Self Reinforcement for Important Passage Retrieval

Ricardo Ribeiro
INESC-ID and ISCTE-IUL
Lisboa, Portugal
rdmr@l2f.inesc-id.pt

Luís Marujo
School of Computer Science,
CMU; INESC-ID; and IST
lmarujo@cs.cmu.edu

David Martins de Matos
INESC-ID and IST
Lisboa, Portugal
david.matos@inesc-id.pt

João P. Neto
INESC-ID and IST
Lisboa, Portugal
joao.neto@inesc-id.pt

Anatole Gershman
School of Computer Science
CMU, Pittsburgh, USA
anatoleg@cs.cmu.edu

Jaime Carbonell
School of Computer Science
CMU, Pittsburgh, USA
jgc@cs.cmu.edu

ABSTRACT

In general, centrality-based retrieval models treat all elements of the retrieval space equally, which may reduce their effectiveness. In the specific context of extractive summarization (or important passage retrieval), this means that these models do not take into account that information sources often contain lateral issues, which are hardly as important as the description of the main topic, or are composed by mixtures of topics. We present a new two-stage method that starts by extracting a collection of key phrases that will be used to help centrality-as-relevance retrieval model. We explore several approaches to the integration of the key phrases in the centrality model. The proposed method is evaluated using different datasets that vary in noise (noisy *vs* clean) and language (Portuguese *vs* English). Results show that the best variant achieves relative performance improvements of about 31% in clean data and 18% in noisy data.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]; I.2.7 [**Natural Language Processing**]: Text analysis

General Terms

Algorithms, Experimentation, Theory

Keywords

Passage Retrieval, Extractive Summarization, Automatic Key Phrase Extraction, AKE, Centrality

1. INTRODUCTION

One of the most popular [6] retrieval models is PageRank [1]. Although born in the web page retrieval research topic, it was the basis for several models proposed in different disciplines [4, 6]. Namely, in the automatic important passage retrieval/summarization area, models like TextRank [18] or LexRank [5] are still used as baselines or as the foundation of state-of-the-art work [3, 28]. However, as noted by Ribeiro and de Matos [20], this kind of models treats all passages of an information source equally. This is a problem because information sources often contain lateral issues, which are hardly as important as the description of the main topic, or are composed by mixtures of topics. One possible solution is to use a biased centrality model [20].

In this work, we propose to use, on top of a biased centrality model, a method to better reinforce the most important passages. We achieved that by devising a two-stage retrieval method, in which the first step consists in the automatic extraction of a collection of key phrases that will further bias the centrality model. To evaluate our method we use both clean and noisy datasets, in different languages.

In this document, the next section describes the related work; section 3 presents the datasets; sections 4 and 5 detail the two stages of our method; the results are included and discussed in section 6. The conclusions close the document.

SIGIR'13, July 28–August 1, 2013, Dublin, Ireland.

2. RELATED WORK

2.1 Automatic Key Phrase Extraction

Automatic Key Phrase Extraction (AKE) is a Natural Language Processing (NLP) task that selects the most important words or phrases from a document. Key phrases are phrases consisting of one or more significant words (keywords). As they capture semantic metadata, search engines can use them to enhance indexing, to help users in queries completion [30] and improve web traffic prediction [12]. Several NLP applications, such as summarization, information retrieval, information extraction, and question answering, can benefit from their extraction as well. Several unsupervised key phrase methods have been proposed, such as language modeling, graph-based ranking and clustering [16]. However, the TF-IDF across different methods remains a strong unsupervised baseline [7]. Both supervised and unsupervised approaches have been explored to perform AKE. Supervised methods formalize this problem as a binary classification problem of two steps [17, 14]: candidate generation and filtering of the phrases selected before.

2.2 Important Passage Retrieval

Assessing the relevant content is the first step of automatic summarization systems. On the one hand, extractive sum-

marization consists of determining the most relevant segments, usually sentences, of one or more information sources. On the other hand, automatic abstractive summarizers also need to identify the most relevant content that, then, will be submitted to transformation and generation stages.

Text and speech information sources influence the complexity of the approaches differently. For text summarization it is common to use complex information, such as syntactic [27], semantic [25], and discourse information [26], either to assess relevance or reduce the length of the output. However, speech summarization approaches have a extra layer of complexity caused by speech-related issues like recognition errors or disfluencies. As a consequence, it is necessary to use speech-specific information (for example, acoustic/prosodic features [15] or recognition confidence scores [31]) or by improving both the assessment of relevance and the intelligibility of the output of an automatic speech recognition system (by using related information [21]). These problems not only increase the difficulty in determining the salient information, but also constrain the applicability of text summarization techniques to speech summarization. Nevertheless, shallow text summarization approaches such as Latent Semantic Analysis (LSA) [9] and Maximal Marginal Relevance (MMR) [2] seem to achieve performances comparable to the ones using specific speech-related features [19]. In addition, discourse features start to gain some importance in speech summarization [15, 33].

2.3 Two-stage methods

Closely related to our work are the unsupervised key phrase extraction approaches that have been explored to reinforce summarization [32, 29, 11, 22, 24]. Namely, Litval and Last [11] and Riedhammer et al. [22] propose the use of key phrases to summarize news articles [11] and meetings [22]. Litval and Last explore both supervised and unsupervised methods to extract key phrases as a first step towards extractive summarization. We use a feature-rich supervised method for key phrase extraction, unlike the ones proposed by Litval and Last which are grounded on ad-hoc and structural features and use a graph-based representation. Moreover, our adaptation of the centrality summarization model plays an important role in the whole process, an inexistent step in their work. In that sense, Riedhammer et al. propose the method closest to ours: the first stage consists in a simple key phrase extraction step, based on part-of-speech patterns; then, these key phrases are used to define the relevance and redundancy components of a MMR summarization model.

3. DATASETS

In order to assess the quality of our method, we analyzed its performance using two different datasets.

3.1 Portuguese (PT) Broadcast News (BN)

The PT BN dataset, used in previous work [21], consists of the automatic transcriptions of 18 BN stories in European Portuguese, which are part of a news program. News stories cover several generic topics like "society", "politics", and "sports", among others. For each news story, there is a human-produced abstract, used as reference. The average word recognition error rate is 16.5% and automatic sentence segmentation attained a slot error rate (SER, commonly used to evaluate this kind of task) of 81.5%. Although this dataset was used in previous work, news story segmentation problems were corrected in one case. However, this does not change the relevant properties of the corpus.

3.2 English (EN) Event Reports (ER)

To evaluate our method using clean input, we used the Concisus Corpus of Event Summaries [23]. The used subcorpus is composed by 78 event reports and respective summaries, distributed across three different types of events: aviation accidents, earthquakes, and train accidents.

4. KEY PHRASE EXTRACTION

In previous work, Marujo et al. [14] expanded the MAUI toolkit with shallow semantic features, such as number of Named Entities, POS tags, 4 n-gram domain model probabilities. These expansion improved the quality of their approach to generate tag clouds of Portuguese Broadcast News. As a result of the domain similarity and the good results, we used the same approach to extract key phrases in our summarization experiments for Portuguese.

In the following year, Marujo et al. [13] adapted the AKE work to English and investigated additional semantic features and pre-processing steps, namely Light Filtering and Co-reference normalization. These new features included the detection of rhetorical devices, Freebase sub-categories, and news articles top categories. Including such new features and pre-processing steps improved the key phrase extraction results beyond the state-of-art. Therefore, we used the methodology of Marujo et al. [13] in our summarization experiments in English corpus. However, since the pre-processing steps, namely Light Filtering, could have impact on the outcome of our experiments, they were removed. This fact lead to the exclusion of the Freebase sub-categories which were only beneficial in combination with pre-processing steps. Unfortunately, the news articles top categories were not available. Therefore, the inclusion of rhetorical devices features is the main difference between the PT and EN AKE.

5. IMPORTANT PASSAGE RETRIEVAL

To determine the most important sentences of an information source, we use the centrality model described by Ribeiro and de Matos [20]. The reasons to choose this model are its adaptability (the authors of the model suggest how to integrate additional information sources), the language independence, and the state-of-the-art performance on both clean and noisy input.

This centrality model is based on the notion of support set: after dealing with the representational aspects, the first step of the method is to compute a set consisting of the most semantically related passages, designated support set. Then, the most important passages are the ones that occur in the largest number of support sets.

Given a segmented information source $I \triangleq p_1, p_2, ..., p_N$, a support set is computed for each passage p_i (Eq. 1, $sim()$ is a similarity function, and ε_i is a threshold).

$$S_i \triangleq \{s \in I : sim(s, p_i) > \varepsilon_i \wedge s \neq p_i\} \quad (1)$$

Passages are ranked in accordance to Eq. 2.

$$\underset{s \in \cup_{i=1}^{n} S_i}{\arg\max} \left|\{S_i : s \in S_i\}\right| \quad (2)$$

For our two-stage important passage retrieval method, we adapted the model in three different ways: **KP-Centrality**

is the approach where key phrases are considered regular passages (augmenting the number of support sets). In method **OKP-Centrality**, passages that do not contain key phrases are removed from the support sets; **CKP-Centrality** weights the passages using the bagged decision tree confidence scores.

6. EVALUATION

To assess the quality of our method we made experiments using the two datasets presented in section 3. To evaluate the detection of the most important sentences, we used ROUGE [10], namely ROUGE-1, which is the most widely used evaluation measure for this scenario. In the experiments using the PT BN dataset, the summary size was determined by the size of the reference human summaries, which consisted in about 10% of the input news story. In the experiments using the EN ER dataset, we generate 3 sentence summaries, commonly found in online news web sites, like Google News.

We compare the new 3 different approaches described in section 5 to the baseline (the centrality-as-relevance raw model), with the number of key phrases ranging from 5 to 40. The metric used to configure the centrality model was the cosine (using IDF). The heuristic used to compute the size of each support set was the one based on the selection of the sentences with less distance to sentence under analysis [20]. LexRank performance was also included for a better understanding of the improvements.

6.1 Results

Figure 1 shows the results for the PT BN dataset. As we can see, the best performing approach is method KP-Centrality. Method OKP-Centrality only outperforms the baseline for 10 key phrases and it is clearly worse for 5 key phrases. However, its performance remains similar to baseline for the other variations. Method CKP-Centrality performance is always similar to the baseline's performance. In figure 2 is possible to observe the performance of the

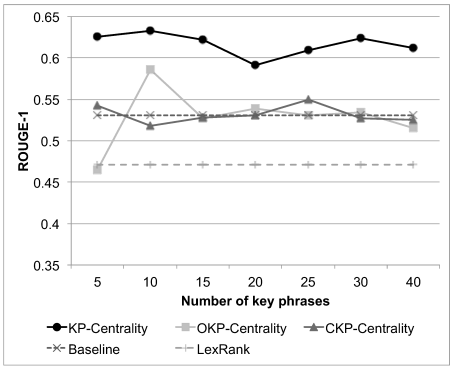

Figure 1: ROUGE-1 scores for the PT BN dataset.

same methods when applied to the EN ER dataset. In this dataset, both methods KP-Centrality and CKP-Centrality have a better performance than the baseline. Similarly to what happens in the PT BN dataset, KP-Centrality achieves the best results. However, in this dataset the performance improves directly with the number of keys phrases (until 60), while in the other dataset the best results are achieved around 10 key phrases. The performance of CKP-Centrality does not vary with the number of key phrases. Method

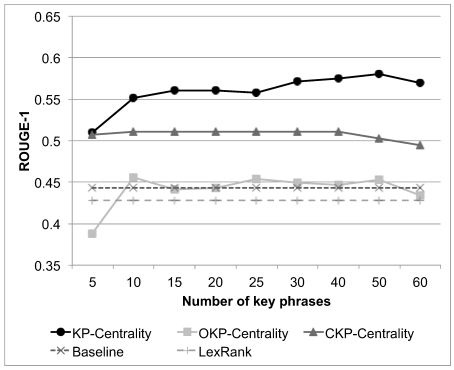

Figure 2: ROUGE-1 scores for the EN ER dataset.

OKP-Centrality achieves a performance similar in both datasets, although in the EN dataset it does not outperform the baseline.

6.2 Discussion

We start by noting that the Portuguese BN dataset is noisy (it is affected by speech-related problems like recognition and segmentation errors), while the English one is clean. Another important aspect is that the performance of the AKE is better for English than Portuguese due to the contribution of the rhetorical signals. The influence of this last aspect can be seen in the performance of KP-Centrality, which keeps improving on the English dataset (until 60 key phrases), what does not happen in the Portuguese BN dataset. The two mentioned aspects can be the justification for the results achieved by CKP-Centrality: it outperforms the baseline in the English dataset, while having a performance similar to the baseline in the Portuguese dataset. The stability of the performance of OKP-Centrality and CKP-Centrality can be justified by their nature: they do not generate more support sets than the base model, while in KP-Centrality, new support sets are also computed for key phrases. The poor performance of OKP-Centrality shows that the removal of passages does not improve the base mode, since it already has the capability of distinguishing between the main topic and lateral issues.

7. CONCLUSIONS

In this work, we introduced a two-stage method for important passage retrieval. Popular centrality-based models treat equally all elements of the retrieval space, impacting the retrieval task negatively. In line with recent work [8, 20], that begins to address this problem, we show that our method can improve the performance of a retrieval model that already addresses this issue. The method we propose starts by extracting a collection of key phrases that will be used to bias a centrality-as-relevance retrieval model. We explore three different approaches to the integration of the key phrases and experiment using noisy (automatic speech transcriptions) and clean (event reports) data. One of the approaches (KP-Centrality) clearly improves the baseline model in both noisy (by 18%) and non-noisy data (by 31%). The rhetorical devices used in the English dataset can be a possible justification for the performance difference between the two datasets. On the other hand, the approach where passages that do not contain key phrases are removed does

not achieve as good results, which means different aspects are captured in the two stages of our method. Key phrases and this centrality model seem to complement each other.

In the future, we plan to explore the use of key phrases in the computation of the similarity between passages, and improve the current methods of integration of the key phrases.

8. ACKNOWLEDGMENTS

We would like to thank FCT for supporting this research through PEst-OE/EEI/LA0021/2011, the Carnegie Mellon Portugal Program, and grant SFRH/BD/33769/2009.

9. REFERENCES

[1] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30:107–117, 1998.

[2] J. Carbonell and J. Goldstein. The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries. In *SIGIR 1998*, 1998.

[3] H. Ceylan, R. Mihalcea, U. Özertem, E. Lloret, and M. Palomar. Quantifying the Limits and Success of Extractive Summarization Systems Across Domains. In *Proc. of NAACL*, pages 903–911. ACL, 2010.

[4] Y. Ding, E. Yan, A. Frazho, and J. Caverlee. Pagerank for ranking authors in co-citation networks. *J. Am. Soc. Inf. Sci. Technol.*, 60(11):2229–2243, 2009.

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

[6] M. Franceschet. PageRank: Standing on the Shoulders of Giants. *Communications of the ACM*, 54(6), 2011.

[7] K. Hasan and V. Ng. Conundrums in unsupervised keyphrase extraction: making sense of the state-of-the-art. In *Proc. of the COLING*, 2010.

[8] O. Kurland and L. Lee. PageRank without Hyperlinks: Structural Reranking using Links Induced by Language Models. *ACM TOIS*, 28(4):1–38, 2010.

[9] T. K. Landauer and S. T. Dumais. A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction and Representation of Knowledge. *Psych. Review*, 104(2):211–240, 1997.

[10] C.-Y. Lin. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summ. Branches Out: Proc. of the ACL-04 Workshop*, 2004.

[11] M. Litvak and M. Last. Graph-Based Keyword Extraction for Single-Document Summarization. In *Coling 2008: MMIES*, pages 17–24, 2008.

[12] L. Marujo, M. Bugalho, J. P. Neto, A. Gershman, and J. Carbonell. Hourly traffic prediction of news stories. In *3rd Int. C.A.R.S. Workshop in ACM RecSys*, 2011.

[13] L. Marujo, A. Gershman, J. Carbonell, R. Frederking, and J. P. Neto. Supervised topical key phrase extraction of news stories using crowdsourcing, light filtering and co-reference normalization. In *Proc. of the LREC*, 2012.

[14] L. Marujo, M. Viveiros, and J. P. Neto. Keyphrase Cloud Generation of Broadcast News. In *Interspeech 2011*. ISCA, September 2011.

[15] S. R. Maskey and J. Hirschberg. Comparing Lexical, Acoustic/Prosodic, Structural and Discourse Features for Speech Summarization. In *Proceedings of the 9^{th} EUROSPEECH - INTERSPEECH 2005*, 2005.

[16] Y. Matsuo and M. Ishizuka. Keyword extraction from a single document using word co-ocurrence statistical information. *Int. Journal on A.I. Tools*, 2004.

[17] O. Medelyan, V. Perrone, and I. H. Witten. Subject metadata support powered by Maui. In *Proceedings of the JCDL '10*, page 407. ACM Press, 2010.

[18] R. Mihalcea and P. Tarau. TextRank: Bringing Order into Texts. In *Proc. of the EMNLP*. ACL, 2004.

[19] G. Penn and X. Zhu. A Critical Reassessment of Evaluation Baselines for Speech Summarization. In *Proc. of ACL-08: HLT*, pages 470–478. ACL, 2008.

[20] R. Ribeiro and D. M. de Matos. Revisiting Centrality-as-Relevance: Support Sets and Similarity as Geometric Proximity. *JAIR*, 42:275–308, 2011.

[21] R. Ribeiro and D. M. de Matos. *Multi-source, Multilingual Information Extraction and Summarization*, chapter Improving Speech-to-Text Summarization by Using Additional Information Sources. Theory and Applic. of NLP. Springer, 2013.

[22] K. Riedhammer, B. Favre, and D. Hakkani-Tür. Long story short – Global unsupervised models for keyphrase based meeting summarization. *Speech Communication*, 52:801–815, 2010.

[23] H. Saggion and S. Szasz. The concisus corpus of event summaries. In *Proc. of the LREC*, 2012.

[24] R. Sipos, A. Swaminathan, P. Shivaswamy, and T. Joachims. Temporal corpus summarization using submodular word coverage. In *Proc. of CIKM*, 2012.

[25] R. I. Tucker and K. Spärck Jones. Between shallow and deep: an experiment in automatic summarising. Technical Report 632, University of Cambridge, 2005.

[26] V. R. Uzêda, T. A. S. Pardo, and M. das Graças Volpe Nunes. A comprehensive comparative evaluation of RST-based summarization methods. *ACM Trans. on Speech and Language Processing*, 6(4):1–20, 2010.

[27] L. Vanderwende, H. Suzuki, C. Brockett, and A. Nenkova. Beyond SumBasic: Task-focused summarization and lexical expansion. *Information Processing and Management*, 43:1606–1618, 2007.

[28] X. Wan, H. Li, and J. Xiao. EUSUM: Extracting Easy-to-Understand English Summaries for Non-Native Readers. In *SIGIR 2010*, 2010.

[29] X. Wan, J. Yang, and J. Xiao. Towards an iterative reinforcement approach for simultaneous document summarization and keyword extraction. In *Proceedings of the 45th ACL*, pages 552–559, 2007.

[30] F. Wei, W. Li, Q. Lu, and Y. He. Query-sensitive mutual reinforcement chain and its application in query-oriented multi-document summarization. In *Proc. of the ACM SIGIR*, pages 283–290. ACM, 2008.

[31] K. Zechner and A. Waibel. Minimizing Word Error Rate in Textual Summaries of Spoken Language. In *Proceedings of the NAACL*, pages 186–193, 2000.

[32] H. Zha. Generic summarization and keyphrase extraction using mutual reinforcement principle and sentence clustering. In *SIGIR 2002*, 2002.

[33] J. J. Zhang, R. H. Y. Chan, and P. Fung. Extractive Speech Summarization Using Shallow Rhetorical Structure Modeling. *IEEE TASLP*, 18(6), 2010.