Low Latency MaxEnt- and RNN-Based Word Sequence Models for Punctuation Restoration of Closed Caption Data

Máté Ákos Tündik(✉), Balázs Tarján, and György Szaszák

Department of Telecommunications and Media Informatics,
Budapest University of Technology and Economics,
Magyar Tudósok körútja 2, Budapest 1117-H, Hungary
{tundik,tarjanb,szaszak}@tmit.bme.hu

Abstract. Automatic Speech Recognition (ASR) rarely addresses the punctuation of the obtained transcriptions. Recently, Recurrent Neural Network (RNN) based models were proposed in automatic punctuation exploiting wide word contexts. In real-time ASR tasks such as closed captioning of live TV streams, text based punctuation poses two particular challenges: a requirement for low latency (limiting the future context), and the propagation of ASR errors, seen more often for informal or spontaneous speech. This paper investigates Maximum Entropy (MaxEnt) and RNN punctuation models in such real-time conditions, but also compares the models to off-line setups. As expected, the RNN outperforms the MaxEnt baseline system. Limiting future context results only in a slighter performance drop, whereas ASR errors influence punctuation performance considerably. A genre analysis is also carried out w.r.t. the punctuation performance. Our approach is also evaluated on TED talks within the IWSLT English dataset providing comparable results to the state-of-the-art systems.

Keywords: Punctuation recovery · Recurrent Neural Network · LSTM · Maximum Entropy · Low latency real-time modeling

1 Introduction

Punctuation insertion into the output of Automatic Speech Recognition (ASR) is a known problem in speech technology. The importance of having punctuations in automatically generated text – transcripts, indexing, closed captions, for metadata extraction etc. – has been outlined several times [1,16], as punctuation helps both human readability, and also eventual subsequent processing with text based tools, which usually require the punctuation marks at the very first step of their operation: the tokenization. In dictation systems, punctuation marks can be explicitly dictated; however, in several other domains where ASR is used, this is not possible.

N. Camelin et al. (Eds.): SLSP 2017, LNAI 10583, pp. 155–166, 2017.
DOI: 10.1007/978-3-319-68456-7_13

Two basic approaches can be distinguished for automatic punctuation, although they are often used in combination: prosody and text based approaches. In general prosody based approaches require less computation, less training data and hence can result in lightweight punctuation models. They are also more robust to ASR errors; recently proposed text based approaches on the other hand provide mostly more accurate punctuation, but are more sensitive to ASR errors and may introduce high latency due to the processing of a wide context, requiring extensive computations and also future context which directly results in high latency.

In this paper we focus on reducing this latency by still maintaining the accuracy provided by text based models. We demonstrate systems intended to be used for punctuation of closed-captioned data. ASR technology is widely used by television companies to produce closed captions especially for live programs [21], which require almost real-time processing with little latency.

Much effort has been devoted to develop reliable punctuation restoration algorithms, early approaches proposed to add punctuation marks to the N-gram language model of the ASR as hidden events [8,23]. These models have to be trained on huge corpora to reduce data sparsity [8]. More sophisticated sequence modeling approaches were also inspired by this idea: a transducer alike approach getting a non-punctuated text as input is capable of predicting punctuation as was presented in numerous works [1,3,11], with frameworks built on top of Hidden Markov Models (HMM), Maximum Entropy (MaxEnt) models or conditional random fields, etc. MaxEnt models allow for any easy combination of textual and prosodic features into a common punctuation model [10]. In a comprehensive study [2], many features were compared in terms of their effect on punctuation accuracy of a MaxEnt model. It was found that the most powerful textual features were the word forms and part-of-speech (POS) tags, whereas the best prosodic feature was the duration of inter-word pauses.

Applying a monolingual translation paradigm for punctuation regarded as a sequence modeling task was also proposed in [5], which also allowed for considerably reducing time latency. Recently, sequence-to-sequence modeling deep neural network based solutions have been also presented: taking a large word-context and projecting the words via an embedding layer into a bidirectional Recurrent Neural Network (RNN) [22], high quality punctuation could be achieved. RNNs are successfully used in many sequence labeling tasks as they are able to model large contexts and to learn distributed features of words to overcome data sparsity issues. The first attempt to use RNN for punctuation restoration was presented in [24], where a one-directional LSTM [9] was trained on Estonian broadcast transcripts. Shortly after, Tilk and Alumäe introduced a bidirectional RNN model using GRU [7] together with attention mechanism, which outperformed previous state-of-the-art on Estonian and English IWSLT datasets [25]. In a recent study [15], capitalization and punctuation recovery are treated as correlated multiple sequence labeling tasks and modeled with bidirectional RNN. In [14], a prosody based punctuation approach was proposed using an RNN on top of phonological phrase sequence modeling.

In this paper, we introduce a lightweight RNN-based punctuation restoration model using bidirectional LSTM units on top of word embeddings, and compare its performance to a MaxEnt model. We pay a special attention to low latency solutions. Both approaches are evaluated on automatic and manual transcripts and in various setups including on-line and off-line operation. We present results on Hungarian broadcast speech transcripts and the IWSLT English dataset [4] to make the performance of our approach comparable to state-of-the-art systems. Apart form the purely prosody based approach outlined in [14], we are not aware of any prior work for punctuation restoration for Hungarian speech transcripts.

Our paper is structured in the following way: first we present the used datasets in Sect. 2, then we move on to presenting the experimental systems in Sect. 3. The results of Hungarian and English Punctuation Restoration tasks are presented and discussed in Sect. 4. Our conclusions and future ideas are drawn in Sect. 5.

2 Data

2.1 The Hungarian Broadcast Dataset

The Hungarian dataset consists of manually transcribed closed captions made available by the Media Service Support and Asset Management Fund (MTVA), Hungary's public service broadcaster. The dataset contains captions for various TV genres enabling us to evaluate the punctuation models on different speech types, such as weather forecasts, broadcast news and conversations, magazines, sport news and sport magazines. We focus on the restoration of those punctuations, which have a high importance for understandability in Hungarian: commas, periods, question marks and exclamation marks. The colons and semicolons were mapped to comma. All other punctuation symbols are removed from the corpora. We reserve a disjunct 20% of the corpus for validation and use a representative test set, not overlapping with training and validation subsets. For further statistics about training and test data we refer the reader to Table 1.

Table 1. Statistics of the Hungarian dataset

Genres	Training & Validation					Test					
	#Words	#Com	#Per	#Que	#Excl	#Words	#Com	#Per	#Ques	#Excl	WER
Weather	478 K	40 K	31.5 K	30	730	2.4 K	250	200	0	20	6.8
Brc.-News	3493 K	279 K	223 K	3.5 K	4.6 K	17 K	1.5 K	1 K	20	50	10.1
Sport News	671 K	55 K	39.5 K	280	2 K	6 K	500	400	2	30	21.4
Brc.-Conv.	4161 K	533 K	225 K	26.5 K	4 K	46.8 K	6.3 K	2.6 K	250	130	24.7
Sport mag.	-	-	-	-	-	22.7 K	2 K	1.4 K	100	50	30.3
Magazine	4909 K	732 K	376 K	72 K	36 K	10.4 K	1.5 K	700	150	70	38.7
Mixed	1526 K	187 K	102 K	11 K	11.4 K	30.7	4 K	1.7 K	280	150	-
ALL	15238 K	1826 K	997 K	113 K	58.8 K	136 K	16 K	8 K	800	500	24.2

The automatic transcription of the test set is carried out with an ASR system optimized for the task (close captioning of live audio) [27]. The language model for the ASR was trained on the same corpus as the punctuation model and was coupled with a Deep Neural Network based acoustic model trained on roughly 500 hours of speech using the Kaldi ASR toolkit [18]. The average word error rate (WER) of the automatic transcripts was around 24%, however showed a large variation depending on genre (see later Table 1). Note, that for Mixed category there was no available audio data in the test database.

2.2 The English IWSLT Dataset

The IWSLT dataset consists of English TED talks transcripts, and has recently became a benchmark for evaluating English punctuation recovery models [4,15, 24,25]. We use the same training, validation and test sets as the studies above, containing 2.1 M, 296 K and 13 K words respectively. This dataset deals with only three types of punctuations: comma, period and question mark.

3 Experimental Setups

3.1 MaxEnt Model

The maximum entropy (MaxEnt) model was suggested by Ratnaparkhi for POS Tagging [19]. In his framework, each sentence is described as a token (word) sequence. Each classified token is described with a set of unique features. The system learns the output labels based on these. In supervised learning, the output labels are hence assigned to the token series. To determine the set of features, the MaxEnt model defines a joint distribution through the available tags and the current context, which can be controlled with a radius parameter. Pre-defined features such as word forms, capitalization, etc. can also be added.

We use the MaxEnt model only with word form-related input features, and all tokens are represented in lower case. To obtain these input features, we use *Huntag*, an open-source, language independent Maximum Entropy Markov Model-based Sequential tagger for both Hungarian and English data [20].

The radius parameter of the MaxEnt tagger determines the size of the context considered. By default, left (past) and right (future) context is taken into account. We will refer to this setup as *off-line mode*. As taking future context into account increases latency, we consider the limit of it, which we will refer to by *on-line mode*. In the experiments we use round brackets to specify left and right context, respectively. Hence (5,1) means that we are considering 5 past and 1 future token actually.

3.2 Recurrent Neural Networks

We split the training, validation and test corpus into short, fixed-length subsequences, called chunks (see the optimized length in Table 2), without overlapping, i.e. such that every token appears once. A vocabulary is built from

the k-most common words from the training set, by adding a garbage collector "*Unknown*" entry to map rare words. Incomplete sub-sequences were padded with zeros. An embedding weight matrix was added based on pre-trained embedding weights and the tokens of the vocabulary.

We investigate the performance of an unidirectional and a bidirectional RNN model in our experiments. Our target slot for punctuation prediction is preceding the actual word. The used architectures are presented in Fig. 1.

Our RNN models (WE-LSTM and WE-BiLSTM, named after using "Word Embedding") are built up in the following way: based on the embedding matrix, the preprocessed sequences are projected into the embedding space (x_t represents the word vector x at time step t). These features are fed into the following layer composed of LSTM or BiLSTM hidden cells, to capture the context of x_t. The output is obtained by applying a softmax activation function to predict the y_t punctuation label for the slot preceding the current word x_t. We chose this simple and lightweight structure to allow for real-time operation with low latency.

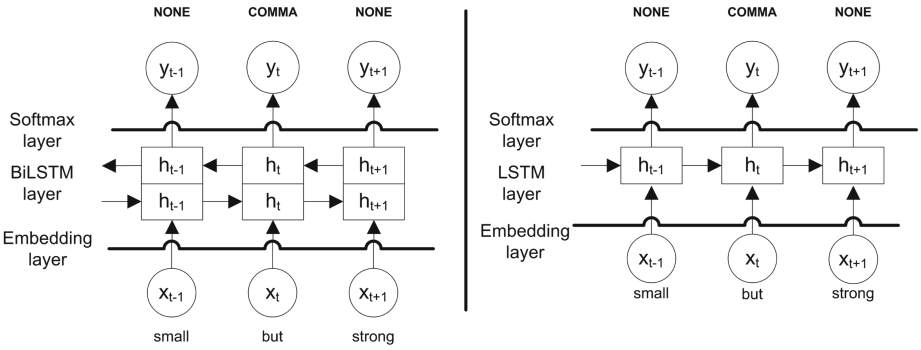

Fig. 1. Structure of WE-BiLSTM (left) and WE-LSTM (right) RNN model

The Hungarian punctuation models were trained on the 100 K most frequent words in the training corpus, by mapping the remaining outlier words to a shared "*Unknown*" symbol. RNN-based recovery models use 600-dimensional pre-trained Hungarian word embeddings [13]. This relative high dimensionality of the embeddings comes from the highly agglutinating nature of Hungarian. In our English RNN-models, a 100-dimensional pre-trained "GloVe" word embedding [17] is used for projection. During training, we use categorical cross-entropy cost function and also let the imported embeddings updated.

We performed a systematic grid search optimization for hyperparameters of the RNNs on the validation set: length of chunks, vocabulary size, number of hidden states, mini-batch size, optimizers. We also use early stopping to prevent overfitting, controlled with patience. Table 2 summarizes the final values of each hyperparameter used in the Hungarian and the English WE-BiLSTM and WE-LSTM models, also including those ones which were inherited from [25], to ensure a partial comparability.

Table 2. Hyperparameters of WE-BiLSTM and WE-LSTM models

Language	Model	Chunk length (#words)	Vocab. Size (#words)	Word embedding dimension	#Hidden states	Batch size	Optimizer	Patience
HUN	WE-BiLSTM	200	100 000	600	512	128	RMSProp	3
HUN	WE-LSTM				256			2
EN	WE-BiLSTM	200	27 244 (by [25])	100 (by [25])	256			2
EN	WE-LSTM	250						

As for the MaxEnt setup, we differentiate low latency and lightweight on-line mode, and robust off-line mode using the future context. All RNN models for punctuation recovery were implemented with the Keras library [6], trained on GPU. The source code of the RNN models is publicly available[1].

We briefly mention that beside word forms, we were considering other textual features too: lemmas, POS-tags (also suggested by [26]) and morphological analysis. The latter were extracted using the *magyarlánc* toolkit, designed for morphological analysis and dependency parsing in Hungarian [28]. Nevertheless, as using word forms yielded the most encouraging results, and also as further analysis for feature extraction increases latency considerably, the evaluated experimental systems rely on word forms features only, input to the embedding layers.

4 Results and Discussion

This section presents the punctuation recovery results for the Hungarian and English tasks. For evaluation, we use standard information retrieval metrics such as Precision (Pr), Recall (Rc), and the F1-Score (F1). In addition, we also calculate the Slot Error Rate (SER) [12], as it is able to incorporate all types of punctuation errors – insertions (Ins), substitutions (Subs) and deletions (Dels) – into a single measure:

$$SER = \frac{C(Ins) + C(Subs) + C(Del)}{C(totalslots)}, \tag{1}$$

for slots considered following each word in the transcription (in (1) $C(.)$ is the count operator).

4.1 Hungarian Overall Results

First, we compare the performance of the baseline MaxEnt sequence tagger (see Subsect. 3.1) to the RNN-based punctuation recovery system (see Subsect. 3.2) on the Hungarian broadcast dataset. Both approaches are presented in two configurations. In the *on-line mode* punctuations are predicted for the slot preceding the current word in the input sequence resulting in a low latency system, suitable for real-time application. In the *off-line mode*, aimed at achieving the best

[1] https://github.com/tundik/HuPP.

result with the given features and architecture, the future word context is also exploited. Please note that hyperparameters of all approaches and configurations were optimized on the validation set as explained earlier (see Sect. 3).

The test evaluations are presented in Table 3 for the reference and in Table 4 for the automatic (ASR) transcripts, respectively. In the notation of MaxEnt models (i, j), i stands for the backward (past), whereas j stands for the forward (future) radius. As it can be seen, the prediction results for comma stand out from the others for all methods and configurations. This can be explained by the fact that Hungarian has generally clear rules for comma usage. In contrast to that, period prediction may also benefit from acoustic information, which assumption is supported by the results in [14], showing robust period recovery with less effective comma restoration.

Table 3. Punctuation restoration results for Hungarian reference transcripts

Reference transcript	Model	Comma			Period			Question			Exclamation			SER
		Pr	Rc	F1	Pr	Rc	F1	Pr	Rc	F1	Pr	Rc	F1	
Off-line mode	MaxEnt-(19, 19)	72.5	59.6	65.5	52.1	40.0	45.2	**55.7**	21.8	31.3	31.1	31.5	31.3	63.5
	WE-BiLSTM	**72.9**	**71.2**	**72.0**	**59.1**	**56.1**	**57.6**	52.4	**38.7**	**44.5**	**51.3**	**36.1**	**42.4**	**50.1**
On-line mode	MaxEnt-(25, 1)	71.8	58.1	64.2	47.5	35.7	40.8	50.4	16.2	24.5	29.3	**33.3**	31.2	66.9
	WE-LSTM	**72.7**	**69.5**	**71.1**	**56.2**	**48.3**	**52.0**	**60.4**	**31.1**	**41.1**	**61.1**	29.4	**39.7**	**53.6**

As Table 3 shows, switching to the RNN-based punctuation restoration for Hungarian reference transcripts reduces SER by around 20% relative compared to the baseline MaxEnt approach. The WE-BiLSTM and WE-LSTM are especially beneficial in restoring periods, question marks and exclamation marks as they are able to exploit large contexts much more efficiently than the MaxEnt tagger. Limiting the future context in on-line configuration causes much less deterioration in results than we had expected. The features from the future word sequence seem to be useful if task requires maximizing recall, otherwise the WE-LSTM is an equally suitable model for punctuation recovery.

Table 4. Punctuation restoration results for Hungarian ASR transcripts

ASR transcript	Model	Comma			Period			Question			Exclamation			SER
		Pr	Rc	F1	Pr	Rc	F1	Pr	Rc	F1	Pr	Rc	F1	
Off-line mode	MaxEnt-(19, 19)	**64.5**	55.8	59.9	41.1	31.2	35.6	**41.2**	8.8	14.4	48.8	17.1	25.4	79.2
	WE-BiLSTM	63.9	**67.7**	**65.7**	**50.5**	**49.0**	**49.8**	37.7	**24.1**	**29.4**	**60.9**	**24.0**	**34.4**	**70.1**
On-line mode	MaxEnt-(25, 1)	**64.3**	54.9	59.2	38.9	29.4	33.5	36.0	7.1	11.9	47.1	20.6	28.6	81.3
	WE-LSTM	63.8	**65.1**	**64.4**	**47.8**	**42.0**	**44.7**	**48.5**	**20.5**	**28.9**	**61.8**	**21.7**	**32.1**	**73.1**

As outlined in the introduction, limiting the future context and propagation of ASR errors into the punctuation recovery pipeline are considered to be the most important factors hindering effective recovery of punctuations in live TV streams. Results confirm that a large future context is less crucial for robust recovery of punctuations, contradictory to our expectations. In contrast, ASR errors seem to be more directly related to punctuation errors: switching from reference transcripts to ASR hypotheses resulted in 15–20% increase in SER (see

Table 4). Although the performance gap is decreased between the two approaches in case of input featuring ASR hypothesis, RNN still outperforms MaxEnt baseline by a large margin.

4.2 Hungarian Results by Genre

The Hungarian test database can be divided into 6 subsets based on the genres of the transcripts (see Table 1). We also analyzed punctuation recovery for these subsets, hypothesizing that more informal and more spontaneous genres are harder to punctuate, in parallel to the more ASR errors seen in these scenarios. Some of the punctuation marks for specific genres were not evaluated (see "N/A" in Table 1), if the Precision or Recall was not possible to be determined based on their confusion matrix.

As the RNN-based approach outperformed the MaxEnt tagger for every genre, we decided to include only results of WE-BiLSTM and WE-LSTM systems in Tables 5 and 6 for better readability.

Table 5. Hungarian reference transcript results by genres

Reference transcript	Genre	Comma			Period			Question			Exclamation			SER
		Pr	Rc	F1	Pr	Rc	F1	Pr	Rc	F1	Pr	Rc	F1	
RNN Off-line mode	Weather	61.2	54.3	57.5	46.7	46.7	46.7	N/A	N/A	N/A	**90.0**	45.0	60.0	69.3
	Brc.-News	**89.9**	**84.4**	**87.1**	**84.3**	**90.7**	**87.3**	**91.7**	**50.0**	**64.7**	83.9	**56.5**	**67.5**	**20.0**
	Sport news	68.3	60.6	64.2	49.4	51.4	50.4	N/A	N/A	N/A	75.0	30.0	42.9	67.0
	Brc.-Conv.	80.4	74.5	77.3	63.9	64.9	64.4	63.0	46.4	53.5	88.9	18.5	30.6	38.7
	Sport mag.	61.2	61.1	61.1	43.9	49.3	46.5	55.2	37.5	44.7	38.5	9.4	15.2	73.1
	Magazine	67.6	67.6	67.6	45.1	46.3	45.7	50.5	29.7	37.5	50.0	5.6	10.1	58.6
RNN On-line mode	Weather	60.2	57.5	58.8	45.7	37.9	41.4	N/A	N/A	N/A	87.5	35.0	50.0	70.6
	Brc.-News	**88.4**	**83.1**	**85.7**	**86.6**	**81.3**	**83.9**	**75.0**	40.9	**52.9**	**100.0**	**67.4**	**80.5**	**24.1**
	Sport news	68.7	57.2	62.4	42.4	37.5	39.8	N/A	N/A	N/A	90.0	60.0	72.0	74.2
	Brc.-Conv.	80.1	74.0	76.9	66.7	54.8	60.1	63.0	**45.6**	**52.9**	77.6	29.2	42.5	40.8
	Sport mag.	60.8	59.7	60.3	42.3	34.8	38.2	53.3	38.3	44.5	20.0	7.5	11.0	77.3
	Magazine	67.6	65.1	66.3	43.5	32.8	37.4	57.3	27.2	36.9	36.4	11.3	17.2	61.5

If we compare the results to the statistics in Table 1, it can be seen that the punctuation recovery system performed best on those genres (broadcast news, broadcast conversations, magazine), for which we had the most training samples. However, the relatively large difference among these three, well-modeled genres suggests that there must be another factor in the background, as well, which is the predictability of the given task. Analogous to language modeling, the more formal, the task is, the better is the predictability of punctuations (see broadcast news results). Obviously, conversational (broadcast conversations) and informal (magazine) speech styles (characterized with less constrained wording and increased number of disfluencies and ungrammatical phrases) make prediction more difficult and introduce punctuation errors compared to more formal styles.

The relatively high SER of the weather forecast and the sport programs genres point out the importance of using a sufficient amount of in-domain training data. Besides collecting more training data, adaptation techniques could be utilized to improve results for these under-resourced genres.

Table 6. Hungarian ASR transcript results by genres

ASR transcript	Genre	Comma			Period			Question			Exclamation			SER	WER
		Pr	Rc	F1	Pr	Rc	F1	Pr	Rc	F1	Pr	Rc	F1		
RNN Off-line mode	Weather	64.7	54.3	59.1	45.9	42.0	43.9	N/A	N/A	N/A	**100.0**	c**50.0**	**66.7**	70.0	6.8
	Brc.-News	**79.5**	**80.0**	**79.7**	**74.7**	**82.6**	**78.4**	**50.0**	14.3	22.2	80.0	46.2	58.5	**37.0**	10.1
	Sport news	47.7	54.8	51.0	32.1	40.5	35.8	N/A	N/A	N/A	**100.0**	**50.0**	**66.7**	107.5	21.4
	Brc.-Conv.	70.6	67.5	69.0	56.8	51.3	53.9	43.8	**28.9**	**34.8**	84.6	13.6	23.4	60.2	24.7
	Sport mag.	55.9	59.8	57.8	39.2	43.5	41.3	48.0	16.2	24.2	N/A	N/A	N/A	87.5	30.3
	Magazine	58.6	60.2	59.5	35.6	28.4	31.6	31.2	14.9	20.2	N/A	N/A	N/A	83.1	38.7
RNN On-line mode	Weather	65.3	58.7	61.8	37.8	34.6	36.1	N/A	N/A	N/A	100.0	12.5	22.2	72.7	6.8
	Brc.-News	**76.5**	**79.3**	**77.9**	**76.8**	**70.7**	**73.6**	N/A	N/A	N/A	**100.0**	50.0	**66.7**	**42.9**	10.1
	Sport news	48.9	54.3	51.5	28.6	30.1	29.3	N/A	N/A	N/A	75.0	**60.0**	**66.7**	108.9	21.4
	Brc.-Conv.	70.3	66.8	68.5	57.4	41.2	48.0	37.0	**24.8**	**29.7**	86.7	16.0	27.1	62.2	24.7
	Sport mag.	53.6	56.4	55.0	37.8	30.4	33.4	**42.1**	21.6	28.6	14.3	4.5	6.9	91.0	30.3
	Magazine	57.6	59.4	58.5	36.5	20.8	26.5	42.1	11.9	18.7	N/A	N/A	N/A	83.9	38.7

By comparing punctuation recovery error of the reference and ASR transcripts, we can draw some interesting conclusions. For the well-modeled genres (Brc.-News, Brc.-Conv., magazine) the increase in SER correlates with the word error rate (WER) of the ASR transcript. However, for the remaining genres (weather, sport news, sport magazine), this relationship between SER and WER is much less predictable. It is particularity difficult to explain the relatively poor results for the sport news genre. Whereas the WER of the ASR transcript is moderate (24.7%), the SER of punctuation is almost doubled for it (67% to 107%). We assume that this phenomenon is related to the high number of named entities in the sport news program, considering that the highest OOV Rate (10%) can be spotted for this genre among all the 6 tested genres.

4.3 English Results

In this subsection, we compare our solutions for punctuation recovery with some recently published models. For this purpose, we use the IWSLT English dataset, which consists of TED Talks transcripts and is a considered benchmark for English punctuation recovery. For complete comparability, we used the default training, validation and test datasets. However, the hyperparameters were optimized for this task (see Table 2). Please note that the IWSLT dataset does not contain samples for exclamation marks.

We present the English punctuation recovery results in Tables 7 and 8. As it can be seen, in on-line mode, the proposed RNN (WE-LSTM) significantly outperformed the so-called T-LSTM configuration presented in [25], which had the best on-line results on this dataset so far to the best of our knowledge. Without using pre-trained word embedding (noWE-LSTM) our results are getting very close to the T-LSTM configuration.

Although in this paper we primarily focused on creating a lightweight, low latency punctuation recovery system, we also compared our WE-BiLSTM system to the best available off-line solutions. As it is shown in Tables 7 and 8, both T-BRNN-pre from [25] configuration and Corr-BiRNN form [15] outperformed

Table 7. Punctuation restoration results for English reference transcripts

Reference transcript	Model	Comma			Period			Question			SER
		Pr	Rc	F1	Pr	Rc	F1	Pr	Rc	F1	
Off-line mode	MaxEnt-(6, 6)	45.6	26.7	33.7	59.4	57.0	58.2	52.4	23.9	32.8	77.2
	WE-BiLSTM	55.5	45.1	49.8	65.9	**75.1**	70.2	57.1	52.2	54.5	59.8
	T-BRNN-pre [25]	**65.5**	47.1	54.8	73.3	72.5	72.9	**70.7**	**63.0**	**66.7**	**49.7**
	Corr-BiRNN [15]	60.9	**52.4**	**56.4**	**75.3**	70.8	**73.0**	70.7	56.9	63.0	50.8
On-line mode	MaxEnt-(10, 1)	44.9	23.7	31.0	53.4	50.1	51.7	50.0	21.7	30.8	83.2
	noWE-LSTM	47.3	**42.7**	44.9	60.9	50.4	55.2	**68.2**	32.6	44.1	76.4
	WE-LSTM	56.3	40.3	**47.0**	**61.2**	**60.5**	**60.8**	55.5	**43.5**	48.8	**68.1**
	T-LSTM [24]	49.6	41.1	45.1	60.2	53.4	56.6	57.1	43.5	**49.4**	74.0

Table 8. Punctuation restoration results for English ASR transcripts

ASR transcript	Model	Comma			Period			Question			SER
		Pr	Rc	F1	Pr	Rc	F1	Pr	Rc	F1	
Off-line mode	MaxEnt-(6, 6)	40.6	23.9	30.1	56.2	53.5	54.8	31.6	17.1	22.2	84.0
	WE-BiLSTM	46.8	39.6	42.9	60.7	70.3	65.1	44.4	45.7	45.0	72.5
	T-BRNN-pre [25]	**59.6**	42.9	49.9	**70.7**	**72.0**	**71.4**	60.7	48.6	54.0	**57.0**
	Corr-BiRNN [15]	53.5	**52.5**	**53.0**	63.7	68.7	66.2	**66.7**	**50.0**	**57.1**	65.4
On-line mode	MaxEnt-(10, 1)	42.6	23.9	30.7	53.2	48.9	51.0	33.3	17.1	23.0	87.0
	noWE-LSTM	40.2	**39.3**	39.7	56.2	46.6	51.0	**76.5**	38.2	**51.0**	86.5
	WE-LSTM	**48.8**	37.1	**42.2**	**57.6**	**57.3**	**57.4**	41.2	41.2	41.2	**78.3**
	T-LSTM [24]	41.8	37.8	39.7	56.4	49.3	52.6	55.6	**42.9**	48.4	83.7

our WE-BiLSTM mainly due to their better performance for commas and question marks. However, these punctuation recovery systems are using much more complex structure and it is questionable whether they would be able to operate in real time scenarios. We consider the high recall of periods by our WE-BiLSTM models as a nice achievement both in reference and ASR transcripts.

5 Conclusions

In this paper, we introduced a low latency, RNN-based punctuation recovery system, which we evaluated on Hungarian and English datasets and compared its performance to a MaxEnt sequence tagger. Both approaches were tested in off-line mode, where textual features could be used from both forward and backward directions; and also in on-line mode, where only backward features were used to allow for real-time operation. The RNN-based approach outperformed the MaxEnt baseline by a large margin in every test configuration. However, what is more surprising, on-line mode causes only a small drop in the accuracy of punctuation recovery.

By comparing results on different genres of the Hungarian broadcast transcripts, we found (analogous to language modeling) that the accuracy of text

based punctuation restoration mainly depends on the amount of available training data and the predictability of the given task. Note, that we are not aware of any prior work in the field of text based punctuation recovery of Hungarian speech transcripts.

In order to compare our models to state-of-the-art punctuation recovery systems, we also evaluated them on the IWSLT English dataset in both on-line and off-line modes. In on-line mode, our WE-LSTM system achieved the overall best result. In off-line mode, however, some more complex networks turned out to perform better than our lightweight solution.

For future work, we are mainly interested in merging of our word-level system and the prosody-based approach outlined in [14] for Hungarian. Extending the English model with further textual or acoustic features is also a promising direction, as we keep our focus on low latency for both languages.

All in all, we consider as important contributions of our work that (1) we use a lightweight and fast RNN model by closely maintained performance; (2) we target real-time operation with little latency; (3) we use the approach for the highly agglutinating Hungarian which has a much less constrained word order than English, as grammatical functions depend much less on the word order than on suffixes (case endings), which makes sequence modeling more difficult due to higher variation seen in the data.

Acknowledgements. The authors would like to thank the support of the Hungarian National Research, Development and Innovation Office (NKFIH) under contract ID *OTKA-PD-112598*; the Pro Progressio Foundation; NVIDIA for kindly providing a Titan GPU for the RNN experiments.

References

1. Batista, F., Moniz, H., Trancoso, I., Mamede, N.: Bilingual experiments on automatic recovery of capitalization and punctuation of automatic speech transcripts. IEEE Trans. Audio Speech Lang. Process. **20**(2), 474–485 (2012)
2. Batista, F.: Recovering capitalization and punctuation marks on speech transcriptions. Ph.D. thesis. Instituto Superior Técnico (2011)
3. Beeferman, D., Berger, A., Lafferty, J.: Cyberpunc: A lightweight punctuation annotation system for speech. In: Proceedings of ICASSP, pp. 689–692. IEEE (1998)
4. Che, X., Wang, C., Yang, H., Meinel, C.: Punctuation prediction for unsegmented transcript based on word vector. In: Proceedings of LREC, pp. 654–658 (2016)
5. Cho, E., Niehues, J., Kilgour, K., Waibel, A.: Punctuation insertion for real-time spoken language translation. In: Proceedings of the Eleventh International Workshop on Spoken Language Translation (2015)
6. Chollet, F.: Keras: Theano-based deep learning library. Code: https://github.com/fchollet. Documentation: http://keras.io (2015)
7. Chung, J., Gulcehre, C., Cho, K., Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling. arxiv preprint arXiv:1412.3555 (2014)
8. Gravano, A., Jansche, M., Bacchiani, M.: Restoring punctuation and capitalization in transcribed speech. In: Proceedings of ICASSP, pp. 4741–4744. IEEE (2009)

9. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
10. Huang, J., Zweig, G.: Maximum entropy model for punctuation annotation from speech. In: Proceedings of Interspeech, pp. 917–920 (2002)
11. Lu, W., Ng, H.T.: Better punctuation prediction with dynamic conditional random fields. In: Proceedings of EMNLP, pp. 177–186. ACL (2010)
12. Makhoul, J., Kubala, F., Schwartz, R., Weischedel, R.: Performance measures for information extraction. In: Proceedings of DARPA broadcast news workshop, pp. 249–252 (1999)
13. Makrai, M.: Filtering wiktionary triangles by linear mapping between distributed models. In: Proceedings of LREC, pp. 2770–2776 (2016)
14. Moró, A., Szaszák, G.: A phonological phrase sequence modelling approach for resource efficient and robust real-time punctuation recovery. In: Proceedings of Interspeech (2017)
15. Pahuja, V., Laha, A., Mirkin, S., Raykar, V., Kotlerman, L., Lev, G.: Joint learning of correlated sequence labelling tasks using bidirectional recurrent neural networks. arxiv preprint arXiv:1703.04650 (2017)
16. Paulik, M., Rao, S., Lane, I., Vogel, S., Schultz, T.: Sentence segmentation and punctuation recovery for spoken language translation. In: Proceedings of ICASSP, pp. 5105–5108. IEEE (2008)
17. Pennington, J., Socher, R., Manning, C.D.: Glove: global vectors for word representation. In: Proceedings of EMNLP, pp. 1532–1543 (2014)
18. Povey, D., et al.: The Kaldi speech recognition toolkit. In: Proceedings of ASRU, pp. 1–4. IEEE (2011)
19. Ratnaparkhi, A., et al.: A maximum entropy model for part-of-speech tagging. In: Proceedings of EMNLP, pp. 133–142 (1996)
20. Recski, G., Varga, D.: A Hungarian NP chunker. Odd Yearb. **8**, 87–93 (2009)
21. Renals, S., Simpson, M., Bell, P., Barrett, J.: Just-in-time prepared captioning for live transmissions. In: Proceedings of IBC 2016 (2016)
22. Schuster, M., Paliwal, K.K.: Bidirectional recurrent neural networks. IEEE Trans. Sig. Process. **45**(11), 2673–2681 (1997)
23. Shriberg, E., Stolcke, A., Baron, D.: Can prosody aid the automatic processing of multi-party meetings? evidence from predicting punctuation, disfluencies, and overlapping speech. In: ISCA Tutorial and Research Workshop (ITRW) on Prosody in Speech Recognition and Understanding (2001)
24. Tilk, O., Alumäe, T.: LSTM for punctuation restoration in speech transcripts. In: Proceedings of Interspeech, pp. 683–687 (2015)
25. Tilk, O., Alumäe, T.: Bidirectional recurrent neural network with attention mechanism for punctuation restoration. In: Proceedings of Interspeech, pp. 3047–3051 (2016)
26. Ueffing, N., Bisani, M., Vozila, P.: Improved models for automatic punctuation prediction for spoken and written text. In: Proceedings of Interspeech, pp. 3097–3101 (2013)
27. Varga, Á., Tarján, B., Tobler, Z., Szaszák, G., Fegyó, T., Bordás, C., Mihajlik, P.: Automatic close captioning for live Hungarian television broadcast speech: a fast and resource-efficient approach. In: Ronzhin, A., Potapova, R., Fakotakis, N. (eds.) SPECOM 2015. LNCS, vol. 9319, pp. 105–112. Springer, Cham (2015). doi:10.1007/978-3-319-23132-7_13
28. Zsibrita, J., Vincze, V., Farkas, R.: magyarlanc: A toolkit for morphological and dependency parsing of Hungarian. In: Proceedings of RANLP, pp. 763–771 (2013)