different strategies for developing our multi-document methods, single-layer and waterfall, the best results were not achieved by the same architecture in the evaluation datasets because waterfall approach seems to be preferable to summarize large number of documents (e.g., 25 documents) and the single-layer seems more suitable for small number of documents (e.g., 10 documents). We confirmed this tendency by reducing the documents per topic to 10 in DUC 2007 and experimenting with waterfall and single-layer architectures. Both architectures achieved better results than the baseline and the reference systems. Analysis of the results also suggests that the waterfall model offers the best trade-off between performance and redundancy.

A possible future research direction is the compression of the sentences selected by our extractive summarizer. The process of compressing sentences should use event information to delete irrelevant words and to shorten long phrases. A solution to adequately compress sentences using event information entails solving multiple subproblems. For example, the identification of the relation between named entities (relationship extraction), identification of sentences mentioning the same event (event co-reference), and extract when the events take place (temporal information extraction), among other problems.

6. Acknowledgements

This work was supported by national funds through FCT under project UID/CEC/50021/2013, the Carnegie Mellon Portugal Program, and grant SFRH/BD/33769/2009.

References

[1] G. Glavaš, J. Šnajder, Event graphs for information retrieval and multi-document summarization, Expert Systems with Applications 41 (15) (2014) 6904–6916. doi:10.1016/j.eswa.2014.04.004.

[2] D. R. Radev, H. Jing, M. Styś, D. Tam, Centroid-based summarization of multiple documents, Information Processing and Management 40 (2004) 919–938. doi:10.1016/j.ipm.2003.10.006.

[3] G. Erkan, D. R. Radev, LexRank: Graph-based Centrality as Salience in Text Summarization, Journal of Artificial Intelligence Research 22 (2004) 457–479. doi:10.1613/jair.1523.

[4] D. Wang, T. Li, S. Zhu, C. Ding, Multi-document summarization via sentence-level semantic analysis and symmetric matrix factorization, in: Proceedings of the 31^{st} Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 2008, pp. 307–314. doi:10.1145/1390334.1390387.

[5] R. Ribeiro, D. M. de Matos, Revisiting Centrality-as-Relevance: Support Sets and Similarity as Geometric Proximity, Journal of Artificial Intelligence Research 42 (2011) 275–308. doi:10.1613/jair.3387.

[6] J. Carbonell, J. Goldstein, The Use of MMR, Diversity-based Reranking for Reordering Documents and Producing Summaries, in: Proceedings of the 21^{st} Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 1998, pp. 335–336. doi:10.1145/290941.291025.

[7] S. Guo, S. Sanner, Probabilistic latent maximal marginal relevance, in: Proceedings of the 33^{rd} International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 2010, pp. 833–834. doi:10.1145/1835449.1835639.

[8] S. Sanner, S. Guo, T. Graepel, S. Kharazmi, S. Karimi, Diverse Retrieval via Greedy Optimization of Expected 1-call@K in a Latent Subtopic Relevance Model, in: Proceedings of the 20^{th} ACM International Conference on Information and Knowledge Management, ACM, 2011, pp. 1977–1980. doi:10.1145/2063576.2063869.

[9] K. W. Lim, S. Sanner, S. Guo, On the Mathematical Relationship Between Expected N-call@K and the Relevance vs. Diversity Trade-off, in: Proceedings of the 35^{th} International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 2012, pp. 1117–1118. doi:10.1145/2348283.2348497.

[10] C.-Y. Lin, E. Hovy, The automated acquisition of topic signatures for text summarization, in: Proceedings of the 18^{th} Conference on Computational Linguistics - Volume 1, ACL, 2000, pp. 495–501. doi:10.3115/990820.990892.

[11] R. Sipos, A. Swaminathan, P. Shivaswamy, T. Joachims, Temporal corpus summarization using submodular word coverage, in: Proceedings of the 21^{st} ACM International Conference on Information and Knowledge Management, ACM, 2012, pp. 754–763. doi:10.1145/2396761.2396857.

[12] E. Filatova, V. Hatzivassiloglou, Event-based extractive summarization, in: Proceedings of ACL Workshop on Summarization, 2004, pp. 104–111.

[13] W. Li, M. Wu, Q. Lu, W. Xu, C. Yuan, Extractive Summarization Using Inter- and Intra- Event Relevance, in: Proceedings of the 21^{st} International Conference on Computational Linguistics and the 44^{th} Annual Meeting of the Association for Computational Linguistics, ACL, 2006, pp. 369–376. doi:10.3115/1220175.1220222.

[14] M. Liu, W. Li, M. Wu, Q. Lu, Extractive Summarization Based on Event Term Clustering, in: Proceedings of the 45^{th} Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions, ACL, 2007, pp. 185–188.

[15] R. Zhang, W. Li, Q. Lu, Sentence ordering with event-enriched semantics and two-layered clustering for multi-document news summarization, in: Proceedings of the 23^{rd} International Conference on Computational Linguistics: Posters, ACL, 2010, pp. 1489–1497.

[16] N. Daniel, D. Radev, T. Allison, Sub-event based multi-document summarization, in: Proceedings of the HLT-NAACL 03 on Text Summarization Workshop - Volume 5, ACL, 2003, pp. 9–16. doi:10.3115/1119467.1119469.

[17] P. Li, Y. Wang, W. Gao, J. Jiang, Generating aspect-oriented multi-document summarization with event-aspect model, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, ACL, 2011, pp. 1137–1146.

[18] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, arXiv preprint arXiv:1301.3781.

[19] N. Homem, J. P. Carvalho, Authorship identification and author fuzzy "fingerprints", in: Proceedings of the Annual Meeting of the North American Fuzzy Information Processing Society (NAFIPS), IEEE, 2011, pp. 1–6. doi:10.1109/NAFIPS.2011.5751998.

[20] L. Marujo, J. P. Carvalho, A. Gershman, J. Carbonell, J. P. Neto, D. M. de Matos, Textual event detection using fuzzy fingerprints, in: Proceedings of the 7^{th} IEEE International Conference Intelligent Systems IS'2014, Springer International Publishing, 2015, pp. 825–836. doi:10.1007/978-3-319-11313-5_72.

[21] C.-Y. Lin, Rouge: A package for automatic evaluation of summaries, in:

S. S. Marie-Francine Moens (Ed.), Text Summarization Branches Out: Proceedings of the ACL-04 Workshop, ACL, 2004, pp. 74–81.

[22] K. Ravi, V. Ravi, A survey on opinion mining and sentiment analysis: Tasks, approaches and applications, Knowledge-Based Systems (2015) – doi:http://dx.doi.org/10.1016/j.knosys.2015.06.015.

[23] R. M. Alguliev, R. M. Aliguliyev, N. R. Isazade, Desamc+docsum: Differential evolution with self-adaptive mutation and crossover parameters for multi-document summarization, Knowledge-Based Systems 36 (2012) 21 – 38. doi:http://dx.doi.org/10.1016/j.knosys.2012.05.017.

[24] W. Luo, F. Zhuang, Q. He, Z. Shi, Exploiting relevance, coverage, and novelty for query-focused multi-document summarization, Knowledge-Based Systems 46 (2013) 33 – 42. doi:http://dx.doi.org/10.1016/j.knosys.2013.02.015.

[25] G. Binh Tran, Structured summarization for news events, in: Proceedings of the 22^{nd} International Conference on World Wide Web Companion, International World Wide Web Conferences Steering Committee, 2013, pp. 343–348.

[26] G. Binh Tran, M. Alrifai, D. Quoc Nguyen, Predicting relevant news events for timeline summaries, in: Proceedings of the 22^{nd} International Conference on World Wide Web Companion, International World Wide Web Conferences Steering Committee, 2013, pp. 91–92.

[27] C. Walker, S. Strassel, J. Medero, ACE 2005 Multilingual training Corpus, LDC.

[28] J. Allan, J. Carbonell, G. Doddington, J. Yamron, Y. Yang, B. Archibald, M. Scudder, Topic Detection and Tracking Pilot Study Final Report, in: Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop, 1998, pp. 194–218.

[29] Y. Yang, T. Pierce, J. Carbonell, A study of retrospective and on-line event detection, in: Proceedings of the 21^{st} Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 1998, pp. 28–36. doi:10.1145/290941.290953.

[30] J. Carbonell, Y. Yang, J. Lafferty, R. D. Brown, T. Pierce, X. Liu, CMU Approach to TDT: Segmentation, Detection, and Tracking, in: Proceedings of the 1999 DARPA Broadcast News Conference, 1999.

[31] Y. Yang, J. G. Carbonell, R. D. Brown, T. Pierce, B. T. Archibald, X. Liu, Learning approaches for detecting and tracking news events, IEEE Intelligent Systems 14 (4) (1999) 32–43. doi:10.1109/5254.784083.

[32] R. Nallapati, A. Feng, F. Peng, J. Allan, Event Threading Within News Topics, in: Proceedings of the 13^{rd} ACM International Conference on Information and Knowledge Management, ACM, 2004, pp. 446–453. doi:10.1145/1031171.1031258.

[33] A. Feng, J. Allan, Finding and Linking Incidents in News, in: Proceedings of the 16^{th} ACM Conference on Conference on Information and Knowledge Management, ACM, 2007, pp. 821–830. doi:10.1145/1321440.1321554.

[34] Y. Hong, J. Zhang, B. Ma, J. Yao, G. Zhou, Q. Zhu, Using Cross-Entity Inference to Improve Event Extraction, in: Proceedings of the 49^{th} Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, ACL, 2011, pp. 1127–1136.

[35] H. Ji, R. Grishman, Knowledge Base Population: Successful Approaches and Challenges, in: Proceedings of the 49^{th} Annual Meeting of the Association for Computational Linguistics, ACL, 2011, pp. 1148–1158.

[36] S. Liao, R. Grishman, Using Document Level Cross-event Inference to Improve Event Extraction, in: Proceedings of the 48^{th} Annual Meeting of the Association for Computational Linguistics, ACL, 2010, pp. 789–797.

[37] M. Naughton, N. Stokes, J. Carthy, Sentence-level event classification in unstructured texts, Information Retrieval 13 (2) (2010) 132–156.

[38] P. Dasigi, E. Hovy, Modeling Newswire Events using Neural Networks for Anomaly Detection, in: Proceedings of the 25^{th} International Conference on Computational Linguistics (COLING 2014), 2014, pp. 1414–1422.

[39] E. Hovy, T. Mitamura, F. Verdejo, J. Araki, A. Philpot, Events are not simple: Identity, non-identity, and quasi-identity, in: Proceedings of the 1^{st} Workshop on EVENTS, ACL, 2013, pp. 21–28.

[40] J. Araki, E. Hovy, T. Mitamura, Evaluation for Partial Event Coreference, in: Proceedings of the 2^{nd} Workshop on EVENTS, ACL, 2014, pp. 68–76.

[41] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, P. Kuksa, Natural language processing (almost) from scratch, Journal of Machine Learning Research 12 (2011) 2493–2537.

[42] R. Socher, C. C. Lin, C. Manning, A. Y. Ng, Parsing natural scenes and natural language with recursive neural networks, in: Proceedings of The 28^{th} International Conference on Machine Learning, 2011, pp. 129–136.

[43] R. Ribeiro, L. Marujo, D. Martins de Matos, J. a. P. Neto, A. Gershman, J. Carbonell, Self reinforcement for important passage retrieval, in: Proceedings of the 36^{th} International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 2013, pp. 845–848. doi:10.1145/2484028.2484134.

[44] L. Marujo, J. Portêlo, D. M. de Matos, J. P. Neto, A. Gershman, J. Carbonell, I. Trancoso, B. Raj, Privacy-preserving important passage retrieval, in: Proceeding of the 1^{st} International Workshop on Privacy-Preserving IR: When Information Retrieval Meets Privacy and Security co-located with 37^{th} Annual International ACM SIGIR conference (SIGIR 2014), CEUR-WS.org, 2014, pp. 7–12.

[45] Q. Le, T. Mikolov, Distributed representations of sentences and documents, in: T. Jebara, E. P. Xing (Eds.), Proceedings of the 31^{st} International Conference on Machine Learning (ICML-14), 2014, pp. 1188–1196.

[46] A. Nenkova, Understanding the process of multi-document summarization: content selection, rewrite and evaluation, Ph.D. thesis, Columbia University (2006).

[47] K. McKeown, R. J. Passonneau, D. K. Elson, A. Nenkova, J. Hirschberg, Do Summaries Help?, in: Proceedings of the 28^{th} Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 2005, pp. 210–217. doi:10.1145/1076034.1076072.

[48] D. Graff, The aquaint corpus of english news text, LDC.

[49] E. Vorhees, D. Graff, Aquaint-2 information-retrieval text, LDC.

[50] J. Wang, J. Zhu, Portfolio theory of information retrieval, in: Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 2009, pp. 115–122. doi:10.1145/1571941.1571963.

[51] A. Haghighi, L. Vanderwende, 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, ACL, 2009, pp. 362–370.

[52] R. Parker, D. Graff, J. Kong, K. Chen, K. Maeda, English gigaword fifth edition, LDC.