Impact of automatic segmentation on the quality, productivity and self-reported post-editing effort of intralingual subtitles

Aitor Álvarez[1], Marina Balenciaga[1], Arantza del Pozo[1], Haritz Arzelus[1], Anna Matamala[2], Carlos-D. Martínez-Hinarejos[3]

[1] Human Speech and Language Technology Group, Vicomtech-IK4, San Sebastian, Spain
[2] Department of Translation, Interpreting and East Asian Studies, UAB, Barcelona, Spain
[3] Pattern Recognition and Human Language Technologies Research Center, Universitat Politècnica de València, Spain
[1]{aalvarez, mbalenciaga, adelpozo, harzelus}@vicomtech.org, [2]anna.matamala@uab.cat, [3]cmartine@dsic.upv.es

Abstract

This paper describes the evaluation methodology followed to measure the impact of using a machine learning algorithm to automatically segment intralingual subtitles. The segmentation quality, productivity and self-reported post-editing effort achieved with such approach are shown to improve those obtained by the technique based in counting characters, mainly employed for automatic subtitle segmentation currently. The corpus used to train and test the proposed automated segmentation method is also described and shared with the community, in order to foster further research in this area.

Keywords: automatic subtitling, subtitle segmentation, machine learning

1. Introduction

Society and the governments are increasingly requesting larger amounts of subtitled TV content (Neto et al., 2008), since subtitles are the most practical technique to guarantee the accessibility of audiovisual material to those who cannot access the audio (AENOR, 2012).

The current demand and promising future of intralingual subtitling has accelerated research into more productive methods that help to cover challenging subtitling situations such as, for example, live broadcasts. In recent years, technological advances in speech recognition have enabled automatic intralingual subtitling to be a reality (Álvarez et al., 2015). However, automatic subtitling technology has limitations, a major one being its inability to segment subtitle text in a logical way. Within the subtitling field, segmentation refers to the division of the original text into sections that viewers can understand immediately (Díaz-Cintas and Remael, 2007), playing a fundamental role in the creation of quality subtitles.

This work analyses the application of a machine learning algorithm to automatically segment the text contained in intralingual subtitles and compares its performance against that of the main technique based in counting characters employed in most automatic subtitling systems currently. Its impact is measured in terms of subtitle quality and regarding the productivity achieved and self-reported effort when integrated in the subtitling process through post-editing. Also, given the little work carried out so far in automatic subtitle segmentation, the corpus employed to train and test the presented machine learning algorithm is described and shared with the aim of fostering further research and technology development.

2. Background

2.1. From traditional to automatic subtitling

Traditional subtitling is carried out by professionals who aim to reproduce in text on screen the original dialogues, the discursive elements in the image and, when addressed to those who cannot hear the original audio, the information contained in the soundtrack of audiovisual contents (Díaz-Cintas and Remael, 2007).

From a linguistic perspective, subtitles can be classified as intralingual, interlingual or bilingual. Depending on the time available for their preparation, they can be prerecorded, live or semi-live. And according to the recipient, subtitles can either be for the hearing or for the deaf and hard-of-hearing, the latter containing additional information to facilitate comprehension, such as contextual information or sound effects (Díaz-Cintas and Remael, 2007).

Several studies collect good subtitling practices (Karamitroglou, 1998; Ivarsson and Carroll, 1998; Díaz-Cintas and Remael, 2007; Ford Williams, 2009; Ofcom, 2015). Also, there are some regulations governing quality subtitling standards such as the UNE 153010 (AENOR, 2012). The main features of good subtitles can be classified into:

- **Spacing features:** distributing the text into one or two lines between 4 and 43 characters.
- **Timing features**: showing subtitles at a speed of 130-170 words per minute keeping them on screen between 1 and 6 seconds, while synchronizing with the audio and inserting short pauses between consecutive subtitles.
- **Linguistic features**: keeping the original terms and avoiding more than two sentences per subtitles, one per line.
- **Orthotypographic features**: following the general guidelines of printed text.

All of the above features impact subtitle segmentation and are taken into account for manual segmentation by

professional subtitlers.

Automatic subtitling was born in response to a high subtitling demand, as a more productive alternative that enabled subtitling in challenging situations, such as live broadcasts, where traditional subtitling was not directly applicable. However, at present, automatic subtitling is yet not capable of creating subtitles that equal human quality and, thus, its focus is on facilitating the generation or post-editing of automatic subtitles by professional subtitlers, both in live and pre-recorded settings. In this context, post-editing can be defined as the process by which a professional edits, modifies and/or corrects the output of an automatic subtitling system. Post-editing is increasingly gaining relevance as it is proving to help achieve subtitles of high quality in a more productive way.
The different types of technology that can be employed for automatic intralingual subtitling are:

- **Stenotyping:** It involves using a shorthand typewriter representing syllables, words, punctuation signs and/or phrases phonetically. Allowing the generation of subtitles at speeds of 220 and 300 words per minute, it is generally used for live subtitling. Its precision reaches 97-98%, the generated delay is low and the severity of errors medium. However, learning this technique requires a long time (around three years) and the cost is high (Romero-Fresco 2011).

- **Respeaking:** This technique involves producing real-time subtitles by means of speech recognition software transcribing a simultaneous reformulation of the source text dictated by the respeaker to the computer (Eugeni, 2008). Being easier to master than stenotyping and similiar in average performance, Respeaking has recently become the most widely used method for live subtitling in countries such as the UK, France or Germany (Mikul, 2014).

- **Automatic transcription:** This technology involves the generation of subtitles directly from the source audio, without the need of human intervention. In state-of-the-art systems, after a pre-processing step to normalize the audio and select the segments with contain speech, a speech recognition software transcribes the speech detected and synchronizes it with the audio. A posterior linguistic processing normalizes the numbers, abbreviations and acronyms, and capitalization and punctuation marks are automatically included. Finally, subtitle segmentation is performed and subtitles are generated in the required format. Various subtitling solutions based on automatic transcription systems have been developed for several languages and domain in recent years (Neto et al., 2008; Ortega et al., 2009; Álvarez et al., 2015). Although they do not perform as well as professional subtitlers, they have achieved promising results with productivity gain experiments suggesting that post-editing automatic subtitles is faster than creating them from scratch (Álvarez et al., 2015).

2.2. Subtitle segmentation

While good segmentation facilitates reading and understanding subtitles (Ford Williams, 2009), bad segmentation can interrupt the natural reading flow, make the audience lose concentration and obscure the subtitle message (Perego et al., 2010).
Good subtitle segmentation involves making each subtitle constitute a complete linguistic unit, according to the main rules of syntax and semantics. In traditional subtitling, the concept of the "highest syntactic node" (Karamitroglou, 1998) is widely employed, establishing that each subtitle line should contain the highest possible level of syntactic information.
The guidelines governing segmentation quality involve the following most relevant criteria (AENOR, 2012; Díaz-Cintas and Remael, 2007):

- Take advantage of silences, grammatical pauses and punctuation signs.
- Do not divide noun-, verb- or prepositional-phrases.
- Do not split compound verb forms, and words.
- In subtitles consisting of two sentences, place each sentence in one line. If compound sentences do not fit into one line, use a line per proposition. Write conjunctions and nexus in the bottom line. If simple sentences require division, put the subject in the top line and the predicate in the bottom line. With question-answers, place the question in the top line and the answer in the bottom line, unless information is exposed too soon this way.

To date, most of the automatic subtitling solutions have not been able to discriminate the natural pauses, syntactic and semantic information relevant for quality segmentation and, thus, automatic segmentation in stenotyping, respeaking or audio transcription applications is mainly performed considering only the maximum number of characters allowed per line or through manual intervention. Machine learning has only recently started to be applied for automatic segmentation. The first reference in the field (Álvarez et al., 2014) focused in the development of Support Vector Machines (SVM) and Logistic Regression (LR) classifiers to automatically segment subtitle text considering the good practices of traditional subtitling.

3. Technical approach

With the aim of improving the results in (Álvarez et al., 2014), a new classification approach was developed in this study. Given that subtitle segmentation can be treated as a text labeling problem, in which each of the words in subtitles carries a specific function, Conditional Random Fields (CRFs) were used as the main machine learning algorithm for the automatic segmentation task. CRFs are often employed for labeling and parsing sequential natural language text (Lafferty et al., 2001) and unlike other classifiers, such as SVM or LR, CRFs consider the surrounding observations to predict the current label. This is an important feature, since predicting the optimal segmentation point depends not only on the current word, but also on the surrounding context.

The feature vectors, which describe the information related to each word during classification, were composed by the following characteristics: (1) the current an the surrounding 2 words, (2) the Part-Of-Speech (POS) information of the current and the surrounding 2 words, (3) the amount of characters per line and subtitle, (4) speaker change information, and (5) the time differences between the current, previous and next words.
The CRFs were trained and evaluated with the CRFSuite software (Naoaki Okazaki, 2007), and the POS information was extracted using the ixa-pipe-pos toolkit (Agerri et al., 2014).

4. Corpus characteristics

The corpus was composed by 23 episodes of the Spanish "Mi Querido Klikowsky" TV series, containing 1,150 minutes and a total amount of 20,154 subtitles, 90% of which was used to train the CRF model, and the rest was kept for testing purposes. The subtitle files, provided in SRT format, were created manually by professionals, and their segmentation was performed following specific predefined rules to keep linguistic and syntactic coherence. The contents include many segments with spontaneous speech, grammatically incorrect sentences, and some words and expressions pronounced in several Spanish dialects, such as Argentinian and Andalusian. Because of this, the POS tagger made more mistakes than desired.
The corpus described above (EiTB_Subt_Corp) will be made available to the research community through the META-SHARE repository[1] under the Creative Commons Attribution-NonCommercial-ShareAlike (CC-BY-NC-SA) license.

5. Evaluation methodology

In order to analyze the impact of the proposed CRF-based segmentation approach described in section 3., we have carried out evaluation at three different levels. First, segmentation quality has been measured through objective metrics. Second, a post-editing experiment has been conducted to test whether the application of the developed algorithm affects productivity. Third, subjective feedback regarding the post-editing task has been collected measuring the self-reported effort of post-editors through a Likert scale questionnaire.

5.1. Quality evaluation

The segmentation quality of subtitles was calculated in terms of precision, recall and F1-score as follows:

$$Precision = \frac{correct\ segmentations}{total\ number\ of\ segmentations}$$

$$Recall = \frac{correct\ segmentations}{total\ number\ of\ correct\ segmentations}$$

$$F1 - score = 2 \times \frac{precision \times recall}{precision + recall}$$

The test set described in section 4. was segmented using both, the proposed machine learning algorithm and the technique based in counting characters. Then, the precision, recall and F1-score achieved with both methods were compared.

5.2. Productivity evaluation

Whether the developed segmentation approach facilitates and streamlines the process of generating quality subtitles was evaluated through a post-editing task.
Nine students of the Subtitling Module included in the UAB's METAV[2] and MTAV[3] Masters Programs volunteered to participate in the test. Eight of them were Translation and Interpretation graduates, while one had a degree in Business Administration. In addition to the subtitling practice acquired through the masters program, several participants had further experience: three of them had been subtitling between one and three years and other four between one and six months. One participant worked as a professional subtitler at the time of the experiment and another participant had a six-month post-editing experience.
The post-editing task was arranged as follows. First, the test set described in section 4. was divided into smaller sets of 50 subtitles, which were segmented using both, the proposed algorithm and the counting characters method. Then, each participant was asked to post-edit the segmentation of two of these sets, each of which had been segmented using one of the two techniques under evaluation. In order not to influence the post-editing task, the evaluation sets assigned to each post-editor contained different subtitles. Subtitling Workshop[4] and the Toggl[5] tools were employed as subtitling and time tracking softwares, respectively. After finishing the task, participants generated a Toggl report including the time required to complete it.

5.3. Self-reported effort

In order to gather additional subjective information regarding the post-editing experience of the volunteers, they were asked to rate:

- the self-reported effort expended post-editing each of the subtitle sets on a 1 to 5 scale (1 being the lowest and 5 the highest)
- their level of agreement/disagreement on a 1 to 5 scale (1 being "strongly disagree" and 5 being "strongly agree") with the statements shown in Table 2.

6. Results

6.1. Quality evaluation

Table 1 shows results of the quality evaluation. As it can be seen, results achieved by the CRF model outperform those of the counting character technique. With the machine learning method, 85.08% of the retrieved cuts are correct and 80.30% of the correct cuts that should be retrieved are generated. Such results go down to 22.08% and 15.43% in the case of the counting character segmentation technique.

[1] http://www.meta-share.eu/

[2] http://metav.uab.cat

[3] http://pagines.uab.cat/mtav

[4] http://subworkshop.sourceforge.net/

[5] http://toggl.com

Algorithm	Precision	Recall	F1-score
Counting Characters	22.08%	15.43%	18.17%
CRF model	85.08%	80.30%	82.62%

Table 1: Quality evaluation results

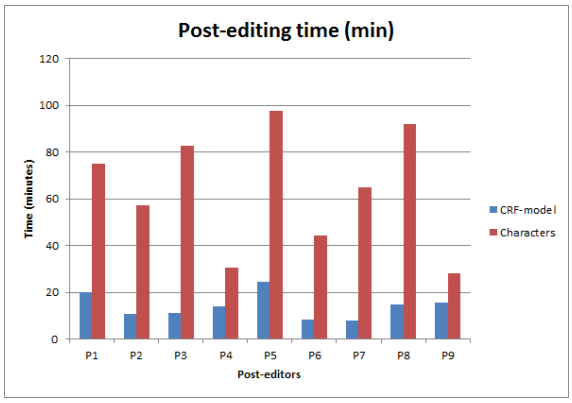

Figure 1: Productivity evaluation results in minutes

6.2. Productivity evaluation

The time in minutes required by each participant to post-edit the subtitle files is presented in Figure 1 for both segmentation methods. This time only includes the post-edition task; that is, the time needed to automatically segment the subtitles was not considered. The participants are ordered considering their previous subtitling experience; P1 being the most experienced post-editor, and P9 the participant with less experience.

As it can be appreciated, all participants needed more time to post-edit a subtitle in the test set segmented with the counting characters technique. Participants needed 14.2 minutes on average to post-edit a subtitle file segmented with the CRF-model, whilst it took them 63.7 minutes to post-edit the subtitles splitted with the counting characters method. In other words, participants needed 49 minutes more to post-edit the same number of subtitles (50) segmented with the counting characters method.

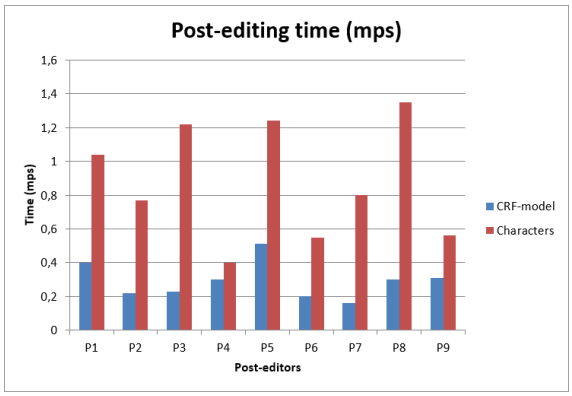

Figure 2: Productivity results in minutes per subtitles (mps)

With the aim of comparing the average time needed per each of participants to post-edit a subtitle in the two sets segmented with the methods under evaluation, the time measured in minutes per subtitle (mps) was computed and presented in Figure 2. On average, it took them 0.3 minutes to post-edit a subtitle segmented with the CRF-model and 0.88 minutes to post-edit a subtitle segmented with the counting characters method, which is 3 times longer overall. Thus, the presented machine learning algorithm allows post-editing segmentation faster, increasing productivity, and making the post-editing task more pleasurable.

6.3. Self-reported effort

Figure 3 shows the self-reported effort results of post-editing the segmentation of the two subtitle sets processed with the methods under evaluation, in a 1 to 5 scale (1 being the lowest and 5 the highest).

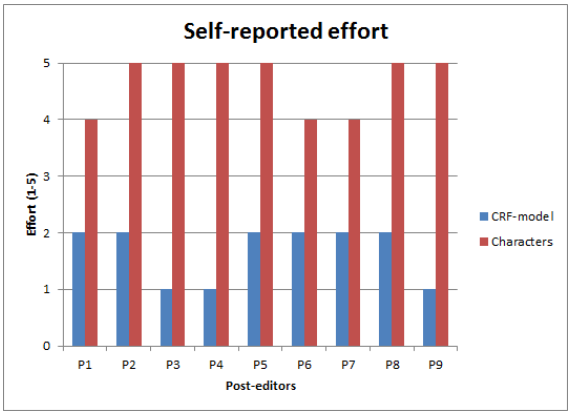

Figure 3: Self-reported post-editing effort results

As it can be seen, there is a clear difference between the self-reported post-editing effort of the two subtitle sets. That segmented using the machine learning algorithm received scores of 1 or 2, with an average of 1.66. On the other hand, the subtitle set segmented using the counting character technique was assessed with scores of 4 or 5, and 4.66 in average. Thus, according to the participants' assessment, post-editing subtitles segmented with the method based in counting characters took them more effort.

Participants also gave their opinion regarding the statements shown in Table 2, in a 1 to 5 scale (1 being "strongly disagree" and 5 being "strongly agree"). From the gathered results, it can be observed that in general most participants found it more enjoyable to post-edit subtitles automatically segmented by the machine learning algorithm, scoring its related statements higher. In particular, participants found it easier and less boring to post-edit, respect and use the guidelines of good segmentation with the machine learning algorithm on average. In addition, segmentations produced by such algorithm were perceived to be of better quality. And the resulting post-edited subtitles were also thought to be better segmented.

7. Conclusions and Future Work

This work has measured the impact of using a machine learning algorithm to automatically segment intralingual

Statement	Counting characters	Machine learning
I found it difficult to post-edit subtitle segmentation	3.88	1.44
I have been able to respect and use the guidelines of good segmentation	3.66	4.55
I found it boring to post-edit this subtitle file	3	2
Subtitles in this file were well segmented before being post-edited	1.11	3.22
I managed to achieve subtitles with good segmentation quality after post-editing this file	3.55	4.33

Table 2: Average subjective assessment results

subtitles in terms of quality, productivity and self-reported post-editing effort. Quality has been evaluated objectively through precision, recall and F1-score metrics; a post-editing task has been carried out to obtain objective measures of productivity; and the self-reported effort has been assessed subjectively through a ranking questionnaire. All evaluations have been performed through comparison with the main technique employed for automatic subtitle segmentation nowadays, which is based in counting characters. The quality achieved by the proposed CRF-based classifier has been shown to outperform that of the counting character technique by far. Post-editing productivity has shown to increase up to three times and the self-reported effort of the post-editing task to decrease three points. In addition, post-editors have found subtitle segmentations generated by the machine learning method to be of better quality, easier and less boring to post-edit respecting the guidelines of good segmentation and their post-edited versions thought to be better segmented. These successful results show the potential of machine learning to model the segmentation rules employed in traditional subtitling from a relatively small corpus of already segmented subtitles.

Future work should involve testing the proposed automatic segmentation approach more extensively on bigger datasets, different languages and speech recognition output. In addition, new classification methods should be tested for the automatic segmentation task. Among others, Recurrent Neural Networks (RNNs) have been proven to be useful for sequence labeling due to their explotable properties, including that they can make use of the past and future contextual information, and that they are robust to possible local distorsions of the input features (Graves, 2012). Finally, more parameters like stop words, syntactic functions or grammatical relations could be explored in order to check their impact on this task. The authors encourage researchers to look into these and other challenges, exploiting the released corpus of quality segmented subtitles as necessary.

8. Acknowledgements

Anna Matamala is member of transMedia Catalonia, which is a research group funded by the Catalan government (2014SGR027).

9. References

AENOR. (2012). Subtitulado para personas sordas y personas con discapacidad auditiva. UNE 153010:2012. Technical report, Madrid.

Agerri, R., Bermudez, J., and Rigau, G. (2014). Multilingual, Efficient and Easy NLP Processing with IXA Pipeline. *EACL 2014*, page 5.

Álvarez, A., Arzelus, H., and Etchegoyhen, T. (2014). Towards customized automatic segmentation of subtitles. In *Advances in Speech and Language Technologies for Iberian Languages*, pages 229–238. Springer.

Álvarez, A., Mendes, C., Raffaelli, M., Luís, T., Paulo, S., Piccinini, N., Arzelus, H., Neto, J., Aliprandi, C., and del Pozo, A. (2015). Automating live and batch subtitling of multimedia contents for several european languages. *Multimedia Tools and Applications*, pages 1–31.

Díaz-Cintas, J. and Remael, A. (2007). *Audiovisual Translation, Subtitling*. St. Jerome Publishing.

Eugeni, C. (2008). Respeaking the news for the deaf: for a real special needs-oriented subtitling. *Studies in English Language and Literature*, 21.

Ford Williams, G. (2009). Online subtitling editorial guidelines v1.1. Technical report, BBC.

Graves, A. (2012). *Supervised Sequence Labelling with Recurrent Neural Networks*. Studies in Computational Intelligence. Springer, Heidelberg, New York.

Ivarsson, J. and Carroll, M. (1998). Code of good subtitling practice. Technical report, ESIST (European Association for Studies in Screen Translation), Berlin.

Karamitroglou, F. (1998). A proposed set of subtitling standards in europe. *Translation Journal*, 2.

Lafferty, J., McCallum, A., and Pereira, F. C. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Mikul, C. (2014). Caption quality: International approaches to standards and measurement. Technical report, Media Access Australia.

Naoaki Okazaki. (2007). CRFsuite: a fast implementation of Conditional Random Fields (CRFs).

Neto, J., Meinedo, H., Viveiros, M., Cassaca, R., Martins, C., and Caseiro, D. (2008). Broadcast news subtitling system in portuguese. In *IEEE International Conference on Acoustics, Speech and Signal Processing, 2008. ICASSP 2008.*, pages 1561–1564, Las Vegas, Nevada. IEEE.

Ofcom. (2015). Code on television access services. Technical report.

Ortega, A., García, J., Miguel, A., and Lleida, E. (2009). Real-time live broadcast news subtitling system for Spanish. In *10th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Brighton.

Perego, E., del Missier, F., Porta, M., and Mosconi, M. (2010). The cognitive effectiveness of subtitle processing. *Media Psychology*, 13(3):243–272.