Interspeech 2018
2-6 September 2018, Hyderabad

Multimodal Name Recognition in Live TV Subtitling

Marek Hrúz[1], *Aleš Pražák*[1], *Michal Bušta*[2]

[1]University of West Bohemia, Faculty of Applied Sciences
NTIS - New Technologies for the Information Society
Univerzitní 8, 306 14 Plzeň, Czech Republic
[2] Czech Technical University, Faculty of Electrical Engineering
Center for Machine Perception, Department of Cybernetics
Karlovo namesti 13, 121 35 Prague, Czech Republic
aprazak@ntis.zcu.cz, mhruz@ntis.zcu.cz, bustam@fel.cvut.cz

1. Abstract

In this paper, we present a method of combining a visual text reader with a system of automatic speech recognition to suppress errors when encountering out-of-vocabulary words – specifically names. The visual text reader outputs detected words that are mapped into a large list of names via the Levenshtein distance. The detected names are inserted into the class-based language model on the fly which improves recognition results. To demonstrate the effect on the real speech recognition task we use data from sports TV broadcasting where a lot of names are present in both the audio and video streams. We superseded manual vocabulary editing in live TV subtitling through respeaking by an automated online process. Further, we show that automatically adding the names to the recognition vocabulary online and with forgetting lowers the WER relatively by 39 % in comparison with the case when names of all sportsmen are added to the vocabulary beforehand and by 15 % when only the relevant names are added beforehand.

2. Introduction

With the rapid development in the field of Automatic Speech Recognition (ASR) over the last decades, real-time Large Vocabulary Continuous Speech Recognition (LVCSR) is being used as a cost-effective alternative to live TV subtitling over manual keyboard transcriptions. Since ASR technology is still not able to automatically transcribe any television broadcast with acceptable accuracy, so-called respeaking – a technique in which a professional respeaker listens to the source audio and dictates it in a quiet environment to the well-tailored speech recognition system – has consolidated as the most widely adopted live subtitling technique. Live subtitling through respeaking was pioneered by British BBC in 2003 [1] and now it is used all over the world. At our department, we have developed our own remote live subtitling solution with many innovations [2]. Now, we provide live subtitles for the Czech television, the public service broadcaster in the Czech Republic, mainly for live sports broadcasting.

Even with large vocabularies, the major problem of live subtitling of sports TV programs is Out-Of-Vocabulary (OOV) words, especially the names of sportsmen and teams, which cannot be covered by any real language model training data. Many OOV related papers were published, for example, using OOV detector for named entities recognition [3] or proposing OOV recovery based on sub-word units [4], but these methods are not usable in the task of live TV subtitling. In our case, the sports event participants are usually known in advance, so we solved this problem with a class-based language model, where its classes should be filled before each live subtitling. Since each sport has its specific terms and phrases that are commonly used during TV commentary of the sport, we manually transcribed TV commentaries of 40 different sports (e.g. baseball, golf, figure skating or shooting). Each name of a player, competitor, team, nationality or sports place in these manual transcriptions was labeled. The names that did not relate to the transcribed match or competition were not labeled, because the commentators may use them freely (e.g. legendary sportsmen such as "Bjoerndalen", "Plyushchenko" or "Jagr"). Different labels were used for the names of players or competitors, for the names of competing teams or countries or participant nationalities and for the designations of sports places. In addition, each label was supplemented by a number representing one of 6 grammatical cases (i.e. different word forms) of the expression. Finally, taking into account the above-mentioned labels instead of the individual words, class-based language model with 18 classes (6 classes for players and competitors, 6 classes for teams and nationalities and 6 classes for sports places) was trained for each sport (see [5] for details). After mixing with other relevant language model training data, resulting trigram language models incorporate about 550k words for each different sports domain.

Before each live subtitling session, the respeaker adds the names and surnames of sportsmen participating in the match or competition to the recognition system based on official start lists published on the web pages or other relevant sources. Language model classes are then filled with names and surnames automatically inflected (rule-based) to 6 grammatical cases. Both a surname only and a combination of name and surname are generated. By default, all names within one class have the same weight, so uniform probability distribution is used. A similar process is used for the names of teams, nationalities and sports places. All items may be prepared in advance, so the respeaker only chooses a predefined selection, items are added to the recognition system (including all class n-grams) and the respeaker immediately starts subtitling.

In this paper, we focus on enhancing live TV subtitling by employing a technique from computer vision for automatic detection and recognition of texts and subsequent names extraction by our own algorithm. We call this process visual name spotting. The aim is to automatically fill language model classes during live subtitling with names from live sports graphics, so to supersede current manual work by an automated online process.

10.21437/Interspeech.2018-1748

Figure 1: *Example of TV broadcast sports event graphics.*

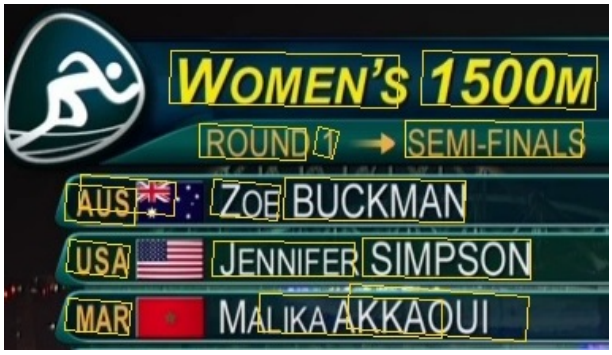

Figure 2: *Example of detected text regions. It demonstrates the errors that sometimes occur during the detection – overlapping regions and missed characters.*

3. Visual name spotting

In recent years there has been a lot of research conducted addressing the problem of reading the text in the wild. In this paper, we are concerned with reading a computer-generated, well-formed text. This problem can be considered as a subset of wild text reading. Several systems were developed in the past including but not limited to text detection methods based on hand-crafted features [6, 7, 8], or deep learning methods [9, 10, 11] and more recently [12]. In this paper, we use the models and algorithms developed in [13]. The algorithm uses a Convolutional Neural Network (CNN) to detect regions that very likely contain text, geometrically transforms the regions into rectangles of fixed height and variable width. Another CNN reads the text inside the transformed rectangles. The net uses CTC loss function [14] during optimization. The output is a sequence of characters that need to be processed further to obtain names.

First, the text sequences are stripped of non-alphabet characters. We are concerned only with Latin characters. Strings of length one are ignored. This yields a list of detections which represent all the text in the image. To filter the non-name texts we use the following algorithm.

- Merge overlapping detections – we do not want the same character(s) to be detected more than once
- Obtain all text detections from the same row – geometric centers of the detections are inside the vertical boundaries of each other
- In the row, find the longest string concatenating neighboring detections, where the maximum allowed gap between neighbors is defined as twice the mean width of detected characters
- Find the concatenated string in a list of all relevant sportsmen using normalized Levenshtein distance

This procedure assumes that the text is written in lines which is quite natural. See for example Figure 1 and Figure 2. Another assumption is that a list of all relevant sportsmen is available. Nowadays, it is common that organizers of sports events provide a list of all competing sportsmen. Theoretically, a list of all sportsmen in the world could be used, but a large number of entries in the list with many similar ones would be more time consuming to process and more error-prone.

The concatenation of the detected strings is necessary due to the failures of the text detector. Sometimes the last or the first letter is missed by the detector but it is included in the following/prior detection. We observed that the concatenation and thus searching in longer strings is beneficial for the whole process. To reduce the number of false alarms we use a threshold on the normalized Levenshtein distance (in our experiments 0.75). The name with the lowest normalized Levenshtein distance is considered to be a match. The matched names are outputted by the algorithm as a list of detected names.

This list is also used for name spotting in previously unmatched detections. One can think of it as a second pass through the detections. These detections are usually outside the sports graphics and thus are considered as wild text. An example is a name of a sportsman on his/her jersey. Without the context, it is very difficult to determine whether the detected string is a name or not, because the first name is missing. That is why we apply name re-spotting – we match all the unmatched strings with the already detected names. The process is very similar to the process of finding names in the graphics – we again use the normalized Levenshtein distance but we search only among the surnames of sportsmen.

4. Name spotting integration

The names detected by the proposed visual name spotting algorithm should be added to the live subtitling system just in time of their detection. This is very simple thanks to the phonetic prefix (lexical) tree structure of our recognition network. Lexical trees are highly efficient for languages with a high degree of inflection (such as the Czech language), where many word forms are derived from the same word stem. A time-synchronous Viterbi search on word-conditioned lexical tree copies is carried out. To enable recognition with a vocabulary containing more than one million words in real-time, the decoding process is highly parallelized by partitioning the vocabulary (and related lexical tree copies) to smaller units, their parallel decoding, and smart data synchronization. New words added to the vocabulary during the recognition are represented by an additional parallel unit which is simply integrated into the decoding process. Since trigram language model probabilities are factorized along the lexical trees on the fly, no pre-computing is required and new words can be recognized starting with the next time frame with full n-gram statistics.

During our first experiment, online detected names in all grammatical forms (both surname only and first name + surname) were added to the corresponding language model classes with the same weights. Consequently, with the increasing number of detected names their weights in each class decreased over time. This is not an optimal strategy because the recognition

system has to distinguish between more and more (often phonetically similar) names based on acoustic model only. Since the focus in individual sports usually moves from one sportsman to another as their performances are over, the probability that some name will be uttered by the commentator/respeaker decreases with time from the detection of that name in live sports graphics. This leads to some sort of "forgetting" so that recently detected names have higher weights in language model classes. We experimented with several forgetting curves, the details are described in the next section.

Although visual name spotting outside the sports graphics (mainly on sportsmen jerseys) is not usable as primary name detector due to the huge amount of false detections of short names, it can be used during forgetting process to reset the weights of previously spotted names in language model classes. This follows the basic TV sports scenario, where the sportsmen are introduced prior to their performance through the sports graphics and then the commentator talks principally about people who are in the shot.

5. Experiments

For experiments, we used part of our audiovisual corpus of TV broadcasting. Due to the high rate of OOV names, we have chosen three hours of live TV broadcasting of athletics from the Olympic Games in Rio de Janeiro in 2016. There are several athletic disciplines (e.g. 1500 Metres, Shot Put, Hammer Throw, Long Jump) accompanied by the commentary containing 12 452 words of which 736 are names of athletes. The visual name spotting process is carried out on one frame every second. The test data contain 832 visual name slots, this means there are 832 boxes with one name in the graphics (see Figure 1). There are 243 unique athlete names in these name slots, which may span several seconds.

Our visual name spotting process was able to detect 821 name slots, where the detection of the name slot was successful if the correct name was detected at least once during its displaying. Missing name slots were caused by trimmed text regions as in Figure 2 for the name "Malika Akkaoui". Considering one false detection, the visual name spotting detection error was 1.4 %.

We used our own speech recognition system optimized for low-latency real-time operation for experiments. Since three hours of test data commentary are respoken by the respeaker, a speaker-specific acoustic model was used. We use common three-state HMMs with output probabilities modeled by a Deep Neural Network (DNN). Language model consists of 527 762 words and 18 classes (as described in previous sections), from which only 6 classes for competitors names were used for experiments.

To show the problem of OOV names on our test data, the first experiment was carried out without any added name, so only names covered by the basic vocabulary could be recognized. The OOV rate on names was 39.35 % and Word Error Rate (WER) on names reached 50.82 % (see Table 1).

In the next step, we added all the names of athletes participating in the Olympic Games in Rio de Janeiro to the language model classes prior to the recognition. In real live subtitling, this could be a quite simple one-time job since this list should be known at the time when the Olympics starts. In total, 2 153 names were added comprising 4 173 items (surname and first name + surname) in each class. The total number of items is not equal to the expected double name count since some surnames of different athletes are the same. All names within one class have the same weight, hence the class perplexity is equal to the number of items in that class. To narrow the list of sportsmen and to lower names error rate in practice, only sportsmen participating in subtitled sports event (i.e. one TV program) are added to the recognition system just before subtitling. This requires a lot of manual work – there are hundreds of sports events at the Olympic Games. In our case, 250 names were added comprising 495 items in each language model class. See results for "All names prior" and "Event names prior" in Table 1.

	Mean class perplexity	**WER overall**	**WER names**
No added names		7.89 %	50.82 %
All names prior	4 173	2.44 %	8.97 %
Event names prior	495	2.13 %	6.39 %
Names from graphics	286	2.18 %	6.66 %
Names from graphics with forgetting	123	2.10 %	5.84 %
Names from graphics with forgetting and re-spotting	119	2.04 %	5.43 %

Table 1: *Experimental results of online LVCSR of three hours of TV broadcasting.*

Finally, the name classes were filled automatically during recognition based on online visual name spotting process described in this paper. The class perplexity in Table 1 is averaged over time. Achieved WER on names (6.66 %) is higher than in case of prior addition of event names (6.39 %) for the following reasons:

- The visual name spotting process is not error-free.
- Some names could be uttered before they are presented in the graphics.
- Seven reported sportsmen did not join the competition, but they could be mentioned by the commentator.

Considering the time saved by the automation of the name addition process this is already a very promising result. The error is relatively just 4 % higher. When we employ the forgetting of names in language model classes, the results are even better (see row "Names from graphics with forgetting" in Table 1) – a relative decrease of error by 8.6 % when compared to the manual addition of relevant names. The shape of the forgetting curve has some impact on the error rate. We experimented with three different curves; $f_1(x) = 30/(x + 30)$, $f_2(x) = 1-(x/300)^5$, $f_3(x) = 0.6-0.4\cdot\tanh(x/20-4)$; denoted exp(x), pow(x), and tanh(x), respectively, in Figure 3. The constants in the functions were set experimentally. In our experiment the curve $f_1(x)$ performed the best with WER 5.84 %, while $f_2(x)$ had error 7.61 %, and $f_3(x)$ had error 6.25 %. As the results show, the curve $f_2(x)$ is the worst due to the fact that at some point in time the name is forgotten completely and cannot be recognized, even though it may be uttered at a later time. By postponing the time of forgetting, the results are better, but they do not achieve the success of $f_1(x)$. Curve $f_3(x)$ performs better than the case when no forgetting is implemented, but not as good as $f_1(x)$. This seems to be due to a constant leftover remembering of the name which raises the total perplexity over time.

The best results are obtained when we support the forgetting technique by detection of names in the wild text. We reset

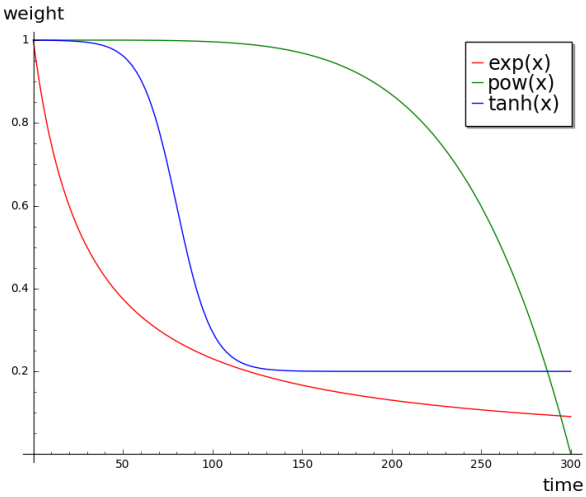

Figure 3: *Illustration of forgetting curves.*

the name weights based on the name re-spotting (see the last row of Table 1). It is natural since the presence of an athlete in the video indicates that the commentator may refer to him/her by name. This approach results in WER 5.43 % on names which is a relative improvement of 15 % when compared to the manual addition of relevant names to the vocabulary prior to the recognition.

6. Conclusion and Future Work

We have presented a visual name spotting process for a multimodal approach to handling out-of-vocabulary words – specifically names. We have demonstrated its impact on results of speech recognition of TV broadcasting of sports event. We have used a successful technique from the field of computer vision based on deep neural networks to detect texts in visual modality of TV broadcasting. We have identified names among these texts by matching them to a large vocabulary of sportsmen names. We have shown that adding the detected names to the recognition system online performs better than when all names are added beforehand (WER on names 6.66 % vs. 8.97 %). Next, we have introduced a forgetting technique which assigns weights to the detected names based on the elapsed time from the last appearance of the name in the video. With this technique, the WER on names drops to 5.43 % which is a relative decrease of error by 15 % when compared to a manual addition of relevant names before the sports event starts. By this, we have shown that the method has capabilities of improving systems of live TV subtitling both in operating costs and performance.

Although some algorithms in this paper were specifically designed for sports TV programs, they can be applied to a larger variety of multimodal data. On the other hand, it is possible to modify them to be even more general. A task-specific information – a list of all relevant sportsmen – can be substituted for example by a Google prompt allowing to validate detected names and even to correct them automatically. Furthermore, the visual text reader can be re-trained to task-specific data to improve visual name spotting results.

7. Acknowledgement

This research was supported by the Grand Agency of the Czech Republic, project no. P103/12/G084. Access to computing and storage facilities owned by parties and projects contributing to the National Grid Infrastructure MetaCentrum, provided under the programme "Projects of Large Research, Development, and Innovations Infrastructures" (CESNET LM2015042), is greatly appreciated.

8. References

[1] WHP and M. Evans, "Whp 065," 2003.

[2] A. Pražák, Z. Loose, J. Trmal, J. Psutka, and J. Psutka, "Novel approach to live captioning through re-speaking: Tailoring speech recognition to re-speaker's needs," in *INTERSPEECH*, 2012.

[3] C. Parada, M. Dredze, and F. Jelinek, "OOV sensitive named-entity recognition in speech," in *INTERSPEECH*, 2011.

[4] L. Qin, M. Sun, and A. Rudnicky, "OOV detection and recovery using hybrid models with different fragments," in *INTERSPEECH*, 2011.

[5] J. V. Psutka, A. Pražák, J. Psutka, and V. Radová, "Captioning of live tv commentaries from the olympic games in sochi: Some interesting insights," in *TSD*, 2014.

[6] B. Epshtein, E. Ofek, and Y. Wexler, "Detecting text in natural scenes with stroke width transform," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, June 2010, pp. 2963–2970.

[7] L. Neumann and J. Matas, "A method for text localization and recognition in real-world images," in *Computer Vision – ACCV 2010*, R. Kimmel, R. Klette, and A. Sugimoto, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 770–783.

[8] C. Yao, X. Bai, W. Liu, Y. Ma, and Z. Tu, "Detecting texts of arbitrary orientations in natural images," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, June 2012, pp. 1083–1090.

[9] M. Jaderberg, A. Vedaldi, and A. Zisserman, "Deep features for text spotting," in *European Conference on Computer Vision*, 2014.

[10] W. Huang, Y. Qiao, and X. Tang, "Robust scene text detection with convolution neural network induced mser trees," in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham: Springer International Publishing, 2014, pp. 497–511.

[11] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman, "Reading text in the wild with convolutional neural networks," *Int. J. Comput. Vision*, vol. 116, no. 1, pp. 1–20, Jan. 2016. [Online]. Available: http://dx.doi.org/10.1007/s11263-015-0823-z

[12] X. Zhou, C. Yao, H. Wen, Y. Wang, S. Zhou, W. He, and J. Liang, "East: An efficient and accurate scene text detector," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2642–2651, 2017.

[13] Y. Patel, M. Busta, and J. Matas, "E2E-MLT - an unconstrained end-to-end method for multi-language scene text," *CoRR*, vol. abs/1801.09919, 2018. [Online]. Available: http://arxiv.org/abs/1801.09919

[14] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd International Conference on Machine Learning*, ser. ICML '06. New York, NY, USA: ACM, 2006, pp. 369–376. [Online]. Available: http://doi.acm.org/10.1145/1143844.1143891