WiSeBE: Window-Based Sentence Boundary Evaluation

Carlos-Emiliano González-Gallardo[1,2(✉)] and Juan-Manuel Torres-Moreno[1,2]

[1] LIA - Université d'Avignon et des Pays de Vaucluse, 339 chemin des Meinajaries, 84140 Avignon, France
carlos-emiliano.gonzalez-gallardo@alumni.univ-avignon.fr, juan-manuel.torres@univ-avignon.fr

[2] Département de GIGL, École Polytechnique de Montréal, C.P. 6079, succ. Centre-ville, Montréal, Québec H3C 3A7, Canada

Abstract. Sentence Boundary Detection (SBD) has been a major research topic since Automatic Speech Recognition transcripts have been used for further Natural Language Processing tasks like Part of Speech Tagging, Question Answering or Automatic Summarization. But what about evaluation? Do standard evaluation metrics like precision, recall, F-score or classification error; and more important, evaluating an automatic system against a unique reference is enough to conclude how well a SBD system is performing given the final application of the transcript? In this paper we propose Window-based Sentence Boundary Evaluation (WiSeBE), a semi-supervised metric for evaluating Sentence Boundary Detection systems based on multi-reference (dis)agreement. We evaluate and compare the performance of different SBD systems over a set of Youtube transcripts using WiSeBE and standard metrics. This double evaluation gives an understanding of how WiSeBE is a more reliable metric for the SBD task.

Keywords: Sentence Boundary Detection · Evaluation
Transcripts · Human judgment

1 Introduction

The goal of Automatic Speech Recognition (ASR) is to transform spoken data into a written representation, thus enabling natural human-machine interaction [33] with further Natural Language Processing (NLP) tasks. Machine translation, question answering, semantic parsing, POS tagging, sentiment analysis and automatic text summarization; originally developed to work with formal written texts, can be applied over the transcripts made by ASR systems [2,25,31]. However, before applying any of these NLP tasks a segmentation process called Sentence Boundary Detection (SBD) should be performed over ASR transcripts to reach a minimal syntactic information in the text.

To measure the performance of a SBD system, the automatically segmented transcript is evaluated against a single reference normally done by a human. But

I. Batyrshin et al. (Eds.): MICAI 2018, LNAI 11289, pp. 119–131, 2018.
https://doi.org/10.1007/978-3-030-04497-8_10

given a transcript, does it exist a unique reference? Or, is it possible that the same transcript could be segmented in five different ways by five different people in the same conditions? If so, which one is correct; and more important, how to fairly evaluate the automatically segmented transcript? These questions are the foundations of Window-based Sentence Boundary Evaluation (WiSeBE), a new semi-supervised metric for evaluating SBD systems based on multi-reference (dis)agreement.

The rest of this article is organized as follows. In Sect. 2 we set the frame of SBD and how it is normally evaluated. WiSeBE is formally described in Sect. 3, followed by a multi-reference evaluation in Sect. 4. Further analysis of WiSeBE and discussion over the method and alternative multi-reference evaluation is presented in Sect. 5. Finally, Sect. 6 concludes the paper.

2 Sentence Boundary Detection

Sentence Boundary Detection (SBD) has been a major research topic science ASR moved to more general domains as conversational speech [17,24,26]. Performance of ASR systems has improved over the years with the inclusion and combination of new Deep Neural Networks methods [5,9,33]. As a general rule, the output of ASR systems lacks of any syntactic information such as capitalization and sentence boundaries, showing the interest of ASR systems to obtain the correct sequence of words with almost no concern of the overall structure of the document [8].

Similar to SBD is the Punctuation Marks Disambiguation (PMD) or Sentence Boundary Disambiguation. This task aims to segment a formal written text into well formed sentences based on the existent punctuation marks [11,19,20,29]. In this context a sentence is defined (for English) by the Cambridge Dictionary[1] as:

> *"a group of words, usually containing a verb, that expresses a thought in the form of a statement, question, instruction, or exclamation and starts with a capital letter when written".*

PMD carries certain complications, some given the ambiguity of punctuation marks within a sentence. A period can denote an acronym, an abbreviation, the end of the sentence or a combination of them as in the following example:

> *The U.S. president, Mr. Donald Trump, is meeting with the F.B.I. director Christopher A. Wray next Thursday at 8 p.m.*

However its difficulties, DPM profits of morphological and lexical information to achieve a correct sentence segmentation. By contrast, segmenting an ASR transcript should be done without any (or almost any) lexical information and a flurry definition of sentence.

[1] https://dictionary.cambridge.org/.

The obvious division in spoken language may be considered speaker utterances. However, in a normal conversation or even in a monologue, the way ideas are organized differs largely from written text. This differences, added to disfluencies like revisions, repetitions, restarts, interruptions and hesitations make the definition of a sentence unclear thus complicating the segmentation task [27]. Table 1 exemplifies some of the difficulties that are present when working with spoken language.

Table 1. Sentence Boundary Detection example

Speech transcript	SBD applied to transcript
two two women can look out after a kid so bad as a man and a woman can so you can have a you can have a mother and a father that that still don't do right with the kid and you can have to men that can so as long as the love each other as long as they love each other it doesn't matter	two // two women can look out after a kid so bad as a man and a woman can // so you can have a // you can have a mother and a father that // that still don't do right with the kid and you can have to men that can // so as long as the love each other // as long as they love each other it doesn't matter//

Stolcke and Shriberg [26] considered a set of linguistic structures as segments including the following list:

- Complete sentences
- Stand-alone sentences
- Disfluent sentences aborted in mid-utterance
- Interjections
- Back-channel responses.

In [17], Meteer and Iyer divided speaker utterances into segments, consisting each of a single independent clause. A segment was considered to begin either at the beginning of an utterance, or after the end of the preceding segment. Any dysfluency between the end of the previous segments and the begging of current one was considered part of the current segments.

Rott and Červa [23] aimed to summarize news delivered orally segmenting the transcripts into *"something that is similar to sentences"*. They used a syntactic analyzer to identify the phrases within the text.

A wide study focused in unbalanced data for the SBD task was performed by Liu *et al.* [15]. During this study they followed the segmentation scheme proposed by the Linguistic Data Consortium[2] on the Simple Metadata Annotation Specification V5.0 guideline (SimpleMDE_V5.0) [27], dividing the transcripts in Semantic Units.

[2] https://www.ldc.upenn.edu/.

A Semantic Unit (SU) is considered to be an atomic element of the transcript that manages to express a complete thought or idea on the part of the speaker [27]. Sometimes a SU corresponds to the equivalent of a sentence in written text, but other times (the most part of them) a SU corresponds to a phrase or a single word.

SUs seem to be an inclusive conception of a segment, they embrace different previous segment definitions and are flexible enough to deal with the majority of spoken language troubles. For these reasons we will adopt SUs as our segment definition.

2.1 Sentence Boundary Evaluation

SBD research has been focused on two different aspects; features and methods. Regarding the features, some work focused on acoustic elements like pauses duration, fundamental frequencies, energy, rate of speech, volume change and speaker turn [10,12,14].

The other kind of features used in SBD are textual or lexical features. They rely on the transcript content to extract features like bag-of-word, POS tags or word embeddings [7,12,16,18,23,26,30]. Mixture of acoustic and lexical features have also been explored [1,13,14,32], which is advantageous when both audio signal and transcript are available.

With respect to the methods used for SBD, they mostly rely on statistical/neural machine translation [12,22], language models [8,15,18,26], conditional random fields [16,30] and deep neural networks [3,7,29].

Despite their differences in features and/or methodology, almost all previous cited research share a common element; the evaluation methodology. Metrics as Precision, Recall, F1-score, Classification Error Rate and Slot Error Rate (SER) are used to evaluate the proposed system against one reference. As discussed in Sect. 1, further NLP tasks rely on the result of SBD, meaning that is crucial to have a good segmentation. But comparing the output of a system against a unique reference will provide a reliable score to decide if the system is good or bad?

Bohac *et al.* [1] compared the human ability to punctuate recognized spontaneous speech. They asked 10 people (correctors) to punctuate about 30 min of ASR transcripts in Czech. For an average of 3,962 words, the punctuation marks placed by correctors varied between 557 and 801; this means a difference of 244 segments for the same transcript. Over all correctors, the absolute consensus for period (.) was only 4.6% caused by the replacement of other punctuation marks as semicolons (;) and exclamation marks (!). These results are understandable if we consider the difficulties presented previously in this section.

To our knowledge, the amount of studies that have tried to target the sentence boundary evaluation with a multi-reference approach is very small. In [1], Bohac *et al.* evaluated the overall punctuation accuracy for Czech in a straightforward multi-reference framework. They considered a period (.) valid if at least five of their 10 correctors agreed on its position.

Kolář and Lamel [13] considered two independent references to evaluate their system and proposed two approaches. The fist one was to calculate the SER for each of one the two available references and then compute their mean. They found this approach to be very strict because for those boundaries where no agreement between references existed, the system was going to be partially wrong even the fact that it has correctly predicted the boundary. Their second approach tried to moderate the number of unjust penalizations. For this case, a classification was considered incorrect only if it didn't match either of the two references.

These two examples exemplify the real need and some straightforward solutions for multi-reference evaluation metrics. However, we think that it is possible to consider in a more inclusive approach the similarities and differences that multiple references could provide into a sentence boundary evaluation protocol.

3 Window-Based Sentence Boundary Evaluation

Window-Based Sentence Boundary Evaluation (WiSeBE) is a semi-automatic multi-reference sentence boundary evaluation protocol which considers the performance of a candidate segmentation over a set of segmentation references and the agreement between those references.

Let $\mathbf{R} = \{R_1, R_2, ..., R_m\}$ be the set of all available references given a transcript $T = \{t_1, t_2, ..., t_n\}$, where t_j is the j^{th} word in the transcript; a reference R_i is defined as a binary vector in terms of the existent SU boundaries in T.

$$R_i = \{b_1, b_2, ..., b_n\} \tag{1}$$

where

$$b_j = \begin{cases} 1 \text{ if } t_j \text{ is a boundary} \\ 0 \text{ otherwise} \end{cases}$$

Given a transcript T, the candidate segmentation C_T is defined similar to R_i.

$$C_T = \{b_1, b_2, ..., b_n\} \tag{2}$$

where

$$b_j = \begin{cases} 1 \text{ if } t_j \text{ is a boundary} \\ 0 \text{ otherwise} \end{cases}$$

3.1 General Reference and Agreement Ratio

A General Reference (R_G) is then constructed to calculate the agreement ratio between all references in. It is defined by the boundary frequencies of each reference $R_i \in \mathbf{R}$.

$$R_G = \{d_1, d_2, ..., d_n\} \tag{3}$$

where

$$d_j = \sum_{i=1}^{m} t_{ij} \quad \forall t_j \in T, \quad d_j = [0, m] \tag{4}$$

The Agreement Ratio ($R_{G_{AR}}$) is needed to get a numerical value of the distribution of SU boundaries over $\mathbf{R}$. A value of $R_{G_{AR}}$ close to 0 means a low agreement between references in $\mathbf{R}$, while $R_{G_{AR}} = 1$ means a perfect agreement ($\forall R_i \in \mathbf{R}, R_i = R_{i+1} | i = 1, ..., m-1$) in $\mathbf{R}$.

$$R_{G_{AR}} = \frac{R_{G_{PB}}}{R_{G_{HA}}} \tag{5}$$

In the equation above, $R_{G_{PB}}$ corresponds to the ponderated common boundaries of R_G and $R_{G_{HA}}$ to its hypothetical maximum agreement.

$$R_{G_{PB}} = \sum_{j=1}^{n} d_j \, [d_j \geq 2] \tag{6}$$

$$R_{G_{HA}} = m \times \sum_{d_j \in R_G} 1 \, [d_j \neq 0] \tag{7}$$

3.2 Window-Boundaries Reference

In Sect. 2 we discussed about how disfluencies complicate SU segmentation. In a multi-reference environment this causes disagreement between references around a same SU boundary. The way WiSeBE handle disagreements produced by disfluencies is with a Window-boundaries Reference (R_W) defined as:

$$R_W = \{w_1, w_2, ..., w_p\} \tag{8}$$

where each window w_k considers one or more boundaries d_j from R_G with a window separation limit equal to R_{W_l}.

$$w_k = \{d_j, d_{j+1}, d_{j+2}, ...\} \tag{9}$$

3.3 *WiSeBE*

WiSeBE is a normalized score dependent of (1) the performance of C_T over R_W and (2) the agreement between all references in $\mathbf{R}$. It is defined as:

$$WiSeBE = F1_{R_W} \times R_{G_{AR}} \quad WiSeBE = [0, 1] \tag{10}$$

where $F1_{R_W}$ corresponds to the harmonic mean of precision and recall of C_T with respect to R_W (Eq. 11), while $R_{G_{AR}}$ is the agreement ratio defined in (5). $R_{G_{AR}}$ can be interpreted as a scaling factor; a low value will penalize the overall *WiSeBE* score given the low agreement between references. By contrast, for a high agreement in $\mathbf{R}$ ($R_{G_{AR}} \approx 1$), $WiSeBE \approx F1_{R_W}$.

$$F1_{R_W} = 2 \times \frac{precision_{R_W} \times recall_{R_W}}{precision_{R_W} + recall_{R_W}} \tag{11}$$

$$precision_{R_W} = \frac{\sum_{b_j \in C_T} 1 \quad [b_j = 1, b_j \in w \quad \forall w \in R_W]}{\sum_{b_j \in C_T} 1 \quad [b_j = 1]} \tag{12}$$

$$recall_{R_W} = \frac{\sum_{w_k \in R_W} 1 \quad [w_k \ni b \quad \forall b \in C_T]}{p} \tag{13}$$

Equations 12 and 13 describe precision and recall of C_T with respect to R_W. Precision is the number of boundaries b_j inside any window w_k from R_W divided by the total number of boundaries b_j in C_T. Recall corresponds to the number of windows w with at least one boundary b divided by the number of windows w in R_W.

4 Evaluating with *WiSeBE*

To exemplify the $WiSeBE$ score we evaluated and compared the performance of two different SBD systems over a set of YouTube videos in a multi-reference environment. The first system (S1) employs a Convolutional Neural Network to determine if the middle word of a sliding window corresponds to a SU boundary or not [6]. The second approach (S2) by contrast, introduces a bidirectional Recurrent Neural Network model with attention mechanism for boundary detection [28].

In a first glance we performed the evaluation of the systems against each one of the references independently. Then, we implemented a multi-reference evaluation with $WiSeBE$.

4.1 Dataset

We focused evaluation over a small but diversified dataset composed by 10 YouTube videos in the English language in the news context. The selected videos cover different topics like technology, human rights, terrorism and politics with a length variation between 2 and 10 min. To encourage the diversity of content format we included newscasts, interviews, reports and round tables.

During the transcription phase we opted for a manual transcription process because we observed that using transcripts from an ASR system will difficult in a large degree the manual segmentation process. The number of words per transcript oscilate between 271 and 1,602 with a total number of 8,080.

We gave clear instructions to three evaluators (ref_1, ref_2, ref_3) of how segmentation was needed to be perform, including the SU concept and how punctuation marks were going to be taken into account. Periods (.), question marks (?), exclamation marks (!) and semicolons (;) were considered SU delimiters (boundaries) while colons (:) and commas (,) were considered as internal SU marks. The number of segments per transcript and reference can be seen in Table 2. An interesting remark is that ref_3 assigns about 43% less boundaries than the mean of the other two references.

Table 2. Manual dataset segmentation

Reference	v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8	v_9	v_{10}	Total
ref_1	38	42	17	11	55	87	109	72	55	16	502
ref_2	33	42	16	14	54	98	92	65	51	20	485
ref_3	23	20	10	6	39	39	76	30	29	9	281

4.2 Evaluation

We ran both systems (S1 & S2) over the manually transcribed videos obtaining the number of boundaries shown in Table 3. In general, it can be seen that S1 predicts 27% more segments than S2. This difference can affect the performance of S1, increasing its probabilities of false positives.

Table 3. Automatic dataset segmentation

System	v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8	v_9	v_{10}	Total
S1	53	38	15	13	54	108	106	70	71	11	539
S2	38	37	12	11	36	92	86	46	53	13	424

Table 4 condenses the performance of both systems evaluated against each one of the references independently. If we focus on F1 scores, performance of both systems varies depending of the reference. For ref_1, S1 was better in 5 occasions with respect of S2; S1 was better in 2 occasions only for ref_2; S1 overperformed S2 in 3 occasions concerning ref_3 and in 4 occasions for *mean* (**bold**).

Also from Table 4 we can observe that ref_1 has a bigger similarity to S1 in 5 occasions compared to other two references, while ref_2 is more similar to S2 in 7 transcripts (underline).

After computing the mean F1 scores over the transcripts, it can be concluded that in average S2 had a better performance segmenting the dataset compared to S1, obtaining a F1 score equal to 0.510. But... What about the complexity of the dataset? Regardless all references have been considered, nor agreement or disagreement between them has been taken into account.

All values related to the $WiSeBE$ score are displayed in Table 5. The Agreement Ratio ($R_{G_{AR}}$) between references oscillates between 0.525 for v_8 and 0.767 for v_5. The lower the $R_{G_{AR}}$, the bigger the penalization $WiSeBE$ will give to the final score. A good example is S2 for transcript v_4 where $F1_{R_W}$ reaches a value of 0.800, but after considering $R_{G_{AR}}$ the $WiSeBE$ score falls to 0.462.

It is feasible to think that if all references are taken into account at the same time during evaluation ($F1_{R_W}$), the score will be bigger compared to an average of independent evaluations ($F1_{mean}$); however this is not always true. That is the case of S1 in $v10$, which present a slight decrease for $F1_{R_W}$ compared to $F1_{mean}$.

Table 4. Independent multi-reference evaluation

Transcript	System	ref_1			ref_2			ref_3			$Mean$		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1
v_1	S1	0.396	0.553	0.462	0.377	0.606	0.465	0.264	0.609	0.368	0.346	0.589	0.432
	S2	0.474	0.474	**0.474**	0.474	0.545	**0.507**	0.368	0.6087	**0.459**	0.439	0.543	**0.480**
v_2	S1	0.605	0.548	**0.575**	0.711	0.643	**0.675**	0.368	0.700	**0.483**	0.561	0.630	**0.578**
	S2	0.595	0.524	0.557	0.676	0.595	0.633	0.351	0.650	0.456	0.541	0.590	0.549
v_3	S1	0.333	0.294	0.313	0.267	0.250	0.258	0.200	0.300	0.240	0.267	0.281	0.270
	S2	0.417	0.294	**0.345**	0.417	0.313	**0.357**	0.250	0.300	**0.273**	0.361	0.302	**0.325**
v_4	S1	0.615	0.571	0.593	0.462	0.545	0.500	0.308	0.667	0.421	0.462	0.595	0.505
	S2	0.909	0.714	**0.800**	0.818	0.818	**0.818**	0.455	0.833	**0.588**	0.727	0.789	**0.735**
v_5	S1	0.630	0.618	**0.624**	0.593	0.593	**0.593**	0.481	0.667	**0.560**	0.568	0.626	**0.592**
	S2	0.667	0.436	0.527	0.611	0.407	0.489	0.500	0.462	0.480	0.593	0.435	0.499
v_6	S1	0.491	0.541	**0.515**	0.454	0.563	0.503	0.213	0.590	0.313	0.386	0.565	0.443
	S2	0.500	0.469	0.484	0.522	0.552	**0.536**	0.250	0.590	**0.351**	0.4234	0.537	**0.457**
v_7	S1	0.594	0.578	**0.586**	0.462	0.533	0.495	0.406	0.566	0.473	0.487	0.559	0.518
	S2	0.663	0.523	0.585	0.558	0.522	**0.539**	0.465	0.526	**0.494**	0.562	0.524	**0.539**
v_8	S1	0.443	0.477	0.459	0.514	0.500	0.507	0.229	0.533	0.320	0.395	0.503	0.429
	S2	0.609	0.431	**0.505**	0.652	0.417	**0.508**	0.370	0.567	**0.447**	0.543	0.471	**0.487**
v_9	S1	0.437	0.564	0.492	0.451	0.627	0.525	0.254	0.621	0.360	0.380	0.603	**0.459**
	S2	0.623	0.600	**0.611**	0.585	0.608	**0.596**	0.321	0.586	**0.414**	0.509	0.598	0.541
v_{10}	S1	0.818	0.450	**0.581**	0.818	0.450	**0.581**	0.455	0.556	**0.500**	0.697	0.523	**0.582**
	S2	0.692	0.450	0.545	0.615	0.500	0.552	0.308	0.444	0.364	0.538	0.4645	0.487
Mean scores	S1	—		0.520	—		0.510	—		0.404	—		0.481
	S2	—		**0.543**	—		**0.554**	—		**0.433**	—		**0.510**

An important remark is the behavior of S1 and S2 concerning v_6. If evaluated without considering any (dis)agreement between references ($F1_{mean}$), S2 overperforms S1; this is inverted once the systems are evaluated with $WiSeBE$.

5 Discussion

5.1 $R_{G_{AR}}$ and Fleiss' Kappa correlation

In Sect. 3 we described the $WiSeBE$ score and how it relies on the $R_{G_{AR}}$ value to scale the performance of C_T over R_W. $R_{G_{AR}}$ can intuitively be consider an agreement value over all elements of $\mathbf{R}$. To test this hypothesis, we computed the Pearson correlation coefficient (PCC) [21] between $R_{G_{AR}}$ and the Fleiss' Kappa [4] of each video in the dataset (κ_R).

A linear correlation between $R_{G_{AR}}$ and κ_R can be observed in Table 6. This is confirmed by a PCC value equal to 0.890, which means a very strong positive linear correlation between them.

5.2 $F1_{mean}$ vs. $WiSeBE$

Results form Table 5 may give an idea that $WiSeBE$ is just an scaled $F1_{mean}$. While it is true that they show a linear correlation, $WiSeBE$ may produce a

Table 5. $WiSeBE$ evaluation

Transcript	System	$F1_{mean}$	$F1_{R_W}$	$R_{G_{AR}}$	$WiSeBE$
v_1	S1	0.432	0.495	0.691	0.342
	S2	**0.480**	0.513		**0.354**
v_2	S1	**0.578**	0.659	0.688	**0.453**
	S2	0.549	0.595		0.409
v_3	S1	0.270	0.303	0.684	0.207
	S2	**0.325**	0.400		**0.274**
v_4	S1	0.505	0.593	0.578	0.342
	S2	**0.735**	0.800		**0.462**
v_5	S1	**0.592**	0.614	0.767	**0.471**
	S2	0.499	0.500		0.383
v_6	S1	0.443	0.550	0.541	**0.298**
	S2	**0.457**	0.535		0.289
v_7	S1	0.518	0.592	0.617	0.366
	S2	**0.539**	0.606		**0.374**
v_8	S1	0.429	0.494	0.525	0.259
	S2	**0.487**	0.508		**0.267**
v_9	S1	0.459	0.569	0.604	0.344
	S2	**0.541**	0.667		**0.403**
v_{10}	S1	**0.582**	0.581	0.619	**0.359**
	S2	0.487	0.545		0.338
Mean scores	S1	0.481	0.545	0.631	0.344
	S2	**0.510**	0.567		**0.355**

Table 6. Agreement within dataset

Agreement metric	v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8	v_9	v_{10}
$R_{G_{AR}}$	0.691	0.688	0.684	0.578	0.767	0.541	0.617	0.525	0.604	0.619
κ_R	0.776	0.697	0.757	0.696	0.839	0.630	0.743	0.655	0.704	0.718

different system ranking than $F1_{mean}$ given the integral multi-reference principle it follows. However, what we consider the most profitable about $WiSeBE$ is the twofold inclusion of all available references it performs. First, the construction of R_W to provide a more inclusive reference against to whom be evaluated and then, the computation of $R_{G_{AR}}$, which scales the result depending of the agreement between references.

6 Conclusions

In this paper we presented WiSeBE, a semi-automatic multi-reference sentence boundary evaluation protocol based on the necessity of having a more reliable way for evaluating the SBD task. We showed how $WiSeBE$ is an inclusive metric which not only evaluates the performance of a system against all references, but also takes into account the agreement between them. According to your point of view, this inclusivity is very important given the difficulties that are present when working with spoken language and the possible disagreements that a task like SBD could provoke.

$WiSeBE$ shows to be correlated with standard SBD metrics, however we want to measure its correlation with extrinsic evaluations techniques like automatic summarization and machine translation.

Acknowledgments. We would like to acknowledge the support of CHIST-ERA for funding this work through the Access Multilingual Information opinionS (AMIS), (France - Europe) project.

We also like to acknowledge the support given by the Prof. Hanifa Boucheneb from VERIFORM Laboratory (École Polytechnique de Montréal).

References

1. Bohac, M., Blavka, K., Kucharova, M., Skodova, S.: Post-processing of the recognized speech for web presentation of large audio archive. In: 2012 35th International Conference on Telecommunications and Signal Processing (TSP), pp. 441–445. IEEE (2012)
2. Brum, H., Araujo, F., Kepler, F.: Sentiment analysis for Brazilian portuguese over a skewed class corpora. In: Silva, J., Ribeiro, R., Quaresma, P., Adami, A., Branco, A. (eds.) PROPOR 2016. LNCS (LNAI), vol. 9727, pp. 134–138. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-41552-9_14
3. Che, X., Wang, C., Yang, H., Meinel, C.: Punctuation prediction for unsegmented transcript based on word vector. In: LREC (2016)
4. Fleiss, J.L.: Measuring nominal scale agreement among many raters. Psychol. Bull. **76**(5), 378 (1971)
5. Fohr, D., Mella, O., Illina, I.: New paradigm in speech recognition: deep neural networks. In: IEEE International Conference on Information Systems and Economic Intelligence (2017)
6. González-Gallardo, C.E., Hajjem, M., SanJuan, E., Torres-Moreno, J.M.: Transcripts informativeness study: an approach based on automatic summarization. In: Conférence en Recherche d'Information et Applications (CORIA), Rennes, France, May (2018)
7. González-Gallardo, C.E., Torres-Moreno, J.M.: Sentence boundary detection for French with subword-level information vectors and convolutional neural networks. arXiv preprint arXiv:1802.04559 (2018)
8. Gotoh, Y., Renals, S.: Sentence boundary detection in broadcast speech transcripts. In: ASR2000-Automatic Speech Recognition: Challenges for the new Millenium ISCA Tutorial and Research Workshop (ITRW) (2000)

9. Hinton, G., et al.: Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. IEEE Signal Process. Mag. **29**(6), 82–97 (2012)
10. Jamil, N., Ramli, M.I., Seman, N.: Sentence boundary detection without speech recognition: a case of an under-resourced language. J. Electr. Syst. **11**(3), 308–318 (2015)
11. Kiss, T., Strunk, J.: Unsupervised multilingual sentence boundary detection. Comput. Linguist. **32**(4), 485–525 (2006)
12. Klejch, O., Bell, P., Renals, S.: Punctuated transcription of multi-genre broadcasts using acoustic and lexical approaches. In: 2016 IEEE Spoken Language Technology Workshop (SLT), pp. 433–440. IEEE (2016)
13. Kolář, J., Lamel, L.: Development and evaluation of automatic punctuation for French and english speech-to-text. In: Thirteenth Annual Conference of the International Speech Communication Association (2012)
14. Kolář, J., Švec, J., Psutka, J.: Automatic punctuation annotation in Czech broadcast news speech. In: SPECOM 2004 (2004)
15. Liu, Y., Chawla, N.V., Harper, M.P., Shriberg, E., Stolcke, A.: A study in machine learning from imbalanced data for sentence boundary detection in speech. Comput. Speech Lang. **20**(4), 468–494 (2006)
16. Lu, W., Ng, H.T.: Better punctuation prediction with dynamic conditional random fields. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. pp. 177–186. Association for Computational Linguistics (2010)
17. Meteer, M., Iyer, R.: Modeling conversational speech for speech recognition. In: Conference on Empirical Methods in Natural Language Processing (1996)
18. Mrozinski, J., Whittaker, E.W., Chatain, P., Furui, S.: Automatic sentence segmentation of speech for automatic summarization. In: 2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings, vol. 1, p. I. IEEE (2006)
19. Palmer, D.D., Hearst, M.A.: Adaptive sentence boundary disambiguation. In: Proceedings of the Fourth Conference on Applied Natural Language Processing, pp. 78–83. ANLC 1994. Association for Computational Linguistics, Stroudsburg, PA, USA (1994)
20. Palmer, D.D., Hearst, M.A.: Adaptive multilingual sentence boundary disambiguation. Comput. Linguist. **23**(2), 241–267 (1997)
21. Pearson, K.: Note on regression and inheritance in the case of two parents. Proc. R. Soc. Lond. **58**, 240–242 (1895)
22. Peitz, S., Freitag, M., Ney, H.: Better punctuation prediction with hierarchical phrase-based translation. In: Proceedings of the International Workshop on Spoken Language Translation (IWSLT), South Lake Tahoe, CA, USA (2014)
23. Rott, M., Červa, P.: Speech-to-text summarization using automatic phrase extraction from recognized text. In: Sojka, P., Horák, A., Kopeček, I., Pala, K. (eds.) TSD 2016. LNCS (LNAI), vol. 9924, pp. 101–108. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-45510-5_12
24. Shriberg, E., Stolcke, A.: Word predictability after hesitations: a corpus-based study. In: Proceedings of the Fourth International Conference on Spoken Language, 1996. ICSLP 1996, vol. 3, pp. 1868–1871. IEEE (1996)
25. Stevenson, M., Gaizauskas, R.: Experiments on sentence boundary detection. In: Proceedings of the sixth conference on Applied natural language processing, pp. 84–89. Association for Computational Linguistics (2000)

26. Stolcke, A., Shriberg, E.: Automatic linguistic segmentation of conversational speech. In: Proceedings of the Fourth International Conference on Spoken Language, 1996. ICSLP 1996, vol. 2, pp. 1005–1008. IEEE (1996)
27. Strassel, S.: Simple metadata annotation specification v5. 0, linguistic data consortium (2003). http://www.ldc.upenn.edu/projects/MDE/Guidelines/SimpleMDE_V5.0.pdf
28. Tilk, O., Alumäe, T.: Bidirectional recurrent neural network with attention mechanism for punctuation restoration. In: Interspeech 2016 (2016)
29. Treviso, M.V., Shulby, C.D., Aluisio, S.M.: Evaluating word embeddings for sentence boundary detection in speech transcripts. arXiv preprint arXiv:1708.04704 (2017)
30. Ueffing, N., Bisani, M., Vozila, P.: Improved models for automatic punctuation prediction for spoken and written text. In: Interspeech, pp. 3097–3101 (2013)
31. Wang, W., Tur, G., Zheng, J., Ayan, N.F.: Automatic disfluency removal for improving spoken language translation. In: 2010 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), pp. 5214–5217. IEEE (2010)
32. Xu, C., Xie, L., Huang, G., Xiao, X., Chng, E.S., Li, H.: A deep neural network approach for sentence boundary detection in broadcast news. In: Fifteenth Annual Conference of the International Speech Communication Association (2014)
33. Yu, D., Deng, L.: Automatic Speech Recognition. Springer, London (2015). https://doi.org/10.1007/978-1-4471-5779-3