Cross-Lingual Speech-to-Text Summarization

Elvys Linhares Pontes, Carlos-Emiliano González-Gallardo, Juan-Manuel Torres-Moreno, Stéphane Huet

Cross-Lingual Speech-to-Text Summarization

Elvys Linhares Pontes[1], Carlos-Emiliano González-Gallardo[1,2], Juan-Manuel Torres-Moreno[1,2], and Stéphane Huet[1]

[1] LIA, Université d'Avignon et des Pays de Vaucluse, Avignon, 84000 France
[2] École Polytechnique de Montréal, Montréal, Canada
[3] Universidade Federal do Ceará, Sobral, Ceará Brazil
elvys.linhares-pontes@alumni.univ-avignon.fr

Abstract. Cross-Lingual Text Summarization generates a summary in a language different from the language of the source documents. We propose a French-to-English cross-lingual transcript summarization framework that automatically segments a French transcript and analyzes the information in the source and the target languages to estimate the saliency of sentences. Additionally, we use a multi-sentence compression method to simultaneously compress and improve the informativeness of sentences. Experimental results show that our framework outperformed extractive methods using automatic sentence segmentation, even with transcription errors.

Keywords: Cross-Lingual Text Summarization · Multi-Sentence Compression · Automatic Speech Recognition

1 Introduction

Nowadays, audio data are part of daily life in the form of news, interviews and conversations, whether it is on the radio or on the Internet. Manual analysis of these data is very difficult because it requires a huge number of persons to analyze this information in the time available. One way to analyze and accelerate the data processing is Automatic Speech Summarization, which differs from the traditional Automatic Text Summarization task [24] because there are other problems to take into account like speech recognition errors, the lack of sentence boundaries, the wide range of sentence sizes, colloquialisms, and uneven information distributions [5, 3, 14, 23].

Cross-Lingual Text Summarization (CLTS) consists in summarizing a text where the summary language differs from the original document language. This application can be split in two sub-applications: Text Summarization (TS) and Machine Translation (MT). Each sub-application generates outcomes with errors and putting them one after the other may reduce the quality of cross-lingual summaries because of the accumulation of errors.

Recent works [25, 27, 30] analyzed the information of a document in both languages (source and target languages) to extract more details and identify the most relevant sentences. Following this idea, we propose a framework to realize French-to-English CLTS of transcript documents. In a nutshell, our approach

first automatically segments a French transcript document and translates it into English using Google Translate. Then, we estimate the sentence relevance based on the information they contain in French and English. Similar English sentences are compressed to generate a unique, short, and informative compression. Finally, the cross-lingual summary is composed of the compression of the most relevant sentences without redundancy.

The rest of this paper is organized as follows: we make an overview of relevant work for CLTS methods and Automatic Speech Summarization in Section 2. Next, we detail our approach in Section 3. The experimental setup and results are discussed in Sections 4 and 5, respectively. Finally, we provide our conclusion and some final comments in Section 6.

2 Related Work

If TS has reached a stage of maturity with well-established methods, Speech-to-Text Summarization and Cross-Lingual Summarization have their own challenges.

Speech-to-text summarization has to face three main problems: documents are not segmented into sentences, they may contain disfluencies, specific to the oral language, or they are subject to misrecognized words when using Automatic Speech Recognition (ASR). Nevertheless, it can benefit from acoustic and prosodic cues, or information about the role of speakers to determine the importance or the structure of an utterance. McKeown *et al.* [16] showed how the summarization approaches used in TS can be adapted to this speech-to-text task. They focused on two types of spoken sources, broadcasts news and meetings, taking advantage of acoustic, prosodic, lexical, and structural features to detect speakers' turns and overcome the difficulties that are present in spoken language.

Mrozinski *et al.* applied an extractive summarization approach over broadcast news stories and conference lectures [17]. In a first step, they performed sentence segmentation of the transcripts using word-based and class-based statistical language models; then during the summarization phase they selected the highest scoring sentences based on a combination of word significance score, confidence score, and linguistic likelihood.

Rott & Cerva divided their summarization system in three steps: automatic speech recognizer, syntactic analyzer, and text summarizer [22]. Sentence Boundary Detection was performed during the syntactic analysis, where they identified phrases in the recognized text using the Syntactic Engineering Tool (SET) [8]. Text summarization was performed using a TF-IDF method which selects the most informative phrases.

With regard to Cross-Lingual Summarization (CLTS), it has to deal with errors introduced by MT. CLTS was originally addressed as two separate tasks, making the information analysis in only one of the two languages [10, 20], which produces an early or a late translation scheme. In the early translation approach, the first step is to translate the source documents into the target language, the second step is to summarize the translated documents using only information of

the translated sentences. The late translation approach does the reverse; first it aims to summarize the documents in its source language and then it translates them to the target language.

Further studies have considered translation quality and the information in both languages in order to generate correct and informative cross-lingual summaries. A Support Vector Machine (SVM) regression method was developed by Wan *et al.* in order to predict the translation quality using parse features to produce English-to-Chinese CLTS [26]. These translation quality scores were used in addition to relevance scores to select the sentences for the summaries. In order to take into account both language sides for establishing the similarity between sentences, Wan introduced two graph-based summarization methods, SimFusion and CoRank [25] . The SimFusion method was inspired by the PageRank algorithm [21] in order to calculate the relevance of sentences, where the weight arcs are defined by the linear combination of the cosine similarity of sentences in English and Chinese. The CoRank method simultaneously ranks English and Chinese sentences by incorporating mutual influences between them. The relevance of a sentence is defined by its similarity with other sentences in each language separately and between languages.

If extractive approaches are mainstream for CLTS, a few studies have been done to propose abstractive summarization. First, Yao *et al.* took advantage of statistical MT systems that are usually phrase-based to define relevance scores at the phrase level [28]. These scores were used to select and compress sentences simultaneously. Zhang *et al.* used Predicate-Argument Structures (PAS) to identify a set of concepts and facts in the source side, and their counterparts in the target side with the help of an alignment method [30]. The relevance of concepts and facts are estimated using the CoRank algorithm [25], while summaries were produced by fusing the most relevant source-side PAS elements considering their translation quality. Finally, a French-to-English cross-lingual abstractive summarization approach was proposed in [12]. This CLTS system combined multi-sentence and sentence compression methods in order to produce informative cross-lingual summaries.

3 Our Proposition

French-to-English CLTS aims to generate an English short summary that describes the main information from a French transcript document. Following the CLTS approach proposed by Linhares Pontes *et al.* [12], we analyze documents in the source and the target languages to select the most relevant sentences. Then, we create clusters of similar sentences to independently analyze the subjects of a document. In order to compress and to improve the informativeness of the summarization, we compress the clusters composed of two or more similar sentences. Finally, the summary is composed of the compression of the most relevant sentences without redundancy.

The following subsections present the architecture of our system.

3.1 Ranking Sentences

The CoRank method jointly ranks sentences in both languages by assimilating mutual influences between them. We first translate the French sentences into English using the Google Translate system, then we use the CoRank method to estimate the informativeness of sentences (more details in [25]).

3.2 Multi-Sentence Compression

Following the idea proposed in [12], we consider the similarity in both languages to create clusters of similar sentences. Then, we use the Stanford CoreNLP tool [15] with jMWE [9] to detect Multi-Word Expressions in the English side, while the corresponding expressions were deduced on the French side with the help of the Giza++ alignment tool [19]. Among several state-of-the-art Multi-Sentence Compression (MSC) methods [4, 1, 18, 13], we use Linhares Pontes *et al.*'s approach [12] to generate a compressed sentence from each cluster of similar sentences. This system builds compressions controlled by the presence of keywords, to increase informativeness, and 3-grams, to ensure grammaticality. Finally, the sentences of each cluster are replaced by their compression in the document.

4 Experimental Setup

We use the early translation, the late translation, and the CoRank methods [25] to evaluate the performance of our system. We adapted the SimFusion method to create the early and late translation methods. The early translation method only considers the similarity in the target language and the late translation method only considers it in the source language.

In order to avoid the generation of short compressions, we only compress sentences with at least 10 words. All systems produce summaries containing a maximum of 250 words and without redundant sentences. We consider two sentences to be similar/redundant if they have a cosine similarity value bigger than 0.5.

4.1 Dataset

The MultiLing Pilot 2011 dataset [6] is a collection of WikiNews English texts that were translated into Arabic, Czech, English, French, Greek, Hebrew and Hindi languages by native speakers. Each language version of this dataset has 10 topics where each topic is composed of 10 source texts and 3 reference summaries. Each summary has a maximum of 250 words. In this work, we use the French version of the MultiLing Pilot 2011 dataset as source language and the corresponding English version as the target language.

To our knowledge, no work has been done regarding cross-lingual summarization of transcripts generated by an ASR system. We believe this to be a good challenge given the difficulties brought by ASR transcripts. For this reason we

wanted to explore this less controlled scenario and analyze the repercussions over the cross-lingual text summarization of two main problems of ASR transcripts: transcription errors and the lack of sentences.

4.2 Transcription error simulation

Automatic transcription performance is normally compared against one or more references using Word Error Rate (WER). This measure considers three different errors and calculates a general value indicating the quality of the transcript; the lower the value (closer to zero), the higher its quality. The three errors considered by WER (Equation 3) are deletions, insertions, and substitutions:

$$\text{WER} = \frac{D + I + S}{N} \tag{1}$$

where D corresponds to the number of deletions, I to the number of insertions, S to the number of substitutions and N to the number of words in the reference. An ASR transcript carries all three errors at different ratios; for this controlled scenario we simulated in an isolated way each error to observe how each of them affects the performance of cross-lingual speech-to-text summarization.

We approximated WER by simulating the errors produced by ASR systems in a straightforward approach. The deletion error dataset (ASR_D) was created by choosing m words of each document randomly and by deleting them. Concerning the substitution error dataset (ASR_S), for each document we first selected a set $Y = \{y_1, ..., y_m\}$ of words randomly, then for each word w_i of the document a randomly generated decision value $v_i \in [0, 1]$ was calculated; if v_i happened to be greater than a given threshold $t = 0.5$, then w_i was replaced by y_j, this cycle was repeated until all words y_j in Y where picked. The insertion error dataset (ASR_I) followed the same procedure as ASR_D but instead of replacing w_i by y_j, y_j was placed after w_i. For all three error datasets m was calculated as:

$$m = \text{WER} \times N \tag{2}$$

where N corresponds to the length (number of words) in each original document and WER was fixed to 0.15.

4.3 Automatic Segmentation

Common ASR transcripts have no punctuation, which further complicates NLP tasks like automatic summarization. We simulated the lack of punctuation by deleting all punctuation signs inside the MultiLing Pilot 2011 French dataset (ASR_NO) and the datasets with induced transcription errors (ASR_D, ASR_S, ASR_I); then we automatically restored them. This task is known as Sentence Boundary Detection (SBD).

To restore the punctuation within the corpus we followed the best model reported by González-Gallardo & Torres-Moreno in [7]. This approach targets the segmentation problem as a classification one. It uses a Convolutional Neural Network (CNN) with Subword-level Information Vectors [2] to predict if the centered

word (w_i) within a window $W = \{w_{i-(m-1)/2}, ..., w_{i-1}, w_i, w_{i+1}, ..., w_{i+(m-1)/2}\}$ corresponds to a sentence border or not.

The hidden architecture of the CNN consists of two convolutional layers with a valid padding and a stride value of one, followed by a max pooling layer and three fully connected layers with a dropout layer attached at the end. The outputs of all convolutional, max pooling and fully connected layers have a RELU activation function. The CNN was trained with a 380M words of the French Wikipedia.

Table 1 presents the automatic evaluation performed over the unpunctuated datasets. As seen from the "no boundary" class (NO_BOUND), the method has a really good performance (over 0.95 for all metrics), no matter of the type of transcription errors. Given the unbalanced nature of the data this is an expected behavior. Nevertheless for the "boundary" class (BOUND) the performance drops when trying to segment the noisy transcripts. The worst scenario corresponds to the dataset with substitution errors (ASR_S), where precision and recall present relative drops of 34% and 17% against ASR_NO.

Table 1. Results of Sentence Boundary Detection over the ASR datasets.

Dataset	Class	Precision	Recall	F1
ASR_NO	NO_BOUND	0.971	0.986	**0.978**
	BOUND	0.840	0.721	**0.776**
ASR_D	NO_BOUND	0.966	0.963	0.965
	BOUND	0.654	0.673	0.663
ASR_I	NO_BOUND	0.960	0.956	0.958
	BOUND	0.592	0.616	0.604
ASR_S	NO_BOUND	0.958	0.950	0.954
	BOUND	0.554	0.600	0.576

4.4 Automatic Evaluation

Automatic evaluation relies on comparing the information contained in the candidate summary against one or more reference summaries or the source document. The ROUGE [11] measure developed by Lin *et al.* compares the differences between the distribution of words of the candidate summary and a set of reference summaries. The comparison is made splitting into n-grams both the candidate and the reference to calculate their intersection. Standard n-grams values for ROUGE are unigrams and bigrams, both expressed as:

$$\text{ROUGE} - n = \frac{\sum_{n-grams} \in \{Sum_{can} \cap Sum_{ref}\}}{\sum_{n-grams} \in Sum_{ref}}, \tag{3}$$

where n is the n-gram order, Sum_{can} the candidate summary and Sum_{ref} the reference summary.

A third common ROUGE-n variation is ROUGE-SUγ. This ROUGE-2 variation takes into account skip units (SU) $\leq \gamma$. We considered the ROUGE-1, -2 and -SU4 measures in order to evaluate and compare our system.

5 Experimental Evaluation

Table 2 shows the ROUGE scores for each version of the MultiLing Pilot dataset. Our method outperformed the other methods for the original, ASR_NO and ASR_S dataset versions, while the CoRank method obtained the best results for other versions. As we expected, the ASR errors, introduced at the word or segmentation levels, reduced the performance of systems.

Table 2. ROUGE f-measure results for French-to-English MultiPilot 2011 dataset.

Dataset	Algorithms	ROUGE-1	ROUGE-2	ROUGE-SU4
Original	Early translation	0.4165	0.1021	0.1607
	Late translation	0.4142	0.1023	0.1589
	SimFusion	0.4173	0.1035	0.1606
	CoRank	0.4628	0.1324	0.1932
	Our proposition	**0.4724**	**0.1369**	**0.1962**
ASR_NO	Early translation	0.4115	0.0967	0.1567
	Late translation	0.4115	0.0992	0.1568
	SimFusion	0.4140	0.0981	0.1589
	CoRank	0.4608	0.1267	0.1891
	Our proposition	**0.4705**	**0.1336**	**0.1922**
ASR_D	Early translation	0.4160	0.0950	0.1566
	Late translation	0.4076	0.0896	0.1504
	SimFusion	0.4142	0.0914	0.1547
	CoRank	**0.4666**	**0.1192**	**0.1860**
	Our proposition	0.4474	0.1053	0.1711
ASR_I	Early translation	0.4027	0.0827	0.1481
	Late translation	0.3933	0.0828	0.1420
	SimFusion	0.3987	0.0814	0.1452
	CoRank	**0.4504**	**0.1089**	**0.1770**
	Our proposition	0.4481	0.1067	0.1744
ASR_S	Early translation	0.4080	0.0847	0.1495
	Late translation	0.4038	0.0834	0.1463
	SimFusion	0.4077	0.0848	0.1505
	CoRank	0.4206	0.0921	0.1584
	Our proposition	**0.4445**	**0.1072**	**0.1718**

We analyzed the original dataset results as a reference to compare the performance of the systems with other dataset versions. The joint analysis of both languages generated better results. The analysis of the similarity in both languages and cross-language increased the results considerably. Finally, the addition of the compression of similar sentences to these multiple analysis of similarities achieved the best results.

The automatic segmentation process may split long sentences in two or more short sentences that can be more or less relevant to the document. In addition, these sentences are more likely to contain grammatical errors. However, the segmentation errors had little impact on the performance of systems (ASR_NO in tables 1 and 2).

The low performance of automatic segmentation process to identify sentence bound combined with automatic speech recognition errors reduced the performance of all systems (ASR_D, ASR_I and ASR_S in Tables 1 and 2). These errors modified the structure of sentences causing large translation errors and changing the meaning of some sentences. The CoRank method achieved the best results for the deletion and insertion dataset versions; however, poor results were obtained for the substitution errors. Our approach was more stable for all kinds of ASR errors by generating cross-lingual summaries with similar ROUGE scores.

All in all, the joint analysis of information in both languages and MSC generate more informative cross-lingual summaries. Our segmentation process kept a good quality of all summaries, i.e. all systems generated summaries with similar ROUGE scores to the original dataset. The addition of ASR errors reduced the quality of summaries of all systems because of translation and meaning errors. Our approach generated cross-lingual summaries with similar ROUGE scores for the dataset with ASR errors while the CoRank method achieved unstable results depending on the kind of errors.

6 Conclusion

We have proposed a compressive method to generate cross-lingual transcript summaries. Our framework analyzes a transcript document in French and English languages to identify the relevant information and compress similar sentences to increase the informativeness of summaries. The simulated ASR errors showed to have an impact on the performance of all systems; nevertheless, our approach achieved the best results for the original, ASR_NO and ASR_S dataset versions. Contrary to the CoRank method, our approach attained stable results for all kinds of ASR errors.

In future work, we plan to realize a manual evaluation to measure the grammaticality and the informativeness of the cross-lingual summaries. We will also use a language model or neural networks to correct grammatical errors [29] generated by ASR in order to improve the quality of transcripts and, consequently, the quality of summaries.

Acknowledgement

This work was granted by the European Project CHISTERA-AMIS ANR-15-CHR2-0001.

References

1. Banerjee, S., Mitra, P., Sugiyama, K.: Multi-document Abstractive Summarization Using ILP Based Multi-sentence Compression. In: 24th International Conference on Artificial Intelligence (IJCAI). pp. 1208–1214. IJCAI'15 (2015)
2. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. arXiv preprint arXiv:1607.04606 (2016)
3. Christensen, H., Gotoh, Y., Kolluru, B., Renals, S.: Are extractive text summarisation techniques portable to broadcast news? In: IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU). pp. 489–494 (2003)
4. Filippova, K.: Multi-sentence compression: Finding shortest paths in word graphs. In: COLING. pp. 322–330 (2010)
5. Furui, S., Kikuchi, T., Shinnaka, Y., Hori, C.: Speech-to-text and speech-to-speech summarization of spontaneous speech. IEEE Transactions on Speech and Audio Processing **12**(4), 401–408 (July 2004)
6. Giannakopoulos, G., El-Haj, M., Favre, B., Litvak, M., Steinberger, J., Varma, V.: TAC2011 multiling pilot overview. In: 4th Text Analysis Conference TAC (2011)
7. González-Gallardo, C.E., Torres-Moreno, J.M.: Sentence Boundary Detection for French with Subword-Level Information Vectors and Convolutional Neural Networks. ArXiv (Feb 2018)
8. Kovář, V., Horák, A., Jakubíček, M.: Syntactic analysis using finite patterns: A new parsing system for czech. In: Language and Technology Conference. pp. 161–171. Springer (2009)
9. Kulkarni, N., Finlayson, M.A.: jMWE: a Java toolkit for detecting multi-word expressions. In: Workshop on Multiword Expressions: from Parsing and Generation to the Real World (MWE). pp. 122–124 (2011)
10. Leuski, A., Lin, C.Y., Zhou, L., Germann, U., Och, F.J., Hovy, E.: Cross-lingual C*ST*RD: English Access to Hindi Information **2**(3), 245–269 (Sep 2003)
11. Lin, C.Y.: ROUGE: A Package for Automatic Evaluation of Summaries. In: Workshop Text Summarization Branches Out (ACL'04). pp. 74–81 (2004)
12. Linhares Pontes, E., Huet, S., Torres-Moreno, J.M., Linhares, A.C.: Cross-language text summarization using sentence and multi-sentence compression. In: Natural Language Processing and Information Systems. pp. 467–479. Springer International Publishing, Cham (2018)
13. Linhares Pontes, E., Huet, S., Gouveia da Silva, T., Linhares, A.C., Torres-Moreno, J.M.: Multi-sentence compression with word vertex-labeled graphs and integer linear programming. In: TextGraphs-12: the Workshop on Graph-based Methods for Natural Language Processing. ACL (2018)
14. Linhares Pontes, E., Torres-Moreno, J.M., Linhares, A.C.: Lia-rag: a system based on graphs and divergence of probabilities applied to speech-to-text summarization. In: Addendum, M..P. (ed.) Multiling CCCS (2015)
15. Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., McClosky, D.: The Stanford CoreNLP natural language processing toolkit. In: 52nd Annual Meeting of the Association for Computational Linguistics (ACL): System Demonstrations. pp. 55–60 (2014)

16. McKeown, K., Hirschberg, J., Galley, M., Maskey, S.: From text to speech summarization. In: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). vol. 5, pp. v–997 (2005)
17. Mrozinski, J., Whittaker, E.W., Chatain, P., Furui, S.: Automatic sentence segmentation of speech for automatic summarization. In: IEEE International Conference on Acoustics Speech and Signal Processing Proceedings (ICASSP) (2006)
18. Niu, J., Chen, H., Zhao, Q., Su, L., Atiquzzaman, M.: Multi-document abstractive summarization using chunk-graph and recurrent neural network. In: IEEE International Conference on Communications, ICC. pp. 1–6 (2017)
19. Och, F.J., Ney, H.: A systematic comparison of various statistical alignment models. Computational Linguistics **29**(1), 19–51 (2003)
20. Orasan, C., Chiorean, O.A.: Evaluation of a Cross-lingual Romanian-English Multi-document Summariser. In: 6th International Conference on Language Resources and Evaluation (LREC) (2008)
21. Page, L., Brin, S., Motwani, R., Winograd, T.: The pagerank citation ranking: Bringing order to the web. In: 7th International World Wide Web Conference. pp. 161–172. Brisbane, Australia (1998)
22. Rott, M., Červa, P.: Speech-to-text summarization using automatic phrase extraction from recognized text. In: International Conference on Text, Speech, and Dialogue (TSD). pp. 101–108. Springer (2016)
23. Taskiran, C.M., Pizlo, Z., Amir, A., Ponceleon, D., Delp, E.J.: Automated video program summarization using speech transcripts. IEEE Transactions on Multimedia **8**(4), 775–791 (2006)
24. Torres-Moreno, J.M.: Automatic Text Summarization. Wiley & Sons (2014)
25. Wan, X.: Using bilingual information for cross-language document summarization. In: ACL. pp. 1546–1555 (2011)
26. Wan, X., Li, H., Xiao, J.: Cross-language document summarization based on machine translation quality prediction. In: ACL. pp. 917–926 (2010)
27. Wan, X., Luo, F., Sun, X., Huang, S., Yao, J.g.: Cross-language document summarization via extraction and ranking of multiple summaries. Knowledge and Information Systems (2018)
28. Yao, J., Wan, X., Xiao, J.: Phrase-based compressive cross-language summarization. In: EMNLP. pp. 118–127 (2015)
29. Yuan, Z., Briscoe, T.: Grammatical error correction using neural machine translation. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 380–386. Association for Computational Linguistics (2016)
30. Zhang, J., Zhou, Y., Zong, C.: Abstractive cross-language summarization via translation model enhanced predicate argument structure fusing. IEEE/ACM Trans. Audio, Speech & Language Processing **24**(10), 1842–1853 (2016)