Revisiting Centrality-as-Relevance: Support Sets and Similarity as Geometric Proximity: Extended Abstract*

Ricardo Ribeiro[1,3] and **David Martins de Matos**[2,3]
[1]Instituto Universitário de Lisboa (ISCTE-IUL)
[2]Instituto Superior Técnico, Universidade Técnica de Lisboa
[3]Spoken Language Systems Laboratory - L[2]F/INESC-ID
{ricardo.ribeiro,david.matos}@inesc-id.pt

Abstract

In automatic summarization, centrality-as-relevance means that the most important content of an information source, or of a collection of information sources, corresponds to the most central passages, considering a representation where such notion makes sense (graph, spatial, etc.). We assess the main paradigms and introduce a new centrality-based relevance model for automatic summarization that relies on the use of support sets to better estimate the relevant content. Geometric proximity is used to compute semantic relatedness. Centrality (relevance) is determined by considering the whole input source (and not only local information), and by taking into account the existence of minor topics or lateral subjects in the information sources to be summarized. The method consists in creating, for each passage of the input source, a support set consisting only of the most semantically related passages. Then, the determination of the most relevant content is achieved by selecting the passages that occur in the largest number of support sets. This model produces extractive summaries that are generic, and language- and domain-independent. Thorough automatic evaluation shows that the method achieves state-of-the-art performance, both in written text, and automatically transcribed speech summarization, even when compared to considerably more complex approaches.

1 Introduction

A summary conveys to the end user the most relevant content of one or more information sources, in a concise and comprehensible manner. Several difficulties arise when addressing this problem, but one of utmost importance is how to assess the significant content. Usually, approaches vary in complexity if processing text or speech. While in text summarization, up-to-date systems make use of complex information, such as syntactic [Vanderwende *et al.*, 2007], semantic [Tucker and Spärck Jones, 2005], and discourse information [Harabagiu and Lacatusu, 2005; Uzêda *et al.*, 2010], either to assess relevance or reduce the length of the output, common approaches to speech summarization try to cope with speech-related issues by using speech-specific information (for example, prosodic features [Maskey and Hirschberg, 2005], or recognition confidence scores [Zechner and Waibel, 2000]) or by improving the intelligibility of the output of an automatic speech recognition system (by using related information [Ribeiro and de Matos, 2008a]). Nonetheless, shallow text summarization approaches such as Latent Semantic Analysis [Landauer *et al.*, 1998; Gong and Liu, 2001] and Maximal Marginal Relevance [Carbonell and Goldstein, 1998] seem to achieve performances comparable to the ones using specific speech-related features [Penn and Zhu, 2008].

A common family of approaches to the identification of the relevant content is the *centrality* family. These methods base the detection of the most salient passages on the identification of the central passages of the input source(s). 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 (x) to be included in the summary the ones that are close to the centroid. Pioneer work (on multi-document summarization) by Radev *et al.* [1999] and Radev *et al.* [2000] creates clusters of documents by representing each document as a *tf-idf* vector; the centroid of each cluster is also defined as a *tf-idf* vector, with the coordinates corresponding to the weighted average of the *tf-idf* values of the documents of the cluster; finally, sentences that contain the words of the centroids are presumably the best representatives of the topic of the cluster, thus being the best candidates to belonging to the summary.

$$centrality(x) = similarity(x, centroid) \quad (1)$$

Another approach to centrality estimation is to compare each candidate passage to every other passage (y) and select the ones with higher scores (the ones that are closer to every other passage). One simple way to do this is to represent passages as vectors using a weighting scheme like the aforementioned *tf-idf*; then, passage similarity can be assessed using, for instance, the cosine, assigning to each passage a centrality score as defined in Eq. 2.

$$centrality(x) = \frac{1}{N} \sum_{y} similarity(x, y) \quad (2)$$

*This paper is an extended abstract of the JAIR publication [Ribeiro and de Matos, 2011].

These scores are then used to create a sentence ranking: sentences with highest scores are selected to create the summary.

A major problem of this relevance paradigm is that by taking into account the entire input source in this manner, either to estimate centroids or average distances of input source passages, we may be selecting extracts that being *central* to the input source are, however, not the most relevant ones. In cognitive terms, the information reduction techniques in the summarization process are quite close to the discourse understanding process [Endres-Niggemeyer, 1998], which, at a certain level, works by applying rules that help uncovering the macrostructure of the discourse. One of these rules, *deletion*, is used to eliminate from the understanding process propositions that are not relevant to the interpretation of the subsequent ones. This means that it is common to find, in the input sources to be summarized, lateral issues or considerations that are not relevant to devise the salient information (discourse structure-based summarization is based on the relevance of nuclear text segments, [Marcu, 2000; Uzêda *et al.*, 2010]), and that may affect centrality-based summarization methods by inducing inadequate centroids or decreasing the scores of more suitable sentences.

As argued by previous work [Gong and Liu, 2001; Steyvers and Griffiths, 2007], we also assume that input sources are mixtures of topics, and propose to address that aspect using the input source itself as guidance. By associating to each passage of the input source a support set consisting only of the most semantically related passages in the same input source, groups of related passages are uncovered, each one constituting a latent topic (the union of the supports sets whose intersection is not empty). In the creation of these support sets, semantic relatedness is assessed by geometric proximity. Moreover, while similar work usually explores different weighting schemes to address specific issues of the task under research [Orăsan *et al.*, 2004; Murray and Renals, 2007; Ribeiro and de Matos, 2008b], we explore different geometric distances as similarity measures, analyzing their performance in context (the impact of different metrics from both theoretical and empirical perspectives in a clustering setting was shown in [Aggarwal *et al.*, 2001]). To build the summary, we select the sentences that occur in the largest number of support sets—hence, the most *central* sentences, without the problem that affects previous centrality-based methods.

Our method produces generic, language- and domain-independent summaries, with low computational requirements. We test our approach both in speech and text data. In the empirical evaluation of the model over text data, we used an experimental setup previously used in published work [Mihalcea and Tarau, 2005; Antiqueira *et al.*, 2009], which enabled an informative comparison to the existing approaches. In what concerns the speech experiments, we also used a corpus collected in previous work [Ribeiro and de Matos, 2008a], as well as the published results. This allowed us to compare our model to state-of-the-art work.

2 Support Sets and Geometric Proximity

The leading concept in our model is the concept of support set: the first step of our method to assess the relevant content is to create a support set for each passage of the input source by computing the similarity between each passage and the remaining ones, selecting the closest passages to belong to the support set. The most relevant passages are the ones that occur in the largest number of support sets.

Given a segmented information source $I \triangleq p_1, p_2, ..., p_N$, support sets S_i associated with each passage p_i are defined as indicated in Eq. 3 ($sim()$ is a similarity function, and ε_i is a threshold).

$$S_i \triangleq \{s \in I : sim(s, p_i) > \varepsilon_i \wedge s \neq p_i\} \tag{3}$$

The most relevant segments are given by selecting the passages that satisfy Eq. 4.

$$\underset{s \in \cup_{i=1}^{n} S_i}{\arg\max} \left|\{S_i : s \in S_i\}\right| \tag{4}$$

A major difference from previous centrality models and the main reason to introduce the support sets is that by allowing different thresholds to each set (ε_i), we let centrality be influenced by the latent topics that emerge from the groups of related passages. In the degenerate case where all ε_i are equal, we fall into the degree centrality model proposed by Erkan and Radev [2004]. But using, for instance, a naïve approach of having dynamic thresholds (ε_i) set by limiting the cardinality of the support sets (a *k*NN approach), centrality is changed because each support set has only the most semantically related passages of each passage. From a graph theory perspective, this means that the underlying representation is not undirected, and the support set can be interpreted as the passages recommended by the passage associated to the support set. This contrasts with both LexRank [Erkan and Radev, 2004] models, which are based on undirected graphs. On the other hand, the models proposed by Mihalcea and Tarau [2005] are closer to our work in the sense that they explore directed graphs, although only in a simple way (graphs can only be directed forward or backward). Nonetheless, semantic relatedness (content overlap) and centrality assessment (performed by the graph ranking algorithms HITS [Kleinberg, 1999] and PageRank [Brin and Page, 1998]) is quite different from our proposal. Although not addressing automatic summarization, considering the *k*NN approach to the definition of the support set size, the work of Kurland and Lee [2005; 2010] is the most similar to our model. However, Kurland and Lee base neighborhood definition on generation probabilities, while we explore geometric proximity. Nevertheless, from the perspective of our model, the *k*NN approach to support set definition is only a possible strategy (others can be used): our model can be seen as a generalization of both *k*NN and εNN approaches, since what we propose is the use of differentiated thresholds (ε_i) for each support set (Eq. 3).

3 Evaluation

Despite the number of approaches to summary evaluation, the most widely used metric is still ROUGE [Lin, 2004] and is the one we use in our study. We chose ROUGE not only owing to its wide adoption, but also because one of the data sets used in our evaluation has been used in published studies, allowing us to easily compare the performance of our model with other

known systems. Namely, we use the ROUGE-1 score, known to correlate well with human judgment [Lin, 2004]. Moreover, we estimate confidence intervals using non-parametric bootstrap with 1000 resamplings [Mooney and Duval, 1993].

Since we are proposing a generic summarization model, we conducted experiments both in text and speech data.

3.1 Experiment 1: Text

In this section, we describe the experiments performed and analyze the corresponding results when using as input source written text.

Data

The used corpus, known as TeMário, consists of 100 newspaper articles in Brazilian Portuguese [Pardo and Rino, 2003]. Although our model is general and language-independent, this corpus was used in several published studies, allowing us to perform an informed comparison of our results. The articles in the corpus cover several domains, such as "world", "politics", and "foreign affairs". For each of the 100 newspaper articles, there is a reference human-produced summary. The text was tokenized and punctuation removed, maintaining sentence boundary information.

Evaluation Setup

Considering the preprocessing step we applied to the corpus and the observed differences in the published results, we found it important to evaluate centrality approaches under the same conditions. Thus, we implemented the following centrality models:

- **Uniform Influx** (corresponds to the non-recursive, unweighted version of the model), proposed by Kurland and Lee for re-ranking in document retrieval (we experimented with several k in graph definition, the sames used for support set cardinality in the kNN strategy, and μ—10, 20, 50, 100, 200, 500, 1000, 2000, 5000, 10000);
- **PageRank**, proposed by both Mihalcea and Tarau and Erkan and Radev (passage similarity metrics differ and Mihalcea and Tarau also explore directed graphs);
- **Degree centrality** as proposed by Erkan and Radev (we experimented with several thresholds δ, ranging from 0.01 to 0.09); and,
- **Baseline**, in which the ranking is defined by the order of the sentences in the news article, with relevance decreasing from the begining to the end.

Concerning summary size, the number of words in the generated summaries directly depends on the number of words of the reference abstracts, which consisted in compressing the input sources to 25-30% of the original size.

Results

Overall, the best performing configuration of the proposed model is significantly better, using the directional Wilcoxon signed rank test with continuity correction, than TextRank Undirected, ($W = 2584$, $p < 0.05$), Uniform Influx ($W = 2740$, $p < 0.05$), and also Continuous LexRank ($W = 2381.5$, $p < 0.1$).[1]

[1]Statistical tests were computed using R [R Development Core Team, 2009].

3.2 Experiment 2: Speech

In this section, we describe the experiments performed and analyze the corresponding results when using as input source automatically transcribed speech.

Data

To evaluate our ideas in the speech processing setting, we used the same data of Ribeiro and de Matos [2008a]: the automatic transcriptions of 15 broadcast news stories in European Portuguese, part of a news program. Subject areas include "society", "politics", "sports", among others. The average word recognition error rate is 19.5% and automatic sentence segmentation attained a slot error rate (SER, commonly used to evaluate this kind of task) of 90.2%.

Evaluation Setup

Given the implemented models, in this experiment we compare the support sets relevance model to the following systems:

- An LSA baseline.
- The following graph-based methods: Uniform Influx [Kurland and Lee, 2005; 2010], Continuous LexRank and Degree centrality [Erkan and Radev, 2004], and TextRank [Mihalcea and Tarau, 2004; 2005].
- The method proposed by Ribeiro and de Matos, which explores the use of additional related information, less prone to speech-related errors (e.g. from online newspapers), to improve speech summarization (Mixed-Source).
- Two human summarizers (extractive) using as source the automatic speech transcriptions of the news stories (Human Extractive).

To be able to perform a good assessment of the automatic models, we conducted two experiments: in the first one, the number of sentence-like units (SUs) extracted to compose the automatic summaries was defined in accordance to the number of sentences of the reference human abstracts (which consisted in compressing the input source to about 10% of the original size); in the second experiment, the number of extracted SUs of the automatic summaries was determined by the size of the shortest corresponding human extractive summary.

Results

Overal results show that the proposed model achieved the best performance when using the abstracts size. Although not achieving the best performance in the experiment using the extracts size, there is no significant difference between the best support sets-based relevance model configuration and the ones achieved by human summarizers: applying the directional Wilcoxon signed rank test with continuity correction, the test values when using the shortest human extracts size are W = 53, p = 0.5.

The proposed model has a better performance with statistical significance than Degree ($W = 53.5$, $p < 0.005$ when using the abstracts size; $W = 54$, $p < 0.005$ when using the shortest human extracts size), TextRank Undirected ($W = 92.5$, $p < 0.05$ when using the abstracts size; $W = 96$,

$p < 0.05$ when using the shortest human extracts size), and Uniform Influx ($W = 60$, $p < 0.01$ when using the abstracts size; $W = 51$, $p < 0.06$ when using the shortest human extracts size).

Further, comparing our model to more complex (not centrality-based), state-of-the-art models [Lin *et al.*, 2010] suggests that at least similar performance is attained: the relative performance increment of our model over LexRank is of 57.4% and 39.8% (both speech experiments), whereas the relative gain of the best variant of the model proposed by Lin *et al.* over LexRank is of 39.6%. Note that this can only be taken as indicative, since an accurate comparison is not possible because data sets differ, Lin *et al.* do not explicit which variant of LexRank is used, and do not address statistical significance.

4 Conclusions

In our work, we assessed the main approaches of the centrality-as-relevance paradigm, and introduced a new centrality-based relevance model for automatic summarization. Our model uses support sets to better characterize the information sources to be summarized, leading to a better estimation of the relevant content. In fact, we assume that input sources comprehend several topics that are uncovered by associating to each passage a support set composed by the most semantically related passages. Building on the ideas of Ruge [1992], *[...] the model of semantic space in which the relative position of two terms determines the semantic similarity better fits the imagination of human intuition [about] semantic similarity [...]*, semantic relatedness was computed by geometric proximity. We explore several metrics and analyze their impact on the proposed model as well as (to a certain extent) on the related work. Centrality (relevance) is determined by taking into account the whole input source, and not only local information, using the support sets-based representation. Moreover, although not formally analyzed, notice that the proposed model has low computational requirements.

We conducted a thorough automatic evaluation, experimenting our model both on written text and transcribed speech summarization. The obtained results suggest that the model is robust, being able to detect the most relevant content without specific information of where it should be found and performing well in the presence of noisy input, such as automatic speech transcriptions. However, it must be taken into consideration that the use of ROUGE in summary evaluation, although generalized, allowing to easily compare results and replicate experiments, is not an ideal scenario, and consequently, results should be corroborated by a perceptual evaluation. The outcome of the performed trials show that the proposed model achieves state-of-the-art performance in both text and speech summarization, even when compared to considerably more complex approaches.

Acknowledgments

This work was partially supported by national funds through FCT – Fundação para a Ciência e a Tecnologia, under project PEst-OE/EEI/LA0021/2011.

References

[Aggarwal *et al.*, 2001] Charu C. Aggarwal, Alexander Hinneburg, and Daniel A. Keim. On the Surprising Behavior of Distance Metrics in High Dimensional Space. In *Database Theory — ICDT 2001, 8th International Conference London, UK, January 4–6, 2001 Proceedings*, volume 1973 of *Lecture Notes in Computer Science*, pages 420–434. Springer, 2001.

[Antiqueira *et al.*, 2009] Lucas Antiqueira, Osvaldo N. Oliveira Jr., Luciano da Fontoura Costa, and Maria das Graças Volpe Nunes. A complex network approach to text summarization. *Information Sciences*, 179(5):584–599, 2009.

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

[Carbonell and Goldstein, 1998] Jaime Carbonell and Jade Goldstein. The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries. In *SIGIR 1998: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 335–336. ACM, 1998.

[Endres-Niggemeyer, 1998] Brigitte Endres-Niggemeyer. *Summarizing Information*. Springer, 1998.

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

[Gong and Liu, 2001] Yihong Gong and Xin Liu. Generic Text Summarization Using Relevance Measure and Latent Semantic Analysis. In *SIGIR 2001: Proceedings of the 24st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 19–25. ACM, 2001.

[Harabagiu and Lacatusu, 2005] Sanda Harabagiu and Finley Lacatusu. Topic Themes for Multi-Document Summarization. In *SIGIR 2005: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 202–209. ACM, 2005.

[Kleinberg, 1999] J. M. Kleinberg. Authoritative Sources in a Hyperlinked Environment. *Journal of the ACM*, 46(5):604–632, 1999.

[Kurland and Lee, 2005] Oren Kurland and Lillian Lee. PageRank without Hyperlinks: Structural Re-Ranking using Links Induced by Language Models. In *SIGIR 2005: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 306–313. ACM, 2005.

[Kurland and Lee, 2010] Oren Kurland and Lillian Lee. PageRank without Hyperlinks: Structural Reranking using Links Induced by Language Models. *ACM Transactions on Information Systems*, 28(4):1–38, 2010.

[Landauer *et al.*, 1998] Thomas K. Landauer, Peter W. Foltz, and Darrel Laham. An Introduction to Latent Semantic Analysis. *Discourse Processes*, 25(2):259–284, 1998.

[Lin *et al.*, 2010] Shih-Hsiang Lin, Yao-Ming Yeh, and Berlin Chen. Extractive Speech Summarization – From the View of Decision Theory. In *Proceedings of the 11^{th} Annual Conference of the International Speech Communication Association (INTERSPEECH 2010)*, pages 1684–1687. ISCA, 2010.

[Lin, 2004] Chin-Yew Lin. ROUGE: A Package for Automatic Evaluation of Summaries. In Marie-Francine Moens and Stan Szpakowicz, editors, *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81. ACL, 2004.

[Marcu, 2000] Daniel Marcu. *The Theory and Practice of Discourse Parsing and Summarization*. The MIT Press, 2000.

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

[Mihalcea and Tarau, 2004] Rada Mihalcea and Paul Tarau. TextRank: Bringing Order into Texts. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411. ACL, 2004.

[Mihalcea and Tarau, 2005] Rada Mihalcea and Paul Tarau. A Language Independent Algorithm for Single and Multiple Document Summarization. In *Proceedings of the Second International Joint Conference on Natural Language Processing: Companion Volume to the Proceedings of Conference including Posters/Demos and Tutorial Abstracts*, pages 19–24. Asian Federation of Natural Language Processing, 2005.

[Mooney and Duval, 1993] Christopher Z. Mooney and Robert D. Duval. *Bootstrapping : a nonparametric approach to statistical inference*. Sage Publications, 1993.

[Murray and Renals, 2007] Gabriel Murray and Steve Renals. Term-Weighting for Summarization of Multi-Party Spoken Dialogues. In *Machine Learning for Multimodal Interaction IV*, volume 4892 of *Lecture Notes in Computer Science*, pages 155–166. Springer, 2007.

[Orăsan *et al.*, 2004] Constantin Orăsan, Viktor Pekar, and Laura Hasler. A comparison of summarisation methods based on term specificity estimation. In *Proceedings of the Fourth International Language Resources and Evaluation (LREC'04)*, pages 1037–1041. ELRA, 2004.

[Pardo and Rino, 2003] Thiago Alexandre Salgueiro Pardo and Lucia Helena Machado Rino. TeMario: a corpus for automatic text summarization. Technical Report NILC-TR-03-09, Núcleo Interinstitucional de Lingüística Computacional (NILC), São Carlos, Brazil, 2003.

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

[R Development Core Team, 2009] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2009. ISBN 3-900051-07-0.

[Radev *et al.*, 1999] Dragomir R. Radev, Vasileios Hatzivassiloglou, and Kathleen R. McKeown. A Description of the CIDR System as Used for TDT-2. In *Proceedings of the DARPA Broadcast News Workshop*, 1999.

[Radev *et al.*, 2000] Dragomir R. Radev, Hongyan Jing, and Malgorzata Budzikowska. Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies. In *NAACL-ANLP 2000 Workshop: Automatic Summarization*, pages 21–30. ACL, 2000.

[Ribeiro and de Matos, 2008a] Ricardo Ribeiro and David Martins de Matos. Mixed-Source Multi-Document Speech-to-Text Summarization. In *Coling 2008: Proceedings of the 2nd workshop on Multi-source Multilingual Information Extraction and Summarization*, pages 33–40. Coling 2008 Organizing Committee, 2008.

[Ribeiro and de Matos, 2008b] Ricardo Ribeiro and David Martins de Matos. Using Prior Knowledge to Assess Relevance in Speech Summarization. In *2008 IEEE Workshop on Spoken Language Technology*, pages 169–172. IEEE, 2008.

[Ribeiro and de Matos, 2011] Ricardo Ribeiro and David Martins de Matos. Revisiting Centrality-as-Relevance: Support Sets and Similarity as Geometric Proximity. *Journal of Artificial Intelligence Research*, 42:275–308, 2011.

[Ruge, 1992] Gerda Ruge. Experiments on linguistically-based term associations. *Information Processing and Management*, 28(3):317–332, 1992.

[Steyvers and Griffiths, 2007] Mark Steyvers and Tom Griffiths. *Handbook of Latent Semantic Analysis*, chapter Probabilistic Topic Models, pages 427–448. Lawrence Erlbaum Associates, 2007.

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

[Uzêda *et al.*, 2010] Vinícius Rodrigues Uzêda, Thiago Alexandre Salgueiro Pardo, and Maria das Graças Volpe Nunes. A comprehensive comparative evaluation of RST-based summarization methods. *ACM Transactions on Speech and Language Processing*, 6(4):1–20, 2010.

[Vanderwende *et al.*, 2007] Lucy Vanderwende, Hisami Suzuki, Chris Brockett, and Ani Nenkova. Beyond SumBasic: Task-focused summarization and lexical expansion. *Information Processing and Management*, 43:1606–1618, 2007.

[Zechner and Waibel, 2000] Klaus Zechner and Alex Waibel. Minimizing Word Error Rate in Textual Summaries of Spoken Language. In *Proceedings of the 1^{st} conference of the North American chapter of the ACL*, pages 186–193. Morgan Kaufmann, 2000.