Extending a Single-Document Summarizer to Multi-Document: a Hierarchical Approach

Luís Marujo[1,2,3]**, Ricardo Ribeiro**[1,4]**, David Martins de Matos**[1,2]**,**
João P. Neto[1,2]**, Anatole Gershman**[3]**, and Jaime Carbonell**[3]
[1]INESC-ID Lisboa, [2]IST/ULisboa, [4]ISCTE-IUL, Lisboa, Portugal
[3]School of Computer Science, CMU, Pittsburgh, USA
{lmarujo,anatoleg,jgc}@cs.cmu.edu
{ricardo.ribeiro,david.matos,joao.neto}@inesc-id.pt

Abstract

The increasing amount of online content motivated the development of multi-document summarization methods. In this work, we explore straightforward approaches to extend single-document summarization methods to multi-document summarization. The proposed methods are based on the hierarchical combination of single-document summaries, and achieves state of the art results.

1 Introduction

The use of the Internet to fulfill generic information needs motivated pioneer multi-document summarization efforts as NewsInEssence (Radev et al., 2005) or Newsblaster (McKeown et al., 2002), online since 2001. In general, multi-document summarization approaches have to address two different problems: passage selection and information ordering. Current multi-document systems adopt, for passage selection, approaches similar to the ones used in single-document summarization, and use the chronological order of the documents for information ordering (Christensen et al., 2013). The problem is that most approaches fail to generate summaries that cover generic topics which comprehend different, equally important, subtopics.

We propose to extend a state-of-the-art single-document summarization method, KP-CENTRALITY (Ribeiro et al., 2013), capable of focusing on diverse important topics while ignoring unimportant ones, to perform multi-document summarization. We explore two hierarchical strategies to perform this extension.

This document is organized as follows: Sect. 2 addresses the related work; Sect. 3 presents our multi-document summarization appproach; experimental results close the paper.

2 Related Work

Most of the current work in automatic summarization focuses on extractive summarization. The most popular baselines for multi-document summarization fall into one of the following general models: Centrality-based (Radev et al., 2004; Erkan and Radev, 2004; Wang et al., 2008; Ribeiro and de Matos, 2011), Maximal Marginal Relevance (MMR) (Carbonell and Goldstein, 1998; Guo and Sanner, 2010; Sanner et al., 2011; Lim et al., 2012), and Coverage-base methods (Lin and Hovy, 2000; Sipos et al., 2012). Additionally, methods such as KP-CENTRALITY (Ribeiro et al., 2013), which is centrality and coverage-based, follow more than one paradigm. In general, Centrality-based models are used to produce generic summaries, while the MMR family generates query-oriented ones. Coverage-base models produce summaries driven by words, topics or events.

Centrality-as-relevance methods base the detection of the most salient passages on the identification of the central passages of the input source(s). One of the main representatives of this family is *Passage-to-Centroid Similarity-based Centrality*. Centroid-based methods build on the idea of a pseudo-passage that represents the central topic of the input source— the *centroid*—selecting as passages to be included in the summary the ones that are close to the centroid. Another approach to centrality estimation is to com-

176

pare each candidate passage to every other passage and select the ones with higher scores (the ones that are closer to every other passage): the *Pair-wise Passage Similarity-based Centrality*.

MMR (Carbonell and Goldstein, 1998) is a query driven relevance model based on the following mathematical model:

$$\arg\max_{S_i} \left[\lambda(Sim_1(S_i, Q)) - (1-\lambda)(\max_{S_j} Sim_2(S_i, S_j)) \right]$$

where $Sim1$ and $Sim2$ are similarity metrics that do not have to be different; S_i are the yet unselected passages and S_j are the previously selected ones; Q is the required query to apply the model; and, λ is a parameter that allows to configure the result to be from a standard relevance-ranked list ($\lambda = 1$) to a maximal diversity ranking ($\lambda = 0$).

Coverage-based summarization defines a set of concepts that need to occur in the sentences selected for the summaries. The concepts are events (Filatova and Hatzivassiloglou, 2004), topics (Lin and Hovy, 2000), salient words (Lin and Bilmes, 2010; Sipos et al., 2012), and word n-grams (Gillick et al., 2008; Almeida and Martins, 2013).

3 Multi-Document Summarization

Our multi-document approach is built upon a centrality and coverage-based single-document summarization method, KP-CENTRALITY (Ribeiro et al., 2013). This method, through the use of key phrases, is easily adaptable and has been shown to be robust in the presence of noisy input. This is an important aspect considering that using as input several documents frequently increases the amount of unimportant content).

When adapting a single-document summarization method to perform multi-document summarization, a possible strategy is to combine the summaries of each document. To iteratively combine the summaries, we explore two different approaches: single-layer hierarchical and waterfall. Given that the summarization method also uses as input a set of key phrases, we extract from each input document the required set of key phrases, join the extracted sets, and rank the key phrases using their frequency. To generate each summary, we use the top key phrases, excluding the ones that do not occur in the input document.

3.1 Single-Document Summarization Method

To retrieve the most important sentences of an information source, we used the KP-CENTRALITY method (Ribeiro et al., 2013). We chose this model for its adaptability to different types of information sources (e.g., text, audio and video), while supporting privacy (Marujo et al., 2014), and offering state-of-art performance. It is based on the notion of combining key phrases with support sets. A support set is a group of the most semantically related passages. These semantic passages are chosen using heuristics based on the passage order method (Ribeiro and de Matos, 2011). This type of heuristics uses the structure of the input document (source) to partition the candidate passages to be included in the support set in two subsets: the ones closer to the passage associated with the support set under construction and the ones further apart. These heuristics use a permutation, $d_1^i, d_2^i, \cdots, d_{N-1}^i$, of the distances of the passages s_k to the passage p_i, related to the support set under construction, with $d_k^i = dist(s_k, p_i)$, $1 \leq k \leq N-1$, where N is the number of passages, corresponding to the order of occurrence of passages s_k in the input source. The metric that is normally used is the cosine distance.

The KP-Centrality method consists of two steps. First, it extracts key phrases using a supervised approach (Marujo et al., 2012) and combines them with a bag-of-words model in a compact matrix representation, given by:

$$\begin{bmatrix} w(t_1, p_1) \dots w(t_1, p_N) \ w(t_1, k_1) \dots w(t_1, k_M) \\ \vdots & \vdots \\ w(t_T, p_1) \dots w(t_T, p_N) \ w(t_T, k_1) \dots w(t_T, k_M) \end{bmatrix},$$
(1)

where w is a function of the number of occurrences of term t_i in passage p_j or key phrase k_l, T is the number of terms and M is the number of key phrases. Then, using a segmented information source $I \triangleq p_1, p_2, \ldots, p_N$, a support set S_i is computed for each passage p_i using:

$$S_i \triangleq \{s \in I \cup K : sim(s, q_i) > \varepsilon_i \land s \neq q_i\}, \quad (2)$$

for $i = 1, \ldots, N + M$. Passages are ranked excluding the key phrases K (*artificial passages*) according to:

$$\arg\max_{s \in (\cup_{i=1}^n S_i) - K} \left| \{S_i : s \in S_i\} \right|. \quad (3)$$

3.2 Single-Layer Hierarchical

In this model, we use KP-CENTRALITY to generate, for each news document, an intermediate summary with the same size of the output summary for the input documents. An aggregated summary is obtained by concatenating the chronologically ordered intermediate summaries. The output summary is again generated by applying KP-CENTRALITY to the aggregated summary as Figure 1 shows.
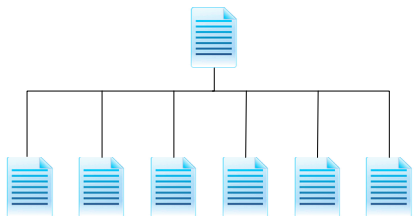

Figure 1: Single-layer architecture.

3.3 Waterfall

This model differs from the previous one in the merging process. The underlying merging of the documents follows a cascaded process: it starts by merging the intermediate summaries, with the same size of the output summary, of the first two documents, according to their chronological order. This document is then summarized and merged with the summary of following document. We iterate this process through all the documents until the most recent one as Figure 2 illustrates.
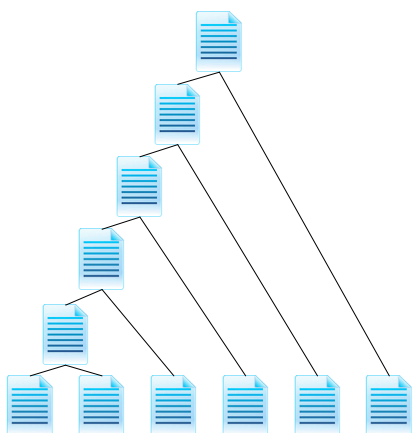

Figure 2: Waterfall architecture.

4 Experimental Results

We compare the performance of our methods against other representative models, namely MEAD, MMR, Expected n-call@k (Lim et al., 2012), and the Portfolio Theory (Wang and Zhu, 2009). MEAD is a centroid-based method and one of the most popular centrality-based methods. MMR is one of the most used query-based methods. Expected n-call@k adapts and extends MMR as a probabilistic model (Probabilistic Latent MMR). The Portfolio Theory also extends MMR based on the idea of ranking under uncertainty. As baseline, we used the straightforward idea of combining all input documents into a single one, and then submit the document to the single-document summarization method. Considering that most coverage-based systems explore event information, we opted for not including them in this comparative analysis.

To assess the informativeness of the summaries generated by our methods, we used ROUGE-1 and ROUGE-2 (Lin, 2004) on DUC 2007 and TAC 2009 datasets. The main summarization task in DUC 2007[1] is the generation of 250-word summaries of 45 clusters of 25 newswire documents (from the AQUAINT corpus) and 4 human reference summaries. The TAC 2009 Summarization task[2] has 44 topic clusters. Each topic has 2 sets of 10 news documents obtained from the AQUAINT 2 corpus.There are 4 human 100-word reference summaries for each set, where the reference summaries for the first set are query-oriented, and for the second set are update summaries. In this work, we used the first set of reference summaries. We evaluate the different models by generating summaries with 250 words. We only present the best results.

The used features include the bag-of-words model representation of the sentences (TF-IDF), the key phrases and the query (obtained from the topics descriptions). Including the query is a new extension to the KP-CENTRALITY method, which, in general, improved the results. We experimented with different numbers of key phrases, obtaining the best results with 40 key phrases. To compare and rank the sentences, we use several distance metrics, namely: Frac133 (generic Minkowski distance,

---

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

Distance	Model	DUC 2007		TAC 2009	
		R1	R2	R1	R2
frac133	baseline	0.3565	0.0744	0.4706	0.1268
cosine		0.3406	0.0670	0.4746	0.1391
frac133	waterfall	0.3569	0.0765	0.4943	0.1441
frac133	single-layer	**0.3775**	0.0882	0.4983	0.1526
cosine	waterfall	0.3701	**0.0904**	**0.5137**	**0.1693**
cosine	single-layer	0.3707	0.0822	0.4993	0.1590
frac133	single-layer (shuffle)	0.3689	0.0807	0.5060	0.1483
cosine	waterfall (shuffle)	0.3626	0.0844	0.5107	0.1630
	MEAD	0.3282	0.0765	0.4153	0.0845
	MMR	0.3269	0.0780	0.3917	0.0801
	E.n-call@k	0.3209	0.0701	0.3873	0.0699
	Portfolio	0.3595	0.0792	0.4292	0.0758
	LexRank	0.2881	0.0534	0.3845	0.0623

Table 1: ROUGE-1 (R1) and ROUGE-2 (R2) scores.

with $N = 1.(3)$), Euclidean, Chebyshev, Manhattan, Minkowski, the Jensen-Shannon Divergence, and the cosine similarity. Table 1 shows that the best results were obtained by the proposed hierarchical models, in both datasets. Overal, the best performing distance metric for our centrality-based method was the cosine similarity and the best strategy for combining the information was the waterfall approach, namely, in terms of ROUGE-2. In DUC 2007, frac133 using the single-layer method achieved the best ROUGE-1 score, although the difference for cosine is hardly noticeable. Single-layer with frac133 shows a performance improvement of 0.0180 ROUGE-1 points (relative performance improvement of 5.0%) over the best of the other systems, Portfolio, in DUC 2007, and of 0.0845 ROUGE-1 points (19.7% relative performance improvement) in TAC 2009. In terms of ROUGE-2, the waterfall method using cosine achieved an improvement of 0.0112 (relative performance improvement of 14.1%) over Portfolio, in DUC 2007, and of 0.0848 (relative performance improvement of 100.4%) over MEAD, the best performing of the reference systems using this metric, in TAC 2009. Note that our baseline obtained results similar to the best reference system in DUC 2007 and better results than all reference systems in TAC 2009 (0.0454 ROUGE-1 points corresponding to a 10.6% relative performance improvement; 0.0546 ROUGE-2

points corresponding to a 64.6% relative performance improvement). The better results obtained on the TAC 2009 dataset are due to the small size of the reference summaries and to the fact that the documents sets to be summarized contain topics with higher diversity of subtopics.

The shuffle results included in Table 1 are averages of 10 trials. They are lower than the other obtained using the documents organized in chronological order. This suggests that the order of the input documents is important to the summarization methods.

Figure 3 shows an example of summary produced by our multi-document method. The figure also includes the respective reference summary for comparison.

5 Conclusions and Future Work

In this work, we explore two different approaches to extend a single-document summarization method to multi-document summarization: single-layer hierarchical and waterfall.

Experimental results show that the proposed approaches perform better than previous state-of-the-art methods on standard datasets used to evaluate this task. In general, the best performing approach is the waterfall approach using the cosine similarity. In fact, this configuration achieves the best results on the TAC 2009 dataset, considering both ROUGE-1

Generated Summary:

President Bill Clinton said Friday he will appeal a federal judge's ruling that struck down a law giving the president the power to veto specific items in bills passed by Congress. The law, passed by Congress last year, allowed the president for the first time to veto particular items in spending bills and certain limited tax provisions passed by Congress. Clinton said the funding that Congress has added to the bill is excessive and threatened to veto some items by using the line-item veto power. The White House said that the president used his authority to cancel projects that were not requested in the budget and would not substantially improve the quality of life of military service members. Judge Thomas Hogan ruled that the law – which gives the president the power to strike items from tax and spending measures without vetoing the entire bill – violates the traditional balance of powers between the various branches of government "The Line-Item Veto Act is unconstitutional because it impermissibly disrupts the balance of powers among the three branches of government," said Thomas Hogan." In its appeal, the Justice Department argues that the new challengers also do not have standing to challenge the law, and that in any case the law is in line with the historic relationship between Congress and the president.

Reference summary:

Congress passed a law authorizing the line item veto (LIV) in 1996 accepting arguments that the measure would help preserve the integrity of federal spending by allowing the president to strike unnecessary spending and tax items from legislation thus encouraging the government to live within its means. It was considered in line with the historic relationship between Congress and the president and would provide a tool for eliminating wasteful pork barrel spending while enlivening debate over the best use of funds. It was argued that the LIV would represent presidential exercise of spending authority delegated by Congress. President Clinton exercised the LIV on 82 items in 1997 saving $1.9 billion in spending projected over five years. The affected items were projects for specific localities, many in the area of military construction, which had been added to the president's budget by Congress. The first court ruling on the LIV act was in U.S. District Court when in February 1998 it was ruled unconstitutional on the grounds that it violated the separation of powers. The Department of Justice appealed that decision and in June 1998 the Supreme Court ruled the LIV act unconstitutional but on the grounds that it violated Article I, 7, Clause 2 (The "presentment clause") of the Constitution that establishes the process by which a bill becomes law. President Clinton expressed his deep disappointment.

Figure 3: Example of summary produced by our summarizer and the reference summary Topic D0730G of DUC 2007

and ROUGE-2 metrics, and, although not achieving the best results in the DUC 2007 dataset, in terms of ROUGE-1, it also achieves a performance improvement over Portfolio of 0.0106 ROUGE-1 points (relative performance improvement of 3%).

In future work, we aim to adapt the proposed multi-document summarization method to perform abstractive summarization.

Acknowledgments

References

Miguel Almeida and Andre Martins. 2013. Fast and robust compressive summarization with dual decomposition and multi-task learning. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 196–206, Sofia, Bulgaria, August. ACL.

Jaime Carbonell and Jade Goldstein. 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*, pages 335–336. ACM.

Janara Christensen, Mausam, Stephen Soderland, and Oren Etzioni. 2013. Towards coherent multi-document summarization. In *Proceedings of the North American Chapter of the Association for Computational Linguistics*. ACL.

Güneş Erkan and Dragomir R. Radev. 2004. LexRank: Graph-based Centrality as Salience in Text Summarization. *Journal of Artificial Intelligence Research*, 22:457–479.

Elena Filatova and Vasileios Hatzivassiloglou. 2004. Event-based extractive summarization. In *Proc. of ACL Workshop on Summarization*, pages 104–111.

Dan Gillick, Benoit Favre, and Dilek Hakkani-Tur. 2008. The icsi summarization system at tac 2008. In *Proceedings of the Text Understanding Conference*.

Shengbo Guo and Scott Sanner. 2010. Probabilistic latent maximal marginal relevance. In *Proc. of the 33rd International ACM SIGIR Conference on Research*

and Development in Information Retrieval, pages 833–834. ACM.

Kar Wai Lim, Scott Sanner, and Shengbo Guo. 2012. On the Math. Relationship Between Expected N-call@K and the Relevance vs. Diversity Trade-off. In *Proc. of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1117–1118. ACM.

Hui Lin and Jeff Bilmes. 2010. Multi-document summarization via budgeted maximization of submodular functions. In *Proceedings of the North American Chapter of the Association for Computational Linguistics*, pages 912–920. ACL.

Chin-Yew Lin and Eduard Hovy. 2000. The automated acquisition of topic signatures for text summarization. In *Proceedings of the 18th Conference on Computational Linguistics - Volume 1*, pages 495–501. ACL.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summ. Branches Out: Proc. of the ACL-04 Workshop*.

Luís Marujo, Anatole Gershman, Jaime Carbonell, Robert Frederking, and Joã P. Neto. 2012. Supervised topical key phrase extraction of news stories using crowdsourcing, light filtering and co-reference normalization. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. ELRA.

Luís Marujo, José Portêlo, David Martins de Matos, João P Neto, Anatole Gershman, Jaime Carbonell, Isabel Trancoso, and Bhiksha Raj. 2014. Privacy-preserving important passage retrieval. In *ACM SIGIR PIR workshop*.

Kathleen R. McKeown, Regina Barzilay, David Evans, Vasileios Hatzivassiloglou, Judith L. Klavans, Ani Nenkova, Carl Sable, Barry Schiffman, and Sergey Sigelman. 2002. Tracking and Summarizing News on a Daily Basis with Columbia's Newsblaster. In *HLT*.

Dragomir R. Radev, Hongyan Jing, Małgorzata Styś, and Daniel Tam. 2004. Centroid-based summarization of multiple documents. *Information Processing and Management*, 40.

Dragomir R. Radev, Jahna Otterbacher, Adam Winkel, and Sasha Blair-Goldensohn. 2005. NewsInEssence: Summarizing Online News Topics. *Communications of the ACM*, 48(10):95–98.

Ricardo Ribeiro and David Martins de Matos. 2011. Revisiting Centrality-as-Relevance: Support Sets and Similarity as Geometric Proximity. *JAIR*, 42:275–308.

Ricardo Ribeiro, Luís Marujo, David Martins de Matos, João P. Neto, Anatole Gershman, and Jaime Carbonell. 2013. Self reinforcement for important passage retrieval. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 845–848. ACM.

Scott Sanner, Shengbo Guo, Thore Graepel, Sadegh Kharazmi, and Sarvnaz Karimi. 2011. Diverse retrieval via greedy optimization of expected 1-call@k in a latent subtopic relevance model. In *Proc. of the 20th ACM International Conference on Information and Knowledge Management*, pages 1977–1980. ACM.

Ruben Sipos, Adith Swaminathan, Pannaga Shivaswamy, and Thorsten Joachims. 2012. Temporal corpus summarization using submodular word coverage. In *Proc. of the 21st ACM International Conference on Information and Knowledge Management*. ACM.

Jun Wang and Jianhan Zhu. 2009. Portfolio theory of information retrieval. In *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 115–122.

Dingding Wang, Tao Li, Shenghuo Zhu, and Chris Ding. 2008. Multi-document summarization via sentence-level semantic analysis and symmetric matrix factorization. In *Proc. of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 307–314. ACM.