2017 IEEE 17th International Conference on Advanced Learning Technologies

Automatic Lecture Subtitle Generation and How It Helps

Xiaoyin Che
Hasso Plattner Institute
Prof.-Dr.-Helmert-Str. 2-3
14482, Potsdam, Germany
xiaoyin.che@hpi.de

Sheng Luo
Hasso Plattner Institute
Prof.-Dr.-Helmert-Str. 2-3
14482, Potsdam, Germany
sheng.luo@hpi.de

Haojin Yang
Hasso Plattner Institute
Prof.-Dr.-Helmert-Str. 2-3
14482, Potsdam, Germany
haojin.yang@hpi.de

Christoph Meinel
Hasso Plattner Institute
Prof.-Dr.-Helmert-Str. 2-3
14482, Potsdam, Germany
christoph.meinel@hpi.de

***Abstract*—In this paper we propose an integrated framework of automatic bilingual subtitle generation for lecture videos, especially for MOOCs. The framework consists of Automatic Speech Recognition (ASR), Sentence Boundary Detection (SBD), and Machine Translation (MT). Then we quantitatively evaluate the auto-generated subtitles, the manually produced subtitles from scratch, and the auto-generated subtitles with manual modification in term of accuracy and time expenditure, in both original and target languages. The result shows that the auto-generated subtitles in the original language (*English*) are fairly accurate already. By using them as the draft, human subtitle producers can save 54% of the working time and simultaneously reduce the error rate by 54.3%, which is a significant improvement. However, the effectiveness of machine-translated subtitles (*English to Chinese*) is limited. In the end, if the proposed framework is applied, the total working time in preparing bilingual subtitles can be shortened by approximately 1/3, with no decline in quality.**

***Keywords*—Automatic Subtitling, Sentence Boundary Detection, Lecture Videos, MOOC**

I. Introduction

With the rapid development of e-Learning technology, geometrical limitation is no longer considered as a major barrier in knowledge spreading. Numerous of lectures are recorded in videos, uploaded to the internet and can be accessed from any corner of this planet. In recent years, a new form of e-Learning, the **M**assive **O**pen **O**nline **C**ourses (MOOCs), successfully turns "learning online" into a fashion and attracts millions of learners into this community, regardless of their ages, genders, nationalities and educational backgrounds [1].

However, language barrier becomes a huge rock in this wave of knowledge globalization [2]. It excludes potential learners who cannot understand the teaching language and downgrades the learning achivements for those who have learnt the language but are not fluent with it. In this case, subtitles are considered as the best breaker [3]. Research shows that learners want not only the subtitles in target language, but also in original language, with bilingual subtitles as the most welcomed option [4]. Some learners claim that bilingual subtitles facilitate them to learn both the course content and the foreign language in the same time [5] − killing two birds with one stone. Thus when provided, the quality of the subtitles in both target and original languages are equally important.

Some MOOC platforms have already implemented the function of subtitle embedding, such as Coursera, edX or openHPI [4]. These subtitles are generally created manually by volunteer groups, course teaching teams or hired staffs. However, manual subtitle production has very high cost in time or/and money. Naturally people would look for the possibility of automatic subtitle generation. Due to the current technical conditions, the auto-generated subtitles inevitably contain errors and it is perhaps inappropriate to offer imperfect subtitles directly to learners. But if the auto-generated subtitles are in decent accuracy, we believe taking them as the draft could be very helpful for the human subtitle producers.

In this paper, we aim to address the above problems by proposing an integrated framework of automatic bilingual subtitle generation and evaluating its output on how the quality is and how much it may help the human subtitle producers. The newest tools in **A**utomatic **S**peech **R**ecognition (ASR) and **M**achine **T**ranslation (MT) will be applied, as well as the state-of-the-art **S**entence **B**oundary **D**etection (SBD) technique which is mainly based on word vectors and **D**eep **N**eural **N**etworks (DNN). Then the framework outputs, the auto-generated subtitles, will be evaluated by comparing with the ground-truth confirmed by multiple experts and manual produced or modified subtitles created by dozens of volunteers with different nationalities and backgrounds. The rest of this paper is arranged as follow: section II discusses the related works, section III introduces the framework we proposed and section IV illustrates the multi-task evaluation in detail. Conclusion comes afterwards.

II. Related Work

Undoubtedly, ASR technology is the foundation of automatic subtitle generation, but in this paper we would not discuss the development of ASR, only using it. Some initial solutions in automatic subtitling simply implement the ASR and take the silence in the audio track as sentence boundaries [6, 7]. In order to achieve better subtitle quality, how to precisely segment the ASR output is gradually getting more emphasized [8]. Quantitative analysis proves that better segmentation result could save time in manual post-editing [9]. Furthermore, the deficiency of punctuation marks in ASR output is also considered as a shortcoming in the subtitles generated [10] and efforts have been made for it [11].

Another focus point in subtitle quality is the way how to implement the subtitles to the video. There are some suggested standards for "good" subtitle [12, 13] and also some

2161-377X/17 $31.00 © 2017 IEEE
DOI 10.1109/ICALT.2017.11

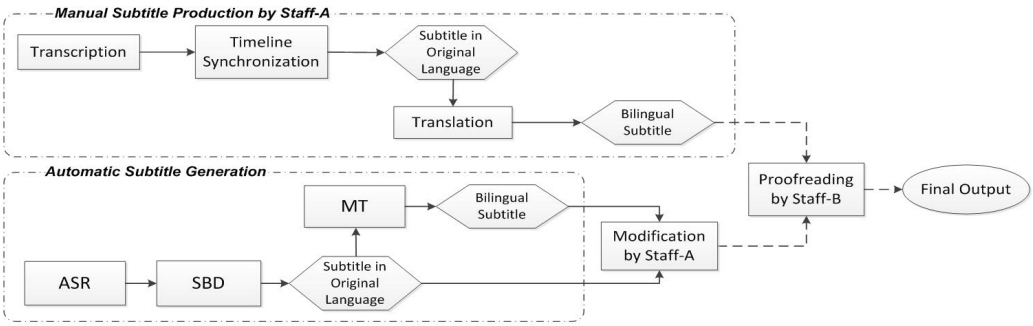

Fig. 1. The procedure of subtitle creation: manual & automatic.

applications towards these standards [14]. Since auto-generated subtitle can still not avoid errors, how to facilitate the human producers in post-editing is also widely discussed [15, 16]. Inspired by these previous efforts, our approach would pursue better segmentation method with punctuation marks restored, standardize the subtitle format in a user-friendly way and quantitatively evaluate how it helps in post-editing.

III. Automatic Lecture Subtitle Generation

When producing subtitles manually, the procedure can be generally divided into three steps: transcription, timeline synchronization and translation. These steps could be done by one person (staff-A) and then proofread by others (staff-B). The automatic subtitle generation framework we proposed is in similar structure, as seen in Fig. 1, which consists of ASR, SBD and MT. Due to technical limitations, a manual modification process by staff-A might be needed in practice before proofreading by staff-B.

In this section, we focus on automatic techniques only. The major technical contribution of our framework is the SBD process, which include a state-of-the-art lexical model based on word vectors and DNN, a simple pause-only acoustic model, a joint decision scheme and a special configuration oriented to the characteristics of subtitle. For both ASR and MT, we implement 3rd-party services. However, these processes would still be introduced by the order in Fig. 1.

A. Automatic Speech Recognition

In our work we choose the IBM Watson Speech-to-Text service as the ASR tool [17]. By submitting the audio file to the ASR server through API, the transcript file could be retrieved within 1.5$\sim$2 times of the audio duration. The transcript contains timestamps for each word and has also been segmented into sentence units. However, we found this default segmentation is far from good enough. Therefore we developed our own SBD toolkit.

B. Sentence Boundary Detection

Here we ignore the default segmentation from ASR and put all the words within the transcript of a lecture in a list according to their timestamps. The words will be represented by pre-trained word vectors in lexical model, while the pauses between adjacent words would be calculated and used in acoustic model. After the joint decision, punctuation

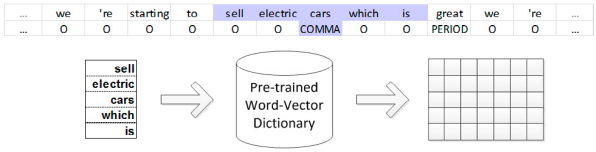

Fig. 2. The structure of the lexical model training.

marks would be restored, which can be considered directly as sentence boundaries. However, in purpose of making the subtitle user-friendly, the grammatical sentence units may not be always appropriate for subtitle items because of the length variation. So a special configuration is also implemented.

1) Lexical Model: In our lexical model, word vectors are used as the only feature. Word vectors are learned through neural language models [18] and proven to be successful in many **N**atural **L**anguage **P**rocessing (NLP) tasks. It is suggested that the semantic distance between words can be measured by the mathematical distance of corresponding word vectors [19]. All these facts make us believe that word vector could be a good choice in detecting sentence boundaries or predicting punctuation marks.

In order to train the lexical model, we prepared a training dataset containing manually created transcripts of 1710 TED talks, approximately 2.1M words. As illustrated in Fig. 2, the text would be first transformed into a long word sequence and then traversed by a m-word sliding window to create samples. Before fed into the DNN for training, each word in the sample would be translated into an n-dimensional word vector by a pre-trained dictionary. The goal is to classify whether there should be a punctuation mark (*comma, period or question mark*) after the k-th word in the sample.

We made a test on the accuracy of the lexical model trained. The ASR transcripts of 8 TED talks are used as the testing set with about 12k words. The parameters are set as follow: m=5, k=3, n=50, and we built the DNN with 3 hidden layers on CAFFE framework [20]. The GloVe.6B.50d vector set is used as the dictionary [21], while "this" is adopted as the default replacement of out-of-vocabulary words. Experiment result shows that when the punctuation mark type is distinguished, the generally accuracy is 49.6%. If we ignore the type, the accuracy could reach 70% [22].

2) Acoustic Model and Joint Decision Scheme: Compared to the lexical model, the acoustic model we applied is quite

simple. It is a heuristic pause-only model, which classifies only 2-classes: punctuated or not. For a word W_i in the ASR transcript, we simply calculate the pause duration p between W_i and W_{i+1} and use a variant of Sigmoid function:

$$P_a = \frac{1 - e^{-4p}}{1 + e^{-4p}}, \quad p \in [0, +\infty) \tag{1}$$

to project p into P_a, while $P_a \in [0, 1)$. Roughly the threshold of predicting a punctuation mark is 0.28 second. With same testing data of lexical model, the 2-classes accuracy of the acoustic model is 60.9%.

By now we have both lexical and acoustic models which can work independently. The next step is to fuse their outputs and further improve the classification result. Therefore we develop a 2-stages joint decision scheme, which works like "segmenting" and "sub-segmenting" [23]. In stage-1, we use the posterior probabilities of both lexical and acoustic models to detect sentence boundary positions. Then the input word sequence can be split by these positions into a bunch of segments. Stage-2 works with each segment in order to pursue any potential sub-segmenting possibility with lexical model output only. Evaluating with same data, the joint accuracy reaches 77.6%, which is the state-of-the-art and obviously better than either lexical model (70.7%) or acoustic model (60.9%) independently.

3) Subtitle-Oriented Configuration: If our purpose is to offer lecture transcript in paragraphs, the SBD process could be ended now, with all possible punctuation marks restored in the text. However, as supplementary material to the video, subtitle could only be displayed in a limited region, commonly in the bottom area of the video, otherwise it may cover the major visual content [13]. Therefore, one single subtitle item should have a maximum length, up to two rows in general principle [12, 13]. Since we would offer bilingual subtitle, each language could only occupy one row.

We set the maximum length per item heuristically to 60 Latin characters when processing English. When a continuous word sequence between two adjacent restored punctuation marks is longer than 60 characters, we apply the lexical model to find the most possible re-segmenting position. This process runs recursively until all segments are shorter than the maximum length. By this approach, we successfully avoid splitting basic grammar units: for example, the line break will never take place between "take" and "place" or after an article "an". Besides, we also set a minimum length per item to 15 Latin characters, because if a subtitle item is too short, its duration is very likely to be short either. When displaying, these short subtitle items may appear as "flashing", which does no good for the learners' watching experiences [12]. In such cases, the short items will be combined with the next item.

Now with proper formatting, the generation of subtitle file in original language could be concluded. Since the lexical model is pre-trained, the whole SBD process of a 10-minutes lecture can be done in several seconds. The time expenditure is approaching zero.

C. Machine Translation

The MT tool we adopt is Microsoft Translator API. The textual content of the previously generated subtitles in original language will be submitted to the translation server item by item. The returning text in target language will be directly added as a second line of the corresponding subtitle item, in order to create bilingual subtitle. In this process, one second is needed for every two items. For the convenience of further manual modification and proofreading, the subtitle file with only target language will not be provided by our automatic subtitle generation framework.

IV. Evaluation

In this section we would like to evaluate the contribution of the auto-generated subtitles in process of online lecture preparation. Naturally the first task in evaluation is about the accuracy. However, since the error cannot be totally avoided by far, perhaps some learners may accept error-containing subtitles, but as e-Learning providers, we always want to provide high quality materials − in this context − accurate subtitles. Based on this idea, we also evaluate how the auto-generated subtitles can help the human editors in subtitle preparation, with both original and target languages.

A. Methodology

As introduced before, manual subtitle production consists of three steps before proofreading: transcription, timeline synchronization and translation. In our auto-generation framework, "ASR+SBD" cover the tasks of transcription and synchronization, with standardized subtitle file in original language as output. MT further creates subtitle file in target language. Therefore, we would also evaluate them separately.

We selected a few clips of lecture video as testing data, with English as teaching language. We then invited 24 volunteers to help producing subtitles for comparison. The volunteers are fluent but non-native English speakers, who come from China, Germany, Russia, Iran and Indonesia respectively. There are two reasons for this arrangement: the first is that based on our experiences, subtitle production is generally taken charge by staffs who speak the target language, not the original language; the second is that we cannot find enough native English speakers at this time. In our experiment, all volunteers worked independently. Ground-truth subtitles are created by several experts, including the corresponding lecturer himself.

For original language evaluation, each volunteer would receive two video clips (V_1 and V_2) and one auto-generated subtitle file, which might be suitable for either V_1 or V_2, addressed as A_1 or A_2. If a volunteer received A_1, he/she was expected to create the modified subtitle S_1' for V_1 based on A_1 and the manual subtitle S_2 for V_2 from scratch, vice versa. Meanwhile, we also demanded the volunteers to report the time spent on individual tasks. In this way, all volunteers were divided into two groups, working on "S_1'+S_2" or "S_1+S_2'" separately. Any volunteer would not work with same video twice, by which we avoid the performance deviation caused by the volunteer's different familiarities with certain video. Since all volunteers have worked with both videos, the difference of English skills between them can also be balanced. So we can evaluate the average error rate of A_i, S_i and S_i' respectively in comparison with ground-truth, and calculating the average time expenditure of S_i and S_i' in a fairly convincing way.

TABLE I
EVALUATION WITH ORIGINAL LANGUAGE

	A_1	S_1	S'_1	A_2	S_2	S'_2	$\bar{A}$	$\bar{S}$	$\bar{S}'$
T	–	1001	489	–	845	360	–	923	424
P-1	.143	.154	.081	.070	.117	.043	.107	.135	.062
P-2	.186	.184	.116	.096	.143	.074	.141	.164	.095

TABLE II
EVALUATION WITH TARGET LANGUAGE

	B_1	C_1	C'_1	B_2	C_2	C'_2	$\bar{B}$	$\bar{C}$	$\bar{C}'$
Time	–	636	675	–	853	671	–	745	673
Items	20	20	20	17	17	17	18.5	18.5	18.5
Modi.	10	2.1	4.9	9	3.9	3.6	9.5	3.0	4.3
E.R.	.500	.106	.244	.529	.228	.203	.514	.162	.230

For target language evaluation, we tested "English to Chinese" translation only. After excluding non-Chinese speakers, the remaining volunteers are also divided into two groups with different translation tasks. Each volunteer in group A would receive two video clips, one error-free English-only transcript E_1 and one machine-translated bilingual subtitle B_2. He/She would be expected to create Chinese subtitle C_1 and C'_2, as well as the time spent. Similarly, volunteers in group B should submit "C'_1+C_2". After collecting all possible C_i and C'_i, average accuracy and time expenditure are measured.

B. Evaluation with Original Language

The video clips used in original language evaluation derive from the welcome video of openHPI platform. V_1 starts at 0:29 and V_2 starts at 2:02, both of which last 64 seconds. It is comparatively short, because we could not give too much burden to the volunteers. Since it might be too short to be measured by WER (*Word Error Rate*), we apply character-level Levenshtein Distance [24]. Accuracy evaluation consists of two phases. Phase 1 focuses on textual content only, which ignores the subtitle item boundaries by connecting all items in a single long string, with one space between every two adjacent subtitle items. Phase 2 takes subtitle item boundary positions into consideration. Each boundary is counted as a 3-character word. By taking ground-truth subtitle file as anchor, a Levenshtein Distance can be calculated. The ratio of this distance and the length of the anchor will be acknowledged as error-rate, which will be address as "P-1" and "P-2" for the two phases. Time spent (T) are counted in seconds.

In Table 1 we could first see in "A" columns that the general accuracy of auto-generated subtitle is fairly acceptable, with the error rate between 10∼15%. However, a more encouraging finding is the significant improvement in subtitle quality and the unneglectable achievement in time saving when taking the auto-generated subtitles as draft in manual subtitle production. The pure textual error-rate drops from 0.135 to 0.062 (54.3%), while the ASR baseline (0.107) is in the middle. In boundary included evaluation, the error-rate also drops around 42.1%. Meanwhile, the working time can be shortened by 54% averagely. More details would be discussed later.

C. Evaluation with Target Language

The video clips used in target language evaluation are selected from lesson 6.3 of the MOOC "Web Technologies" in 2015, talking about search engines. V_1 starts at 0:05 and V_2 starts at 5:22, with the same duration: 103 seconds. Different from original language evaluation, there is no ground-truth for translation task in this context. Therefore we have to evaluate the accuracy manually with following principles:

- ◇ The human translators of the MOOC "Web Technologies" are invited as reviewers.
- ◇ Subtitle items are taken as the units in evaluation.
- ◇ Units are marked as "modification needed" or not, based on its meaning, grammar and fluency.
- ◇ The error-rate is defined as how many units need to be modified.

The statistics of target language evaluation can be found in Table 2, in which time spent is still counted in seconds. Different from what achieved in original language, the performance of machine translation in subtitle generation is less convincing. Especially for V_1, the quality of modified Chinese subtitles based on MT is apparently worse than directly translating from English. On the other hand, the working time saved is also very limited, around only 9.6%. We will discuss potential reasons in next chapter.

D. Result Analysis

First based on the original language evaluation, the auto-generated subtitle has been proven to be very helpful. Another contribution of auto-generated subtitle, which does not appear in Table 1, is that there is no need for the human producers to care about the timeline synchronization, which may take the human editor quite some time, at least the same as video duration. Besides, we made a lot of efforts in SBD, but in Table 1 there is no hint of the gain. However, if we directly apply the default ASR segmentation without the SBD process, the P-1 and P-2 error-rates of the baseline ASR subtitle would be 0.117 and 0.149 respectively, both of which are worse than what we offered. Since the translation tasks in our experiments are based on the error-free transcripts, the effects of SBD in MT unfortunately cannot be measured.

It is not ideal that the auto-generated subtitle in target language is much less helpful to the human producers. We believe one major reason is that the quality of machine translated text is far from good enough. From Table 2 we can find that the error-rate of baseline is 0.516, which is way higher than in original language. Another possibility is that when there is no technical error, the human producers tend to give a green light to a machine translated text line, but it may not be accepted by reviewers because of the remaining language disfluency. Particularly, the extra difficulty in translation between English and Chinese is widely acknowledged.

Finally we would like to estimate the total time expenditure throughout the whole procedure of subtitle creation before proofreading, step by step according to Figure 1. Suppose the duration of the input lecture video is d, based on the stats in Table 1 and the corresponding 64 seconds testing videos, we can calculate the transcription time as $14.4d$. Similarly with Table 2, the translation time is $7.2d$. In additional with $1.5d$ for timeline synchronization, the total expenditure of time is $23.1d$ in manual subtitle production. If automatic subtitle generation is activated, with $2d$ for ASR, $0.2d$ for SBD and MT, $6.6d$ for modification in original language and $6.5d$ in target language,

the total expenditure of time is $15.3d$. Thus, we can roughly save 1/3 of the total working time if the proposed automatic subtitle generation framework is applied while keeping the output in same quality level.

V. Conclusion

In this work, we first introduced a complete framework of automatic subtitle generation for lecture videos, which consists of ASR, SBD and MT accordingly. The major technical contribution is the SBD we implemented, which include a DNN-based lexical model, a pause-only acoustic model, a joint decision scheme and a special configuration oriented to user-friendly bilingual subtitle format. Then we evaluated the auto-generated subtitles in comparison with experts-confirmed ground-truth and volunteers-created manual subtitles. Result shows that in original teaching language, English in our case, the auto-generated subtitles can significantly help the volunteers with both accuracy and efficiency. Although the effectiveness in target language (Chinese) is below expectation, the implementation of the automated framework may save up to 1/3 of the total working time in subtitle production.

Acknowledgment

We would like to express our sincere gratitude to all the volunteers: Aragats Amirkhanyan, Christian Bartz, Cao Jie, Chen Sicheng, Fan Yafei, Gao Ting, Tatiana Gayvoronskaya, He Jinbo, Jia He, Lin Xiao, Meng Yao, Qi Na, Ran Qidi, Mina Rezaei, Johannes Sianipar, Song Peng, Ihsan Sukmana, Sun Yue, Christian Tietz, Wang Huanzhong, Zhao Chen, Zhou Hao, Zhu Weijia, Zuo Zhe.

References

[1] L. Pappano, "The year of the mooc," *The New York Times*, vol. 2, no. 12, p. 2012, 2012.

[2] M. Schell, "How to globalize online course content," in *Globalized e-learning cultural challenges*. IGI Global, 2007, pp. 155–167.

[3] T. Beaven, A. Comas-Quinn, M. Hauck, B. de los Arcos, and T. Lewis, "The open translation mooc: creating online communities to transcend linguistic barriers," *Journal of Interactive Media in Education*, vol. 2013, no. 3, 2013.

[4] X. Che, S. Luo, C. Wang, and C. Meinel, "An attempt at mooc localization for chinese-speaking users," *International Journal of Information and Education Technology*, vol. 6, no. 2, p. 90, 2016.

[5] S. Wu, A. Fitzgerald, and I. H. Witten, "Second language learning in the context of moocs," in *CSEDU 2014*, vol. 1. SCITEPRESS, 2014, pp. 354–359.

[6] A. Pražák, J. V. Psutka, J. Hoidekr, J. Kanis, L. Müller, and J. Psutka, "Automatic online subtitling of the czech parliament meetings," in *International Conference on Text, Speech and Dialogue*. Springer, 2006, pp. 501–508.

[7] A. Ortega, J. E. G. Laínez, A. Miguel, and E. Lleida, "Real-time live broadcast news subtitling system for spanish." in *INTERSPEECH*, 2009, pp. 2095–2098.

[8] A. Álvarez, H. Arzelus, and T. Etchegoyhen, "Towards customized automatic segmentation of subtitles," in *Advances in Speech and Language Technologies for Iberian Languages*. Springer, 2014, pp. 229–238.

[9] A. Álvarez Muniain, M. Balenciaga, A. d. Pozo Echezarreta, H. Arzelus Irazusta, A. Matamala, and C. D. Martínez Hinarejos, "Impact of automatic segmentation on the quality, productivity and self-reported post-editing effort of intralingual subtitles," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, 2016, pp. 3049–3053.

[10] G. Eizmendi *et al.*, "Automatic speech recognition for live tv subtitling for hearing-impaired people," *Challenges for Assistive Technology: AAATE 07*, vol. 20, p. 286, 2007.

[11] C. Aliprandi, C. Scudellari, I. Gallucci, N. Piccinini, M. Raffaelli, A. del Pozo, A. Álvarez, H. Arzelus, R. Cassaca, T. Luis *et al.*, "Automatic live subtitling: state of the art, expectations and current trends," in *Proceedings of NAB Broadcast Engineering Conference: Papers on Advanced Media Technologies, Las Vegas*, 2014.

[12] J. Ivarsson and M. Carroll, "Code of good subtitling practice," *Language Today, April*, 1998.

[13] F. Karamitroglou, "A proposed set of subtitling standards in europe," *Translation journal*, vol. 2, no. 2, pp. 1–15, 1998.

[14] S. PIPERIDIS, I. DEMIROS, and P. PROKOPIDIS, "Infrastructure for a multilingual subtitle generation system," *Linguistics in the Twenty First Century*, p. 369, 2009.

[15] A. Álvarez, A. del Pozo, and A. Arruti, "Apyca: Towards the automatic subtitling of television content in spanish," in *Computer Science and Information Technology (IMCSIT), Proceedings of the 2010 International Multiconference on*. IEEE, 2010, pp. 567–574.

[16] H. Sawaf, "Automatic speech recognition and hybrid machine translation for high-quality closed-captioning and subtitling for video broadcast," *Proceedings of Association for Machine Translation in the Americas*, 2012.

[17] G. Saon, H.-K. J. Kuo, S. Rennie, and M. Picheny, "The ibm 2015 english conversational telephone speech recognition system," *arXiv:1505.05899*, 2015.

[18] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin, "A neural probabilistic language model," *Journal of machine learning research*, vol. 3, no. Feb, pp. 1137–1155, 2003.

[19] T. Mikolov, W.-t. Yih, and G. Zweig, "Linguistic regularities in continuous space word representations." in *Hltnaacl*, vol. 13, 2013, pp. 746–751.

[20] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014, pp. 675–678.

[21] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation." in *EMNLP*, vol. 14, 2014, pp. 1532–1543.

[22] X. Che, C. Wang, H. Yang, and C. Meinel, "Punctuation prediction for unsegmented transcript based on word vector," in *The 10th International Conference on Language Resources and Evaluation (LREC)*, 2016.

[23] X. Che, S. Luo, H. Yang, and C. Meinel, "Sentence boundary detection based on parallel lexical and acoustic models," *Interspeech 2016*, pp. 2528–2532, 2016.

[24] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," in *Soviet physics doklady*, vol. 10, no. 8, 1966, pp. 707–710.