Deep Learning for Punctuation Restoration in Medical Reports

Wael Salloum, Greg Finley, Erik Edwards, Mark Miller, and David Suendermann-Oeft
EMR.AI Inc
90 New Montgomery St #400
San Francisco, CA 94105, USA
david@emr.ai

Abstract

In clinical dictation, speakers try to be as concise as possible to save time, often resulting in utterances without explicit punctuation commands. Since the end product of a dictated report, e.g. an out-patient letter, does require correct orthography, including exact punctuation, the latter need to be restored, preferably by automated means. This paper describes a method for punctuation restoration based on a state-of-the-art stack of NLP and machine learning techniques including B-RNNs with an attention mechanism and late fusion, as well as a feature extraction technique tailored to the processing of medical terminology using a novel vocabulary reduction model. To the best of our knowledge, the resulting performance is superior to that reported in prior art on similar tasks.

1 Introduction

Medical dictation has been a major instrument in clinical settings to minimize the administrative burden on physicians (Johnson et al., 2014; Hammana et al., 2015; Hodgson and Coiera, 2016). Rather than having to fill forms in electronic medical record systems (EMRs) or typing out-patient letters, such labor is often outsourced to medical transcription providers, many of which make use of automated speech recognition (ASR), coupled with a manual correction step, to increase effectiveness and speed of transcription (Salloum et al., 2017). Despite the fact that medical dictation reduces time physicians spend on clinical documentation substantially, an average dictation still takes about three minutes (Edwards et al., 2017). In an attempt to dictate as efficiently as possible, often physicians (a) speak extremely fast, (b) use pre-dictated paragraphs (so-called *physician normals*), (c) make massive use of abbreviations, and (d) include very limited (if any) instructions regarding formatting and punctuation.

While the ASR system is in charge of turning spoken words into their textual representation, a sophisticated NLP unit, the post-processor, takes care of formatting and structuring the output to produce a draft resembling the out-patient letter as well as possible. Among other responsibilities (such as formatting numerical expressions, dates, section headers, etc.), the post-processor is also charged with restoring punctuation in the letter's narrative. This paper focuses on the automated punctuation restoration in clinical reports, drawing on the latest advances in the NLP sector.

To achieve best possible results in this study, we paid particular attention to the specific challenges faced in medical texts. Foremost among these is a large domain-specific vocabulary, which makes it difficult if not impossible to apply tools developed for general-domain text. When building a system from scratch, however, several factors conspire to make it hard to obtain enough training data: the large medical vocabulary increases problems related to data sparsity and the handling of out-of-vocabulary (OOV) terms; the data often contain sensitive information and have restricted access or availability; and modern methods, such as neural networks as used here, typically require large amounts of data.

We overcame these issues by developing a text pre-processing strategy to reduce vocabulary size, collapsing particular roots and exploiting the fact that many medical terms are built from relatively few morphemes. Our method, which we call the *vocabulary reduction model*, effectively allows the punctuation restoration neural network to focus on morphosyntactic features of words rather than their full semantic representation, as usually cap-

	Normalized Text			Reduced Text				
Set	**types**	**OOVs**	**tokens**	**types**	**tokens**	PERIOD	COLON	COMMA
Training	57,046	n/a	15,886,158	11,766	15,933,901	1,803,626	631,452	760,444
Dev	28,509	1,561	2,243,187	10,321	2,248,305	268,374	89,647	111,571
Blind Test	31,806	3,108	2,944,787	10,767	2,952,873	325,549	103,693	127,895

Table 1: Corpus statistics after normalization and vocabulary reduction. No OOVs are reported on the reduced text since the vocabulary reduction algorithm will map OOVs to classes. The last three columns show the counts of each punctuation tag per set.

tured by word embeddings, being less important to the placement of punctuation.

After reviewing the prior art in the field of punctuation restoration in Section 2, we describe the corpus used in this study in Section 3. The system's general architecture based on bidirectional recurrent neural networks with attention mechanism and late fusion is discussed in Section 4, followed by Section 5 providing details on the vocabulary reduction model. Evaluation results are covered in Section 6, and conclusion and future outlook in Section 7.

2 Related Work

Early efforts in this field used hidden-event n-gram language modeling to predict where punctuation should be inserted (Stolcke et al., 1998; Beeferman et al., 1998). Numerous other strategies have also been devised: combining n-grams with constituency parse information (Shieber and Tao, 2003); maximum entropy using n-gram and part-of-speech features (Huang and Zweig, 2002); conditional random fields (CRFs) (Ueffing et al., 2013); feed-forward neural networks and CRFs on n-gram and lexical features (Cho et al., 2015); even reframing the problem as monolingual machine translation (Peitz et al., 2011).

Most recently, it has been demonstrated that recurrent neural networks can restore punctuation very effectively (Tilk and Alumäe, 2015, 2016). Such methods are promising because they should be able to handle long-distance dependencies that are missed by other methods.

There has been little work on punctuation restoration in the medical domain specifically. While using pauses showed to help in punctuation restoration for rehearsed speech such as TED Talks (Tilk and Alumäe, 2016), Deoras and Fritsch (2008) note that medical dictations pose a particular challenge because the speech is often delivered rapidly and without typical prosodic cues, such as pauses where one would write commas or other punctuation. Thus, although acoustic information has been successfully incorporated for other domains (Huang and Zweig, 2002; Christensen et al., 2001), the same may not be feasible for medical text, so it is especially desirable to have a reliable text-only method.

3 Corpus

The corpus we are using in this study is composed of 32,275 medical reports (i.e., out-patient letters), which we converted into a sequence of tokens with punctuation as tags (since they are the most relevant to medical dictations, we focused on three punctuation marks: colon, comma, and period, represented in the tag set {COLON, COMMA, PERIOD}). We randomly split our corpus into training set, development set, and blind test set. Detailed corpus statistics are given in Table 1.

To reduce the size of the vocabulary, we performed two layers of text preprocessing. First, we performed several text normalization steps such as converting all digits to "D", normalizing numbers, dates, and times into familiar formats (e.g., "D.D", "DD/DDDD", "DD/DD", "DD/DD/DDDD", "DD:DD"), as well as other tokens of the medical domain into normalized formats (e.g., "DDD/DD" for blood pressure, "lD-lD" for lumbar spinal discs, and "q.D+h" meaning "every D+ hours"). Normalization also included lowercasing, unifying abbreviations (e.g., "p.r.n" and "p.r.n."), and performing simple segmentation (e.g., splitting "'s" from a word). Second, we ran a vocabulary reduction algorithm, as detailed in Section 5, that maps infrequent and OOV words to word classes. The combination of these two layers dramatically reduced the vocabulary size, as shown in Table 1.

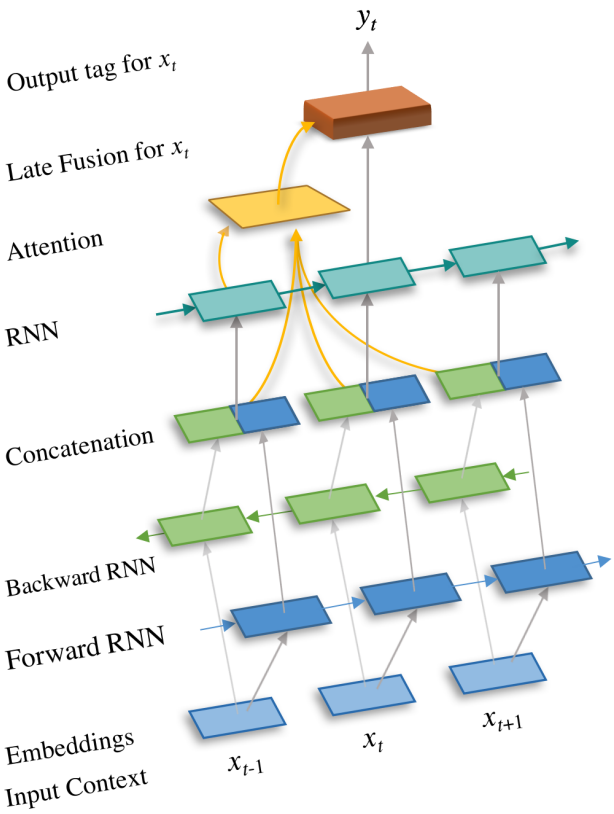

Figure 1: Neural network design for punctuation restoration. The diagram shows an input context for the word x_t and the stack of layers that result in the tag y_t representing the punctuation decision for x_t.

4 The Neural Network Model

We define punctuation restoration as a tagging problem. We try to tag every word in the input sequence with one of four tags: {NONE, COLON, COMMA, PERIOD}. Tagging a word by a punctuation means that the punctuation should be inserted after this word, while tagging with NONE means that the word does not have a punctuation after it. Our neural network approach is based on the work of Tilk and Alumäe (2016). Inspired by Bahdanau et al. (2016), our deep neural network model uses a bidirectional recurrent neural network (B-RNN) (Schuster and Paliwal, 1997) with gated recurrent units (Cho et al., 2014). B-RNNs help in learning long range dependencies on the left and right of the current input word. The B-RNN is composed of a forward RNN and a backward RNN that are preceded by the same word embedding layer. A sliding window of 256 words are passed to the shared embedding layer as one-hot vectors.

On top of the B-RNN, we stack a unidirectional RNN with an attention mechanism (Bahdanau et al., 2016) to assist in capturing relevant contexts that support punctuation restoration decisions. Finally, we use late fusion (Wang and Cho, 2015) to combine the output of the attention mechanism with the current position in the B-RNN without interfering with its memory.

5 The Vocabulary Reduction Model

To improve the modeling of rare words and to deal with OOV words in the test and development sets, we implemented a step that maps rare words to common word classes, reducing the overall size of the vocabulary. This vocabulary reduction allows us to reduce the number of model parameters, which is crucial for fast decoding in a live recognizer.

Table 2 shows examples of prefixes and suffixes that capture the semantic and morpho-syntactic information of infrequent words in our training data such as medical terminology and proper names. For every input word consisting of alphabetical characters only, our vocabulary reduction algorithm goes through the prefix and suffix lists starting from the longer affixes to the shorter ones and tries to match them to the beginning or end of the word, while ensuring that the stem is at least four letters long. If the word starts with a prefix p+ of the prefix list we replace it with "pAAAA" (where "AAAA" represents an alphabetical stem). If it starts with a suffix +q, we replace it with "AAAAq". Finally, if the word matches a prefix p+ and a suffix +q, we split it into two tokens "pAA+" and "+AAq", respectively, to ensure that the information in them gets modeled separately. Every rare word consisting of alphabetical characters only that does not match any suffix or prefix is replaced with a token that represents its length range. The length range is computed with a step of five characters resulting in tokens like AAAA_5 for words shorter than five characters, AAAA_10 for words shorter than ten characters, etc. For example, "angiotensinconvertingenzyme" is replaced with AAAA_30. All other rare words (e.g., "t1cn0m0") are replaced with the token "RARE". These handcrafted rare classes allow us to increase the threshold for considering a word rare. This technique not only significantly reduces the size of the vocabulary, but also allows us to better model rare classes with a higher number of tokens.

Size	Prefix	Suffix
4	inte+, anti+, post+, tran+, over+, intr+, peri+, hype+, para+, neur+, hypo+, micr+, rein+, mult+, card+, comp+, retr+, reco+, self+, gran+, extr+, medi+, hemi+, well+, semi+, endo+, radi+, hemo+, fibr+, oste+, elec+	+tion, +ions, +type, +ness, +ized, +date, +able, +gery, +tive, +sult, +tomy, +ated, +tory, +sion, +ates, +ular, +ical, +osis, +ment, +nary, +rate, +ings, +arge, +onal, +itis, +ents, +like, +lity, +ance, +berg
3	non+, pre+, per+, pro+, mar+, sub+, sch+, str+, tri+, ben+	+ing, +ion, +ted, +ate, +lly, +ive, +tic, +ers, +ble, +ies, +ity, +cal, +man, +sis, +son, +ial, +ous, +ell, +ary, +lar, +tes, +ton, +dez
2	re+, de+, mc+, un+, le+, la+, vi+	+ed, +er, +es, +al, +ry, +te, +ic, +ly, +le

Table 2: Examples of affixes of medical terminology and proper names that capture the semantic and/or morpho-syntactic information of infrequent words in our training data.

Punctuation	Precision	Recall	F-Score
COLON	98.6%	98.6%	98.6%
COMMA	84.0%	82.2%	83.1%
PERIOD	96.1%	96.4%	96.3%
Overall	94.2%	94.0%	94.1%

Table 3: Evaluation of punctuation restoration performance on the blind test set.

We replace a word with its rare class whenever we find it 20 or fewer times in the training data, and we perform the affix-based replacement described above whenever the word occurred less than 100 times. These thresholds were tuned on a held-out development set. Running this algorithm on top of the normalized text results in lowering the vocabulary size in our training data to 11,766 types, meaning that four out of five types are replaced with a class.

6 Evaluation

For the present study, we used Keras with TensorFlow backend (Chollet, 2015; Abadi et al., 2016; Chollet, 2017). We evaluated on the blind test set by passing the whole set to our system as a sequence of about three million tokens without any indication of beginning or end of sentence, paragraph, or report. All words were lowercased, as described earlier, to avoid giving out any hint of sentence or section header start or end. We report the results in Table 3.

We achieve 96.3% F-Score on periods, which we consider the most important as they define sentence boundaries. The latter are crucial for virtually any subsequent NLP process, such as automatic coding of medical reports (Suendermann-Oeft et al., 2016).

The second most important punctuation type in medical reports is colons, as they define section headers and, thus, help format the report structure. We achieve 98.6% F-Score on colons.

Finally, we get 83.1% F-Score on commas, the hardest tag to predict due to human inconsistency in using them. This inconsistency affects the accuracy of the training data as well as the fairness in the evaluation against the test set. The overall performance of the system on all tags is 94.1% in terms of F-Score. Refer to Table 4 for examples of our system's output.

7 Conclusion and Future Work

Although prior work on punctuation restoration has used different corpora from the work presented in this paper, our result (F-Score 94.1%) compares very favorably with previous publications. For example, Cho *et al.* (2015) achieve an F-Score of 61.8% on a meeting and lecture corpus, Tilk and Alumäe (2016) produce 64.4% on TED talk transcripts, and Ueffing *et al.* (2013) report an F-Score of 66.8% on one of Nuance's in-house dictation corpora.

While we have tested the performance of the presented punctuation restoration algorithm on naturalistic medical dictations, we have not yet measured the impact the speech recognizer's word error rate has on the F-Score, a task we plan to address in the near future. We are also interested to learn whether analyzing the speech waveform and characteristic pauses and prosodic patterns in medical dictations can be exploited in a hybrid speech/text punctuation restoration system to enhance accuracy even further. We also plan to replace the vocabulary reduction model by fusing a morphology-aware neural network such as a

Input	... review of systems general positive for fatigue excessive perspiration feeling sick ...
Gold	... review of systems: general: positive for fatigue, excessive perspiration, feeling sick. ...
Punctuated	... review of systems COLON general COLON positive for fatigue COMMA excessive perAA+ +AAtion COMMA feeling sick PERIOD ...
Input	... chronic pruritus dermatology felt that this was neurodermatosis and neurotic excoriations ...
Gold	... chronic pruritus. dermatology felt that this was neurodermatosis and neurotic excoriations. ...
Punctuated	... chronic pruritus PERIOD deAAAA felt that this was neurAA+ +AAosis and neurAA+ +AAtic AAAAions PERIOD ...
Input	... it is available review of systems positive for still some ongoing lower extremity weakness tremulousness and unsteadiness otherwise review of ...
Gold	... it is available. review of systems: positive for still some ongoing lower extremity weakness, tremulousness and unsteadiness. otherwise, review of ...
Punctuated	... it is available PERIOD review of systems COLON positive for still some ongoing lower extremity weakness COMMA AAAAness and unAA+ +AAness PERIOD otherwise COMMA review of ...
Input	... severe clinical depression including hopelessness helplessness worthlessness difficulty focusing concentration and a lot of thoughts of death and dying ...
Gold	... severe clinical depression including hopelessness, helplessness, worthlessness, difficulty focusing, concentration, and a lot of thoughts of death and dying. ...
Punctuated	... severe clinical depression including AAAAness COMMA AAAAness COMMA AAAAness COMMA difficulty AAAAing COMMA concentration COMMA and a lot of thoughts of death and dying PERIOD ...
Input	... is reasonable we will optimize his medications by adding low dose angiotensinconvertingenzyme inhibitors which he currently is not on if the ...
Gold	... is reasonable. we will optimize his medications by adding low dose angiotensinconvertingenzyme inhibitors, which he currently is not on. if the ...
Punctuated	... is reasonable PERIOD we will optimize his medications by adding low dose AAAA_30 inhibitors COMMA which he currently is not on PERIOD if the ...

Table 4: Examples of the output of our system on word sequences of the input. The first example shows the correct handling of consecutive colons indicating a section header and a subsection header. The second example shows the preprocessing of infrequent medical terminology like "neurodermatosis", "neurotic", and "excoriations" by capturing their semantic and part-of-speech information. The third and fourth examples emphasize the case of parallelism captured by mapping "tremulousness and unsteadiness" to "AAAAness and unAA+ +AAness" and "hopelessness helplessness worthlessness" to "AAAAness AAAAness AAAAness", thus predicting commas when needed since the meaning is irrelevant to the punctuation task. The fourth example also shows the correct prediction of coordinated lists, separating them with commas. The final example presents the mapping of a very long word, "angiotensinconvertingenzyme", into "AAAA_30", which reduces the confusion of the network and results in the correct prediction.

character-based convolutional network.

References

M Abadi, A Agarwal, P Barham, E Brevdo, Z Chen, C Citro, GS Corrado, A Davis, J Dean, M Devin, S Ghemawat, I Goodfellow, A Harp, G Irving, M Isard, Y Jia, R Jozefowicz, L Kaiser, M Kudlur, J Levenberg, D Mané, R Monga, S Moore, D Murray, C Olah, M Schuster, J Shlens, B Steiner, I Sutskever, K Talwar, P Tucker, V Vanhoucke, V Vasudevan, F Viégas, O Vinyals, P Warden, M Wattenberg, M Wicke, Y Yu, and X Zheng. 2016. Tensorflow: large-scale machine learning on heterogeneous distributed systems. *arXiv* 1603(04467):1–19.

D Bahdanau, K Cho, and Y Bengio. 2016. Neural machine translation by jointly learning to align and translate [conference paper at iclr 2015]. *arXiv* 1409(0473):1–15.

D Beeferman, A Berger, and JD Lafferty. 1998. Cyberpunc: a lightweight punctuation annotation system for speech. In *Proc ICASSP*. IEEE, volume 2, pages 689–692.

E Cho, K Kilgour, J Niehues, and A Waibel. 2015. Combination of nn and crf models for joint detection of punctuation and disfluencies. In *Proc Interspeech*. ISCA, pages 3650–3654.

K Cho, B van Merriënboer, D Bahdanau, and Y Bengio. 2014. On the properties of neural machine translation: encoder-decoder approaches. *arXiv* 1409(1259):1–9.

F Chollet. 2015. Keras: deep learning library for theano and tensorflow. https://keras.io/.

F Chollet. 2017. *Deep learning with Python*. Manning, Shelter Island, NY.

H Christensen, Y Gotoh, and S Renals. 2001. Punctuation annotation using statistical prosody models. In *Proc ITRW on Prosody in Speech Recognition and Understanding*. ISCA, paper 6, pages 1–6.

A Deoras and J Fritsch. 2008. Decoding-time prediction of non-verbalized punctuation. In *Proc Interspeech*. ISCA, pages 1449–1452.

E Edwards, W Salloum, GP Finley, J Fone, G Cardiff, M Miller, and D Suendermann-Oeft. 2017. Medical speech recognition: reaching parity with humans. In *Proc SPECOM*. Springer, pages 1–10.

I Hammana, L Lepanto, T Poder, C Bellemare, and M-S Ly. 2015. Speech recognition in the radiology department: a systematic review. *HIM J* 44(2):4–10.

T Hodgson and EW Coiera. 2016. Risks and benefits of speech recognition for clinical documentation: a systematic review. *J Am Med Inform Assoc* 23(e1):169–179.

J Huang and G Zweig. 2002. Maximum entropy model for punctuation annotation from speech. In *Proc ICSLP*. ISCA, pages 917–920.

M Johnson, S Lapkin, V Long, P Sanchez, H Suominen, J Basilakis, and L Dawson. 2014. A systematic review of speech recognition technology in health care. *BMC Med Inform Decis Mak* 14(94):1–14.

S Peitz, M Freitag, A Mauser, and H Ney. 2011. Modeling punctuation prediction as machine translation. In *Proc IWSLT*. pages 238–245.

W Salloum, E Edwards, S Ghaffarzadegan, D Suendermann-Oeft, and M Miller. 2017. Crowdsourced continuous improvement of medical speech recognition. In *Proc AAAI Wrkshp Crowdsourcing*. AAAI, San Francisco, CA.

M Schuster and KK Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Trans Signal Process* 45(11):2673–2681.

SM Shieber and X Tao. 2003. Comma restoration using constituency information. In *Proc HLT-NAACL*. ACL, volume 1, pages 142–148.

A Stolcke, E Shriberg, R Bates, M Ostendorf, D Hakkani, M Plauche, G Tür, and Y Lu. 1998. Automatic detection of sentence boundaries and disfluencies based on recognized words. In *Proc ICSLP*. ISCA, paper 0059, pages 1–4.

D Suendermann-Oeft, S Ghaffarzadegan, E Edwards, W Salloum, and M Miller. 2016. A system for automated extraction of clinical standard codes in spoken medical reports. In *Proc Wrkshp SLT*. IEEE, San Diego, CA.

O Tilk and T Alumäe. 2015. Lstm for punctuation restoration in speech transcripts. In *Proc Interspeech*. ISCA, pages 683–687.

O Tilk and T Alumäe. 2016. Bidirectional recurrent neural network with attention mechanism for punctuation restoration. In *Proc Interspeech*. ISCA, pages 3047–3051.

N Ueffing, M Bisani, and P Vozila. 2013. Improved models for automatic punctuation prediction for spoken and written text. In *Proc Interspeech*. ISCA, pages 3097–3101.

T Wang and K Cho. 2015. Larger-context language modelling. *arXiv* 1511(03729):1–14.