Neural Networks for Joint Sentence Classification in Medical Paper Abstracts

Neural Networks for Joint Sentence Classification in Medical Paper Abstracts

Franck Dernoncourt* MIT francky@mit.edu

Ji Young Lee* MIT jjylee@mit.edu

Peter Szolovits MIT psz@mit.edu

Abstract

Existing models based on artificial neural networks (ANNs) for sentence classification often do not incorporate the context in which sentences appear, and classify sentences individually. However, traditional sentence classification approaches have been shown to greatly benefit from jointly classifying subsequent sentences, such as with conditional random fields. In this work, we present an ANN architecture that combines the effectiveness of typical ANN models to classify sentences in isolation, with the strength of structured prediction. Our model achieves state-of-the-art results on two different datasets for sequential sentence classification in medical abstracts.

1 Introduction

Over 50 million scholarly articles have been published (Jinha, 2010), and the number of articles published every year keeps increasing (Druss and Marcus, 2005; Larsen and Von Ins, 2010). Approximately half of them are biomedical papers. While this repository of human knowledge abounds with useful information that may unlock new, promising research directions or provide conclusive evidence about phenomena, it has become increasingly difficult to take advantage of all available information due to its sheer amount. Therefore, a technology that can assist a user to quickly locate the information of interest is highly desired, as it may reduce the time required to locate relevant information.

When researchers search for previous literature, for example, they often skim through abstracts in order to quickly check whether the papers match their criteria of interest. This process is easier when abstracts are *structured*, i.e., the text in an abstract is divided into semantic headings such as objective, method, result, and conclusion. However, a significant portion of published paper abstracts is *unstructured*, which makes it more difficult to quickly access the information of interest. Therefore, classifying each sentence of an abstract to an appropriate heading can significantly reduce time to locate the desired information.

We call this the *sequential sentence classification task*, in order to distinguish it from general text classification or sentence classification that does not have any context. Besides aiding humans, this task may also be useful for automatic text summarization, information extraction, and information retrieval.

In this paper, we present a system based on ANNs for the sequential sentence classification task. Our model makes use of both token and character embeddings for classifying sentences, and has a sequence optimization layer that is learned jointly with other components of the model. We evaluate our model on the NICTA-PIBOSO dataset as well as a new dataset we compiled based on the PubMed database.

2 Related Work

Existing systems for sequential sentence classification are mostly based on naive Bayes (NB) (Ruch et al., 2007; Huang et al., 2013), support vector machines (SVMs) (McKnight and Srinivasan, 2003; Yamamoto and Takagi, 2005; Hirohata et al., 2008), Hidden Markov models (HMMs) (Lin et al., 2006), and conditional random fields (CRFs) (Kim et al., 2011; Hassanzadeh et al., 2014; Hirohata et al., 2008). They often require numerous hand-engineered features based on lexical (bag-of-words, n-grams, dic-

* These authors contributed equally to this work.

tionaries, cue words), semantic (synonyms, hyponyms), structural (part-of-speech tags, headings), and sequential (sentence position, surrounding features) information.

On the other hand, recent approaches to natural language processing (NLP) based on artificial neural networks (ANNs) do not require manual features, as they are trained to automatically learn features based on word as well as character embeddings. Moreover, ANN-based models have achieved state-of-the-art results on various NLP tasks. For short-text classification, many ANN models use word embeddings (Socher et al., 2013; Kim, 2014; Kalchbrenner et al., 2014), and most recent works are based on character embeddings (Zhang et al., 2015; Conneau et al., 2016; Xiao and Cho, 2016). Dos Santos and Gatti (2014) use both word and character embeddings.

However, most existing works using ANNs for short-text classification do not use any context. This is in contrast with *sequential* sentence classification, where each sentence in a text is classified taking into account its context. The context utilized for the classification could be the surrounding sentences or possibly the whole text. One exception is a recent work on dialog act classification (Lee and Dernoncourt, 2016), where each utterance in a dialog is classified into its dialog act, but only the preceding utterances were used, as the system was designed with real-time applications in mind.

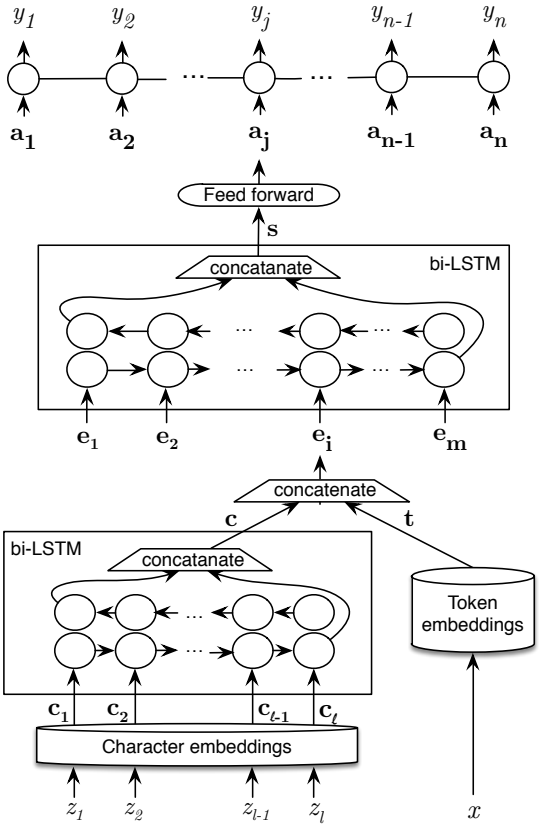

Figure 1: ANN model for sequential sentence classification. x: token, $\mathbf{t}$: token embeddings (300), z_i: i^{th} character of x, $\mathbf{c}_i$: character embeddings (25), $\mathbf{c}$: character-based token embeddings (50), $\mathbf{e}_i$: hybrid token embeddings (350), $\mathbf{s}$: sentence vector (200), $\mathbf{a}_j$: sentence label vector (number of classes), y_j: sentence label. The numbers in parenthesis indicate the dimensions of the vectors. Token embeddings are initialized with GloVe (Pennington et al., 2014) embeddings pretrained on Wikipedia and Gigaword 5 (Parker et al., 2011).

3 Model

In the following, we denote scalars in italic lowercase (e.g., k, b_f), vectors in bold lowercase (e.g., $\mathbf{s}$, $\mathbf{x}_i$), and matrices in italic uppercase (e.g., W_f) symbols. We use the colon notations $x_{i:j}$ and $\mathbf{v}_{i:j}$ to denote the sequences of scalars $(x_i, x_{i+1}, \ldots, x_j)$ and vectors $(\mathbf{v}_i, \mathbf{v}_{i+1}, \ldots, \mathbf{v}_j)$, respectively.

3.1 ANN model

Our ANN model (Figure 1) consists of three components: a hybrid token embedding layer, a sentence label prediction layer, and a label sequence optimization layer.

3.1.1 Hybrid token embedding layer

The hybrid token embedding layer takes a token as an input and outputs its vector representation utilizing both the token embeddings and as well as the character embeddings.

Token embeddings are a direct mapping $\mathcal{V}_T(\cdot)$ from token to vector, which can be pre-trained on large unlabeled datasets using programs such as word2vec (Mikolov et al., 2013b; Mikolov et al., 2013a; Mikolov et al., 2013c) or GloVe (Pennington et al., 2014). Character embeddings are also defined in an analogous manner, as a direct mapping $\mathcal{V}_C(\cdot)$ from character to vector.

Let $z_{1:\ell}$ be the sequence of characters that comprise a token x. Each character z_i is first mapped to its embedding $\mathbf{c}_i = \mathcal{V}_C(z_i)$, and the resulting sequence $\mathbf{c}_{1:\ell}$ is input to a bidirectional LSTM, which outputs the character-based token embedding $\mathbf{c}$.

The output $\mathbf{e}$ of the hybrid token embedding layer for the token x is the concatenation of the character-based token embedding $\mathbf{c}$ and the token embedding $\mathbf{t} = \mathcal{V}_T(x)$.

3.1.2 Sentence label prediction layer

Let $x_{1:m}$ be the sequence of tokens in a given sentence, and $\mathbf{e}_{1:m}$ be the corresponding embedding output from the hybrid token embedding layer. The sentence label prediction layer takes as input the sequence of vectors $\mathbf{e}_{1:m}$, and outputs $\mathbf{a}$, where the k^{th} element of $\mathbf{a}$, denoted $\mathbf{a}[k]$, reflects the probability that the given sentence has label k.

To achieve this, the sequence $\mathbf{e}_{1:m}$ is first input to a bidirectional LSTM, which outputs the vector representation $\mathbf{s}$ of the given sentence. The vector $\mathbf{s}$ is subsequently input to a feedforward neural network with one hidden layer, which outputs the corresponding probability vector $\mathbf{a}$.

3.1.3 Label sequence optimization layer

The label sequence optimization layer takes the sequence of probability vectors $\mathbf{a}_{1:n}$ from the label prediction layer as input, and outputs a sequence of labels $y_{1:n}$, where y_i is the label assigned to the token x_i.

In order to model dependencies between subsequent labels, we incorporate a matrix T that contains the transition probabilities between two subsequent labels; we define $T[i, j]$ as the probability that a token with label i is followed by a token with the label j. The score of a label sequence $y_{1:n}$ is defined as the sum of the probabilities of individual labels and the transition probabilities:

$$s(y_{1:n}) = \sum_{i=1}^{n} \mathbf{a}_i[y_i] + \sum_{i=2}^{n} T[y_{i-1}, y_i].$$

These scores can be turned into probabilities of the label sequences by taking a softmax function over all possible label sequences. During the training phase, the objective is to maximize the log probability of the gold label sequence. In the testing phase, given an input sequence of tokens, the corresponding sequence of predicted labels is chosen as the one that maximizes the score.

4 Experiments

4.1 Datasets

We evaluate our model on the sentence classification task using the following two medical abstract datasets, where each sentence of the abstract is annotated with one label. Table 1 presents statistics on each dataset.

NICTA-PIBOSO This dataset was introduced in (Kim et al., 2011) and was the basis of the ALTA 2012 Shared Task (Amini et al., 2012).

PubMed 20k RCT We assembled this corpus consisting of randomized controlled trials (RCTs) from the PubMed database of biomedical literature, which provides a standard set of 5 sentence labels: objectives, background, methods, results and conclusions.

Dataset	$\|C\|$	$\|V\|$	Train	Validation	Test
PubMed	5	68k	15k (195k)	2.5k (33k)	2.5k (33k)
NICTA	6	17k	722 (8k)	77 (0.9k)	200 (2k)

Table 1: Dataset overview. $|C|$ denotes the number of classes, $|V|$ the vocabulary size. For the train, validation and test sets, we indicate the number of number of abstracts followed by the number of sentences in parentheses.

4.2 Training

The model is trained using stochastic gradient descent, updating all parameters, i.e., token embeddings, character embeddings, parameters of bidirectional LSTMs, and transition probabilities, at each gradient step. For regularization, dropout with a rate of 0.5 is applied to the character-enhanced token embeddings and before the label prediction layer.

5 Results and Discussion

Table 2 compares our model against several baselines as well as the best performing model (Lui, 2012) in the ALTA 2012 Shared Task, in which 8 competing research teams participated to build the most accurate classifier for the NICTA-PIBOSO corpus.

The first baseline (LR) is a classifier based on logistic regression using n-gram features extracted from the current sentence: it does not use any information from the surrounding sentences. The second baseline (Forward ANN) uses the model presented in (Lee and Dernoncourt, 2016): it computes sentence embeddings for each sentence, then classifies the current sentence given a few preceding sentence embeddings as well as the current sentence embedding. The third baseline (CRF) is a CRF that uses n-grams as features: each output variable of the CRF corresponds to a label for a sentence, and the sequence the CRF considers is the entire abstract. The CRF baseline therefore uses both preceding and succeeding sentences when classifying the current sentence. Lastly, the model presented in (Lui, 2012) developed a new

Model	PubMed 20k	NICTA
LR	83.0	71.6
Forward ANN	86.1	75.1
CRF	89.3	81.2
Best published	–	82.0
Our model	**89.9**	**82.7**

Table 2: F1-scores on the test set with several baselines, the best published method (Lui, 2012) from the literature, and our model. Since PubMed 20k was introduced in this work, there is no previous best published method for this dataset. The presented results for the ANN-based models are the F1-scores on the test set of the run with the highest F1-score on the validation set.

approach called feature stacking, which is a meta-alearner that combines multiple feature sets, and is the best performing system on NICTA-PIBOSO published in the literature.

The LR system performs honorably on PubMed 20k RCT (F1-score: 83.0), but quite poorly on NICTA-PIBOSO (F1-score: 71.6): this suggests that using the surrounding sentences may be more important in NICTA-PIBOSO than in PubMed 20k RCT.

The Forward ANN system performs better than the LR system, and worse than the CRF: this is unsurprising, as the Forward ANN system only uses the information from the preceding sentences but does not use any information from the succeeding sentences, unlike the CRF.

Our model performs better than the CRF system and the (Lui, 2012) system. We hypothesize that the following four factors give an edge to our model.

No human-engineered features: Unlike most other systems, our model does not rely on any human-engineered features.

No n-grams: While other systems heavily rely on n-grams, our model maps each token to a token embedding, and feeds it as an input to an RNN. This helps combat data scarcity: for example, "chronic tendonitis" and "chronic tendinitis" are two different bigrams, but their token embeddings should be very similar since they share the same meaning.

Structured prediction: The labels for all sentences in an abstract are predicted jointly, which improves the coherence between the predicted labels in a given abstract.

Joint learning: Our model learned the features and token embeddings jointly with the sequence optimization.

Label	PubMed 20k RCT			
	Precision	Recall	F1-score	Support
Background	71.8	88.2	79.1	3621
Conclusion	93.5	92.9	93.2	4571
Methods	93.7	96.2	94.9	9897
Objectives	78.2	48.1	59.6	2333
Results	94.8	93.1	93.9	9713
Total	90.0	89.8	89.9	30135

Table 3: Detailed results of our model on the PubMed 20k RCT dataset.

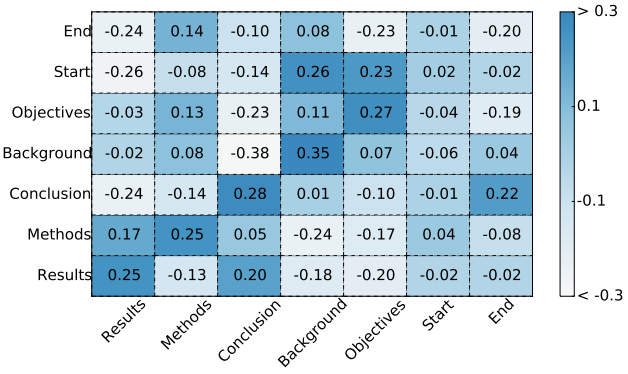

Figure 2: Transition matrix learned on PubMed 20k. The rows represent the label of the previous sentence, the columns represent the label of the current sentence.

Figure 2 presents an example of a transition matrix after the model has been trained on PubMed 20k RCT. We can see that it effectively reflects transitions between different labels. For example, it learned that the first sentence of an abstract is most likely to be either discussing objective (0.23) or background (0.26). By the same token, a sentence pertaining to the methods is typically followed by a sentence pertaining to the methods (0.25) or the results (0.17).

Table 3 details the result of our model for each label in PubMed 20k RCT: the main difficulty the classifier has is distinguishing background sentences from objective sentences.

6 Conclusions

In this article we have presented an ANN architecture to classify sentences that appear in sequence. We demonstrate that jointly predicting the classes of all sentences in a given text improves the quality of the predictions and yields better performance than a CRF. Our model achieves state-of-the-art results on two datasets for sentence classification in medical abstracts.

References

[Amini et al.2012] Iman Amini, David Martinez, and Diego Molla. 2012. Overview of the ALTA 2012 Shared Task. In *Australasian Language Technology Association Workshop 2012*, volume 7, page 124.

[Conneau et al.2016] Alexis Conneau, Holger Schwenk, Loïc Barrault, and Yann Lecun. 2016. Very deep convolutional networks for natural language processing. *arXiv preprint arXiv:1606.01781*.

[dos Santos and Gatti2014] Cícero Nogueira dos Santos and Maira Gatti. 2014. Deep convolutional neural networks for sentiment analysis of short texts. In *International Conference on Computational Linguistics (COLING)*, pages 69–78.

[Druss and Marcus2005] Benjamin G Druss and Steven C Marcus. 2005. Growth and decentralization of the medical literature: implications for evidence-based medicine. *Journal of the Medical Library Association*, 93(4):499.

[Hassanzadeh et al.2014] Hamed Hassanzadeh, Tudor Groza, and Jane Hunter. 2014. Identifying scientific artefacts in biomedical literature: The evidence based medicine use case. *Journal of biomedical informatics*, 49:159–170.

[Hirohata et al.2008] Kenji Hirohata, Naoaki Okazaki, Sophia Ananiadou, Mitsuru Ishizuka, and Manchester Interdisciplinary Biocentre. 2008. Identifying sections in scientific abstracts using conditional random fields. In *International Joint Conference on Natural Language Processing (IJCNLP)*, pages 381–388.

[Huang et al.2013] Ke-Chun Huang, I-Jen Chiang, Furen Xiao, Chun-Chih Liao, Charles Chih-Ho Liu, and Jau-Min Wong. 2013. PICO element detection in medical text without metadata: Are first sentences enough? *Journal of biomedical informatics*, 46(5):940–946.

[Jinha2010] Arif E Jinha. 2010. Article 50 million: an estimate of the number of scholarly articles in existence. *Learned Publishing*, 23(3):258–263.

[Kalchbrenner et al.2014] Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. 2014. A convolutional neural network for modelling sentences. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics.

[Kim et al.2011] Su Nam Kim, David Martinez, Lawrence Cavedon, and Lars Yencken. 2011. Automatic classification of sentences to support evidence based medicine. *BioMed Central (BMC) Bioinformatics*, 12(2):1.

[Kim2014] Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751. Association for Computational Linguistics (ACL).

[Larsen and Von Ins2010] Peder Olesen Larsen and Markus Von Ins. 2010. The rate of growth in scientific publication and the decline in coverage provided by science citation index. *Scientometrics*, 84(3):575–603.

[Lee and Dernoncourt2016] Ji Young Lee and Franck Dernoncourt. 2016. Sequential short-text classification with recurrent and convolutional neural networks. In *Human Language Technologies 2016: The Conference of the North American Chapter of the Association for Computational Linguistics, NAACL HLT*.

[Lin et al.2006] Jimmy Lin, Damianos Karakos, Dina Demner-Fushman, and Sanjeev Khudanpur. 2006. Generative content models for structural analysis of medical abstracts. *BioNLP06 Linking Natural Language Processing and Biology: Towards Deeper Biological Literature Analysis*, 6:65–72.

[Lui2012] Marco Lui. 2012. Feature stacking for sentence classification in evidence-based medicine. In *Australasian Language Technology Workshop 2012: ALTA Shared Task*, page 134.

[McKnight and Srinivasan2003] Larry McKnight and Padmini Srinivasan. 2003. Categorization of sentence types in medical abstracts. In *American Medical Informatics Association (AMIA)*.

[Mikolov et al.2013a] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

[Mikolov et al.2013b] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

[Mikolov et al.2013c] Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013c. Linguistic regularities in continuous space word representations. In *HLT-NAACL*, pages 746–751.

[Parker et al.2011] Robert Parker, David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2011. English Gigaword fifth edition. Technical report, Linguistic Data Consortium, Philadelphia.

[Pennington et al.2014] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. GloVe: global vectors for word representation. *Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP 2014)*, 12:1532–1543.

[Ruch et al.2007] Patrick Ruch, Celia Boyer, Christine Chichester, Imad Tbahriti, Antoine Geissbühler, Paul Fabry, Julien Gobeill, Violaine Pillet, Dietrich Rebholz-Schuhmann, Christian Lovis, et al. 2007. Using argumentation to extract key sentences from biomedical abstracts. *International journal of medical informatics*, 76(2):195–200.

[Socher et al.2013] Richard Socher, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*, volume 1631, page 1642. Citeseer.

[Xiao and Cho2016] Yijun Xiao and Kyunghyun Cho. 2016. Efficient character-level document classification by combining convolution and recurrent layers. *arXiv preprint arXiv:1602.00367*.

[Yamamoto and Takagi2005] Yasunori Yamamoto and Toshihisa Takagi. 2005. A sentence classification system for multi biomedical literature summarization. In *21st International Conference on Data Engineering Workshops (ICDEW'05)*, pages 1163–1163. IEEE.

[Zhang et al.2015] Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems (NIPS)*, pages 649–657.