A Multi Neural Networks based Approach to Complex Chinese Medical Named Entity Recognition

Bin Ji[1], Shasha Li[1], Jie Yu[1], Jun Ma[1], Jintao Tang[1], Dongyang Liang[1], Huijun Liu[2,*]

[1] National University of Defense Technology, Changsha, Hunan
[2] China Academy of Engineering Physics, Mianyang, Sichuan
*lhj12uestc@163.com

Abstract. More and more researchers are paying attention to how to efficiently extract high-value scientific research information from electronic medical records recently. The 2019 China Conference on Knowledge Graph and Semantic Computing (CCKS 2019) takes extraction of medical entity and attribute as an open challenge, specifically, extracts three malignant tumor-related entities or attributes from the Chinese electronic medical records. In this open challenge, we propose a multi neural networks based approach, which consists of a couple of BiLSTM-CRF models and a CNN model for sentence classification. The F1-score obtained on the official test data set by our approach is 76.35%, which rank first in this open challenge. In addition, generalization validation is performed on another data set released by the open challenge organized by the 4th China Health Information Processing Conference (CHIP 2018).

Keywords: neural network; electronic medical record; tumor; medical named entity;

Introduction

With the rapid spread of electronic medical records and the arrival of the medical big data era, the application and development of Natural Language Processing (NLP) technology in the medical field has become a hot research topic. Named entity recognition (NER) is a fundamental task in NLP [1], which aims to identify naming referential items from the text, paving the way for tasks such as relationship extraction. The medical named entity recognition is the most basic task of medical information extraction, and a mass of influential academic conferences take it as an open challenge, e.g., CCKS [2] has organized medical named entity recognition open challenge for Chinese electronic medical records (CEMRs) for three consecutive years since 2017. These open challenges also provide a batch of high-quality annotated data sets for subsequent research.

The 2019 China Conference on Knowledge Graph and Semantic Computing (CCKS 2019) takes medical entity and attribute extraction in CEMR as an open challenge [3], which aims to extract three malignant tumor-related named entities from CEMRs, i.e. tumor primary site, primary tumor size, and tumor metastatic site. In this open challenge, 900 entries of manually annotated CEMRs are released as training data, and 400 entries of raw CEMRs are released as test data, which are

identified by CCKS TR and CCKS TE respectively in this paper. For this open challenge, we proposed a multi neural networks based approach, which consists of a couple of BiLSTM-CRF models and a CNN model for sentence classification and help us to won the champion with a F1-score of 76.35% on the official test data set (CCKS TE).

Related Works

Medical NER refers to the determination of the boundaries of technical terms in the medical field text, and then classifies them based on domain information [4]. At present, the main methods of medical NER can be divided into shallow machine learning based and deep neural network model based methods. Shallow machine learning based methods [5] mainly include CRF, SVM, etc. In 2015, Wang [6] et al. verified the CRF-based Gimli method, and achieved a 72.23% F1-score on the JNLPBA 2004 data set; in 2014, Tang [7] et al. adopted the CRF model for biomedical named entity recognition, added different artificial word vector features and obtained 71.39% F1-score on the JNLPBA 2004 data set; in 2015, Chang [8] et al. combined a small number of artificial features and word vector to construct a CRF model and added post-processing procedure, and obtained a F1-score of 71.77% on JNLPBA 2004 dataset.

Neural network model based approaches are widely applied to the study of medical NER. In 2015, Yao [9] et al. first leveraged neural networks models to generate word embeddings of unlabeled biomedical texts, and then established multi-layer neural networks models, which obtained 71.01% F1-score on the JNLPBA 2004 dataset. In 2016, Li [10] et al. utilized the BiLSTM model to achieve an F1-score of 88.6% on the BioCreative II GM dataset, and a 72.76% F1-score on the JNLPBA 2004 dataset. In 2018, Li [11] et al. proposed a CNN-BLSTM-CRF based neural network model, which achieved the state-of-the-art performance on the Biocreative II GM and the JNLPBA 2004 dataset. The BiLSTM-CRF based approach won the championship of CCKS open challenges in 2017 and 2018, respectively.

In addition, rule-based approaches are a very efficient way to NER, using hand-written rules to match text to recognize named entities [12]. Also, rules can play a crucial role in data preprocessing and post-processing. However, rule-based approaches require professional personnel to write rules and domain expertise, and both the generalization ability and crossing domain transfer ability are poor.

Method

Task Analysis

The definition of tumor primary site, primary tumor size, and tumor metastatic site in the CCKS 2018 open challenge are shown below.

1. **Tumor primary site:** the original body part of the tumor, which is different from tumor metastatic site. Usually, the following context of tumor primary site is “癌”(cancer), “恶性肿瘤”(malignant tumor), “MT”, “CA”, etc., and tumor primary site tends to the specific body part. A case study is shown below, “左肺癌”(left lung cancer) and “左肺下叶癌”(left lung lobe cancer). When the above two appear in an electronic medical record at the same time, "左肺下叶"(left lung lobe) is selected as the tumor primary site.
2. **Primary tumor size:** A measurement of the length, area or volume of the primary tumor, common measure units includes MM, CM, etc. In essence, primary tumor size is an attribute of tumor primary site. In this paper, for unified description, we classify primary tumor size into entity category.
3. **Tumor metastasis site:** The metastatic site of the primary tumor. In theory, any other part of the body can be metastatic site of the primary tumor.

From above definition, we can conclude that as a measurement of primary tumor, primary tumor size depends on tumor primary site. In general, primary tumor size and tumor primary site coexist at the sentence level in the CEMR. In addition, as a measurement, expressions of primary tumor size are highly abstracted, which means that we can achieve efficient extraction of primary tumor size with a rule-based method.

From a medical point of view, there is no necessary intrinsic link between tumor metastatic site and tumor primary site. In theory, any body part or tissue can become a tumor metastatic site, so the extraction of tumor metastatic can be achieved as an independent sub-task.

Both tumor primary site and tumor metastatic site belong to the body part or tissue, and there is no essential difference in form between the two. For tumor metastatic site, only a few feature descriptors can be used to distinguish whether an anatomy is tumor metastatic site or not, but this discriminating ability is weakened as the length of the sentence increases.

Based on the above analysis, we decomposed the CCKS 2019 open challenge into three sub-tasks, which are extraction of tumor primary site, extraction of primary tumor size and extraction of tumor metastatic site.

Method Design

Figure 1 shows the architecture of the multi neural networks based Approach, including a couple of BiLSTM-CRF models for NER and a CNN model for sentence classification[1]. In the following section, we will introduce our approach in detail from aspect of the stated three sub-tasks.

Extraction of tumor primary site

The extraction of tumor primary site is a typical NER process. As stated, there are obvious feature descriptors for malignant tumor in CEMRs, e.g. “癌” (cancer), “恶性肿瘤”(malignant tumor), “MT”, “CA”, etc., so we can conveniently tag the training

[1] Codes for this open challenge can be found: https://github.com/nudt-yh/ccks2019_subtask2

data to train the neural network model. In our approach, the BiLSTM-CRF model was used for the extraction of the tumor primary site, as shown in Figure 2.

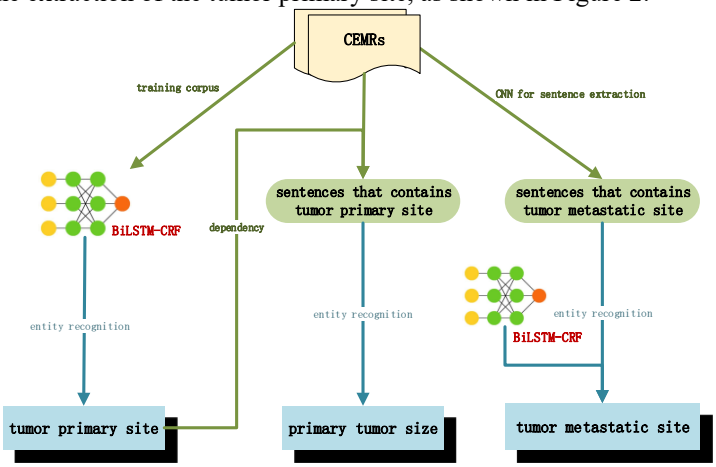

Figure 1 Framework of medical named entity and attribute extraction method

The first layer of BiLSTM-CRF model is the embedding layer, which distribute a word embedding (randomly initialized in this open challenge) to each token of the sentence, and finally obtain the sentence representation, i.e. $x=(x_1,x_2,...,x_n)$, where $x_i \in R^d$, and d is word embedding dimension.

The second layer of the model is the bidirectional LSTM layer, which automatically extracts sentence features. The word embedding sequence $(x_1, x_2, \ldots, x_n)$ of a sentence is taken as input of each time step of the bidirectional LSTM, and the hidden state output sequence $(\overrightarrow{h_1}, \overrightarrow{h_2} \ldots\ldots, \overrightarrow{h_n})$ of forward LSTM and the hidden state output sequence $(\overleftarrow{h_1}, \overleftarrow{h_2}, \ldots, \overleftarrow{h_n})$ of backward LSTM are concatenated by position $h_t = [\overrightarrow{h_t}; \overleftarrow{h_t}] \in R^m$, and get the complete hidden state sequence $(h_1, h_2, \ldots, h_n) \in R^{n \times m}$.

Then a linear layer is set to map the hidden state vector from m-dimension to k-dimension (k is the number of tags defining in the tagging set), and then the automatically extracted sentence features are obtained, which are recorded as the matrix $P = (p_1, p_2, \ldots, p_n) \in R^{n \times k}$. Each element p_{ij} of $p_i \in R^k$ can be regarded as a score that tag the word x_i with the j^{th} tag. Next, a CRF layer is set to tag words.

The third layer of the model is CRF, which performs sequence-level word tagging. The parameter of CRF layer is a transition matrix A with a dimension of $(k+2) \times (k+2)$, where k is tag number, and A_{ij} represents the transition score from the i^{th} tag to the j^{th} tag, so tags that have been previously tagged can be utilized when tagging a new word. If a tag sequence is represented by $y = (y_1, y_2, \ldots, y_n)$, while n equals sentence length, the formula used to measure that tag of sentence X equal to tag sequence y is shown in following formula.

$$\text{score}(X, y) = \sum_{i=1}^{n+1} A_{y_{i-1}, y_i} + \sum_{i=1}^{n} P_{i, y_i}$$

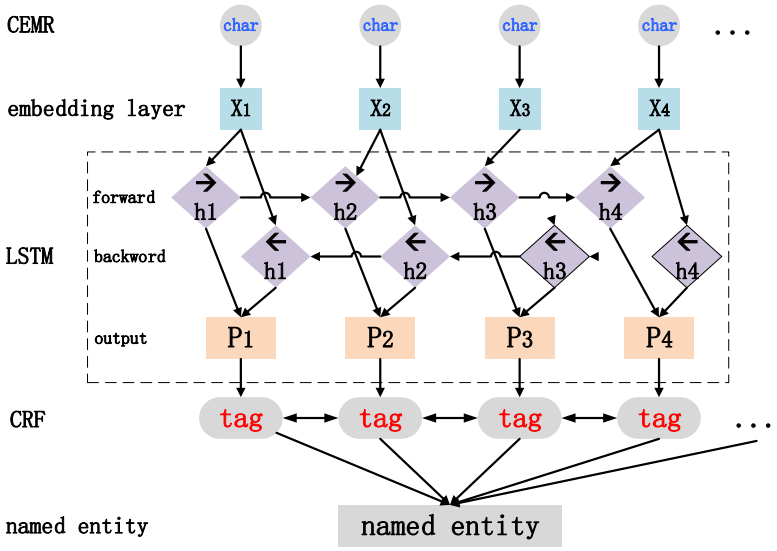

Figure 2 The framework of the BiLSTM-CRF neural network

It can be seen that the $\mathrm{score}(X, y)$ equals the sum of scores of all words in sentence and each score consist of two parts, the first part is from the transition matrix A, and the second part is from the matrix P described above. Softmax is used to normalize probability, which is shown in following formula.

$$P(y|X) = \frac{\exp(\mathrm{score}(X, y))}{\sum_{y'} \exp(\mathrm{score}(X, y'))}$$

While training, for training sample (X, y^x) the following formula is taken to maximize the log probability of tag sequence.

$$\log P(y^x|X) = \mathrm{score}(X, y^x) - \log\left(\sum_{y'} \exp(\mathrm{score}(X, y'))\right)$$

During the encoding process, Viterbi algorithm is used to calculate the optimal tag path with dynamic planning, as the following formula shows.

$$y^* = \arg\max_{y'} \mathrm{score}(X, y')$$

The BIO [13] tagging schema is used to process the CCKS TR into a suitable format for model training based on manual annotation information. Among the three tags B-TU and I-TU are used to tag the first and non-first token of tumor primary site, respectively. O tags the words do not belong to any named entity. A case study of tagging schema is shown in Figure 3.

结	合	临	床	，	右	乳	腺	癌	并	右	腋	窝	淋	巴	结	肿..
0	0	0	0	0	B-TU	I-TU	I-TU	0	0	0	0	0	0	0	0	0

Figure 3 Case study of tagging schema

As stated, when coarse-grained anatomy and fine-grained anatomy are simultaneously extracted as tumor primary site, the fine-grained anatomy is selected as the final tumor primary site, which is performed by post-processing steps.

Extraction of primary tumor size

As stated, primary tumor size is a measurement of the primary tumor. In CEMRs, such metrics are composed of numbers, length units (MM or CM), and binary symbols (*, x, X, etc.) representing multiplication according to certain rules. Since expressions of primary tumor size are highly abstract, so in this open challenge, a rule-based approach is utilized to extract primary tumor size. Detail extraction procedures are shown below.

Step one: For each CEMR, replace the punctuation marks such as "?", "？ ", ";", "； " with ".", and split the CEMR according to "," to obtain a sentence set.

Step two: For each sentence in the sentence set, judge whether it contains corresponding tumor primary site. If not, remove the sentence from the sentence set. Finally combine remaining sentences in the set into a short text.

Step three: Edit a regular expression and utilize it to extract metrics in the short text obtained above. In this open challenge, regular expression is shown below.

RE = '?\d?\d?\d?\.?\d?\d.?(([CcMm][mM]?)|(.?.?[*×X~].?\d?\d?\d?\.?\d?\d?))*.?[CcMm|[mM])'

Step four: Post-processing steps are performed to remove noise metrics, primarily measurements of lymph node size and measurements of distance.

Extraction of primary tumor size depends on the extraction results of tumor primary site. As a result, error propagation may occur from the latter to the former.

Extraction of tumor metastatic site

As stated above, there is no significant intrinsic relationship between tumor metastatic site and tumor primary site. And we have discussed that with a few of feature descriptors, i.e. "转移"(metastatic), "骨质破坏"(bone destruction), "代谢活跃 "(Metabolic activity) etc., it is difficult to distinguish multiple tumor metastatic sites in long sentences. A heuristic extraction approach is to first pre-process the electronic medical record, and then obtain sentences containing tumor metastatic site; second, extract all anatomies in the sentences obtained above; third, post-processing steps are performed to obtain final tumor metastatic sites. Detailed processing procedures are shown below.

Step one: In another academy paper [14] of our research group, we have a detail description on how to obtain sentences that contains tumor metastatic site with CNN for sentence classification. And we directly transfer the pre-process steps into this open challenge.

Step two, a BiLSTM-CRF model is taken to extract the anatomy entity in the sentences obtained in the first step, of which the model architecture is the same to the one shown in Figure 2.

Step three: According to entity format requirements, post-processing steps, which includes entity de-duplication, and entity special format processing, are performed to obtain the final tumor metastatic site.

Training data used to train the BiLSTM-CRF model is from CCKS TR, and the BIO tagging schema is used to tag CEMRs. For there is no manually annotated information that can be directly used to tag CEMRs into suitable training corpus, so

we manually annotated 400 entries CEMRs of CCKS TR with the tagging tool[2] developed by our research group.

Evaluation Metrics and Experiments

Evaluation Metrics

In this open challenge, entity recognition results are submitted online, and the F1-scores can be found in **biendata competitions**[3]. In addition, standard precision (P), recall (R), F1-score are utilized to evaluate the F1-score of our approach on CHIP 2018 data set. Definitions are shown below.

Standard precision (P), recall (R) and F1-score are computed by

$$\mathrm{P} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}},$$

$$\mathrm{R} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}},$$

$$\mathrm{F1} = \frac{2 \times \mathrm{P} \times \mathrm{R}}{\mathrm{P} + \mathrm{R}},$$

A recognized entity mention is regarded as True-Positive (TP) if it is identical to a gold entity mention regardless of its boundary or category. A recognized entity mention is regarded as False-Positive (FP) as long as it cannot satisfy the aforementioned conditions. The False-Negative (FN) value is the number of entity mentions that cannot be recognized by the NER system.

Experiments

The top 5 F1-scores of this open challenge are shown in Table 1, in which the first column shows team names and the second column shows F1-scores. The 76.35% F1-score, which is achieved by our research group, ranks first in this open challenge.

Table 1. The top 5 F1-scores of CCKS 2019 open challenge.

Team name	F1-score (%)
NUDT-YH	**76.350**
THU_MSIIP	76.165
DUTIR	70.490
ZU_NLP	70.167
SCNU_TAMlab	59.906

CHIP 2018 releases an open challenge in the same form as the CCKS 2019 open challenge, and provided 800 entries of annotated CEMRs as training and test data sets. In order to verify the effectiveness and generalization ability of our appraoch, we

[2] Source codes for the tagging tools can be found: https://github.com/nudt-yh/tagging_tool

[3] https://biendata.com/competition/ccks_2019_1_2/final-leaderboard/

transfer our approach to the 800 entries CEMRs. The test results are shown in Table 2, of which the first column is the statistical result of corresponding manually annotated entities.

Table 2. Testing results of our approach on the dataset released by CHIP 2018.

category	entity number	P (%)	R (%)	F1-score (%)
tumor primary site	827	71.58	74.00	72.77
primary tumor size	499	89.91	82.16	85.86
tumor metastatic site	2178	60.55	58.22	59.36
Overall	3504	67.25	65.35	66.29

As can be seen from Table 2, our approach has obtained a F1-score of 66.29% on the CHIP 2018 dataset, which is ~10% lower than the F1-score that achieved on CCKS TE dataset. It was found that criterion inconsistency of the two open challenges is responsible for the decline in the performance. There are two main reasons.

The first reason is the recognition of tumor primary site. In the CHIP 2018 open challenge, the tumor primary site is without a position word, but it is required to have a position word in the CCKS2019 open challenge. A case study is shown below, “左肺上叶后段癌”(post-left lung upper lobe cancer). In CHIP 2018, “左肺上叶”(left lung upper lobe) is the correct tumor primary site, but in CCKS 2019, “左肺上叶后段”(post-left lung upper lobe) is the correct one.

The second reason is that in the CHIP 2018 open challenge, it is necessary to decompose lymph node-related entities and add numerous bone metastatic categories，which can be distinguished by “骨质破坏”(bone destruction). For decomposing lymph node-related entities, a case study is shown below, "左侧腮腺、双颈、右侧锁骨上区间隙多发淋巴结,考虑转移" (The left parotid gland, double neck, right supraclavicular space, multiple lymph nodes, considering metastatic). In the CCKS 2019 open challenge, "左侧腮腺、双颈、右侧锁骨上区间隙多发淋巴结" (The left parotid gland, double neck, right supraclavicular space, multiple lymph nodes) is extracted as tumor metastatic site, while in the CHIP 2018 open challenge "左侧腮腺淋巴结"(left parotid lymph nodes), "双颈淋巴结"(double neck lymph nodes), "右侧锁骨上区淋巴结"(right supraclavicular lymph nodes) should be the correct tumor metastatic sites.

We add a decomposition algorithm to our approach to decompose lymph node-related entities into suitable format. And the test result of our modified approach on CHIP dataset is shown in Table 3.

Table 3. Testing result of our modified approach on the dataset released by CHIP 2018.

catgegory	entity number	P (%)	R (%)	F1-score (%)
tumor primary site	827	71.58	74.00	72.77
primary tumor size	499	89.91	82.16	85.86
tumor metastatic site	2178	**75.92**	**87.56**	**81.32**
Overall	3504	**76.61**	**83.59**	**79.95**

From Table 3, we can get that for tumor metastatic site, an absolute F1-score improvement of 21.96% is obtained, and for the overall F1-score, an absolute improvement of 13.66% is obtained. Another interesting phenomenon is that the overall F1-score of CHIP 2018 open challenge is about 3.6% higher than the one of CCKS 2019 open challenge, on which we will have an in-depth study in the future.

In conclusion, the effectiveness of the proposed approach is validated on the data set released by CHIP 2018 open challenge, and our method has powerful generalization ability.

Conclusions

In this paper, we propose a multi neural networks based approach to complex Chinese medical NER, which effectively completes the open challenge released by CCKS 2019, and achieved an F1-score of 76.35% on the official test data set, which help us to get the first place in the open challenge. Also we transfer our approach to the 800 entries of annotated CEMRs to validate effectiveness and generalization ability, which confirms that our approach is effective and has powerful generalization ability.

But there is still a lot of effect needs to do to perfect our approach. First of all, the extraction of primary tumor site is achieved with a rule-based approach, which we will substitute with neural network based way; secondly, the two BiLSTM-CRF models used in our method are based on randomly initialized character embeddings. It has been proved that domain-related pre-trained character embeddings can effectively improve NER performance [15, 16]. So our second work in the future is to pre-train domain-specific character embeddings with ELMo[17] or BERT[18] to further improve the performance of our approach.

References

[1] Buzhou Tang, Xiaoling Wang, Jun Yan and Qingcai Chen. Entity recognition in Chinese clinical text using attention-based CNN-LSTM-CRF. BMC Medical Informatics and Decision Making. 19(supply 3):74.

[2] CCKS 2018 NER of CEMR, https://www.biendata.com/competition/CCKS2018_1/.

[3] Open challenge 1: CEMR clinical named entity and attribute extraction. http://icrc.hitsz.edu.cn/chip2018/Task.html.

[4] Xiao Sun, Zhongyuan Sun, Fuji Ren. Biomedical named entity recognition based on deep conditional random fields. PR & AI. 2016, 29(11):997-1008.

[5] Dong X S, Qian L J, Guan Y. A multiclass classification method based on deep learning for NER in electronic medical record. Proceedings of the International 2016 New York Scientific Data Summit (NYSDS), 2016:1-10.

[6] Wang X, Yang C, Guan R. A comparative study for biomedical NER. International Journal of Machine Learning & Cybernetics, 2015:1-10.

[7] Tang B, Cao H, Wang X. Evaluating word representation features in biomedical NER tasks. BioMed Research International, 2014:1-6.

[8] Chang F, Guo J, Xu W. Application of word embeddings in biomedical NER tasks.

Digital Inf. Manage. 2015, 13(5):321-327.
[9] Yao L, Liu H, Liu Y. Biomedical NER based on deep natural network. International Journal of Hybrid Information Technology. 2015,8(8):279-288.
[10] Li L, Jin L, Jiang Y. Recognizing biomedical named entities based on sentence vector/twin word embeddings conditioned bidirectional LSTM. Proceedings of China National Conference on Chinese Computational Linguistics. Springer International Publishing, 2016:165-176.
[11] Li,L.S. et al. (2018) Biomedical named entity recognition with CNN-BLSTM-CRF. Journal of Chinese Information Processing, 32(1), 116-122.
[12] Vikas Yadav, Steven Bethard. A survey on recent advances in NER from deep learning models. Proceedings of the 27th international conference on computational linguistics. 2018: 2145-2158.
[13] Bin Ji, Rui Liu, Shasha Li, Jie Yu, Qingbo Wu, Yusong Tan and Jiaju Wu. A hybrid approach for NER in CEMR. BMC Medical Informatics and Decision Making. 19(supply 2):64.
[14] Bin Ji, Shasha Li, Jie Yu, Jun Ma, et al. Research on Chinese Medical Named Entity Recognition Based on Collaborative Cooperation of Multiple Neural Network Models. Journal of Medical Bioinformatics, unpublished.
[15] Yanran Li, Wenjie Li, Fei Sun, and Sujian Li. Component-enhanced Chinese character embeddings. Proceedings of the 2015 Conference on empirical methods in natural language processing. 2015: 829-834.
[16] Shao Yan, Christian Hardmeier, and Joakim Nivre. Multilingual NER using hybrid neural networks. The sixth Swedish language technology conference. 2016.
[17] Matthew E. Peters, Mark Neumann, Mohit Iyyer, et al. Deep contextualized word representations. arXiv:1802.05365 [cs.CL].
[18] Jacob Devlin, Ming-wei Chang, Kenton Lee, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 [cs.CL].