IDIAP RESEARCH REPORT

EMPIRICAL EVALUATION AND COMBINATION OF PUNCTUATION PREDICTION MODELS APPLIED TO BROADCAST NEWS

Alexandre Nanchen Philip N. Garner

Idiap-RR-01-2019

Version of FEBRUARY 06, 2019

Centre du Parc, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

EMPIRICAL EVALUATION AND COMBINATION OF PUNCTUATION PREDICTION MODELS APPLIED TO BROADCAST NEWS

Alexandre Nanchen, Philip N. Garner

Idiap Research Institute, Martigny, Switzerland

ABSTRACT

Natural language processing techniques are dependent upon punctuation to work well. When their input is taken from speech recognition, it is necessary to reconstruct the punctuation; in particular sentence boundaries. We define a range of features from low level acoustics to those with high level lexical semantics, including deep and recurrent models; these in turn are representative of a broad range of approaches used by previous authors for punctuation prediction. We combine the features using a gradient boosting machine that is also capable of indicating the relative importance of each feature. In an empirical study, we show that features from different semantic levels are in fact complementary, that combining statistical and deep learning methods yields better prediction results, and that generalization across different speaking styles is difficult to achieve without adaptation. Our best model achieves an F-Measure of 82.8 on a challenging broadcast news dataset.

Index Terms— sentence boundaries, features importance, models combination, statistical language model, recurrent neural network

1. INTRODUCTION

We are interested in general in automatic speech recognition (ASR), and in particular in cases where the output is subsequently used in tasks such as translation and summarization. Such natural language processing (NLP) tasks are normally trained on text data; use of speech recognition introduces mismatches. Of course, one mismatch arises from the errors owing to the limitations of the speech recognition. However, another important mismatch is that speech recognition output does not contain punctuation. This is in contrast to the text case, where punctuation is present, and is an important cue for NLP tasks. The situation is interesting because ASR and NLP groups tends to work separately, and because punctuation solutions arise in both fields. This latter point raises questions about an obvious solution, which would be to build punctuation into either the ASR or NLP module. Before attempting to do so, it is necessary to understand the relative contribution of the two approaches.

Sentence boundaries are an essential part of text. Tasks like human proofreading or machine translation heavily rely on them [1, 2]. The topic of sentence boundary prediction in ASR is not new; rather it has been investigated for more than two decades. Lexical features capture information about the coherence of word sequences and hence give some cues about sentence boundaries. Their influence has been studied in various manners: lexical rules [3], hidden event language models (HELM) [4, 5] and more recently deep learning methods like word vectors [6] and sequences of words as features [7, 8]. Acoustic related features have also been shown to help in the prediction of sentence boundaries: pauses and word durations are good indicators of end of sentences but can also appear in unnatural places [5, 9]. Pitch and energy help in the disambiguation of full stops from question and exclamation marks [10, 11]. Some robustness of prediction has been obtained by combining different sources of information at different levels of abstraction: at the feature level using decision trees [12], boosting algorithms [13], conditional random fields [14] or at prediction level using probabilistic methods [15] or statistical methods such as voting, linear and logistic regression [16]. More recently, a successful combination of features has been done using deep learning embedding space representations [17].

For comparison purposes, in this study, we focus on the German language only and we try to choose features that are at least representative of most of the models described above, but without attempting to duplicate them exactly. These include low level acoustic features describing pauses, and prosodic features indicative of intent. They are complemented by N-gram language model features, and deep neural networks (DNNs) of the type used in NLP applications.

Although the current trend is towards neural network models, in this paper we evaluate whether such models really lead to improvements when tested on the same data. We test the (intuitive) hypothesis that features from different semantic levels are in fact complementary. We also study the benefit of combining statistical models with DNN models either at the lexical or acoustic level. Finally we investigate whether a combination of models is better for punctuation prediction and, if so, whether the relative contribution of techniques is consistent across datasets. To this end we have chosen to use a machine learning ensemble method known as a “gradient boosting machine” (GBM) [18]. GBMs have several advantages in this case w.r.t. DNNs: They don't need a lot of data to figure out the importance of each model, their training comes up with a relative importance metric that highlights the contribution of each feature, they are robust to over-fitting and they deal well with unbalanced datasets.

In the remainder of the paper, we first describe the primary features in detail, then go on to describe the databases on which we evaluate, and the evaluation itself.

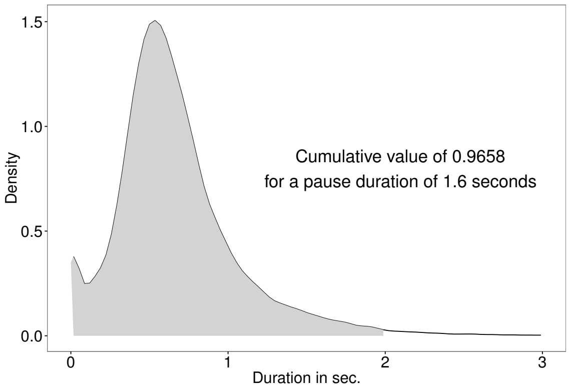

Fig. 1. Standarization of pause durations w.r.t. the training distribution

2. METHOD

The study is primarily experimental. Here we describe the features of interest and how they are handled. Our approach is to evaluate predictions of single feature classifiers, called primary classifiers, and then study the impact of their combination.

2.1. Primary classifier features

Pauses and word durations are good indicators of end of sentences but can also appear in unnatural places [5, 9]. In this study we choose to model pause duration as the cumulative value of a given silence w.r.t. the training silence distribution. The number is a value between 0.0 and 1.0 and represents the significance of a given silence w.r.t. what was observed during training. This method allows for pause duration normalization. See Fig. 1 for a visual explanation.

Pitch and energy help in the disambiguation of full stops from question and exclamation marks by giving cues about the intent of the speaker [10, 11]. In our experiments we use the prosody features obtained from the Kaldi toolkit [19]. They include a probability-of-voicing estimate, pitch and delta-pitch features. The training data consist of negative and positive prosody paths. A prosody path is defined by the last 100 frames (one second) before the end of a word. If there are not enough speech frames, the path is padded with zeros to reach 100 values. If the word is the last of a sentence, the prosody path is labelled as positive otherwise as negative. The prosody probability feature is estimated with a bidirectional long short term memory network (BiLSTM) that has learned to classify positive prosody paths vs. negative prosody paths using a cross entropy loss.

Lexical features capture information about the coherence of word sequences and hence give some cues about sentence boundaries. In this study three lexical models are estimated: two statistical language models and one deep learning model. The intuition behind this approach is that N-grams are good at modeling short term dependencies and dealing with unobserved sequence of words by backing-off, while deep neural networks (DNN) are able to capture longer contexts. This complementarity has been studied in the field of statistical language modeling [20] and we believe it could also help in the task of sentence boundary prediction. The N-gram probability features are two-fold: a forward and a backward probability. The forward probability is defined as the conditional probability of having a sentence boundary after a sequence of words: $\mathrm{Pr}(\texttt{</s>} \mid \texttt{w}_{\texttt{i-1}}, \texttt{w}_{\texttt{i-2}}, \texttt{w}_{\texttt{i-3}})$ where $\texttt{</s>}$ is the end of sentence symbol and $w_{i-1}, w_{i-2}, w_{i-3}$ are the words before the end of the sentence. The backward probability is defined as the conditional probability that a word sequence is preceded by a sentence boundary: $\mathrm{Pr}(\texttt{</s>} \mid \texttt{w}_{\texttt{i+1}}, \texttt{w}_{\texttt{i+2}}, \texttt{w}_{\texttt{i+3}})$ where $w_{i+1}, w_{i+2}, w_{i+3}$ are the words following the sentence boundary. It is estimated by reversing the word order in the sentences. The lexical deep learning model is obtained by training a BiLSTM sentence boundary classifier. The lexical probability feature is the posterior probability that a sentence boundary was emitted given a sequence of words. Only the forward probability is estimated for long context modelling as we believed that the long term influence of the next sentence is of lesser importance.

2.2. Combination of predictions

The final score is computed by combining the probability features predicted by the primary classifiers using a gradient boosting machine. A GBM is a machine learning ensemble method that combines the predictions of weak decision trees [18]. The weak learners are applied in cascade and each one tries to correct the errors of the previous one. The main difference with other boosting methods, like Adaboost, is that the error from the previous classifier is corrected using a gradient descent algorithm. In the past, such methods have been used in the task of Punctuation Prediction to combine discrete features with continuous ones, i.e. lexical with prosody features [12].

3. EXPERIMENTAL SETUP

3.1. Data

The Europarl text corpus [21] is a multilingual parallel corpus that is freely available and used in many natural language processing tasks (NLP) such as machine translation (MT) and language modelling (LM). One of its key characteristic is that it is closely related to spoken speech as it was collected from the European Parliament proceedings. The German part of the corpus (55M of words) is prepared using the Asrt[1] open source toolkit. Example of transformations include the conversion of numbers and dates to their written equivalent. The resulting text is then partitioned into training, validation, and testing sets using respectively 70%, 20% and 10% of the data.

Two German acoustic datasets are used in this paper: the "Technische Universität München broadcast news" (TUM BCN) [22] and the "Scalable Understanding of Multilingual Media" (SUMMA) datasets. The TUM BCN corpus consists of over 160 hours of high quality, manually segmented and annotated radio broadcast news, most of which is pronounced like read speech. Annotation is performed at the paragraph level. Each paragraph contains one or more sentences. Only full stops have been annotated. The dataset is further partitioned into training, validation and testing sets using the same 70%, 20% and 10% proportions as for the text data. The SUMMA dataset consist of 8 hours of manually transcribed data including full stops, commas, exclamation marks and question marks. It originates from two broadcast news channels, Deutsche Welle (DW) and IRIB. The speech content is less structured than the German BCN corpus and sentence boundaries are more difficult to predict. For adaptation and evaluation purposes, the SUMMA dataset is split into two parts: a training and a testing set. Both sets share an equal amount of news channel types (DW and IRIB) and audio recording time.

3.2. Training procedure

We have used three open source toolkits to perform our experiments: the MIT Language Modeling toolkit for statistical N-grams language model estimation[2], the PyTorch toolkit [23] for DNN training and the scikit-learn toolkit [24] for the GBM training.

The German Europarl training set text is used to train two statistical 4-grams language models and the BiLSTM lexical sentence boundary classifier. For the latter one, the following parameters have been empirically chosen to optimize the binary cross entropy (BCE) loss on the German Europarl validation set: word embedding of 100 dimensions, 50 hidden dimensions, 3 bidirectional hidden layers and a maximum sequence size of 7 words. The vocabulary list is the combined words of both 4-gram models.

The BiLSTM prosody sentence boundary classifier is trained on the TUM BCN training set. Its parameters are chosen empirically to optimize the BCE loss of the TUM BCN validation set. The optimal network contains 3 hidden bidirectional layers with three hidden features each and a maximum sequence length of 100 samples. The features are standardized (centered and divided by the standard deviation) before training.

Finally, the training of the GBM classifier is performed two times; one time on the TUM BCN validation set and one time on the SUMMA training set. In total, 14 classifiers are trained.

3.3. Feature importance metric

The GBM training yields a relative importance metric that highlights the contribution of each feature. It is defined as "The number of times a variable is selected for splitting, weighted by the squared improvement to the model as a result of each split, and averaged over all trees" [25, 18]. More formally, for a given tree T:

$$\hat{I}_j^2(T) = \sum_{t=1}^{J-1} \hat{i}_t^2 \mathbb{1}(v_t = j) \tag{1}$$

The right term sums over all non terminal node of tree T. The indicator function is set to one when the splitting variable v_t is equal to the selected variable j. $\hat{i}_t^2$ is the improvement in squared error as a result of the split.

4. EXPERIMENTS

Our formulated hypotheses are tested by performing experiments on the two described acoustic datasets. On the TUM BCN test set, we run and evaluate punctuation prediction on both reference and ASR transcripts; on the SUMMA evaluation set, only on the ASR output. The F-measure retrieval metric is used to be robust to the class imbalance inherent in the sentence boundary prediction task. For each evaluation, we first evaluate the primary classifiers and then their different combinations.

The results on the TUM BCN evaluation set (Table 1, first row, ASR column) show that pause duration contains most of the information about sentence boundaries with an F-measure of 84.1. The prosody feature comes in second position with an F-measure of 77.8 (Table 1 fourth row, ASR column). We did not expect such a high score for prosody. Our explanation is that the speaking style is close to read speech and hence the prosody paths can be learned effectively. The lexical results confirm the hypothesis that N-grams and LSTM features are complementary for the task of sentence boundaries prediction. When combined, they perform better than when used alone. The ASR F-measure is also very close to the reference one, 92.2 vs. 94.2 (Table 1, last row), meaning that the best setup is somehow robust to ASR errors (WER of 8.66%). Looking at the GBM relative training feature importance for the best setup (Table 2, third row) we can observe that the combined lexical features have more influence on the decision process than the combined acoustic related features. This make sense because of the well defined grammatical structure of sentences in the TUM BCN dataset.

[1] https://github.com/idiap/asrt

[2] https://github.com/mitlm/mitlm

Model type	F-measure		Precision		Recall	
	Ref	ASR	Ref	ASR	Ref	ASR
Pause duration (PD)	85.4	84.1	91.5	89.3	80.1	79.4
LM	52.2	51.2	73.5	70.4	40.9	40.2
LSTM	56.4	54.0	76.3	72.6	44.8	43.0
PROSODY (P)	78.7	77.8	79.8	75.9	77.6	79.9
LM + LSTM	63.8	61.3	78.6	75.0	53.6	51.9
LM + LSTM + PD	89.5	87.9	93.0	90.4	86.3	85.5
LM + LSTM + PD + P	94.2	92.2	94.7	91.7	93.7	92.8

Table 1. Sentence boundary prediction both in ASR and reference text for the TUM BCN evaluation set. We can see that the best model is a combination of all features

Feature type Model type	PD	Left LM	Right LM	LSTM	P
LM + LSTM		38.2	29.9	31.9	
LM + LSTM + PD	40.1	21.1	21.6	17.2	
LM + LSTM + PD + P	26.3	14.3	21.0	18.2	20.2

Table 2. GBM relative training features importance. The pause duration feature is the most decisive

The SUMMA evaluation set is evaluated with two GBM models: one trained with ”out of domain” data (TUM BCN) and one trained with ”in domain” data (SUMMA). The ”in domain” training can be viewed as an adaptation process to the new data characteristics. Its goal is to optimize the relative feature importance according to the properties of the ”in domain” data. The results show that, despite a high WER (41.18), the best F-measure has a high value of 82.8 (Table 3, last row). However, the prosody gain observed on the TUM BCN evaluation set is now very small, meaning that the learned prosody paths do not generalize well. The difference in the prosody path structure per dataset is highlighted in Fig. 2. Table 4 (third row) presents the redistribution of the feature importance. On a dataset with less structured speech, the combined acoustic related features have more importance in the decision process than the combined lexical features.

Model type	F-measure		Precision		Recall	
	BCN	SUM	BCN	SUM	BCN	SUM
Pause duration (PD)	72.3	76.5	74.0	84.7	70.6	69.8
LM	40.8	29.7	56.5	73.6	32.0	18.6
LSTM	49.8	39.0	64.4	70.9	40.6	26.9
PROSODY (P)	28.5	2.8	39.7	53.6	22.3	1.5
LM + LSTM	52.8	46.8	65.5	72.9	44.2	34.4
LM + LSTM + PD	79.2	82.2	83.0	87.8	75.8	77.3
LM + LSTM + PD + P	74.6	**82.8**	**88.7**	88.5	64.4	**77.9**

Table 3. Sentence boundary prediction in ASR transcripts for the SUMMA evaluation set. Two GBMs are evaluated; the first one is trained on the same domain as the primary classifiers, the second one is adapted to the SUMMA domain

Feature type Model type	PD	Left LM	Right LM	LSTM	P
LM + LSTM		29.6	28.2	42.2	
LM + LSTM + PD	54.1	12.4	15.4	18.1	
LM + LSTM + PD + P	53.5	11.9	10.2	12.8	11.6

Table 4. Adapted GBM training feature importance. In comparison with Table 2, part of the decision mass has been shifted from the prosody feature to the pause duration feature

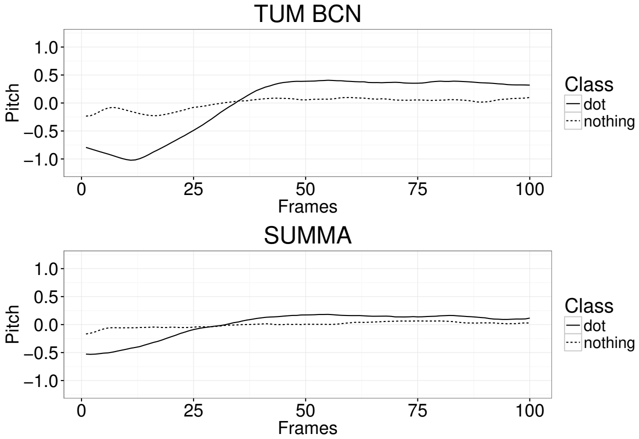

Fig. 2. A comparison of the standardized pitch prosody paths. The curves are the average of one hundred paths. They have a different structure from one dataset to another

5. CONCLUSION

In this paper we have proposed a technique for combining punctuation prediction models and reported on an empirical evaluation on the individual models in comparison to the combined systems. Experiments on Broadcast News have confirmed the hypothesis that individual sentence boundary prediction models are not robust enough to beat an ensemble algorithm. The information from different semantic levels is necessary to be robust to different grammatical and speaking styles. We have also confirmed that, either on the lexical or the acoustic related side, traditional statistical models and more recent deep neural network approaches are complementary. Finally we have observed that the relative importance of combined features is dependent on the characteristics of the evaluation set and hence some adaptation is needed to reach the best results.

6. ACKNOWLEDGEMENTS

This work was supported by the European Union under the EU Horizon 2020 Research and Innovation Action SUMMA [3] and by the Intelligence Advanced Research Project Activity under the project SARAL.

[3] http://summa-project.eu

7. REFERENCES

[1] Sharath Rao, Ian R. Lane, and Tanja Schultz, "Optimizing sentence segmentation for spoken language translation," in *INTERSPEECH 2007, 8th Annual Conference of the International Speech Communication Association, Antwerp, Belgium, August 27-31, 2007*, 2007, pp. 2845–2848.

[2] Eunah Cho, Jan Niehues, Kevin Kilgour, and A. Waibel, "Punctuation insertion for real-time spoken language translation," 2015.

[3] Jeffrey C. Reynar and Adwait Ratnaparkhi, "A maximum entropy approach to identifying sentence boundaries," in *Proceedings of the Fifth Conference on Applied Natural Language Processing*, Stroudsburg, PA, USA, 1997, ANLC '97, pp. 16–19, Association for Computational Linguistics.

[4] Yang Liu, Elizabeth Shriberg, Andreas Stolcke, Dustin Hillard, Mari Ostendorf, and Mary P. Harper, "Enriching speech recognition with automatic detection of sentence boundaries and disfluencies," *IEEE Trans. Audio, Speech & Language Processing*, vol. 14, no. 5, pp. 1526–1540, 2006.

[5] Nicola Ueffing, Maximilian Bisani, and Paul Vozila, "Improved models for automatic punctuation prediction for spoken and written text," in *INTERSPEECH*, 2013.

[6] Xiaoyin Che, Cheng Wang, Haojin Yang, and Christoph Meinel, "Punctuation prediction for unsegmented transcript based on word vector," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may 2016, European Language Resources Association (ELRA).

[7] Ottokar Tilk and Tanel Alumäe, "LSTM for punctuation restoration in speech transcripts," in *INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, September 6-10, 2015*, 2015, pp. 683–687.

[8] Ottokar Tilk and Tanel Alumäe, "Bidirectional recurrent neural network with attention mechanism for punctuation restoration," in *INTERSPEECH*, 2016.

[9] C Xu, L Xie, G Huang, Xiong Xiao, Eng Chng, and Haizhou Li, "A deep neural network approach for sentence boundary detection in broadcast news," pp. 2887–2891, 01 2014.

[10] Elizabeth Shriberg, Andreas Stolcke, Dilek Z. Hakkani-Tür, and Gökhan Tür, "Prosody-based automatic segmentation of speech into sentences and topics," *CoRR*, vol. cs.CL/0006036, 2000.

[11] Ondrej Klejch, Peter Bell, and Steve Renals, "Punctuated transcription of multi-genre broadcasts using acoustic and lexical approaches," in *2016 IEEE Spoken Language Technology Workshop, SLT 2016, San Diego, CA, USA, December 13-16, 2016*, 2016, pp. 433–440.

[12] Jáchym Kolár and Lori Lamel, "Development and evaluation of automatic punctuation for french and english speech-to-text," in *INTERSPEECH*, 2012.

[13] Sébastien Cuendet, Elizabeth Shriberg, J. Fung, Dilek Hakkani-Tur, and dilek, "An analysis of sentence segmentation features for broadcast news , broadcast conversations , and meetings," 2007.

[14] Wei Lu and Hwee Tou Ng, "Better punctuation prediction with dynamic conditional random fields," in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. 2010, EMNLP '10, pp. 177–186, Association for Computational Linguistics.

[15] Andreas Stolcke, Elizabeth Shriberg, Rebecca Bates, Mari Ostendorf, Dilek Hakkani, Madelaine Plauche, Gokhan Tur, and Yu Lu, "Automatic detection of sentence boundaries and disfluencies based on recognized words," 1998.

[16] M. Magimai-Doss, D. Hakkani-Tur, O. Cetin, E. Shriberg, J. Fung, and N. Mirghafori, "Entropy based classifier combination for sentence segmentation," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, April 2007, vol. 4, pp. IV–189–IV–192.

[17] Ondrej Klejch, Peter Bell, and Steve Renals, "Sequence-to-sequence models for punctuated transcription combining lexical and acoustic features," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2017, New Orleans, LA, USA, March 5-9, 2017*, 2017, pp. 5700–5704.

[18] Jerome H. Friedman, "Greedy function approximation: A gradient boosting machine," *Annals of Statistics*, vol. 29, pp. 1189–1232, 2000.

[19] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Nagendra Goel, Mirko Hannemann, Yanmin Qian, Petr Schwarz, and Georg Stemmer, "The kaldi speech recognition toolkit," in *In IEEE 2011 workshop*, 2011.

[20] Tomas Mikolov, Anoop Deoras, Stefan Kombrink, Lukás Burget, and Jan Cernocký, "Empirical evaluation and combination of advanced language modeling techniques," in *INTERSPEECH*, 2011.

[21] Philipp Koehn, "Europarl: A Parallel Corpus for Statistical Machine Translation," in *Conference Proceedings: the tenth Machine Translation Summit*, Phuket, Thailand, 2005, AAMT, pp. 79–86, AAMT.

[22] Felix Weninger, Björn W. Schuller, Florian Eyben, Martin Wöllmer, and Gerhard Rigoll, "A broadcast news corpus for evaluation and tuning of german LVCSR systems," *CoRR*, vol. abs/1412.4616, 2014.

[23] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer, "Automatic differentiation in pytorch," in *NIPS-W*, 2017.

[24] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine Learning in Python ," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[25] J. Elith, J. R. Leathwick, and T. Hastie, "A working guide to boosted regression trees," 2008.