Speech Communication 88 (2017) 83–95

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

Improving the automatic segmentation of subtitles through conditional random field

Aitor Álvarez[a,*], Carlos-D. Martínez-Hinarejos[b], Haritz Arzelus[a], Marina Balenciaga[a], Arantza del Pozo[a]

[a] Human Speech and Language Technology Group, Vicomtech-IK4, San Sebastian, Spain
[b] Pattern Recognition and Human Language Technologies Research Center, Universitat Politècnica de València, Spain

ARTICLE INFO

Article history:
Received 18 January 2016
Revised 5 December 2016
Accepted 23 January 2017
Available online 29 January 2017

Keywords:
Automatic subtitling
Subtitle segmentation
Pattern recognition
Machine learning

ABSTRACT

Automatic segmentation of subtitles is a novel research field which has not been studied extensively to date. However, quality automatic subtitling is a real need for broadcasters which seek for automatic solutions given the demanding European audiovisual legislation. In this article, a method based on Conditional Random Field is presented to deal with the automatic subtitling segmentation. This is a continuation of a previous work in the field, which proposed a method based on Support Vector Machine classifier to generate possible candidates for breaks. For this study, two corpora in Basque and Spanish were used for experiments, and the performance of the current method was tested and compared with the previous solution and two rule-based systems through several evaluation metrics. Finally, an experiment with human evaluators was carried out with the aim of measuring the productivity gain in post-editing automatic subtitles generated with the new method presented.

1. Introduction

Subtitles have acquired great relevance within the audiovisual community during the last years, mainly after the adoption of the new audiovisual directives (Article 7 of the Audiovisual Media Services Directive[1]) of the European Parliament and of the Council in March of 2010. This legislation regulates the rights of people with a visual or hearing disability, and moved member states to take the necessary measures to guarantee that the services of audiovisual providers under their jurisdiction are gradually more and more accessible by means of sign-language, audio-description, easily menu navigation and subtitling.

Given the new audiovisual legislation, broadcasters and subtitling companies are seeking automatic solutions to be more productive than with the traditional manual subtitling. At the same time, disability organisations are pushing for both quantity and quality of subtitles, in order to not only increment the percentage of subtitling in the TV and the Internet, but also request quality subtitles. As a result, the demand of automatic solutions for quality subtitling has grown fast in the audiovisual community.

Several parameters take part in the definition of what the quality of subtitles is (Álvarez et al., 2015). Apart from features related to subtitle layout, duration and text editing, subtitling segmentation is one of the most relevant, as it was demonstrated in Rajendran et al. (2013), a study whose aim was to verify whether a correct text chunking in subtitles had an impact on both comprehension and reading speed using human evaluators. Even though important differences were not found in terms of comprehension, they demonstrated that a correct segmentation by phrase or by sentence significantly reduced the time needed to read subtitles. Furthermore, the strong need for proper segmentation is supported by the psycholinguistic literature on reading (D'Ydewalle and Rensbergen, 1989), where the consensual view is that subtitle lines should end at natural linguistic breaks to improve readability and reduce cognitive effort produced by poorly segmented text lines (Perego et al., 2010).

In this article, a new method based on probabilistic Conditional Random Field is applied to the field of automatic subtitling segmentation for Basque and Spanish languages. This work is a continuation of the previous research presented in Álvarez et al. (2014a), in which Support Vector Machine and Logistic Regression classifiers were employed for the subtitling segmentation task in the Basque language. In the present study, the same Basque corpus was used in order to compare the performance using the new classification method. In addition, the work has been extended to the

* Corresponding author.
E-mail address: aalvarez@vicomtech.org (A. Álvarez).

[1] http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32010L0013&from=EN.

http://dx.doi.org/10.1016/j.specom.2017.01.010
0167-6393/

Spanish language. It allowed us to confirm that the new classification method employed was valid for different types of corpora and languages. Given that the results obtained in Álvarez et al. (2014a) by the Support Vector Machine and Logistic Regression classifiers were very close due to its similar nature, in this work the performance of Support Vector Machine and Conditional Random Field were compared for both languages, leaving out the Logistic Regression classifier. Besides these statistical techniques, two rule-based methods were selected as baseline systems, such as the Chink-Chunk and Counting Character methods. Both rule-based methods were modified and adapted to the subtitle segmentation task by including additional information related to the maximum amount of characters allowed per line, speaker change and timing issues.

The results achieved proved that Conditional Random Fields outperformed clearly the Support Vector Machine based technique in terms of accuracy and computation time for both languages, whilst the rule-based Chink-Chunk and Counting Character methods obtained the worst results.

The article is structured as follows. Section 2 describes existing work on automatic subtitling and segmentation. Section 3 presents the rule-based baseline systems, whilst Section 4 looks at the Conditional Random Field method and how it fits the subtitle segmentation task. Section 5 describes the methodology we designed and implemented to build the new classification approach. Section 6 presents the experimental framework and the evaluation metrics. Section 7 summarizes the evaluation results and the performance comparison between the methods based on Conditional Random Field and Support Vector Machine. Section 8 shows the human performance results in segmentation correction for two options of obtaining draft segmentations. Finally, Section 9 draws conclusions and describes future work.

2. Related work in subtitle segmentation

Automatic segmentation of subtitles is a novel line of research which has not been studied extensively up to the present. To date, most of the automatic subtitling solutions have not been capable of generating syntactic and semantically coherent breaks for quality segmentation and, thus, segmentation is mainly performed considering the maximum number of characters allowed per line or through manual intervention.

The subtitle segmentation is similar to other segmentation techniques that are necessary for many Natural Language Processing tasks: Dialogue Act segmentation (Warnke et al., 1997), sentence boundary detection for Text-to-Speech (Read and Cox, 2007), or punctuation mark enriched speech recognition output (Beeferman et al., 1998). These applications can be used to improve the source data for other tasks such as summarisation (Mrozinski et al., 2006) or Machine Translation (Matusov et al., 2006). Most of these works employ lexical features such as the word sequence (obtained from the speech transcription or recognition) and Part-Of-Speech (POS) labels (Read and Cox, 2007); apart from that, signal features obtained from the speech (e.g., pause durations (Gotoh and Renals, 2000)), or prosodic features (Shriberg et al., 2000) are frequently employed to complement the data obtained from lexical features. It is usual to employ combination of these different features to obtain a more accurate result.

With respect to the models used for the segmentation, there is a wide variety of them applied for this task: Hidden Markov Models (HMM) (Liu et al., 2004), Hidden Event Language Models (HELM) (Güz et al., 2009), Maximum Entropy models (Roark et al., 2006), Neural Networks (Warnke et al., 1997), or Conditional Random Field (CRF) (Oba et al., 2006). Many works employ different models for different sets of features and combine the results (Liu et al., 2005; Oba et al., 2006; Güz et al., 2009) to obtain a more accurate segmentation. Finally, although majority of research has been done on English corpora, other languages have been used in the segmentation problem, like Japanese (Kawahara et al., 2007), German (Gallwitz et al., 2002), or Portuguese (Batista et al., 2010), among others.

When looking at subtitle segmentation, few works have been carried out in the field of automatic segmentation, like the study presented in Álvarez et al. (2014a), where automatic subtitle segmentation was treated as a machine learning problem. In this previous work, Support Vector Machine and Logistic Regression classifiers were built over a Basque corpus consisting of TV cartoon programs and subtitles generated by professional subtitlers. To this end, subtitles with correct or incorrect segmentation were divided into two classes. Positive (correct) feature vectors were extracted from professionally-created subtitle data and contained the segmentation marks found in the corpus, whilst negative (incorrect) vectors were generated by manually inserting improper segmentation marks. Classifiers where then trained on balanced sets formed with these two types of vectors and employed for the segmentation task. The feature vectors were composed by 4 types of characteristics related to timing, number of characters, speaker change and a perplexity value given by a language model built over the training data. During decoding, an iterative algorithm was in charge of generating all the possible candidates for a break at each iteration. These candidates included sequences of consecutive words that did not exceed the maximum allowed length in characters before and after segmentation points. Feature vectors were then computed from these candidates and measured against the machine-learned classifiers and optimal candidates selected according to the obtained score. Similar performance was obtained for the two classifiers under evaluation. In the case of the SVM, it achieved a precision of 82.0% and a recall of 69.0%, with an average F1-score of 74.7%.

However, through this method, only the possible segmentation points were estimated, without distinguishing between different types of breaks. This implies considering line-breaks and subtitle-breaks, which is a critical information, to automatically generate the final subtitles correctly. It has to be noted that not all the subtitles have to have the same number of lines; there can be subtitles with just one line combined with others with two lines depending on the content and the segmentation rules. It is therefore critical to differentiate between line-breaks and subtitle-breaks. Finally, the computation time needed to generate all the candidates and select the optimal ones in the method presented in Álvarez et al. (2014a) was inefficient for a real application. Computing the iterative algorithm for one hour of content with 900 subtitles it took four hours of processing time on an Intel(R) Xeon(R) 2.00GHz and 32GB based server.

The rest of works in the literature regarding subtitle segmentation are focused on comparing the comprehension and reading speech in live-respoken subtitles segmented in a correctly and poorly manner (Rajendran et al., 2013), measuring the impact of arbitrary segmented subtitles on readers (Perego et al., 2010), and on studying the way line-breaking is commonly performed (Perego, 2008). None of these three last works include technology to automatically create and segment subtitles.

3. Rule-based methods for segmentation

This section details the rule-based Counting Character and Chink-Chunk methods adapted to the task of automatic subtitle segmentation.

3.1. Counting Character method

Nowadays, automatic segmentation is mainly performed considering only the maximum number of characters allowed per