Learning to Detect Verbose Expressions in Spoken Texts

Qingbin Liu[1,2], Shizhu He[1], Kang Liu[1], Shengping Liu[3], and Jun Zhao[1,2]

[1] National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
[2] University of Chinese Academy of Sciences, Beijing 100049, China
[3] Beijing Unisound Information Technology, Beijing 100028, China
{qingbin.liu,shizhu.he,kliu,jzhao}@nlpr.ia.ac.cn,
liushengping@unisound.com

Abstract. The analysis and understanding of spoken texts is an important task in artificial intelligence and natural language processing. However, there are many verbose expressions (such as mantras, nonsense, modal particle, etc.) in spoken texts, which brings great challenges to subsequent tasks. This paper devote to detect verbose expressions in spoken texts. Considering the correlation of verbose words/characters in spoken texts, we adapt sequence models to detect them with an end-to-end manner. Moreover, we propose a model with the long-short term memory (LSTM) and modified restrict attention (MRA) mechanism which are able to utilize the mutual influence between long-distance and local words in sentences. In addition, we propose a compare mechanism to model the repetitive verbose expressions. The experimental result shows that compared with the rule-based and direct classification methods, our proposed model increases F1 measure by 54.08% and 18.91%.

Keywords: Spoken texts · Verbose Expressions · Text transformation · Modified restricted attention mechanism · Compare mechanism

1 Introduction

Spoken language understanding and processing are important tasks in artificial intelligence (AI) and natural language processing (NLP)[20, 4, 11, 7, 16]. In addition, the processing of spoken texts is very important for subsequent tasks such as generation tasks[13, 6, 10]. There are many verbose expressions in the spoken texts such as mantras , nonsense words, repetitions like 'this this (这个这个)', and modal particle like 'Ah (啊)' as shown in Fig. 1 that bring great challenges in spoken language processing. In the practical spoken systems such as reservation system, spoken context need to be converted into texts by speech recognition. The errors in the speech recognition also aggravate the above problems.

In this paper, we propose the detection task to detect the verbose expressions such as 'ah (啊)', 'this (这个)' in Fig 1. Deleting these expressions will get normal texts. As far as we know, there is little work in the task, especially for Chinese.

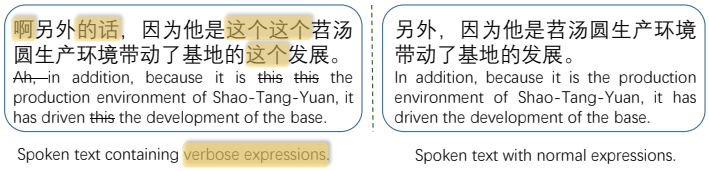

Fig. 1: The spoken texts and normal texts

Then we construct a dataset based on the interview texts and manually annotate the verbose expressions. This dataset drives on the task. The direct methods in the task are the rule-based or direct classification methods. Rule-based methods usually count the frequency of verbose expressions and use rules to detect the expressions. The direct classification detects the verbose expressions based on word embedding. However, these methods ignore the relationship between different words. Actually, in a sentence, different expressions can affect each other. For example, some pronouns are not verbose words as sentence components, but as mantras are verbose words. In addition, the above methods cannot directly detect the repetitive verbose expressions like 'this this (这个这个)' in texts.

Recently, the recurrent neural networks (RNN) and its variant (LSTM)[8], have been applied extensively in many tasks. LSTM can obtain long-distance information in a sentence. Moreover, attention mechanism has been introduced to get "soft" correlated information to many tasks[3, 12]. Even in the machine translation field, just using the attention mechanism can get the best performance in some languages[19, 14]. The LSTM combined with a conditional random field (CRF) achieved best performance in many sequence-labeling tasks[18, 9].

Although the above approaches can improve performance in the task, they still suffer from three problems: 1) Chinese word segmentation is inaccurate for spoken texts. Therefore, we need to incorporate proper word information in character-vector level. 2) The global attention mechanisms extract many irrelevant information in character level and degrade performance. 3) We need an explicit compare mechanism to detect the repetitive verbose characters such as 'this this (这个这个)' in Fig. 1. We propose a new model with LSTM and M-RA to address these problems. MRA utilizes multiply restricted mask matrixes to extract local relevant information in Chinese-character level. Compared with global attention mechanisms, our proposed attention mechanism reduces a lot of irrelevant information. Furthermore, a gate is used to filter irrelevant information between different mask matrixes. We also propose the compare mechanism in our model to explicit recognize the repetitive cases.

Our main contributions are as follows:

1. We propose a new task which devotes to detect verbose expressions in spoken texts. It is very useful in understanding and processing spoken texts. As far as we know, there is little work in the task, especially for Chinese.
2. We propose a model with LSTM and the modified restricted attention mechanism to extract more accurate information for verbose expressions.

3. We constructed and published a dataset based on the interview dialogue, and we believe that it promotes the research progress of the task.
4. The experimental result shows that our proposed method can increase 54.08% and 18.91% F1 measure compared with the rule-based and direct classification methods.

Table 1: Examples of spoken texts from the annotated dataset.

Sentences	Verbose types
{那么[1]}我们近些年以来{啊[2]}，利用这个宝贵的文化资源，打造了三个文化传承的[品牌 + 平台[3]]。 {Then [1]} in recent years, we have {ah [2]} used this precious cultural resources, which has created three cultural inherited [brand + platform [3]].	1, Needless conjunction. 2, Modal particle. 3, Replacement.
像我们罗五的{这个[1]}苔丝糖{的话[2]}， 它一年的收入就要接近[上{这个[1]} + 上[3]]百万。 Such as {this[1]} Shao-Silk sugar {uh[2]}, its annual income is [more {this[1]} + more[3]] than one million dollars a year.	1, Needless pronoun. 2, Meaningless word. 3, Simple repetition.

2 Task Definition

2.1 The Task

The task is to detect the verbose characters in the spoken texts. Deleting these verbose characters will generate a more fluent text that preserves the original meaning.

Formally, we represent each sentence of the interview text as $(S,\ Y)$, where $S = (s_1, ..., s_i, s_n)$ is a sentence with a length n and the labels $y_j \in Y$ indicates the verbose characters in the sentence. The task is to estimate a conditional probability $P(y_j|S)$ from the dataset. The normal sentences can be obtained from the prediction result.

2.2 Data

The dataset is constructed based on 207 Chinese interview texts. Each interview contains two participants: a presenter and an interviewee. The presenter will introduce the main topic firstly and ask questions to the interviewee. The interviewee answers the questions. Usually, the answer is a long paragraph with many verbose characters. We manually annotated the verbose characters as shown in Table 1. We classify the verbose characters into two main categories. The first category includes modal particle like 'ah', needless conjunction like 'then', pronoun like 'this' and meaningless characters like 'uh' in Table 1. They are marked with curly brackets. Another category includes the simple repetition like 'more

more', replacement like 'brand platform' in Table 1. The second category is marked with square brackets and the plus sign. The characters before the plus sign is the verbose characters. Removing all labeled verbose characters will not affect the original meaning and fluency of the sentence.

These 207 texts are directly converted by speech recognition and contains many verbose expressions. Under the premise of following the original meaning, we require the labeling person to mark the verbose expressions as much as possible.

2.3 Challenges

The interview text contains various verbose characters. In dialog, people usually have special habits of speech that cause the verbose characters. For example, some people like to say specific modal characters such as 'ah', 'um'. In addition, people may realize that they have said something wrong and will correct it immediately. Those wrong characters are also converted into text by speech recognition. These all caused the diversification of verbose characters. Another challenge is that characters in different contexts may belong to different labels. For example, some pronouns like 'this' are not verbose characters when used as a sentence component, but as a mantra is verbose characters.

3 Method

In this section, we will firstly introduce the overall architecture of our model in subsection 3.1. Then, we will introduce the modified restrict attention (MRA) in detail.

3.1 Model Overview

We propose a neural networks with the MRA and compare mechanism to predict the probability distribution $P(y_j|S)$. Fig.2 shows our model's architecture. Due to the poor performance of the Chinese word segmentation on this dataset, the model is based on Chinese character units. The context layer extracts on the long-distance relevant information and gets the context focused (CF) representation. The MRA Layers with the normalize layers in the left part of Fig.2 can gather relevant local focused (LF) information into the Chinese characters' representation. We augment the local information with these densely connected layers. The other layers in the right part of the model in Fig.2 compose the compare mechanism. The compare mechanism takes the CF and LF representation as inputs and generates the local focused information behind the characters (BLF) to obtain the rear information.

Context layer is to incorporate long-distance information into the characters' vector. The habits of speech can be modeled in long-distance information. We utilize LSTM in bi-direction to encode each character.

$$\overrightarrow{h_i} = \overrightarrow{LSTM}(\overrightarrow{h}_{i-1}, s_i) \qquad i = 1, ..., N \tag{1}$$

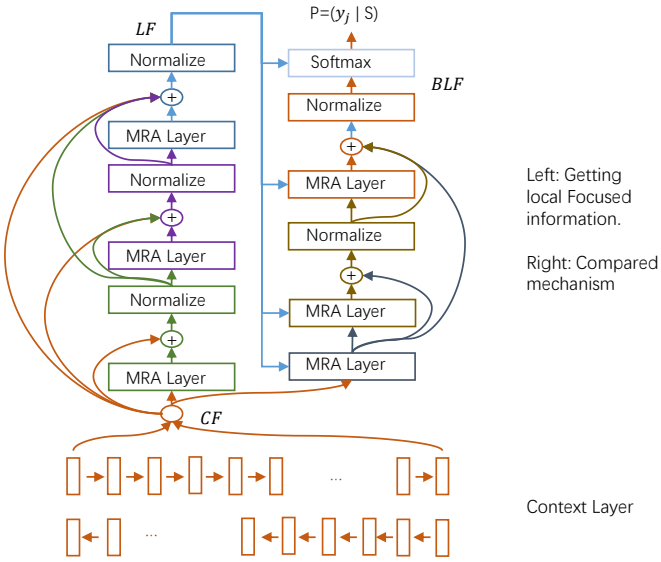

Fig. 2: The proposed model's architecture. CF: Context-Focused Representation. LF: Local-Focused Representation. BLF: Behind the Local-Focused Representation.

$$\overleftarrow{h}_i = \overleftarrow{LSTM}(\overleftarrow{h}_{i+1}, s_i) \qquad i = N, ..., 1 \tag{2}$$

After encoding, we concatenate the forward and reverse vectors together to represent the contextual information. It will be transmitted to next layers with the original characters' vector.

The MRA layer combined with the normalize layer in the left part of Fig.2 can gather relevant local information. We will introduce the MRA Layer in next subsection. The output of the MRA Layer is densely connected with the output of the context layer and front normalize Layers. Compared with residual connection, the densely connected can focus more on the own information. The connection function F is additive operation.

$$LF_i = F(LF_1, ..., LF_{i-1}, CF) \qquad i = 1, ..., L \tag{3}$$

The normalize layer ensures that the data does not become too large during the additive operation. This normalize layer also can accelerate model training. We can think of a MRA Layer, the additive operation and a normalize layer as a block. We use three blocks to encode the local information. Compared with other global attention mechanisms, the restrict attention focuses on the local information by the densely connect and restrict mask matrixes. The global attention usually gather too many contextual information which masking the original information in Chinese characters' vectors. Because the meaningful characters is much more than the verbose characters, the global attention on the long sentence may always predict the non-verbose label for all units. The output LF of the three blocks is transmitted to the compare mechanism.

The compare mechanism in the right part of Fig.2 takes the CF and LF as inputs. The principle of the compare mechanism is to obtain the local focused information BLF which behind the characters. As shown in Table 1, the simple repetition or replacement characters are very relevant to the characters behind. The compare mechanism gets the rear information with different mask matrixes in the MRA layers from CF according to LF. The front information and other global information is masked. Thought different matrixes, the front and rear local information are respectively obtained from different layers. Then, we use a linear layer and softmax to predict the probability based the information.

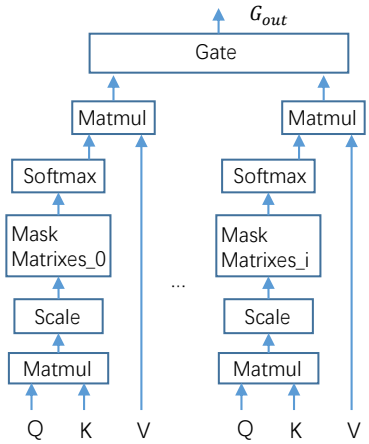

Fig. 3: The architecture of the MRA.

3.2 Modified Restrict Attention

Fig. 3 shows the architecture of the MRA layer. It is based on the multi-head attention proposed by [19]. Q, K and V represent query, keys and values vector[19]. In the self-attention, Q, K and V was obtained from applying different mapping matrixes to the same sentence. In the mutual attention, Q comes from one sentence and K/V comes from another. $\sqrt{d_k}$ is the scaling factor[19]. Compared with multi-head attention, the most difference is the mask matrixes and gate in our MRA. The computation of the a single restrict attention A_i is as follow.

$$A_i = softmax(M_i(\frac{Q * K^T}{\sqrt{d_k}})) * V \tag{4}$$

The mask matrixes M_i replace the attention weight of the needless information with a small number like $(-2^{32}+1)$. Then, through softmax, it will become a very small weight. One mask matrix will be applied to many heads in multi-head attention. As shown in Table 2, the simply repetition exists in Utterance_1. We

can easily identify this repetition in the fourth case of LF. The other characters will be masked by the matrixes. In Utterance_2, we may need the first cases in LF and BLF together to identify the verbose characters. Other information besides LF and BLF is mask.

The length of focused characters is a hyper-parameter and we set 4 in our model. Therefore, there are four cases of each Chinese character as shown in Table 2. The length should not be too large because we mainly focus on the local information. As we can see, there are only one or two valuable local information in LF and BLF. We add a gate for automatic learning to reduce the impact of irrelevant cases. We connect all the A_i and the input sentence representation S_e together to estimate a weight for each cases. S_e is the input sentence representation that will be converted to K/V as mentioned above. r is a tunable parameter that is to sharpen the weight. We tried r in [3, 5, 7] and set $r = 5$ in our model. The output of the gate pass through a feed forward layer and a normalize layer to as the MRA layer's output.

$$G_w = (softmax(W * [A_1, ..., A_i, S_e]))^r \qquad i = 1, ..., 4 \tag{5}$$

Table 2: Examples of different focused information. x: Difficult to express in English.

Utterance_1:	通过举办[泳 + 泳]博会，您觉得给城市面貌带来了哪些变化？ By holding the [Swimming + Swimming] Fair, what are the changes that you feel to our city?
LF:	1):过举办泳 (By holding the Swimming) 2):举办泳泳 (holding the Swimming Swimming) 3):办泳泳博 (x Swimming Swimming x) 4):泳泳博会 (the Swimming Swimming Fair)
BLF:	1):泳博会， (the Swimming Fair，) 2):博会，您 (Swimming Fair, what) 3):会，您觉 (Fair, what are) 4):，您觉得 (, what are the)
Utterance_2:	未来有没有[一些方向 + 新的方向]？ Are there [some directions + new directions] in the future?
LF:	1):一些方向 (some directions) 2):些方向新 (x directions new) 3):方向新的 (directions new) 4):向新的方 (x new x)
BLF:	1):新的方向 (new directions) 2):的方向？ (x directions ?) 3):方向？ (directions ?) 4):向？ (x ?)

4 Experiment

4.1 Dataset and Implementation Details

The dataset contains 207 spoken texts. We randomly cut the training set, development set and test set in a ratio of (8:1:1) and truncate all the sentences longer than 100 characters. The meaningful characters are labeled 0 and the verbose characters are labeled 1. We mainly focus on the performance on label 1. The label 0 characters account for 80.94% of the total Chinese characters. The label 1 only account for 19.06%. In our experiment, the character-embedding dimension is 128. The character embedding is pre-trained in Wiki corpus and fine-tuned in the train set. We use the Adam Optimizer with a fix learning rate 0.002. The model is trained on NVIDIA GTX 1080Ti GPU with the batch size of 128. We use F1 measure to evaluate all the models.

Table 3: Performance of baselines and our models. Bold data: Best Data.

Models:	Label 0			Label 1			Avg.
	P	R	F1	P	R	F1	F1
Rule-based	91.71	49.47	64.27	31.35	**83.76**	45.62	54.95
Direct classification	94.09	87.45	90.65	50.96	70.38	59.11	74.88
Our model$_1$	**96.19**	88.00	**92.45**	56.83	80.41	66.59	79.52
Our model$_2$	94.49	90.17	92.28	62.59	75.79	68.56	80.42
Our model$_3$	94.69	90.09	92.33	62.61	76.68	68.93	80.63
Our model	93.34	**91.19**	92.25	**67.28**	73.56	**70.29**	**81.27**

4.2 Compared with Baseline Models and Ablation Study

Experimental results of baselines and our models are listed in Table 3. Rule-based method counts the verbose words frequency and uses some rules to predict the label for sentences. The rule-based method can get highest recall in label 1 but the lowest precision. Therefore, their F1 measure 45.62 is quite low. The direct classification uses two-layer Convolutional neural network (CNN) to recognize verbose characters. It can utilize the character embedding to obtain useful features. However, due to the lack of contextual information, it also achieves lower performance. Our model$_1$ is the model only with the densely connected MRA. Its F1 measure is 66.59 and 92.45. It prove the effectiveness of our attention mechanism to extract the local information. Our model$_2$ is the densely connected MRA with the compare mechanism. When the compare mechanism is added to our model, its performance is further improved with the F1 measure is 68.56 and 92.28. Model the front and rare local-focused information can detect semantic duplication well. Our model$_3$ only contains the context layer and achieves better performance for extracting long-distance information. Our model with all the layers achieves the best performance with F1 measure is 70.29 and 92.25.

It proved that combining long-distance and local information is important for improving the performance. Compared with the baselines, our proposed model increased F1 measure of label 1 by 54.08% and 18.91%.

4.3 Compared with Different Attention Models and CRF Models

We compare our model with many different global attention models. The attention models in Table 4 are bi-direction LSTM with self-attention. The difference between these models is the computation methods of the self-attention. Q and K in self-attention are from the same sentence. Att_0 directly concatenate the Q and K vectors together with a linear mapping to get the attention information. Att_1 firstly mapping the Q and K with different mapping matrixes and computes the point-wise multiplication. Att_2 only computes the point-wise multiplication between vectors of Q and K. The point-wise multiplication divides the scaling factor $\sqrt{d_z}$[19]. We also tried the BI-LSTM together with cosine similarity attention and multi-head attention[19] but they do not work. Gathering too much global contextual information makes the two model only predict the label 0, because label 0 account for 80.94% in all characters.

As we can see, every other kind of attention mechanisms achieves lower performance compared with our model. The more complex the other attention mechanisms is, the more performance is lost. The most complex attention mechanisms are the cosine similarity attention and multi-head attention that makes the model only predict label 0 for all characters. The Att_0, Att_1 and Att_2 are in a simpler order. The Att_0 only achieves 56.83 F1 value which is 19.15% lower than our model and is lower than CNN baseline. The most simple attention mechanism, the Att_2, only affects a little performance with F1 value is 66.47. All the above attention mechanisms prove that adding global information has no benefit to our model.

Table 4: Performance of different attention mechanisms and CRF models. Bold data: Best Data.

Settings:	Label 0			Label 1			Avg.
	P	R	F1	P	R	F1	F1
Bi-LSTM+Att_0	89.19	87.69	88.44	55.08	58.69	56.83	72.64
Bi-LSTM+Att_1	93.72	88.55	91.06	56.52	71.49	63.13	77.10
Bi-LSTM+Att_2	**96.75**	88.66	**92.53**	55.59	**82.65**	66.47	79.50
Bi-LSTM+CRF	94.71	90.05	92.32	62.45	76.67	68.83	80.58
Bi-LSTM+CRF+Att_0	92.06	84.70	88.23	40.32	58.59	47.77	68.00
Bi-LSTM+CRF+Att_1	95.25	87.66	91.3	51.89	75.28	61.44	73.37
Bi-LSTM+CRF+Att_2	93.23	89.87	91.52	62.30	71.95	66.78	79.15
Our model	93.34	**91.19**	92.25	**67.28**	73.56	**70.29**	**81.27**

The LSTM combined with a conditional random field (CRF) has achieved impressive performance in many sequence-labeling tasks[9, 18]. In Table 4, the

best performance of CRF models is 68.83. Therefore, the CRF has little influence on the performance. It proved that the verbose expressions are very diverse and have no obvious transfer relationship. Compared with CRF models, our model also achieved the best performance with the F1 measure is 70.29. It proves that our model could integrate local-focused information and long-distance information well to achieve better performance.

Table 5: Examples of the error prediction by our model.

Utterance	Original sentences	Gold truth	Predicted sentences
U_1	所以呢要形成这三家啊联动的这么一个一个机制，这么一种氛围。 Therefore, ah, it is necessary to form such a a linkage mechanism and such an atmosphere of the three sides, ah.	要形成三家联动的一个机制，一种氛围。 It is necessary to form a linkage mechanism and an atmosphere of the three sides.	要形成三家联动的机制，氛围。 It is necessary to form a linkage mechanism and atmosphere of the three sides.
U_2	他有一个恒定的规律。 He has a constant rule.	他有一个恒定的规律。 He has a constant rule.	他有恒定的规律。 He has a constant rule.
U_3	啊他他有这方面的考虑。 Ah he he has a consideration in this aspect.	他有这方面的考虑。 He has a consideration in this aspect.	他他有这方面的考虑。 He he has a consideration in this aspect.
U_4	而且我们是通过问题导向进行创新。 Moreover, we are innovating through problem orientation.	而且我们是通过问题导向进行创新。 Moreover, we are innovating through problem orientation.	通过问题导向进行创新。 Innovating through problem orientation.

4.4 Qualitative Analysis

Table 5 shows many error generated by our model. U_1 and U_2 is caused by inconsistencies annotations in the dataset. We can see that, the predicted sentences U_1 and U_2 also are normalized sentences. In English language, U_2 even has the same target translation. The inconsistent annotations in the dataset account for many error predictions. Therefore, our proposed model will performance better without the inconsistent annotations. U_3 is an error prediction of simple repetition. The compare mechanism does not capture this pattern. U_4 is a misrecognition of meaningful conjunction and noun because these conjunction and noun are meaningless in many other cases. Our proposed model does not properly understand contextual information in the sentence.

5 Related Work

The task most similar to ours is text normalization. It's very useful in many texts such as cell phone messages[2, 5, 4], social media texts[1, 20, 11, 7, 21] and broadcast transcription[1]. In cell phone messages normalization, they mainly focus on translate the brief and colloquial words into standard forms. In [2], they treated this as a Machine translation tasks. They achieved good performance using the phrase-based statistical model. In [5], they used an unsupervised model and achieved good performance. [4] proposed a method to utilize the rule-based and machine translation approaches to achieve better performance. However, we only need to delete the verbose characters without translation in the spoken texts.

In the field of social media texts, there are also a lot of text normalization work. [15] utilized many unsupervised features to choose candidate words and used graph-based approach to normalize sentences. [21] weighted different unsupervised features and employed a new training algorithm to search in the large space. In [17], they utilized distributed representations of words to gather the contextual relevant information into vectors and get good performance on a Twitter dataset. In social media texts, they mainly focus on using candidate words to replace the colloquial words. However, we focus on the verbose and noise expressions in spoken texts. The spoken texts have no abbreviated words and have many different cases.

The dataset proposed in [20], is similar to ours. They normalize the chat texts on the Internet chat texts. They proposed a phonetic mapping model to map the chat terms to a standard word via phonetic transcription[20]. They mainly solve the dynamic problem in the Internet chat. [16] used RNN to do text normalization. However, they also mainly focus on the transformation between different words' forms.

In a short word, we focus on the verbose and noise characters in the spoken texts. Meanwhile, deleting these characters will normalize the sentence and keep the original meaning of the sentence.

6 Conclusion

In this work, we propose a new task to detect verbose characters in spoken tests and construct a new dataset. The dataset drives the task of transforming texts with verbose and noise characters into normalized texts. We propose an attention mechanism that use the different mask matrixes and a gate to get relevant local focused information. We also propose a compare mechanism to leverage the front and rear local focused information. Experimental results on the dataset show that our proposed model performances better than many other models and achieves the state-of-the-art performance. In future work, we want to increase data filtering method in our model to reduce the influence of inconsistent annotations and apply our model in different tasks.

References

1. Adda-Decker, M., Adda, G., Lamel, L.: Investigating text normalization and pronunciation variants for german broadcast transcription. In: Sixth International Conference on Spoken Language Processing (2000)
2. Aw, A., Zhang, M., Xiao, J., Su, J.: A phrase-based statistical model for sms text normalization. In: Proceedings of the COLING/ACL on Main conference poster sessions. pp. 33–40. Association for Computational Linguistics (2006)
3. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. CoRR **abs/1409.0473** (2014)
4. Beaufort, R., Roekhaut, S., Cougnon, L.A., Fairon, C.: A hybrid rule/model-based finite-state framework for normalizing sms messages. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. pp. 770–779. Association for Computational Linguistics, Uppsala, Sweden (July 2010)
5. Cook, P., Stevenson, S.: An unsupervised model for text message normalization. In: Proceedings of the workshop on computational approaches to linguistic creativity. pp. 71–78. Association for Computational Linguistics (2009)
6. Dong, L., Mallinson, J., Reddy, S., Lapata, M.: Learning to paraphrase for question answering. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. pp. 875–886. Association for Computational Linguistics, Copenhagen, Denmark (September 2017)
7. Hassan, H., Menezes, A.: Social text normalization using contextual graph random walks. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1577–1586. Association for Computational Linguistics, Sofia, Bulgaria (August 2013)
8. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (Nov 1997)
9. Huang, Z., Xu, W., Yu, K.: Bidirectional LSTM-CRF models for sequence tagging. CoRR **abs/1508.01991** (2015)
10. Khashabi, D., Khot, T., Sabharwal, A., Roth, D.: Learning what is essential in questions. In: Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017). pp. 80–89. Association for Computational Linguistics, Vancouver, Canada (August 2017)
11. Liu, F., Weng, F., Wang, B., Liu, Y.: Insertion, deletion, or substitution? normalizing text messages without pre-categorization nor supervision. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. pp. 71–76 (2011)
12. Liu, Y., Sun, C., Lin, L., Wang, X.: Learning natural language inference using bidirectional LSTM model and inner-attention. CoRR **abs/1605.09090** (2016)
13. Qin, K., Wang, L., Kim, J.: Joint modeling of content and discourse relations in dialogues. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 974–984. Association for Computational Linguistics, Vancouver, Canada (July 2017)
14. Shaw, P., Uszkoreit, J., Vaswani, A.: Self-attention with relative position representations. CoRR **abs/1803.02155** (2018)
15. Sonmez, C., Ozgur, A.: A graph-based approach for contextual text normalization. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 313–324. Association for Computational Linguistics, Doha, Qatar (October 2014)

16. Sproat, R., Jaitly, N.: RNN approaches to text normalization: A challenge. CoRR **abs/1611.00068** (2016)
17. Sridhar, V.K.R.: Unsupervised text normalization using distributed representations of words and phrases. NAACL HLT 2015 pp. 8–16 (2015)
18. Sun, W., Sui, Z., Wang, M., Wang, X.: Chinese semantic role labeling with shallow parsing. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3 - Volume 3. pp. 1475–1483. EMNLP '09, Association for Computational Linguistics, Stroudsburg, PA, USA (2009)
19. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. CoRR **abs/1706.03762** (2017)
20. Xia, Y., Wong, K.F., Li, W.: A phonetic-based approach to chinese chat text normalization. In: Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics. pp. 993–1000. Association for Computational Linguistics, Sydney, Australia (July 2006)
21. Yang, Y., Eisenstein, J.: A log-linear model for unsupervised text normalization. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. pp. 61–72. Association for Computational Linguistics, Seattle, Washington, USA (October 2013)