Proceedings of Symposium in Information and Human Language Technology. Uberlândia, MG, Brazil, October 2–5, 2017.

Evaluating Word Embeddings for Sentence Boundary Detection in Speech Transcripts

Marcos V. Treviso[1], **Christopher D. Shulby**[1,2], **Sandra M. Aluísio**[1]

[1] Institute of Mathematics and Computer Science, University of São Paulo (USP)

[2]CPqD

marcostreviso@usp.br {cshulby,sandra}@icmc.usp.br

Abstract. *This paper is motivated by the automation of neuropsychological tests involving discourse analysis in the retellings of narratives by patients with potential cognitive impairment. In this scenario the task of sentence boundary detection in speech transcripts is important as discourse analysis involves the application of Natural Language Processing tools, such as taggers and parsers, which depend on the sentence as a processing unit. Our aim in this paper is to verify which embedding induction method works best for the sentence boundary detection task, specifically whether it be those which were proposed to capture semantic, syntactic or morphological similarities.*

1. Introduction

The concept of a sentence in written or spoken texts is important in several Natural Language Processing (NLP) tasks, such as morpho-syntactic analysis [Kepler and Finger 2010, Fonseca and Aluísio 2016], sentiment analysis [Brum et al. 2016], and speech processing [Mendonça et al. 2014], among others. However, punctuation marks that constitute a sentence boundary are ambiguous The Disambiguation of Punctuation Marks (DPM) task analyzes punctuation marks in texts and indicates whether they correspond to a sentence boundary. The purpose of the DPM task is to answer the question: *Among the tokens of punctuation marks in a text, which of them correspond to sentence boundaries?*

The Sentence Boundary Detection (SBD) task is very similar to DPM, both of which attempt to break a text into sequential units that correspond to sentences, where DPM is text-based and SBD can be applied to either written text or audio transcriptions and often for clauses, which do not necessarily end in final punctuation marks but are complete thoughts nonetheless. However, performing SBD in speech texts is more complicated due to the lack of information such as punctuation and capitalization; moreover text output is susceptible to recognition errors, in case of Automatic Speech Recognition (ASR) systems are used for automatic transcriptions [Gotoh and Renals 2000]. SBD from speech transcriptions is a task which has gained more attention in the last decades due to the increasing popularity of ASR software which automatically generate text from audio input. This task can also be applied to written texts, like online product reviews [Silla Jr and Kaestner 2004, Read et al. 2012, López and Pardo 2015], in order to better their intelligibility and facilitate the posterior use of NLP tools.

It is important to point out that the differences between spoken and written texts are notable, mainly when we take into consideration the size of the utterances and the number of disfluencies provided in speech. Disfluencies include filled pauses, repetitions,

modifications, repairs, partial utterances, nonword vocalizations and false starts. These phenomena are very common in spontaneous speech. [Liu 2004].

Figure 1 shows the result of a transcript from a neuropsychological retelling task that does not include either capitalization or sentence segmentation, preventing the direct application of NLP methods that rely on these marks for their correct use, such as taggers and parsers. One can easily note that this type of text differs greatly in style and form from written/edited text (on which most NLP tools are trained), such as text found in novels or a newspaper.

> cinderela a história da cinderela... ela:: encontra um cavaleiro com com um cavalo dai ela fica amiga desse cavalo tudo isso é próximo de um castelo e ela vai pro castelo pro castelo na verdade ela vai trabalhar no castelo né e ela começa a fazer lá...

Figure 1. Narrative excerpt transcribed using the NURC annotation manual[1]

These tests are applied by clinicians who tell a story to patients who are instructed to try and remember as many details as possible so that they may retell it. The evaluation of language in discourse production, mainly in narratives, is an attractive alternative because it allows the analysis of linguistic microstructures and phonetic-phonological, morpho-syntactic, semantic-lexical components, as well as semantic-pragmatic macrostructures. Neuropsychological tests are used in clinical settings for detection, progress monitoring and treatment observation in patients with dementias. In an ideal scenario we would like to automate the application of neuropsychological tests and the discourse analysis of the retellings.

NLP applications generally receive text as input; therefore, words can be considered the basic processing unit. In this case, it is important that they are represented in a way which carries the load of all relevant information. In the current approach used here, words are induced representations in a dense vector space. These representations are known as word embeddings; able to capture semantic, syntactic and morphological information from large unannotated corpora [Mikolov et al. 2013, Ling et al. 2015, Lai et al. 2015]. Various studies show that textual information is important for SBD [Gotoh and Renals 2000, Batista et al. 2012, Che et al. 2016]. Even though textual information is a strong indicator for sentence delimitation, boundaries are often associated with prosodic information [Shriberg et al. 2000, Batista et al. 2012], like pause duration, change in pitch and change in energy. However, the extraction of this type of information requires the use of high quality resources, and consequently, few resources with prosodic information are available. On the other hand, textual information can easily be extracted in large scale from the web. Textual information can also be represented in various ways for SBD, for example, n-gram based techniques have presented good results for SBD [Gotoh and Renals 2000, Kim and Woodland 2003, Favre et al. 2008]; however, in contrast to word embeddings, they are induced representations in a sparse vector space.

Our aim in this paper is to verify which embedding induction method works best for the SBD task, specifically whether it be those which were proposed to capture seman-

[1] http://www.letras.ufrj.br/nurc-rj/

tic, syntactic or morphological similarities. For example, we imagine that methods that capture morphological similarities may benefit the SBD performance for impaired speech, since a large number of words produced in this type of speech are out-of-vocabulary words.The paper is organized as follows. Section 2 presents related work on SBD using word embeddings; Section 3 describes the word embedding models evaluated in this paper; Section 4 presents our experimental setup, describing the datasets, method, and preprocessing steps used; Section 5 presents our findings and discussions. Finally, Section 6 concludes the paper and outlines some future work.

2. Related Work

The work of [Che et al. 2016] and [Tilk and Alumäe 2015] use word embeddings to detect boundaries in prepared speech sentences, more specifically in the corpus from 2012 TED talks[2]. [Che et al. 2016] propose a CNN (Convolution Neural Network)-based method with 50 dimensions using GloVe [Pennington et al. 2014]. In [Klejch et al. 2016, Klejch et al. 2017] the authors show that that textual information influences the retrieval of punctuation marks more than prosodic information, even without the use of word embeddings.

The work in [Tilk and Alumäe 2015] is expanded in [Tilk and Alumäe 2016], using bidirectional neural networks with attention mechanisms to evaluate a spontaneous telephone conversation corpus. The authors point out that the bidirectional vision of the RNN (Recurrent Neural Network) is a more impacting feature than the attention mechanism for SBD; with only the use of word embeddings, the achieved results yielded only 10% less than when prosodic information was used together. In [Hough and Schlangen 2017] a system that uses RNNs with word embeddings is proposed for the SBD task in conjunction disfluencies, where results are competitive with the state of the art are achieved on the Switchboard corpus [Godfrey et al. 1992], showing that the simultaneous execution of these tasks is superior to when done individually.

Recently, the work of [Treviso et al. 2017] proposed an automatic SBD method for impaired speech in Brazilian Portuguese, to allow a neuropsychological evaluation based on discourse analysis. The method uses RCNNs (Recurrent Convolutional Neural Networks) which independently treat prosodic and textual information, reaching state-of-the-art results for impaired speech. Also, this study showed that it is possible to achieve good results when comparing them with prepared speech, even when practically the same quantity of text is used. Another interesting evidence was that the use of word embeddings, without morpho-syntactic labels was able to present the same results as when they were used; this indicates that word embeddings contain sufficient morpho-syntactic information for SBD. It was also shown that the method gains the better results than the state-of-the-art method used by [Fraser et al. 2015] by a great margin for both impaired and prepared speech (an absolute difference of $\sim$0.20 and $\sim$0.30, respectively). Beyond these findings, the method showed that the performance remains the same when a different story is used.

3. Word Embeddings Models

The generation of vector representations of words (or word embeddings) is linked to the induction method utilized. The work of [Turian et al. 2010] divides these representations

[2] https://www.ted.com/talks

into three categories: cluster-based, distributional and distributed methods. In this paper, we focus only on distributed representations, because generally they are computationally faster to be induced. These representations are based on real vectors distributed in a multidimensional space induced by unsupervised learning. In the following paragraphs, we describe the three induction methods for word embeddings utilized in our evaluations.

A well-used NLP technique, Word2vec [Mikolov et al. 2013] follows the same principle as the natural language model presented in [Collobert and Weston 2008], with the exception that it does not use a hidden layer, generating a computationally faster log-linear model. This technique is divided into two modeling types: (i) Continuous Bag-of-Words (CBOW), which given a window of words as input, the network tries to predict the word in the middle as output and (ii) the Skip-gram model, which tries to predict the window given the center word as input.

As Word2vec does not consider the word order in the window, this make the process less syntactic in nature, since word order is an essential phenomenon for syntax. In order to deal with this, a modification of Word2vec was proposed which is able to deal with word order by concatenating inputs in the CBOW model (instead of using the sum) and incremental weighting for Skip-gram. This technique is known as Wang2vec [Ling et al. 2015].

A recent induction technique called FastText [Bojanowski et al. 2016, Joulin et al. 2016] uses n-grams of characters of a given word in the hope of capturing morphological information. In order to do this, the Skip-gram Word2vec model was modified so that that the scoring function of the network's output is calculated basing itself on the character n-gram vectors, which are summed with the context vectors in order to represent a word.

4. Experimental Setup

4.1. Corpora/Datasets

The datasets were divided into two categories: impaired speech and prepared speech. Impaired speech is not only spontaneous, but also noisy. The noise is produced internally due to the impaired neuropsychological condition of the participants studied. When people participate in neuropsychological tests, they produce the following phenomena: syntactically malformed sentences; mispronounced words (modifying the original morphology); low quality prosody (due to the shallow voices of the participants and/or abnormal fluctuations in vocal quality); and in general a great quantity and variety of types of disfluencies.

The first dataset of discourse tests is a set of impaired speech narratives, based on a book of sequenced pictures from the well-known Cinderella story. This dataset consists of 60 narrative texts told by Brazilian Portuguese speakers; 20 healthy subjects, called controls (CTL), 20 patients with Alzheimer's disease (AD), and 20 patients with Mild Cognitive Impairment (MCI), diagnosed at Medical School of the University of São Paulo (FMUSP) and also used in [Aluísio et al. 2016]. The second dataset was made available by the FalaBrasil project, and its contents are structured in the same way as the Brazilian Constitution from 1988 [Batista et al. 2013]. The speech in this corpus can be categorized as prepared and also as read speech. To use these files in our scenario a preprocessing step was necessary, which removed lexical tips which indicate the beginning of articles,

sections and paragraphs. This removal was carried out on both the transcripts and the audio. In addition, we separated the new dataset organized by articles, yielding 357 texts in total. Both datasets' properties are presented in Table 1.

Property	Cinderela	Constitution
# Transcipts	60	357
# Words	23807	63275
# Sentences	2066	2698
Duration	4h 11m	7h 39m

Table 1. Summary of corpora utilized in the experiments.

The corpus used to induce the vectors is made up of text from Wikipedia in 2016, a news crawler which collected articles from the G1 portal[3] and the PLN-BR [Bruckschen et al. 2008] corpus. We also executed some basic preprocessing steps on this corpus, being that we forced all of the text to lowercase forms and separated each token from punctuation marks and tokenized the text using whitespace. We do not remove stopwords. After these steps, the embedding induction on the corpus returned ∼356M tokens, of which ∼1.7M were distinct.

4.2. Method

In order to automatically extract new features from the input and at the same time deal with the long dependency problems between words, we propose a method based on RCNNs which was inspired by the sentence segmentation work done by [Tilk and Alumäe 2015] and [Che et al. 2016], and also by the work on text classification utilizing RCNNs by [Lai et al. 2015], where we made some adaptations so that the basic unit of classification was a data sequence. The architecture of our RCNN is the same as the one used in [Treviso et al. 2017] and can be seen in Figure 2.

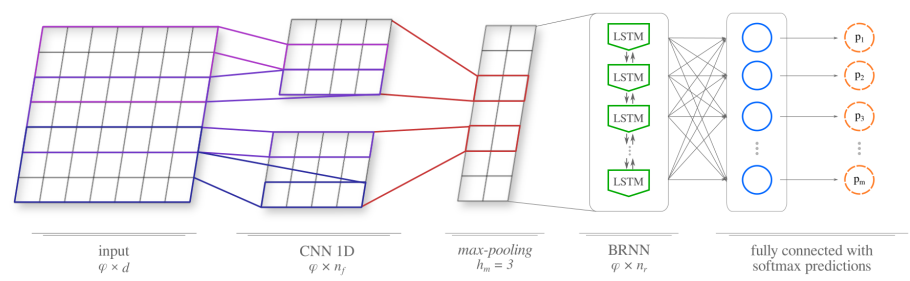


Figure 2. Architecture adapted from [Treviso et al. 2017]

The final model in [Treviso et al. 2017] consists of a linear combination between a model which deals only with lexical information and another which treats only prosodic information. In this paper, we ignore the prosodic model and focus only on the textual information provided by the word embeddings. The strategy to utilize only this information is based on the idea that one can train a text-based model with a large amount of data, since text is readily found on the web.

[3] http://g1.globo.com/

The model's input is a tridimensional word embedding matrix $\mathbf{E} \in \mathbb{R}^{m \times \varphi \times d}$, where m is equal to the vocabulary size used for training the embeddings. Once we have an input matrix composed by word embeddings, the convolutional layer extracts n_f new features from a sliding window with the size h_c, which corresponds to the size of the filter applied to the concatenated vectors $[e_1, ..., e_{h_c}]$ corresponding to a region of h_c neighboring embeddings [Kim 2014].

The convolutional layer produces features for each t-th word as it applies the shared filter for a window of h_c embeddings $\mathbf{e}_{t-h_c+1:t}$ in a sentence with the size φ. Our convolutional layer moves in a single vertical dimension (CNN 1D), one step at a time, which results in a quantity of filters q_f equal to $\varphi - h_c + 2 * p + 1$. And since we want to classify exactly φ elements, we added $p = \lfloor h_c/2 \rfloor$ elements of padding to both sides of the sentence. In addition, we applied a max-pooling operation on the temporal axis focusing on a region of h_m words, with the idea of feeding only the most important features to the next layer.

The features selected by the max-pooling layer are fed to a recurrent layer. The values of the hidden units are computed utilizing n_r LSTM cells [Hochreiter and Schmidhuber 1997] defined as activation units. As in [Tilk and Alumäe 2016], our recurrent layer is based on anterior and posterior temporal states using the bidirectional recurrent mechanism (BRNN). With the use of a bidirectional layer which treats convolutionized features, the network is adept at exploring the principal that nearby words usually have a great influence, while considering that distant words, either to the left or right, can also have an impact on the classification. This frequently happens in the SBD task, for example, in the case of interrogatives, question words like "quem" ("*who*"), "qual" ("*what*") and "quando" ("*when*") can define a sentence.

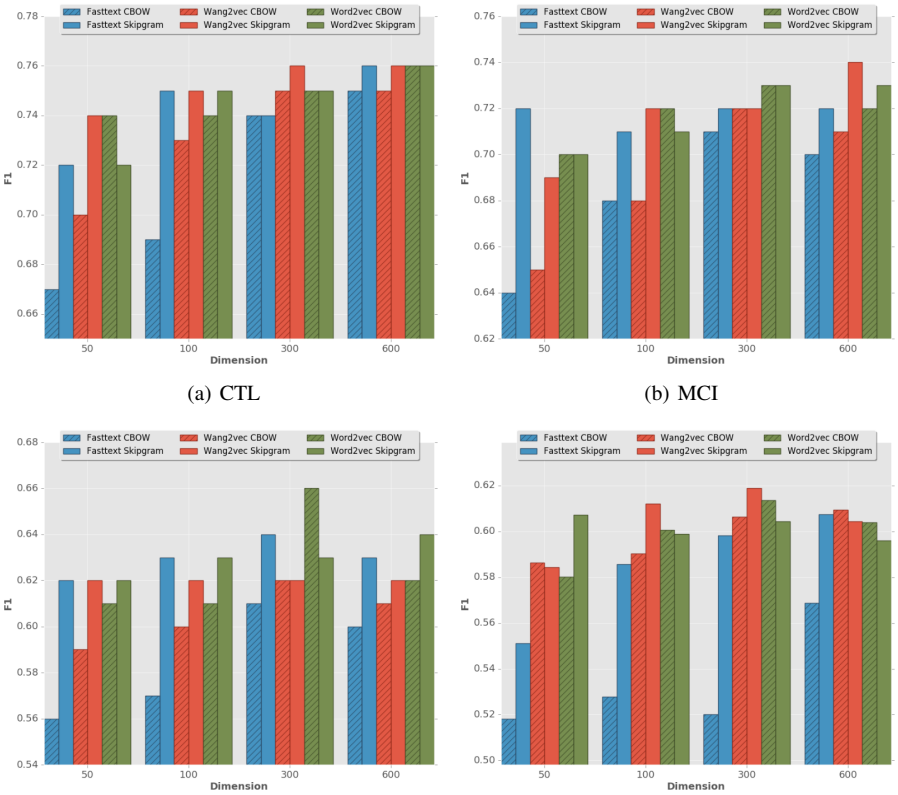
After the BRNN layer, we use dropout as a regularization strategy, which attempts to prevent co-adaptation between hidden units during forward and back-propagation, where some neurons are ignored with the purpose of reducing the chance of overfitting [Srivastava et al. 2014]. Finally, the last layer, receives the output of the BRNN for each instance t, and feeds each into a simple fully connected layer which produces predictions using the softmax activation function, which gives us the final probability that a word precedes a sentence boundary (B) or not (NB).

The word embeddings matrix $\mathbf{E}$ was adjusted during training. Our RCNN uses the same hyperparameters described in [Treviso et al. 2017] and the same training strategy, which consists of cost-function minimization utilizing the RMSProp procedure [Tieleman and Hinton 2012] with back-propagation, considering the unbalanced task of sentence segmentation by penalizing errors from the minority class harsher (B).

5. Results and Discussion

We ran a 5-fold cross-validation for each group analyzed (CLT, MCI or AD), which left about 10% of the data for testing, the rest for training.

The performance results of the RCNN in terms of F_1 on each type of patient and on the Constitution dataset are shown in Figure 3, for which we vary the embedding methods and its training strategies along with the induced vector dimensions between the values of: $d \in \{50, 100, 300, 600\}$.



(a) CTL (b) MCI

(c) AD (d) Constitution

Figure 3. Results for different embedding methods and dimensions

In most cases Word2vec achieved better performance than other methods. Specifically, for CTL with Skip-gram and FastText with CBOW, yielding an F_1 of 0.76, On the other hand, we see that for MCI patients, Wang2vec with Skip-gram was the best technique, yielding an F_1 of 0.74. For the AD subjects the best technique was Word2vec with CBOW, returning an F_1 of 0.66. As expected, results for CTL were higher than for MCI and AD, since the CTL narratives contain less noise. For Constitution data our method performs better using Wang2vec with Skip-gram strategy: F_1 of 0.62.

It is possible to see in Figure 3 that our method tends to better its performance with increasing dimension size. Furthermore, the Skip-gram strategy generally returned better results than CBOW for the FastText and Wang2vec methods, whereas for Word2Vec there were some variations when strategies were switched. Still, the Word2vec Skip-gram with 600 dimensions and CBOW with 300 dimensions were those which returned the best results for spontaneous and/or impaired speech (CTL, MCI and AD). In the case of the Constitution dataset, which is characterized as prepared and read speech, the best results were achieved by Wang2vec Skip-gram with 300 dimensions.

Contrary to the results reported in [Treviso et al. 2017] using textual and prosodic

information, our method obtained better performance for impaired speech transcriptions than for prepared speech. This is probably due to the fact that the Constitution includes more impacting prosodic clues, whereas for spontaneous/impaired speech, the lexical clues are of greater influence for classification. This difference between lexical and prosodic features for prepared and spontaneous speech is consistent with the finding reported in other studies [Kolár et al. 2009, Fraser et al. 2015, Treviso et al. 2017].

6. Conclusion and Future Work

Our objective in this work was to identify the embedding with the best performance for SBD, specifically whether it would be one which captures semantic information, like Word2vec; syntactic, like Wang2vec; or morphological, like FastText. Still, we were not able to discern which type was most influential in general, since the differences from one to another are very small. Also even when one technique was superior to another for a particular set, we still need to investigate whether this was actually the fault of the technique or due to secondary factors, like hyperparameters, random initialization, or even the conditions of the data used.

In general, our results show that using only embeddings the RCNN method achieved similar results (difference of 1%) to the state of the art in terms of F_1, using the same method published in [Treviso et al. 2017] for both classes: CTL and MCI using embeddings with 600 dimensions and prosodic information[4]. However, the results for the Constitution dataset were considerably lower (a difference of 17%) than the results of the model which uses both lexical and prosodic information in conjunction, but the difference is less (4%) for the models which used only prosodic information. Summing up, this indicates that by using a good word embedding model to represent textual information it is possible to achieve similar results with the state-of-the-art for impaired speech.

Future work will include some investigation of the lexical and prosodic clues which impact the classification. Also, we would like to investigate whether disfluency detection in conjunction with SBD can yield better results. Since the method presented in this paper can easily be applied to any language, we plan to evaluate it using English language corpora in order to directly compare the results with the related work.

References

Aluísio, S., Cunha, A., and Scarton, C. (2016). Evaluating progression of alzheimer's disease by regression and classification methods in a narrative language test in portuguese. In *PROPOR*, pages 109–114.

Batista, F., Moniz, H., Trancoso, I., and Mamede, N. (2012). Bilingual experiments on automatic recovery of capitalization and punctuation of automatic speech transcripts. *IEEE Transactions on Audio, Speech, and Language Processing*, pages 474–485.

Batista, P. d. S. et al. (2013). Avanços em reconhecimento de fala para português brasileiro e aplicações: ditado no libreoffice e unidade de resposta audível com asterisk. Master's thesis, Universidade Federal do Pará.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2016). Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

[4]Results using embeddings with 600 dimensions were obtained from the authors of the original article

Bruckschen, M., Muniz, F., Souza, J., Fuchs, J., Infante, K., Muniz, M., Gonçalves, P., Vieira, R., and Aluısio, S. (2008). Anotaçao lingüıstica em xml do corpus pln-br. *Série de relatórios do NILC, ICMC-USP*.

Brum, H., Araujo, F., and Kepler, F. (2016). Sentiment analysis for brazilian portuguese over a skewed class corpora. In *PROPOR*, pages 134–138.

Che, X., Wang, C., Yang, H., and Meinel, C. (2016). Punctuation prediction for unsegmented transcript based on word vector. *LREC*, pages 654–658.

Collobert, R. and Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In *ICML*, pages 160–167.

Favre, B., Hakkani-Tür, D., Petrov, S., and Klein, D. (2008). Efficient sentence segmentation using syntactic features. *Spoken Language Technology Workshop*.

Fonseca, E. R. and Aluísio, S. M. (2016). Improving pos tagging across portuguese variants with word embeddings. In *PROPOR*, pages 227–232.

Fraser, K. C., Ben-david, N., Hirst, G., Graham, N. L., and Rochon, E. (2015). Sentence segmentation of aphasic speech. *NAACL*, pages 862–871.

Godfrey, J. J., Holliman, E. C., and McDaniel, J. (1992). Switchboard: Telephone speech corpus for research and development. In *ICASSP*, pages 517–520. IEEE.

Gotoh, Y. and Renals, S. (2000). Sentence boundary detection in broadcast speech transcripts. *ISCA Workshop*, pages 228–235.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, pages 1735–1780.

Hough, J. and Schlangen, D. (2017). Joint, incremental disfluency detection and utterance segmentation from speech. In *EACL*, pages 326–336.

Joulin, A., Grave, E., Bojanowski, P., and Mikolov, T. (2016). Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Kepler, F. N. and Finger, M. (2010). Variable-length markov models and ambiguous words in portuguese. In *NAACL*, pages 15–23.

Kim, J.-H. and Woodland, P. C. (2003). A combined punctuation generation and speech recognition system and its performance enhancement using prosody. *Speech Communication*, pages 563–577.

Kim, Y. (2014). Convolutional neural networks for sentence classification. In *EMNLP*, pages 1746–1751.

Klejch, O., Bell, P., and Renals, S. (2016). *Punctuated Transcription of Multi-genre Broadcasts Using Acoustic and Lexical Approaches*.

Klejch, O., Bell, P., and Renals, S. (2017). *Sequence-to-Sequence Models for Punctuated Transcription Combing Lexical and Acoustic Features*.

Kolár, J., Liu, Y., and Shriberg, E. (2009). Genre effects on automatic sentence segmentation of speech: A comparison of broadcast news and broadcast conversations. In *ICASSP*, pages 4701—-4704.

Lai, S., Xu, L., Liu, K., and Zhao, J. (2015). Recurrent convolutional neural networks for text classification. In *AAAI*, pages 2267–2273.

Ling, W., Dyer, C., Black, A., and Trancoso, I. (2015). Two/too simple adaptations of word2vec for syntax problems. In *NAACL*.

Liu, Y. (2004). *STRUCTURAL EVENT DETECTION FOR RICH TRANSCRIPTION OF SPEECH*. PhD thesis, Purdue University.

López, R. and Pardo, T. A. S. (2015). Experiments on sentence boundary detection in user-generated web content. In *CICLing*, pages 227–237.

Mendonça, G., Candeias, S., Perdigão, F., Shulby, C., Toniazzo, R., Klautau, A., and Aluísio, S. (2014). A method for the extraction of phonetically-rich triphone sentences. In *Telecommunications Symposium (ITS), 2014 International*, pages 1–5. IEEE.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543.

Read, J., Dridan, R., Oepen, S., and Solberg, L. J. (2012). Sentence boundary detection: A long solved problem? *COLING*, pages 985–994.

Shriberg, E., Stolcke, A., Hakkani-Tür, D., and Tür, G. (2000). Prosody-based automatic segmentation of speech into sentences and topics. *Speech Communication*, pages 127–154.

Silla Jr, C. N. and Kaestner, C. A. (2004). An analysis of sentence boundary detection systems for english and portuguese documents. In *CICLing*, pages 135–141.

Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, pages 1929–1958.

Tieleman, T. and Hinton, G. (2012). Rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural Networks for Machine Learning*.

Tilk, O. and Alumäe, T. (2015). LSTM for punctuation restoration in speech transcripts. In *INTERSPEECH*, pages 683–687.

Tilk, O. and Alumäe, T. (2016). Bidirectional recurrent neural network with attention mechanism for punctuation restoration. In *INTERSPEECH*.

Treviso, M. V., Shulby, C., and Aluísio, S. M. (2017). Sentence segmentation in narrative transcripts from neuropsychological tests using recurrent convolutional neural networks. In *EACL*, pages 315–325.

Turian, J., Ratinov, L., and Bengio, Y. (2010). Word representations: A simple and general method for semi-supervised learning. In *ACL*, pages 384–394.