Bimodal Corpora Terminology Extraction: Another Brick in the Wall

Claudiu Mihăilă*
University of Wolverhampton
Wulfruna Street
Wolverhampton WV1 1LY
claudiu.mihaila@info.uaic.ro

Dalila Mekhaldi
University of Wolverhampton
Wulfruna Street
Wolverhampton WV1 1LY
dalila.mekhaldi@wlv.ac.uk

Abstract

This paper presents a new study on automatic terminology extraction in the context of bimodal corpora that were generated from lectures and meetings. More specifically, the study aims to observe to which extent written text (discussed documents) and spoken text (dialogue transcript) share keywords. Using a hybrid terminology extraction approach, experiments have been performed on a collection of bimodal English corpora, including one scientific conference presentations corpus and two decision-making meetings corpora respectively. The evaluation results highlight a difference between keywords extracted from written text and from spoken text. Moreover, the obtained results emphasise the importance of considering text obtained from different modalities in order to generate rich and consistent keyword lists for bimodal corpora.

Keywords

bimodal corpus, keyword extraction, keyword similarity, spoken document, written document.

1 Introduction

Corpora, which are defined as large bodies of linguistic evidence composed of attested language use [9], are increasingly used in natural language processing field(NLP), such as in machine translation, automatic summarisation, etc. However, the diversity of NLP tasks that are based on corpora is proportional to the variousness of types of the latter. Therefore many corpora types have been distinguished [9], mainly monolingual corpora, parallel corpora and comparable corpora. Whilst most of the existing studies are based on these corpora types, bimodal corpora, which are originally built in the context of multimodal and multimedia applications, have recently appeared as a new corpora type that needs to be studied. Bimodal corpora are defined as pairs of texts either used or generated during multimedia events (e.g. meetings or lectures), each of them obtained through a different modality (e.g. text of written documents, speech transcript, text extracted from video recording frames).

From another side, terminology extraction (TE) consists in the identification and retrieval of the important lexical units, i.e. keywords, from a corpus. One of the challenges in TE is the definition of a keyword, which greatly depends on the context of the application. Other challenges are morphological and lexical variation of keywords, the consideration of single-word and multi-word keywords, etc. Once these challenges are addressed, the extracted keywords are particularly useful for conceptualising a knowledge domain or for supporting the creation of domain ontologies, due to their high specificity and low ambiguity. The accuracy in the identification of keywords may influence NLP tasks like analysis, understanding, generation, translation, automatic summarisation [12], and multiple-choice test item generation [5] of textual documents.

Terminology extraction from bimodal corpora, which is studied in this paper, represents the basis of many NLP and NLP-related fields of study. The task of multimodal document alignment framework, defined in [10], is based on the semantic similarity between spoken and written documents, which might benefit from their commonly extracted keywords. Moreover, bimodal terminology extraction might be useful for the disambiguation of words in the spoken corpora when they contain noise. Another use case of bimodal terminology extraction is the disambiguation of words between written and spoken corpora, e.g. the meaning of an abbreviation in a written document may be fully understood when its corresponding explicit phrase is used in speech, and vice-versa. Additionally, the bimodal terminology extraction could be exploited for the creation of digital libraries and multimedia archives, as well as management tasks related to their resources, including indexing, topic segmentation, automatic summarisation, etc.

This paper is structured as follows: section 2 highlights some of the recent works in terminology extraction field. In section 3, our system and the corpora on which this work was performed are described. Finally, the results of the evaluation are presented and discussed in section 4.

2 Existing studies

Research methods in TE are usually classified as linguistic, statistical or hybrid. Linguistic and statistical methods can be further subdivided into term-based (intrinsic) and context-based (extrinsic) [4].

*The author is also affiliated to the Faculty of Computer Science, "Al.I. Cuza" University of Iaşi, Romania

International Conference RANLP 2009 - Borovets, Bulgaria, pages 236–240

Linguistic approaches use the linguistic information associated to words at different levels to identify the keywords. Part-of-speech sequences [7] or morphological features [1], as well as boundary markers [3], are used in the recognition of terms.

Statistical approaches are based on statistical features such as word frequency, inverse document frequency, mutual information, etc. Statistical frameworks have been explored in the context of mutual bilingual terminology extraction from textual documents, showing that probabilistic models are a viable approach for incorporating alignment scores in automatic terminology [6].

3 Bimodal Corpora TE

Our work is based on three bimodal English corpora in different domains obtained from the recordings of one scientific conference and two decision-making meetings respectively. Each corpus is composed of multiple pairs of spoken text (speech transcripts) and written text. The spoken text corresponds to the manual transcription of the dialogue recording (with a total number of 59 805 words), whilst the written text is manually extracted from the printed documents that were discussed or presented (articles, slideshows, posters, etc.), with a total of 42 427 words. Our assumption is that, during meetings and lectures, speakers use roughly the same keywords that appear in the written documents.

Our bimodal terminology extraction system is based on a hybrid approach, combining both linguistic and statistical features. Thus, the main component of the system is statistical, enriched with shallow linguistic information. The system developed for this task comprises two main parts: corpus parsing and a keyword extraction module.

The written and spoken text files are first preprocessed using Machinese, a publicly available lemmatiser and POS-tagger[1]. The extracted linguistic features will be used by our system for the identification of candidate keywords, in the form of a morphological filter which only permits phrases having certain morphological structures to be considered by the user (noun, adjective, verb, etc.). However, in the current study the morphological category was not restricted and all types of candidates were considered. Moreover, an additional filter based on a stopword list is used to restrict the selection of semantically insignificant words.

The extraction module offers to the user a fully parametrised interface: it is possible to change easily the corpus, the files to analyse, the stopword list, the morphological categories to which the keywords belong, the methods used for scoring the keywords candidates and the number of extracted keywords. In addition to the TF method (term frequency), three other statistical scoring methods (TF·IDF, TF·IDF$_L$, and TF·IDF$_A$) are computed according to the formulae [8]:

$$TF \cdot IDF(t) = \frac{tf(t)}{df(t)} \quad (1)$$

$$TF \cdot IDF_L(t) = tf(t) \log\left(\frac{N}{df(t)}\right) \quad (2)$$

$$TF \cdot IDF_A(t) = 0.5 + \frac{0.5tf(t)}{max(tf(t))} \log\left(\frac{N}{df(t)}\right) \quad (3)$$

where N is the total number of documents in the corpus, $tf(t)$ is the term frequency in the current file, and $df(t)$ is the frequency of the term t in the corpus.

These metrics were chosen as starting point for this preliminary study due to their extensive use and simplicity. However, more complex formulae will be used in the future, which might generate better results.

4 Evaluation

The evaluation approach consists of three steps. First, a common keyword golden standard was manually created by a domain expert for each pair of spoken and written texts in each corpus. In addition to the main author, an expert has been involved in the task of creating the golden corpus. Second, an automatic extraction of keywords from each written and spoken text was performed and then generated keyword lists are evaluated against the golden standard. Finally, the obtained spoken and written keyword lists for each pair were compared in order to measure their overlap.

The four scoring methods (TF, TF·IDF, TF·IDF$_A$, and TF·IDF$_L$) were experimented in order to extract keywords from the test data. However, the TF·IDF$_L$metric has generated the best results, and thus was used for further evaluation. This is mainly due to the fact that this method normalises the score by using only the logarithm. On the other hand, TF·IDFand TF·IDF$_A$divide the term frequency by the document frequency or by the maximum term frequency in the texts of the corpus, which helps increase the score for the candidates which appear only in the current text. Since written texts in evaluated corpora have different contents one from the other, many keyword candidates obtain the maximum score. However, in this study we have limited the number of selected keywords to 10 per text (the same number of keywords has been extracted for the golden standard).

For each spoken/written text pair, a pair of spoken/written keyword lists is created by the system. Based on the manually extracted list of keywords (golden standard), the precision, recall, and F-measure are computed for the respective keyword lists. A match is counted if a term from the list returned by the system is present in the golden standard in the exact variation form. However, problems may occur in the case of orthographic variations (*oxidization* vs. *oxidisation*) or syntactic variations (*lung cancer* vs. *cancer of the lung*).

In order to measure the overlap between the extracted written and spoken keyword lists, the Jaccard similarity coefficient, in equation 4, has been used.

$$J = \frac{|K_w \cap K_s|}{|K_w \cup K_s|} \quad (4)$$

[1] http://www.connexor.eu/technology/machinese/

where K_w and K_s represent the written and, respectively, spoken keyword lists.

However, Jaccard coefficient does not show the rate of the actual keywords in the overlap between spoken/written keywords. Thus, an additional similarity coefficient was implemented, which is based on Jaccard index and considers the golden standard. This new coefficient, called KSim, is defined in equation 5.

$$KSim = \frac{|K_w \cap K_s \cap K_g|}{|K_w \cup K_s|} \quad (5)$$

where K_w, K_s, and K_g represent the written, spoken and golden corpus keyword sets respectively. This similarity coefficient gives the percentage of actual common keywords between the written and spoken texts.

In the following paragraphs, the results of the evaluation obtained for the three corpora are presented and analysed.

4.1 Scientific conference corpus

The first evaluated corpus corresponds to the material of a scientific conference in the domain of physics of particles, CHEP'04. More specifically, the corpus contains eight pairs of scientific papers (i.e. written text) with a total of 34 047 words, and their corresponding transcribed presentations (i.e. spoken text) with a total of 38 206 words. The duration of this corpus is 237 minutes. Each pair of written/spoken texts was cleaned from the data that might be noisy for our evaluation. Thus, the references section was removed from the written text due to the lack of relevant information (e.g. names of authors and editors, publishers, years), as well as non-textual information comprised in the documents (e.g. images, equations, charts). In the spoken text, the question-answering section was removed due to the non-clarity of the speech.

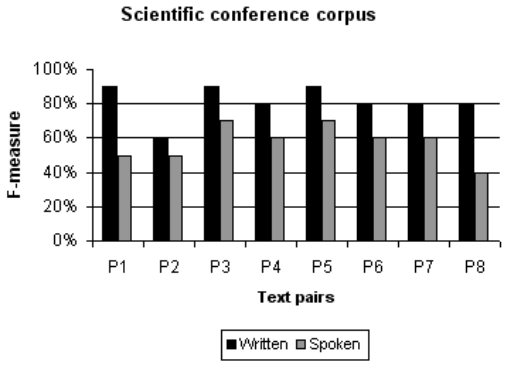

Fig. 1: *F-measures for written and spoken keywords using* $\text{TF}\cdot\text{IDF}_L$

The obtained values of the F-measure for this corpus are presented in Fig. 1. As shown, there is a divergence between the F values of the written and spoken keywords for each pair, with an average F value of 81.25% for the written texts, and 57.5% for spoken texts.

One possible reason that leads to a lower average score for the spoken texts is that they are more spontaneous and less formal than written texts. An additional reason is that synonym words and phrases are more likely to be used in speech, replacing the actual keywords in the written texts. Furthermore, additional factors such as the misunderstanding, spelling errors, or lack of expertise of the human transcriber in the domain have affected the quality of the transcription, and thus the accuracy of retrieved keywords. After a human annotator performed a manual correction to remove inaccuracies, the score for the spoken text increases considerably from 57.5% up to 62%.

After computing the individual scores for each spoken and written text, a comparison of the two lists of extracted keywords is performed in order to measure their similarity. Fig. 2 shows the results of the overlap using Jaccard and KSim coefficients.

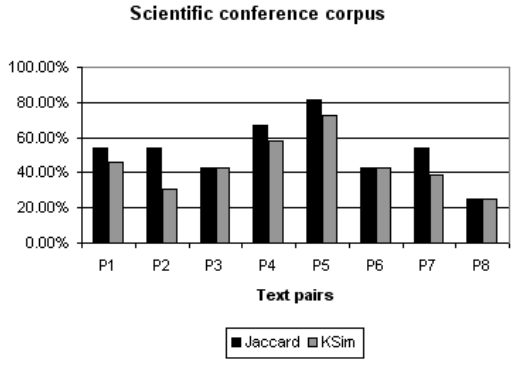

Fig. 2: *Comparison between written and spoken keywords using Jaccard and KSim coefficients*

As seen in the figure, Jaccard index has a relatively high value with an average of 52.59% (+ 5% if the effects caused by speech noise are ignored), showing that the two automatically extracted sets of keywords are quite similar [2]. The obtained average using KSim coefficient is 44.64% (which increases to 49.85% if the speech noise effect is ignored).

Amongst the eight pairs of texts, three have the same value for Jaccard and KSim coefficients, which means that the intersection for those three pairs contains only actual keywords. The average percentage of actual keywords inside the intersection between spoken and written texts, which is obtained by dividing the KSim value by Jaccard value, is 84.88% (86.25% if speech noise is ignored).

4.2 Movies corpus

The second corpus corresponds to a movie-club, a decision-making meeting about the movie to display, which lasts 48 minutes. This meeting has been recorded at the IDIAP Smart Meeting Room [11]. The generated corpus from this meeting is composed of one spoken text (12 563 words) and eight written documents including three articles, two slideshows and three posters (3 576 words). In order to have more than one spoken/written text pair, the eight written documents were categorised into three main sets according to their topics. Similarly, the spoken text has been divided into three topics. Subsequently,

our movie corpus was decomposed into three written/spoken text pairs, M1, M2 and M3.

The scores of the F-measure of extracted keywords for the movie corpus pairs in this corpus are shown in Fig. 3, where the average F value is 26.66% (+6.66% if ignoring speech noise) for both written and spoken texts, which is less than the scores achieved in the previous corpus.

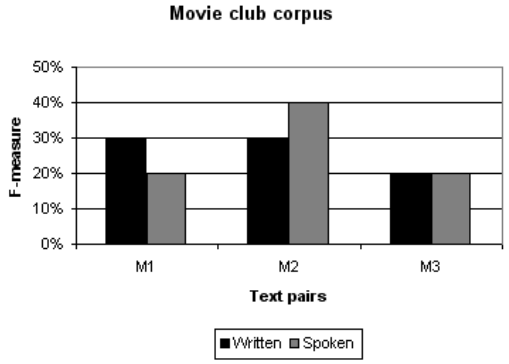

Fig. 3: *F-measures for written and spoken keywords using* $\text{TF}\cdot\text{IDF}_L$

Fig. 4 shows the results of the comparison of written and spoken keywords, where the average is 18% using Jaccard coefficient and 8% using KSim coefficient. Thus, the percentage of actual keywords in the intersection is 44.5%. Even with the manual correction, the scores for these two similarities remained unchanged, due to the fact that the correction did not add any common keywords for written and spoken texts.

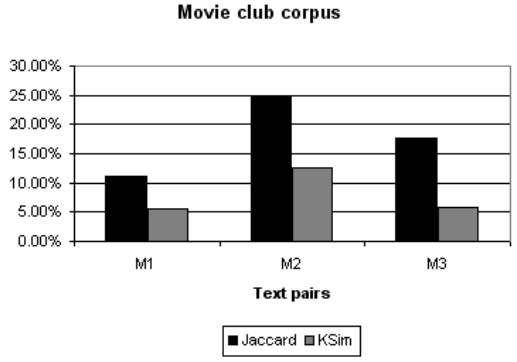

Fig. 4: *Comparison between written and spoken keywords using Jaccard and KSim coefficients*

The main reason for these low scores, either for individual keywords extraction or for keyword lists overlap for written/spoken text, is due to the nature of this corpus and the meeting scenario adopted by speakers. The speech is more spontaneous and informal compared to the previous academic corpus, and speakers interact more and rely less on the written documents.

4.3 Furniture corpus

The last corpus used in our evaluation is obtained from a furniture proposal meeting which aims at determining the type of furniture to buy. This meeting has been registered by ISSCO Research Group [13] and has a duration of 37 minutes. The corpus generated from this meeting is composed of one spoken text (9 036 words) and six written documents, three articles and three slideshows (4 804 words). Similarly to the previous corpus, this corpus was decomposed into three pairs, according to the topics of the written documents (F1, F2 and F3).

The F-measure values for this corpus (presented in Fig. 5) generated lower scores when compared to the scientific conference corpus, with an average of 43.33% (+ 6.66% when correcting noise effects) for written texts and 26.66% (+ 3.33% if correcting speech noise) for spoken texts.

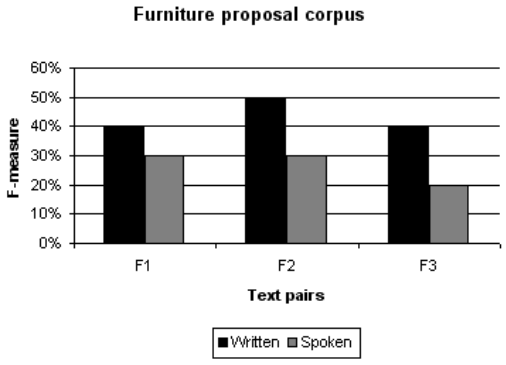

Fig. 5: *F-measures for written and spoken keywords using* $\text{TF}\cdot\text{IDF}_L$

From written texts perspective, the content of discussed documents is less formal than the academic content in the scientific corpus, containing more informal vocabulary such as pronouns and phrasal verbs. From spoken texts perspective, some possible causes of the low scores are due to the spontaneity of speech and high interaction between speakers.

When comparing both spoken and written keywords (Fig. 6), the similarity obtained is 30.56% (+ 3.17% without speech noise effect) using Jaccard coefficient, and 19.72% (+ 2.8% without speech noise effect) using KSim coefficient, with a percentage of actual keywords in the intersection of 64.54% (+ 2.4% without speech noise effect). These lower results compared to the scientific conference are due to the fact that half of the written documents in this corpus were not discussed by speakers.

4.4 Discussion

The evaluation performed on the various bimodal corpora has highlighted several important aspects about the relationship between the nature of the corpora and the used language style (academic, spontaneous, etc.) from one side, and the accuracy of their terminology extraction from the other side. As seen in the evaluation section, the written texts provide more rele-

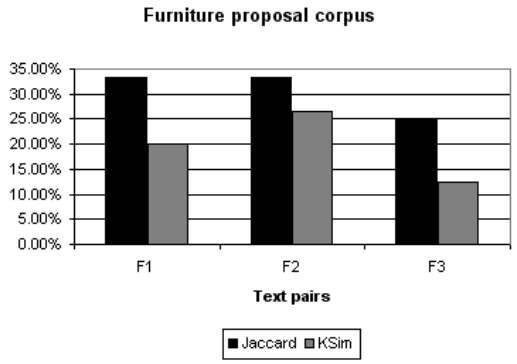

Fig. 6: *Comparison between written and spoken keywords using Jaccard and KSim coefficients*

vant keywords, when compared to spoken text. This is caused by the nature of spoken language which is more informal and rich with a variety of features such as colloquialisms, synonyms, phrasal verbs, variation in syntax and orthography (e.g. in the scientific conference corpus *GRID3* is sometimes referred to as *GRID*), in addition to other speech features such as repairs and word fragments, external noise, etc. Other factors that affected the accuracy of keywords extracting are the expertise of the transcriber in the domain. For instance, in the scientific conference corpus phrases like (*C++*, *C ++*, *Cplusplus*, and *C plus plus*) or (*daq*, *das*, and *data acquisition*) refer to the same concept respectively, but were transcribed differently.

From another side and by computing the symmetric difference for written/spoken pairs, it was revealed that there are cases where the spoken text provides additional keywords that are not present in the written text, and vice-versa. This emphasises the importance of using text extracted from different modalities for bimodal corpora terminology extraction.

5 Conclusions and future work

In this paper, we present a study about the extraction of keywords from bimodal corpora, generated from multimedia events mainly meetings and lectures, and composed of speech transcripts and the written texts of documents being discussed. The extracted written/spoken keyword lists for the respective corpora were evaluated against a manual golden standard. Later on, the overlap of each written/spoken keyword pair was measured by comparing the lists one to the other using specific similarity measures.

According to the obtained results, the system has generated more accurate keywords for written text compared to spoken text, mainly due to the nature of the latter and its ill-formed structure. This leads to the conclusion that written texts are more conducive to obtaining relevant keywords when compared to dialogue transcripts. Nevertheless, in some cases the latter has provided additional keywords that were not identified in the written text.

The extracted bimodal keywords might be used in the multimodal document alignment framework [10] defined in the context of lectures and meetings, in order to prune discovered thematic links between written documents and speech transcript of meetings. These bimodal keywords might be also useful for other tasks such as meeting indexing, searching and retrieval, etc.

As future work, other methods for terminology extraction will be used, which consider other features such as deeper linguistic information and other statistical features. From another side, the speech in the studied bimodal corpora was manually transcribed in order to avoid the negative effects of automatic speech recognition systems on this preliminary study (mainly due to the noise and error rate). However, in the future, automatically transcribed speech should be considered. Finally, focus will be put on bilingual bimodal corpora so as to verify if documents in different languages and obtained through different modalities can be aligned at keyword level.

6 Acknowledgements

We would like to express our gratitude to Alexandra Dobrinescu for her expertise and valuable comments provided during the golden standard compilation.

References

[1] S. Ananiadou. A methodology for automatic term recognition. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING94)*, pages 1034–1038, Kyoto, Japan, 1994.

[2] M. R. Anderberg. *Cluster Analysis for Applications*. Academic Press, London, 1973.

[3] D. Bourigault. *LEXTER, un Logiciel d'EXtraction de TERminologie. Application à l'acquisition des connaissances à partir de textes*. PhD thesis, École des Hautes Études en Sciences Sociales, 1994.

[4] D. Bourigault, C. Jacquemin, and M.-C. L'Homme. *Recent Advances in Computational Terminology*. John Benjamins, Amsterdam, 2001.

[5] L. A. Ha. *Advances In Automatic Terminology Processing: Methodology And Application In Focus*. PhD thesis, University of Wolverhampton, 2007.

[6] L. A. Ha, R. Mitkov, and G. Corpas. Mutual terminology extraction using a statistical framework. In *Proceedings of the 24th edition of the conference of the Spanish Society for Natural Language Processing (SEPLN 2008)*, Madrid, Spain, September 2008.

[7] J. S. Justeson and S. L. Katz. Technical terminology: some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, 3(2):259–289, 1996.

[8] C. D. Manning and H. Schütze. *Foundations Of Statistical Natural Language Processing*. MIT Press, Cambridge, Massachusetts, 1999.

[9] T. McEnery. *Corpus-based Language Studies: an Advanced Resource Book*. Routledge, London, 2005.

[10] D. Mekhaldi. *A Study on Multimodal Document Alignment: Bridging the Gap between Textual Documents and Spoken Language*. PhD thesis, University of Fribourg, 2006.

[11] D. Moore. The IDIAP smart meeting room. Technical report, IDIAP-Com, 2002.

[12] C. Orăsan, V. Pekar, and L. Hasler. A comparison of summarisation methods based on term specificity estimation. In *Proceedings of Fourth International Conference on Language Resources and Evaluation (LREC-04)*, pages 1037–1041, Lisbon, Portugal, 2004.

[13] A. Popescu-Belis and D. Lalanne. Detection and resolution of references to meeting documents. In S. Renals and S. Bengio, editors, *Proceedings of MLMI'06, Machine Learning for Multimodal Interaction'06*, pages 64–75, Berlin, 2006. Springer-Verlag.