Keyphrase Cloud Generation of Broadcast News

Luís Marujo[1,2,3], *Márcio Viveiros*[4], *João P. Neto*[2,3,4]

[1] Language Technologies Institute / Carnegie Mellon University, Pittsburgh, USA
[2] Spoken Language Laboratory / INESC-ID Lisboa, Portugal
[3] Instituto Superior Técnico, Lisboa, Portugal
[4] VoiceInteraction, Lisboa, Portugal

Luis.Marujo@inesc-id.pt, Marcio.Viveiros@voiceinteraction.pt, Joao.Neto@inesc-id.pt

Abstract

This paper describes an enhanced automatic keyphrase extraction method applied to Broadcast News. The keyphrase extraction process is used to create a concept level for each news. On top of words resulting from a speech recognition system output and news indexation and it contributes to the generation of a tag/keyphrase cloud of the top news included in a Multimedia Monitoring Solution system for TV and Radio news/programs, running daily, and monitoring 12 TV channels and 4 Radios.

Index Terms: keyphrase extraction, tag cloud generation, Broadcast News, speech browsing, speech recognition

1. Introduction

There are large amounts of daily produced video and audio information, from TV and Radio channels, that are not searchable and indexed. Nowadays, despite all the available amount of information, users spend an inordinate amount of time in zapping in a constant search for relevant information. The traditional channels with a fixed and generic program alignment are not useful any more. Video-on-demand (VOD) for series, films, and music are becoming the standard access method. The IPTV and the generalization of set-top-box concept are changing the TV experience. Also the concept of search potentiated by Google will be a future request also for TVs users, since is starting to be the standard method to access information. However the user interface is a problem in terms of access. The way to provide users with a good TV experience will be based on recommendation systems. However to build recommendation systems for TV and Radio we need additional work. The availability of a video on the web is problematic because there is no way to access it through content. A search engine is blind to this kind of data.

The advances of Automatic Speech Recognition Tools (ASR) enabled the automatic transcription of Broadcast News (BN) [1]. Standard speech recognition systems generate raw single-case words, without punctuation marks, with numbers written as text, and with many different types of disfluencies. The generation of the missing information makes this representation format easier to read and understand, and mitigates problems to further automatic processing [2]. Capitalization, also known as truecasing, improves human readability, parsing, and NER (Named Entity Recognition). Punctuation marks are useful for parsing, information extraction, machine translation, extractive summarization, and NER.

The correct identification of the main concepts is one of the bases for automatic Indexation and Segmentation of BN process [3]. There is a growing demand for rich interfaces. Tag cloud, sometimes called word cloud, has become a hallmark of Web 2.0 design. Tag clouds are weighted renditions of collections of words (tags) that can be used to represent the concepts, in a visually appealing way to summarize vast amounts of information[4]. ASR clouds are known to be viable ways of individually representing podcasts [5]. The cloud can reflect a high level description of the news. We hope to create a high level description of news using keyphrase extraction. A keyphrase is a set of relevant words or phrases that appear verbatim in a document and that give a brief summary of its content.

Several keyphrase extraction methods have been proposed; these methods can be categorized into simple statistics, linguistic, machine learning, and hybrid. N-Grams [6], word frequency [7], TF*IDF [8], PAT-tree [9] fall into the simple statistics. The linguistic approaches include both lexical and syntactic analysis [10]. The machine learning techniques are usually supervised – the algorithms learn a model from training data containing manually identified keyphrases and use this model to classify – SVM [11], Nave Bayes [12], CRF [13], C4.5 [14] are known examples. Hybrids [14] are a combination of two or more approaches.

This paper describes the development of keyphrase extraction method used to create the middle layer of an hierarchical 3 layers representation of news and it is used to generate tag clouds from the top news of the last 6 hours of Portuguese BNs included in a Media Monitoring Solution (MMS) system. It is also explored several features and classifiers for the keyphrase extraction.

This paper is organized as follows: Section 2 presents the System Architecture; the description of the new modules included in the MMS system is the nucleus of the Section 3; the results are describe in Section 4, and Section 5 concludes and suggests future work.

2. System Architecture

The figure 1 shows the component view of the MMS system architecture presented in [15] where the gray blocks are introduced by this work. The blocks represent software components and the arrows represent the data flow. The sliced blocks represent several instances of the modules running in parallel. Furthermore, XML is the default internal data representation used between components.

The Media Receiver is responsible for capturing and recording the broadcast news program on TV and Radio in a 24/7 setup. The next component is an ASR that generates the audio transcription based on electronic program guide (EPG) information. Then, the transcription is enriched with Punctu-

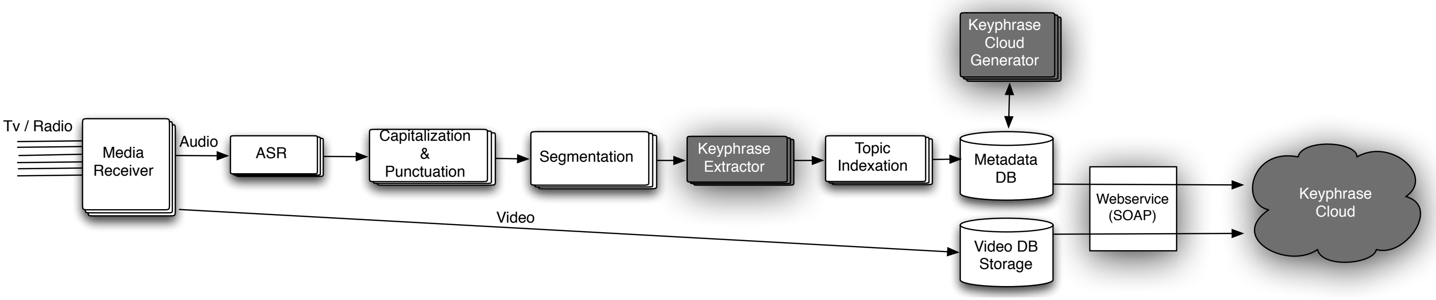

Figure 1: Component view of the system architecture.

ation and Capitalization. Subsequently, each BN program is segmented into several news and the Keyphrases are extracted – the extraction process is described in Section 3.1. Each news is topic indexed or topic classified. Finally each news is stored in the Metadata DB being constituted by the word transcription, keyphrases, and index, besides program/channel and timing information. The Keyphrase Cloud Generator interacts with the Metadata DB creating/updating hourly the keyphrase cloud and storing again the cloud representation in the database. The 3D keyphrase cloud linked to the videos are shown when a user access the system and interact with the databases using a webservice that abstracts the database access.

3. Description of the Keyphrase Cloud Generation

The motivation to develop an automatic keyphrase extraction process is to create an hierarchical 3 layers representation of news. At the base level are the words provenient from ASR output, at the middle layer are the concepts (keyphrases), and at the top the index (topic). Our focus on this paper is on the middle or concept layer where every keyphrase represents a concept. Wherefore our concern is to identify the most relevant concepts without losing substantial precision. The cloud is generated because it provides a fast overview over the concepts of the most recent news.

3.1. Keyphrase Extraction

Our Keyphrase extraction implementation was built over Maui-indexer toolkit [14], state-of-art keyphrase extraction toolkit, which is respectively an extended version of the KEA [16]. The first step consisted of an adaptation of the toolkit for Portuguese, i.e., the incorporation of the Portuguese stemmer and the list of the stopwords based on the KEA porting work for Portuguese [17]. The minimum number of occurrences of phrase in a news document to be consider as a keyphrase candidate is one.

Malformed keyphrases are candidate phrases starting or ending or both in stop words. Hence, they are excluded. Maui-indexer extracts several features: TF, IDF, TFxIDF, position of first occurrence, position of the last occurrence, distance between last and first occurrence, and number of words in the phrase. These are the baseline features (base) and the words are steamed. Then, we enriched the feature extraction process with the following features:

- number of characters (f1) – empirically noun words that are long tend to be relevant;
- number of Named Entities (f2) – recurrently named entities are important keyphrases; the number of named entities per keyphrase is obtained decomposing it into words and labelling them using the MorphoAdorner Name Recognizer[1];
- number of Capital letters (f3) – the identification of acronyms is the main motivation to include this feature.
- Part of Speech (POS) tags (f4) – keyphrases are usually nouns or noun phrases, verbs or verb phrases are less frequent, and the remaining POS tags are rare;
- probability of the keyphrase in a 4-ngram domain model (f5) – feature included to capture how frequent it is to find in a larger BN corpus.

The 4-ngram domain model is interpolation of back-off n-gram models of BNs generated for AUDIMUS ASR language model [18]. It contains about 58,000 unigrams, 7,000,000 bigrams, 15,000,000 trigrams, and 10,000,000 4-grams. The model was compressed based on the Minimal Perfect Hash method developed for language models [19] to allow both faster access to the model and lower memory footprint. The smoothnlp toolkit[2] was used for this purpose. The compress model is about 12% of the original size.

The training phrase consists of extracting features for every candidate keyphrases identified in the train set described in Section 4.1 and include a binary label stating whether it is a keyphrase. Then, the machine learning classifier, e.g. decision tree, establishes a association between the features values and the binary class. This association, that can measured in terms of probability, is designated as classification model. The candidate phrases are ranked based on the probability generated by the model and it is selected top ranked phrases.

The machine learning classifier used by the Maui-indexer is the Bootstrap Aggregating (bagging) [20] applied to C4.5 decision tree algorithm [21]. Bagging is usually applied to decision tree models to reduce variance and mitigate overfitting by combining classification of randomly generated sample training sets.

We decided to use the CART algorithm, that stands for Classification and Regression Tree, because it usually has better performance than the C4.5 [22] and WEB[3] . In addition, we wanted to improve the performance classification for small keyphrases extraction ($<=10$ keyphrases) – to our knowledge the large majority of keyphrases extraction experiments fall into this category – and discover for large keyphrases extraction ($>=30$ keyphrases) how well they perform.

[1] `http://morphadorner.northwestern.edu/`
[2] `http://tinyurl.com/MphfCompres`
[3] `http://tinyurl.com/C4-5VsCart`

3.2. Cloud Generation

The cloud is generated using the 10 top ranked keyphrases from about 10 top news from all tv/radio channels. Each cloud encapsulates the 20 most frequent keyphrases, i.e., each news is represented on average by 2 keyphrases. This procedure is used to generate both the all topics and the 12 topic based clouds. The only difference between the all topics cloud and the topic based clouds is the restriction of the top news that need belonging to the cloud topic.

At these stage of development, the top news are selected based on 2 heuristics:

- time order – news broadcast in beginning of the program are usually more relevant, i.e., they are usually order by their relevance.
- number of occurrences - the same relevant news are frequently transmitted in several channels simultaneously or within a temporal windows, e.g. 6 hours.

3.3. Cloud Visualization

When a user logins into the system, he visualizes the generic 3D cloud (Figure 2 on the right). The numbers in parenthesis reflect the total number of occurrences of the keyphrase in the news and it is an provisional way to show the relevance of the keyphrase in a 3D cloud.

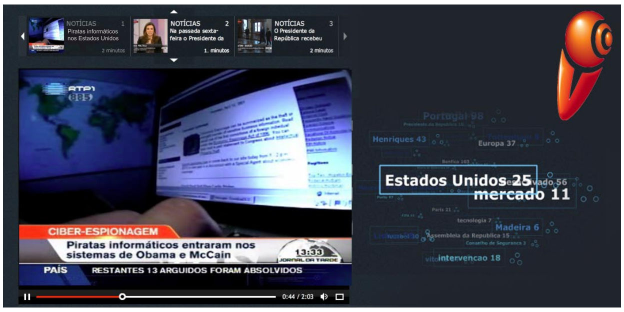

Figure 2: MMS All Topic Cloud interface.

When the user clicks in one keyphrase of the cloud, the most recent news video containing the keyphrase is played. While the news video is being played, the user can also navigate horizontally through the playlist of news containing the keyphrase over time in all tv/radio channels. In addition, the user can also navigate vertically and visualize the different concepts.

4. Results

4.1. Gold Standard Corpus

The gold standard corpus BN annotated with the respective keyphrases was developed because we did not find available any annotated corpus of the domain . Since we have to create a new corpus, we started to build for Portuguese, but in the future we will extend our approach to English. Having this corpus was crucial to train and test our supervised extraction approach.

The gold standard is made of 8 BN programs, containing a total of 110 news subset, from the European Portuguese ALERT BN database [23]. The news transcription were produced by AUDIMUS, the in-house ASR for Portuguese, with low WER (14,56 % on average); and punctuated and capitalize automatically using an in-house Capitalization and Punctuation module [24]. The programs were segmented automatically. Later, each news was manually reexamined to fix any segmentation error.

Afterward, one annotator was asked to extract all keyphrases that summarize each news, that is, to mark all keyphrases that he assessed as needed to represent all concepts shown for each news. The maximum number of words in a keyphrase was limited to 5 words. This is another substantial difference from the large majority of keyphrase extraction work and stimulated the creation of gold standard.

Type	#News	Total #Words	Avg.#Keyphrases
Train	100	29225	24
Test	10	3896	29

Table 1: Corpus description.

4.2. Keyphrase Extraction Evaluation

To evaluate the performance of our keyphrase extraction method, the standard Information Retrieval metrics, i.e., precision (P), recall (R) and F1 measure were used.

The Tables 2 and 3 shows the average results obtained in the test set. The leftmost column shows the maximum number of keyphrases extracted per news. The average number of keyphrases correctly identified appears in 3^{rd} column in the middle.

The first rows of the Tables 2 and 3 are our baseline results that are inline with reported state of art values for supervised approaches [12]. They were calculated for using the top 30 keyphrase candidates extracted. The following 10 rows are included to hightlight the performance gains per each new feature. They are followed by the combination of features. Only the best features combinations were included.

# Keyphrases Extracted	Features	#Keyphrases Identified	P	R	F1
30	base	8.5	28.33	31.71	29.93
30	base+f1	8.6	28.67	32.02	30.25
30	base+f2	9	30	33.69	31.74
30	base+f3	9.2	30.67	35.4	32.86
30	base+f4	8.1	27	30.56	28.67
30	base+f5	8.9	29.67	33.73	31.57
30	base+f1+f2	8.6	28.67	32.02	30.25
30	all prev.+f3	9.4	31.33	35.81	33.42
30	all prev.+f4	9.7	32.33	36.86	**34.45**
10	all prev.+f4	5	50	19.39	27.95
10	All	5.3	**53**	20.63	29.7
20	All	7.4	37	28.21	32.01
30	All	9.2	30.67	35.12	32.74
35	All	10	28.57	37.79	32.54
40	All	10.3	25.75	38.87	30.98

Table 2: Keyphrase Extraction results using Bagging + C4.5 decision tree

5. Conclusions

This paper presents a novel method to extract (keyphrase extraction), represent (hierarchical 3 layer representation) and visualize (tag cloud) the semantic content of BNs transmitted in the last 6 hours, in a way that allows users to skim its content and to jump to the top relevant news.

Although our best keyphrase extraction results (higher F1 measure) are obtained using the C4.5 decision tree, in the large

# Keyphrases Extracted	Features	#Keyphrases Identified	P	R	F1
30	base	8.6	28.67	32.32	30.38
30	base+f1	9.4	31.33	35.48	33.28
30	base+f2	9.3	31	35.34	33.03
30	base+f3	9	30	33.9	31.83
30	base+f4	9.1	30.33	34.55	32.3
30	base+f5	8.9	29.67	33.19	31.33
30	base+f1+f2	9.4	31.33	35.48	33.28
30	all prev.+f3	9.4	31.33	35.78	33.41
30	all prev.+f4	9.4	31.33	35.39	33.24
10	all prev.+f4	4.4	44	16.76	24.27
10	All	4.6	46	17.33	25.18
20	All	7.1	35.5	26.86	30.58
30	All	9.4	31.33	36.19	33.59
35	All	9.8	28	37.68	32.13
40	All	**10.4**	26	**40.18**	31.57

Table 3: Keyphrase Extraction results using Bagging + CART decision tree

majority of the results, the CART outperforms the C4.5 using, for instance, the baseline features and extracting 30 keyphrases.

We also observed that the CART model is more robust to the inclusion of additional features because it gradually improves its performance, while the C4.5 classifier performance fluctuates. Another conclusion that can be draw from the results is that the C4.5 perform better in the extraction of few keyphrases and CART produces better results during the extraction of large number of keyphrases.

The news anchor names, that are said when a studio reporter talks with a field reporter was identified as an extra source of noise to the keyphrase extraction for large keyphrases extraction ($>=$ 30 keyphrases). Thus, future work should address this issue. Another enhancement to the keyphrase extraction that we hope to include in the future is to determine automatically how many keyphrase candidates should be retrieved per news document, using a linear interpolation of the keyphrase candidate confidence values with the news size.

6. Acknowledgements

The authors would like to thank David Matos, Hugo Meinedo, and Isabel Trancoso for their help. Support for this research by FCT through the Carnegie Mellon Portugal Program and under FCT grant SFRH/BD/33769/2009. This work was also partially funded by European Commission under the contract FP7-SME-262428 EuTV, QREN SI IDT 2525, and SI IDT 5108. Support by FCT (INESC-ID multiannual funding) through the PIDDAC Program funds.

7. References

[1] H. Meinedo, A. Abad, T. Pellegrini, J. Neto, and I. Trancoso, "The L2F Broadcast News Speech Recognition System," in *Fala2010*, Vigo, Spain, 2010.

[2] D. Jones, F. Wolf, E. Gibson, E. Williams, E. Fedorenko, D. Reynolds, and M. Zissman, "Measuring the Readability of Automatic Speech-to-Text Transcripts," in *Proceedings of Eurospeech*, 2003, pp. 1585–1588.

[3] R. Amaral and I. Trancoso, "Topic Segmentation and Indexation in a Media Watch System," in *Interspeech 2008*. Brisbane, Australia: ISCA, 2008, pp. 2183–2186.

[4] B. Y.-L. Kuo, T. Hentrich, B. M. . Good, and M. D. Wilkinson, "Tag clouds for summarizing web search results," in *Proceedings of the 16th international conference on World Wide Web - WWW '07*. New York, USA: ACM Press, 2007, pp. 1203–1204.

[5] M. Tsagias, M. Larson, and M. de Rijke, "Term clouds as surrogates for user generated speech," *Proceedings of the 31st annual international ACM SIGIR'08*, p. 773, 2008.

[6] J. D. Cohen, "Highlights : Language- and Domain-Independent Indexing Terms for Abstracting Automatic," *English*, vol. 46, no. 3, pp. 162–174, 1995.

[7] H. P. Luhn, "A statistical approach to mechanized encoding and searching of literary information," *IBM J. Res. Dev.*, vol. 1, pp. 309–317, October 1957.

[8] G. Salton, C. S. Yang, and C. T. Yu, "A theory of term importance in automatic text analysis," Ithaca, NY, USA, Tech. Rep., 1974.

[9] L. Chien, "Pat-tree-based keyword extraction for chinese information retrieval," in *Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval*, ser. SIGIR '97. New York, NY, USA: ACM, 1997, pp. 50–58.

[10] G. Ercan and I. Cicekli, "Using lexical chains for keyword extraction," *Information Processing & Management*, vol. 43, no. 6, pp. 1705 – 1714, 2007, text Summarization.

[11] K. Zhang, H. Xu, J. Tang, and J. Li, "Keyword extraction using support vector machine," *Advances in Web-Age Information Management*, pp. 85–96, 2006.

[12] O. Medelyan and I. H. Witten, "Thesaurus based automatic keyphrase indexing," *Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries - JCDL '06*, p. 296, 2006.

[13] C. Zhang, H. Wang, Y. Liu, D. Wu, Y. Liao, and B. Wang, "Automatic Keyword Extraction from Documents Using Conditional Random Fields," *Information Systems*, vol. 3, 2008.

[14] O. Medelyan, V. Perrone, and I. H. Witten, "Subject metadata support powered by Maui," in *Proceedings of the 10th annual joint conference on Digital libraries - JCDL '10*. New York, New York, USA: ACM Press, 2010, p. 407.

[15] J. Neto, H. Meinedo, and M. Viveiros, "A media monitoring solution," in *ICASSP 2011 - Int. Conf. on Acoustics, Speech, and Signal Processing*, Prague, Czech Republic, 2011.

[16] I. Witten, G. Paynter, E. Frank, C. Gutwin, and C. Nevill-Manning, "KEA: Practical automatic keyphrase extraction," in *Proceedings of the fourth ACM conference on Digital libraries*. ACM, 1999, pp. 254–255.

[17] M. Abadia, L. Dias, and M. Malheiros, "Automatic Extraction of Keywords for the Portuguese Language," *Computational Processing of the Portuguese Language*, pp. 204–207, 2006.

[18] C. Martins, A. Teixeira, and J. Neto, "Dynamic language modeling for European Portuguese," *Computer Speech & Language*, vol. 24, no. 4, pp. 750–773, Oct. 2010.

[19] D. Guthrie, M. Hepple, and W. Liu, "Efficient minimal perfect hash language models," in *Proceedings of LREC'10*. Valletta, Malta: European Language Resources Association (ELRA), May 2010.

[20] L. Breiman, "Bagging predictors," *Machine learning*, vol. 24, no. 2, pp. 123–140, 1996.

[21] J. Quinlan, *C4. 5: programs for machine learning*. Morgan Kaufmann, 1993.

[22] L. Marujo, "Voting Combination of Sentences Splitting Classifiers Applied to Several Types of Texts," *Technical report*, 2010.

[23] J. Neto, H. Meinedo, R. Amaral, and I. Trancoso, "A system for selective dissemination of multimedia information resulting from the ALERT project," in *Proc. ISCA ITRW on Multilingual Spoken Document Retrieval*, Hong Kong, China, March 2003.

[24] F. Batista, I. Trancoso, and N. Mamede, "Automatic Recovery of Punctuation Marks and Capitalization Information for Iberian Languages," in *I Joint SIG-IL/Microsoft Workshop on Speech An Language Technologies for Iberian Languages*, Porto Salvo, Portugal, 2009, pp. 99–102.