The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Multi-Task Medical Concept Normalization Using Multi-View Convolutional Neural Network

Yi Luo,[1] **Guojie Song,**[2*] **Pengyu Li,**[2] **Zhongang Qi**[3]
[1]Department of Computer Science and Engineering, University of California, San Diego, USA
[2]Key Laboratory of Machine Perception, Ministry of Education, Peking University, China
[3]School of Electrical Engineering and Computer Science, Oregon State University, USA
yil901@eng.ucsd.edu, gjsong@pku.edu.cn, lipengyu@pku.edu.cn, qiz@oregonstate.edu

Abstract

Medical concept normalization is a critical problem in biomedical research and clinical applications. In this paper, we focus on normalizing diagnostic and procedure names in Chinese discharge summaries to standard entities, which is formulated as a semantic matching problem. However, non-standard Chinese expressions, short-text normalization and heterogeneity of tasks pose critical challenges in our problem. This paper presents a general framework which introduces a tensor generator and a novel multi-view convolutional neural network (CNN) with multi-task shared structure to tackle the two tasks simultaneously. We propose that the key to address non-standard expressions and short-text problem is to incorporate a matching tensor with multiple granularities. Then multi-view CNN is adopted to extract semantic matching patterns and learn to synthesize them from different views. Finally, multi-task shared structure allows the model to exploit medical correlations between disease and procedure names to better perform disambiguation tasks. Comprehensive experimental analysis indicates our model outperforms existing baselines which demonstrates the effectiveness of our model.

Named Entity Disambiguation (NED), which links mentions in text to entities in knowledge bases, is an important research topic in natural language processing (Bunescu and Pasca 2006; Alhelbawy and Gaizauskas 2014). Chinese medical concept normalization is a typical problem of NED in biomedical domain, which aims to normalize ambiguous medical mentions into concepts in a controlled vocabulary such as International Classification of Disease 10th revision (ICD-10). This problem has wide-ranging applications in clinical research (Leaman, Khare, and Lu 2015), statistical analysis for hospitals, epidemiological research (Pakhomov, Buntrock, and Chute 2006), and diagnosis-related group (DRG) (Hoelzer, Schweiger, and Dudeck 2003), and consequently, the solutions to it benefit a wide range of people and generate great societal and technical impacts.

The major issue of Chinese medical concept normalization is the variety of the data sources, which brings multiple different expressions of the same entity due to diverse writing habits, experiences of physicians, requirements of medical institutions, etc. For example, the dataset used in this paper comes from the discharge summaries, an overview of hospitalization for patients, from 153 triple-A hospitals in 31 provinces and regions in China. Each record in the dataset includes a primary diagnosis of disease and a corresponding procedure both in ambiguous short texts. For clarity, we denote the short-text statement written by doctors as *mention* and the medical concept from standard library as *entity*. This paper focuses on normalizing disease mentions and procedure mentions to standard entities, respectively, which is formulated as a semantic matching task.

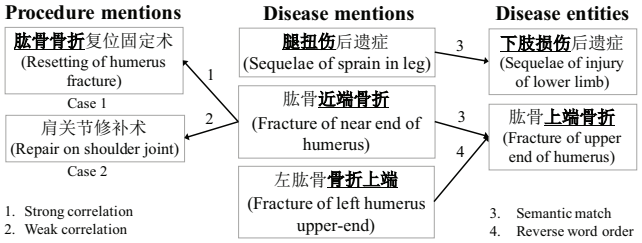

Figure 1: Illustration of three challenges

In general, there exist three major challenges in the two tasks: 1) **Short text normalization**. Unstructured medical texts such as electronic medical records (EMRs) can provide critical context information in entity linking. However, for our problem, Chinese diagnostic statements and procedure names are mention-level short texts whose information is limited. By statistic analysis, there are only 7.8 and 15 characters in average for disease and procedure mentions, respectively. Figure 1 shows some examples. To alleviate this issue, it's necessary to concentrate on how to expand the content of short texts and how to utilize the subtle distinctions between semantic matching patterns. 2) **Non-standard expressions**. Different from English medical text which has been widely researched (Friedman et al. 2004; Aronson et al. 2007), non-standard expressions in Chinese have its own linguistic features which needs to be considered in our task. In Figure 1, two different disease mentions in the second column match to the same disease entity, where reversing word orders occurs such as 'C1a'[1] (Fracture of left humerus upper-end) and 'C1b' (Fracture of upper-

*Corresponding author. Email: gjsong@cis.pku.edu.cn

[1]See Chinese statements in Table 1.

end humerus) with same meaning. Besides, words in Chinese sentences are not separated by delimiters and many words textually different may share similar meanings such as 'C2a' (leg) and 'C2b' (lower limbs) in Figure 1 which adds difficulties to semantic analysis. Hence, how to incorporate Chinese linguistic features to help to evaluate matching similarity is critical. 3) **Correlations between heterogeneous tasks**. Previous works which study encoding diagnosis into standard coding systems (Ning, Yu, and Zhang 2016; Pakhomov, Buntrock, and Chute 2006) fail to pay attention to possible correlations between disease and procedure. We claim that the two heterogeneous tasks possess either strong relation of keyword overlapping such as 'C1a-1' (humerus) of case 1 or weak relation involving the designated mapping such as case 2 in Figure 1. Further, a clear name of one task may facilitate disambiguating the other task. For instance, procedure mention 'C3a' (partial removal of stomach) indicates that the injured body part is stomach, which helps disambiguate diagnostic statement 'C3b' (malignant lymphoma) to 'C3c' (stomach malignant lymphoma). Thus, the correlations between the two tasks can be utilized to better normalize both disease and procedure mentions.

To address the three challenges, we propose a general framework which introduces a tensor generator and a novel multi-view deep architecture with multi-task scheme for disease/procedure concept normalization. First, to tackle the short-text problem, a tensor generator is introduced to expand the short-text comparison to a matching tensor with 4 granularities, including string, character, word, and sentence matchings. String matching provides superficial information; character, word, and sentence matchings provide semantic information, which are like 4 views for one object from different angles. Second, we originally employ the multi-view CNN to extract meaningful matching signals from different angles of the tensor separately, and then aggregate them with view-pooling strategy. Thus, the superficial and semantic information of the matching patterns is integrated sufficiently. Finally, a multi-task learning scheme is introduced to the deep architecture, which utilizes the correlations between disease and procedure to disambiguate and normalize them simultaneously. We propose a multi-task layer on top of the model with shared structures, and train the model for two tasks jointly. The disease and procedure mentions can be better normalized by leveraging the underlying commonality and prior knowledge between them. We conduct extensive experiments, and the results show that our model outperforms existing baselines including both traditional string-matching methods and deep models.

Overall, our main contributions are as follows:

- We propose a multi-task framework in the clinical setting to normalize the disease and procedure mentions jointly.
- We design a tensor generator and a novel multi-view deep architecture, which capture and integrate the meaningful matching signals from different views to solve the short-text normalization and non-standard Chinese expression problems.
- We conduct detailed experimental analysis on comparing our model against single-task/single-view models and existing baselines to validate the superiority of our model.

Related Work

For **normalizing medical concepts**, traditional methods involve dictionary lookup and string matching (Aronson 2005; Kang et al. 2012; Dogan and Lu 2012). However, these approaches cannot tackle mention-entity pairs with different semantic meanings but closer forms. Works that apply machine learning methods to this task are limited due to sparsity of available annotated clinical datasets (Leaman, Khare, and Lu 2015). DNorm (Leaman, Doan, and Lu 2013) is the first to propose pairwise learning to rank method to learn the similarity from mentions and entities in training data. However, this model does not sufficiently take context information into consideration. And it ignores unknown tokens, which may not apply to noisy texts containing many misspellings (Leaman and Lu 2014). In our problem, semantic matching within context could be achieved by interactions in matching tensor and multi-view CNN. The character level matching in our model could handle out-of-vocabulary words since Chinese characters possess its own basic meanings. Recent research introduces semantic matching idea to normalization. CNN and RNN have been applied to model concept representation to classify medical terms in social media texts into concepts in ontology (Limsopatham and Collier 2010). This work uses deep models to learn a simple mapping from social media message to formal entities, but it does not dig the complicated interaction features from multiple semantic levels.

Another area related to our work is **text matching**, which has been researched in many settings, including query document matching (Li and Xu 2014), question answering (Xue, Jeon, and Croft 2008) and paraphrase identification (Dolan, Quirk, and Brockett 2004). Many researchers have recently focused on exploiting deep learning models to capture semantic matching patterns with embeddings in matching texts, either from single representation or multiple granularities, including DeepMatch (Lu and Li 2013), CLSM (Shen et al. 2014), LSTM-RNN (Palangi et al. 2015; Mueller and Thyagarajan 2016), MatchPyramid (Pang et al. 2016), MV-LSTM (Wan et al. 2015). The most similar models to ours are MatchPyramid (MP) and MV-LSTM. MP adopted CNN on matching matrix of words, and convolutions on it layer by layer can extract higher level matching patterns in phrases and sentences. MV-LSTM used positional sentence representation from Bi-LSTM states to capture contextualized local information in semantic matching process. The two models studied matching texts but did not apply to clinical data for normalization purpose and thus did not consider related tasks. Besides, MP relying on hierarchical matching structure fails to capture long distant dependency (Wan et al. 2016). Our model, however, could take long-term memory into consideration by incorporating sentence-matching view in matching tensor. Additionally, both models may not sufficiently tackle short-text problem whereas our model enriches the content and forms a more comprehensive match from different perspectives.

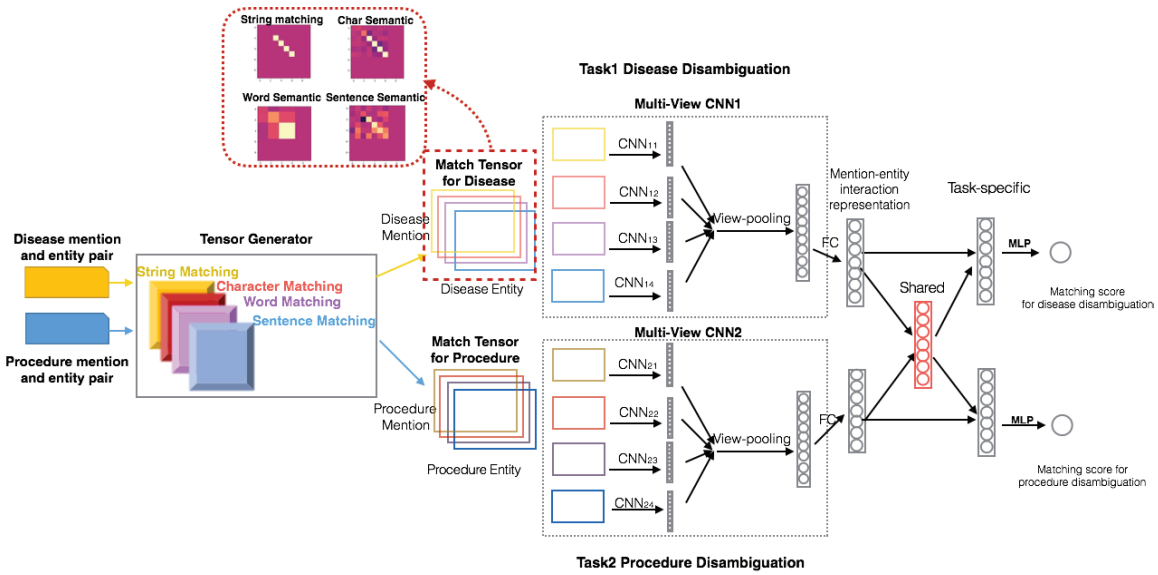

Figure 2: Our proposed architecture for multi-task medical concept normalization. Two tasks have their own matching tensor and multi-view CNN to extract features. Shared structure is used to reinforce the normalization result of both tasks.

ID	Chinese Words	Translation	ID	Chinese Words	Translation
C1a	左肱骨骨折上端	Fracture of left humerus upper-end	C6a	间质肾炎	Interstice nephritis
C1b	肱骨上端骨折	Fracture of upper-end humerus	C6b	间质性肾炎	Interstitial nephritis
C2a	腿	Leg	C7a	慢性乳腺炎	Chronic mastitis
C2b	下肢	Lower limbs	C7b	急性乳腺炎	Acute mastitis
C3a	胃部分切除术	Partial removal of stomach	C7c	乳腺炎	Mastitis
C3b	恶性淋巴瘤	Malignant lymphoma	C1a-1	肱骨	Humerus
C3c	胃恶性淋巴瘤	Stomach malignant lymphoma	C4a-1	扭伤	Sprain
C4a	腿扭伤后遗症	Sequelae of sprain in leg	C4b-1	损伤	Injury
C4b	下肢损伤后遗症	Sequelae of injury of lower limb	C4b-2	伤后遗症	Sequelae with pain
C5a	骨的良性肿瘤	Benign tumor of bone	C5d-1	指骨	Finger
C5b	指骨病损切除术	Excision of lesion of finger	C7a-1	慢性	Chronic
C5c	肋骨良性肿瘤	Benign tumor of rib	C7b-1	急性	Acute
C5d	指骨良性肿瘤	Benign tumor of finger			

Table 1: Chinese words and translations in our paper

Model Formulation

The main idea of this work is based on introducing a tensor generator, and then embedding the multi-view architecture and the multi-task framework to a deep network. Given two mention-entity pairs for disease and procedure, the tensor generator yields two matching tensors separately. Then for each task, interaction representation vector is produced by multi-view CNN. Finally a matching score for normalization is generated in the multi-task module utilizing both shared information and task-specific features. The proposed framework is shown in Figure 2 and all the Chinese words in this paper are translated in Table 1. The details of the model are shown as follows:

- **Matching Tensor**. To tackle short-text problem, for both tasks, a matching tensor is formulated to model interaction between mention-entity pair from both string and semantic aspects in character, word and sentence levels. Particularly, to incorporate context information and solve word-order problem, Bi-LSTM is utilized to integrate sentence level semantics into character vectors.
- **Multi-view CNN model**. We aim to do semantic matching to address non-standard expression problem. CNN is capable of capturing higher level of meaningful matching patterns such as n-grams when convolving across matching matrix (Pang et al. 2016). In our model, four matrices in matching tensor represent different views of matching patterns rather than channels of a picture, where a single CNN can hardly capture all the information sufficiently. Therefore, we adopt multi-view CNN idea to first extract and then effectively aggregate matching signals from four views with a view-pooling strategy.
- **Multi-task learning framework**. Disease and corresponding procedure name for each patient could provide useful information such as body parts to the two related tasks which single task learning may fail to capture. To gain insights from heterogeneous data sources, we design multi-task architecture with constraints to combine the commonalities and differences between medical names in the clinical record.

Matching tensor

The design of matching tensor aims to enrich short-text comparison into string and semantic matching. It resembles human judgement when matching text pairs. Intuitively string matching relying on morphological features is the first to consider. Besides, in Chinese, the meaning of a word is correlated with its composing characters and for unknown token, we may even infer its meaning from the meanings of its

characters (Chen et al. 2015). Internal rich structures within Chinese short texts are useful in matching pairs. Therefore, three semantic levels from basic to contextual are constructed in semantic matching hierarchy, namely character level, word level and sentence level.

Both mention and entity are truncated to maximum length l for the convenience of computation and the dimension of matching tensor is $l \times l \times k$ where k is the number of views. In our problem, we set k to be 4. It is worthy to note that our architecture could be easily extended in other settings, not limited to four views. Matching tensor could include multiple views according to the needs of different tasks, and multi-view CNN scheme could also be adjusted to achieve the objective of learning and aggregating.

Each element in matrix is computed by the similarity of corresponding character pairs of mention and entity respectively. In string level matching, the value is binary reflecting whether character pairs are textually the same. In character level matching, it is the cosine similarity of character embedding vectors. In word level matching, for word level matrix M, we assign the semantic similarity of two words (w in mention m, v in entity n) to the interaction of character pairs constituting those words.

$$M_{pq} = \phi(w) \otimes \phi(v) \quad m_p \in w, n_q \in v \tag{1}$$

where m_p, n_q stand for p-th character in mention m, q-th character in entity n, ϕ is embedding dictionary look-up procedure and $\otimes$ is cosine operation.

In sentence level matching, the goal is to encode context semantics into each local positional vector and resolve word-order problem. Due to the ability to capture long-term memory in two directions and deal with variable-length sequences, Bi-LSTM is employed to do sentence embedding. We pre-train a Siamese Bi-LSTM model and use two Bi-LSTMs to encode mentions and entities into semantic vectors. Sentences are represented as a series of positional vector which is concatenation of hidden states forward and backward for each character. In constructing the matrix, cosine similarity of character vectors from Bi-LSTM states is adopted to evaluate the relevance of character pairs in the context of corresponding sentences.

When producing sentence vectors (see Figure 3), Bi-LSTM takes each sentence (a sequence of character embeddings) as input and updates hidden vectors by rules. Here we design an IDF-weighted strategy to dynamically control the influence of character in sentence modeling. Intuitively, meaningful words such as body parts are expected to be more important in later matching process and thus should be assigned with bigger weights in sentence modeling. The final vector s_i for character m_i is computed as follows:

$$s_i = h_i \times idf_{mi} \tag{2}$$

where h_i stands for the output of Bi-LSTM at i-th timestamp and idf_{mi} is normalized to be in [0,1].

Multi-view CNN

Many works (Hu et al. 2014; Pang et al. 2016) demonstrate the power of CNN in extracting semantic features in interaction matrices of text pairs. Thus CNN is utilized

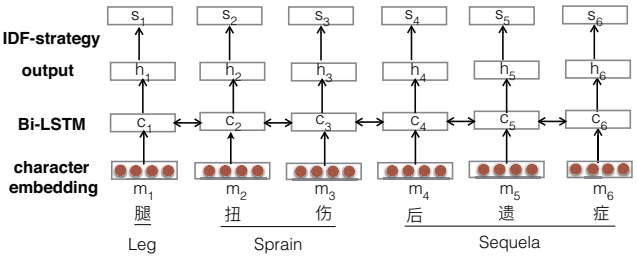

Figure 3: Illustration of sentence modeling

in our model to facilitate semantic matching but traditional CNN can hardly capture matching signals from four views sufficiently. The need to consider the dissimilarity among four different view-spaces motivates us to employ multi-view learning into our problem. We obtain distinct views of matching signals by using four CNNs on interaction matrices. This method functions like multiple-kernel-learning approach, a typical paradigm in multi-view learning. Each CNN performs like a kernel with respect to certain view, and in later stage multiple information sources are integrated by combining outputs of kernels (Xu, Tao, and Xu 2013).

In convolution layers, several filters w convolve over interaction matrix x to generate multiple activation maps and semantic matching patterns within certain windows are captured. Each feature $c_{i,j}$ of activation map is computed by:

$$c_{i,j} = g(\sum_{s=0}^{r-1}\sum_{t=0}^{r-1} w_{s,t} \cdot x_{i+s,j+t} + b) \tag{3}$$

where g is an activation function, r is the size of kernel w and b is a bias. Then a max-pooling layer is used to obtain the most important matching signals. The final fully connected layer reduces the dimensionality of high-level feature vector yielding a fixed-length semantic vector $p_i \in R^d, i = 1, 2, 3, 4$ where d is the dimension of vector.

To achieve a unified representation from perspectives, view pooling layer aggregates four semantic vectors $p = [p_1, p_2, p_3, p_4] \in R^{d \times 4}$ of views by weighting strategy.

$$q = p \times w \tag{4}$$

The weights are updated during training and can be adjusted with different datasets. This allows the model to automatically learn a weight vector $w \in R^4$ of these four views and then the final vector $q \in R^d$ is produced synthesizing those features. In this way, matching patterns from string-matching and character, word and sentence level semantic matching can be captured more sufficiently.

Multi-task learning framework

Multi-task learning (MTL) is capable of learning several tasks simultaneously for mutual benefit. It is typically achieved by common representations of related tasks. Though disease and procedure mentions have different contents, two tasks are highly correlated: 1) both tasks normalize non-standard medical expressions in short texts; 2) both share similar matching patterns; 3) the procedure designed

for a patient is to cure his/her disease in diagnosis, and this relationship may lead to massive shared information, such as identical words in diagnosis and procedure. Therefore, introducing MTL is promising in our architecture.

As shown in Figure 2, the input of our model is heterogeneous and after multi-view CNN extracts matching features for tasks, a shared layer is designed to allow the model to exploit meaningful information from other task to predict the matching score more accurately. Given two semantic vectors R_1, R_2 from multi-view CNN for two tasks, the mutual information is gained with weighting strategy (see Equation 5) yielding a shared representation vector R_s.

$$R_s = \alpha R_1 + (1 - \alpha) R_2 \tag{5}$$

where α is the coefficient regulating the contributions of two source vectors. Then, in task-specific layer, given shared vector R_s and the semantic vector from multi-view CNN, the model uses non-linear transformation to combine the two vectors into a fixed-length task-specific vector R_t. Finally, utilizing information within R_t, a multi-layer perceptron (MLP) generates final matching score, indicating how much the mention and entity are relevant.

In MTL architecture, two tasks are trained together and at each step, all parameters for two tasks are updated. The global joint loss $Loss$ (see Equation 6) consists of the binary cross-entropy loss L for two tasks, a regularization term and a constraint term, which leverages prior knowledge, the probability of co-occurrence of a disease type and related procedure on a patient among all medical records.

$$\begin{aligned} Loss = \frac{1}{n} \sum_{i=1}^{n} (L(f_d(x_i), y_{di}) + L(f_o(x_i), y_{oi})) + \\ \lambda \sum_{i=1}^{n} \delta(e_i^d, e_i^o)(f_d(x_i) - f_o(x_i))^2 + \beta ||\xi||^2 \end{aligned} \tag{6}$$

where $f_d(x_i), f_o(x_i)$ are softmax output for disease and procedure respectively, y_{di}, y_{oi} are labels for two tasks, e_i^d, e_i^o are disease entity and procedure entity for normalizing record x_i, δ denotes co-occurrence probability of disease and procedure entities, λ is regularization parameter, β is the parameter controlling sparsity and $||\xi||$ denotes the L2-norm of all weights in fully-connected layers.

The assumption is that if disease and procedure entities co-occur in high probability, in both normalization tasks, it is more likely that both (or neither) mention-entity pairs for the disease and the procedure are normalized correctly (or incorrectly). This leads to similar matching scores of the disease and the procedure, i.e. the outputs of two tasks. Otherwise, if the disease entity e_i^d and procedure entity e_i^o rarely co-occur in a record which indicates low $\delta(e_i^d, e_i^o)$, dissimilar matching scores are expected.

Experiment

The dataset used in this paper comes from Information Center of a cooperative Chinese hospital including discharge summaries from 153 triple-A hospitals in 31 provinces and regions. It has 3125 types of disease and 3154 of procedures.

Task	Conv1	Pool1	Conv2	Pool2
Disease	4	3	3	2
Procedure	3	2	3	3

Table 2: Kernel sizes of CNN for both tasks

Each record contains a primary diagnosis and main procedure for a patient. Annotations are performed by medical experts producing initial dataset of 7000 entries. Each diagnosis/procedure text is paired with several medical concepts from ICD-10 and annotators mark the normalization correct or incorrect. Thus, each entry consists of two mention-entity pairs for two tasks and two labels.

Due to limited annotated data, we adopt data augmentation techniques to enlarge the dataset. The best way to augment text data is to paraphrase sentences, but is usually unrealistic and labor-intensive due to large volume of samples. It is appropriate to replace words with their synonyms in text augmentation (Zhang, Zhao, and LeCun 2015). Here we use this method with synonyms from medical expert dictionary and Synonym Word Forest of HIT-SCIR[2]. Moreover, as mentioned above, reversing order of certain words does not interfere with the expression of written diagnosis. Consequently, replacing words with their synonyms and reversing words orders are utilized in our text augmentation. To validate the reliability of generated entries, the final dataset with 58031 records for training and 6899 for testing is carefully examined by two medical undergraduates.

In the experiments, we implement four baselines:

- Edit-distance: a basic method focusing on the number of procedures transforming one string to another.
- BM25 (Robertson et al. 1995): a popular baseline for information retrieval.
- MatchPyramid (Pang et al. 2016): applies CNN on matching tensor to capture matching patterns in a hierarchy.
- MV-LSTM (Wan et al. 2015):uses Bi-LSTM to get positional sentence representations forming interaction matrix and adopts k-max pooling and multi-layer perceptron to get similarity score.

Experimental Settings

We trained word2vec model (Mikolov et al. 2013) on over 10 million Chinese clinical narrative corpora with word and character vector dimension of 100. For tensor size l, we set it to be 10, 20 for disease and procedure respectively by experimental statistics. For sentence modeling we conduct a Siamese Bi-LSTM model with 15-dimensional hidden vectors h_t and memory cells c_t. The kernel sizes for CNN model are summed up in Table 2 and 8, 16 feature maps are produced in two convolutional layers. We choose rectifier linear unit (ReLU) (Nair and Hinton 2010) as activation functions and apply droupout (Srivastava et al. 2014) strategy.

For training, we use stochastic gradient descent (SGD) Adam (Kingma and Ba 2014) method with shuffled mini-batches of size 128 and adopt early-stopping strategy. The

[2]http://www.ltp-cloud.com/download/

Data	Method	Accu	F1
Disease	Edit-distance	60.00	48.11
	BM25	59.00	45.36
	MP	80.16	83.89
	MV-LSTM	80.34	83.76
	MTMV-CNN	**85.59**	**87.96**
Procedure	Edit-distance	79.00	75.76
	BM25	75.00	79.47
	MP	83.51	87.38
	MV-LSTM	85.78	88.40
	MTMV-CNN	**91.69**	**93.09**

Table 3: Performance comparison on normalization tasks

learning rate is 0.001 and all trainable parameters are initialized randomly with truncated normalization. We use coefficient α of 0.5. In joint loss, the regularization term λ and β controlling sparsity are set to be 0.1 and 0.001 respectively.

For experimental design, we compare our multi-task multi-view CNN model, namely MTMV-CNN, with four baselines. BM25 and Edit-distance are chosen to represent classic methods, and MatchPyramid (MP) and MV-LSTM are popular methods in semantic matching. Besides, to validate the effect of our model, single-task multi-view (STMV) CNN is used to prove the usefulness of multi-task learning. To discover impacts of various views, multi-task single-view (MTSV) models and MTMV-CNN combining certain views are also implemented as contrast experiments. In our study, accuracy and F-score are chosen as evaluation metrics.

Experimental Results

We compare the performance of our proposed model with other baselines. Several findings can be obtained from the experiment results listed in Table 3.

1) Traditional methods perform worse than deep models due to its incapability of tackling semantic matching cases. Specifically, edit-distance performs worst since it relies on string form too much and can not deal with word order problem. BM25, based on word-bag model, capable of tackling order problem has relatively better results.

2) For baselines using deep models, MP and MV-LSTM have similar results in matching disease pairs but MV-LSTM outperforms MP for procedure by near 2%. It may because that procedure names usually possess more characters with more context information and MV-LSTM can capture this contextualized local information (Wan et al. 2015) to better facilitate the task.

3) Our MTMV-CNN model provides the best accuracy and F1 score among existing baselines. Specifically, our model achieves about 5% and 8% improvement in accuracy for disease and operation respectively over MP. This suggests the power of matching tensor over matching matrix in obtaining rich information and the superiority of multi-view CNN over single CNN in MP. And, in sentence-level matching, contextual semantics could be encoded into characters of the mention and overcome the word-order problem which often occurs in our dataset. Moreover, our model obtains improvements over MV-LSTM. Part of the reason lies in the superiority of matching tensor over interaction matrix. Additionally, more complete and complicated matching signals are extracted by multi-view CNN than k-max pooling strategy in MV-LSTM.

Data	Method	View	Accu	F1
Disease	STMV	All	83.61	86.13
	MTSV	String	81.08	84.95
		Char	82.99	86.00
		Word	75.76	80.35
		Sentence	82.83	85.61
	MTMV	Str+Char	84.20	86.82
		Str+Char+Word	85.00	87.57
		All	**85.59**	**87.96**
Procedure	STMV	All	90.75	92.25
	MTSV	String	87.11	89.82
		Char	87.63	90.20
		Word	83.11	86.76
		Sentence	90.17	91.89
	MTMV	Str+Char	87.30	89.97
		Str+Char+Word	87.83	90.10
		All	**91.69**	**93.09**

Table 4: Performances of model variations

Model Analysis

MTL effect We examine the results of single-task multi-view (STMV) and MTMV model with all views in Table 4. It demonstrates that our MTMV model reaches 85.59% in accuracy with an improvement over STMV by about 2%, 1 % for disease and procedure. This validates our assumption that massive shared information of diseases and procedures can be leveraged in the disambiguation. We also note that MTL improves disease normalization task more. This may because disease mentions are more non-standard with more term variations than the counterpart and the relatively clear procedure mention can provide useful information through shared structure in our model.

Multi-view effect To verify the effect of incorporating multiple views, we compare single view model with multi-task (MTSV) model and MTMV in Table 4. MTSV with character view achieves 82.99% for disease better than MTSV with other views, whereas for procedure MTSV-Sentence reaches the best. This suggests that the character and sentence views are essential for matching Chinese text pairs of disease and procedure respectively. Additionally, for both tasks, word view perform worst and it is likely because the non-standard and short-text attributes lead to frequent occurrence of incorrect word segmentation causing inaccurate embeddings. Furthermore, when combining certain views, integration of more views results in better performance in both tasks, which validates the strength of matching tensor and multi-view CNN. For procedure, however, the enhancement with more views is limited in string, character and word levels but it is significant when adding sentence view. It may because procedure normalization already performs relatively well in the three levels but sentence level information with IDF-weighted strategy enables the model to notice certain important words.

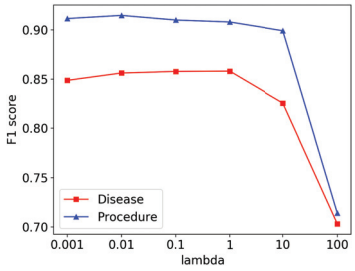

Figure 4: Sensitivity analysis on constraint parameter λ

Parameter Sensitivity Analysis We investigate the influence of constraint parameter λ in the MTMV-CNN model when varying λ from 0.001 to 100 in Figure 4. Procedure normalization remains relatively stable with the variation of λ between 0.001 to 10 whereas for disease F1 scores fluctuate more. We may infer that because disease names contain more complicated cases, the constraint term in MTL has more influence on it than the counterpart. Besides, the model achieves best for disease when λ is close to 1, indicating prominent results at the point.

Cases Analysis We exemplify the results of our model and baselines. All methods can deal with lexically similar cases such as mention-entity pair 'C6a' (interstice nephritis) and 'C6b' (interstitial nephritis). However, for complicated cases such as 'C7a' (chronic mastitis), traditional methods cannot accurately match to entity 'C7c' (mastitis). Instead, string-based approaches will output 'C7b' (acute mastitis) which has higher string similarity. Our model can tackle this problem from semantic matching since 'C7a-1' (chronic) and 'C7b-1' (acute) are opposite in semantics.

MTL model outperforming STL model proves that shared structure provides more information useful in normalization. For instance, for a record with disease 'C5a' (benign tumor of bone) and procedure 'C5b' (excision of lesion of finger), other baselines tend to map disease mention to 'C5c' (benign tumor of rib) due to similar forms and semantic meaning. But it is supposed to be mapped to 'C5d' (benign tumor of finger) since it contains key word 'C5d-1' (finger) as the procedure mention does. Our MTL model is capable of exploiting this information to do the normalization.

Visualization

We visualize the matching matrices of four views, the representative kernels, and related feature maps of the first convolutional layer in Figure 5. The input is the pair of disease mention 'C4a' (sequelae of sprain in leg) and entity 'C4b' (sequelae of injury of lower limb). In this case, 'C4a-1' (sprain) and 'C4b-1' (injury) are semantic related and the two names both contain 'C4b-2' (sequelae with pain). From three views of semantic matching, we can observe that the white square areas indicate strong matching signals between 'C4a-1' (sprain) and 'C4b-1' (injury). Additionally, in four views, the larger blue square areas reflect raw matching signals between two same segments 'C4b-2' (sequelae with pain) in the brightest color.

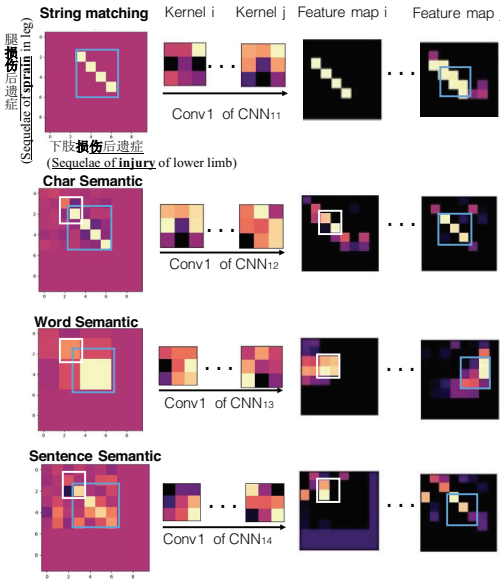

Figure 5: Visualization of matching matrices of four views, kernels, and feature maps

Convolution kernels are served as feature extractors to detect the matching signals along the comparison map. Here for each CNN of multi-view CNN, two representative kernels in the first convolutional (Conv 1) layer are shown modeling matching similarity with different evaluation strategies. The semantic relationship between 'C4a-1' (sprain) and 'C4b-1' (injury) is captured by different kernels for each view producing activation maps with emphasis on different parts. In this way, the rich and comprehensive matching signals of various views can be utilized in matching tasks.

Conclusion

In this paper, we present a general multi-task framework for medical concept normalization. It overcomes three challenges in our setting and can be extended to incorporate multiple views in different settings. Our model designs a matching tensor to enrich short-text comparisons, and incorporates multi-view CNN to well capture and synthesize semantic matching signals from various views. Multi-task learning structure is employed to incorporate shared information within heterogeneous data sources. Experimental results demonstrate the superiority of our model over baselines and verifies the rationality of using multi-task learning and multi-view CNN. Normalizing medical concepts are meaningful and fundamental in many clinical applications, such as information sharing between medical institutions and policy making. Thus we hope that our work could provide insights in relative domains for researchers. In future, we hope to discover more correlations between homogeneous tasks, namely primary and associate diagnostic statements/procedures in normalization, which may shed light on hidden links between diseases/procedures.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (61572041), and the Beijing Natural Science Foundation (4152023).

References

Alhelbawy, A., and Gaizauskas, R. 2014. Graph ranking for collective named entity disambiguation. In *Meeting of the Association for Computational Linguistics*, 75–80.

Aronson, A. R.; Bodenreider, O.; Demner-Fushman, D.; Fung, K. W.; Lee, V. K.; Mork, J. G.; Névéol, A.; Peters, L.; and Rogers, W. J. 2007. From indexing the biomedical literature to coding clinical text: experience with mti and machine learning approaches. In *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, 105–112. Association for Computational Linguistics.

Aronson, A. R. 2005. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In *Proceedings of the AMIA Symposium*, 1721.

Bunescu, R. C., and Pasca, M. 2006. Using encyclopedic knowledge for named entity disambiguation. In *Eacl 2006, Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference, April 3-7, 2006, Trento, Italy*, 9–16.

Chen, X.; Xu, L.; Liu, Z.; Sun, M.; and Luan, H.-B. 2015. Joint learning of character and word embeddings. In *IJCAI*, 1236–1242.

Dogan, R., and Lu, Z. 2012. An inference method for disease name normalization. *Aaai Fall Symposium*.

Dolan, B.; Quirk, C.; and Brockett, C. 2004. Unsupervised construction of large paraphrase corpora: exploiting massively parallel news sources. In *International Conference on Computational Linguistics*, 350.

Friedman, C.; Shagina, L.; Lussier, Y.; and Hripcsak, G. 2004. Automated encoding of clinical documents based on natural language processing. *Journal of the American Medical Informatics Association* 11(5):392–402.

Hoelzer, S.; Schweiger, R. K.; and Dudeck, J. 2003. Transparent icd and drg coding using information technology: linking and associating information sources with the extensible markup language. *Journal of the American Medical Informatics Association* 10(5):463–469.

Hu, B.; Lu, Z.; Li, H.; and Chen, Q. 2014. Convolutional neural network architectures for matching natural language sentences. In *Advances in neural information processing systems*, 2042–2050.

Kang, N.; Singh, B.; Afzal, Z.; van Mulligen, E. M.; and Kors, J. A. 2012. Using rule-based natural language processing to improve disease normalization in biomedical text. *Journal of the American Medical Informatics Association* 20(5):876–881.

Kingma, D. P., and Ba, J. 2014. Adam: A method for stochastic optimization. *Computer Science*.

Leaman, R., and Lu, Z. 2014. Automated disease normalization with low rank approximations. *ACL 2014* 593(5145):24.

Leaman, R.; Doan, R. I.; and Lu, Z. 2013. Dnorm: disease name normalization with pairwise learning to rank. *Bioinformatics* 29(22):2909.

Leaman, R.; Khare, R.; and Lu, Z. 2015. Challenges in clinical natural language processing for automated disorder normalization. *Journal of biomedical informatics* 57:28–37.

Li, H., and Xu, J. 2014. Semantic matching in search. *Foundations & Trends in Information Retrieval* 7(5):343–469.

Limsopatham, N., and Collier, N. 2010. Normalising medical concepts in social media texts by learning semantic representation. In *Meeting of the Association for Computational Linguistics*, 1014–1023.

Lu, Z., and Li, H. 2013. A deep architecture for matching short texts. In *International Conference on Neural Information Processing Systems*, 1367–1375.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, 3111–3119.

Mueller, J., and Thyagarajan, A. 2016. Siamese recurrent architectures for learning sentence similarity. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2786–2792.

Nair, V., and Hinton, G. E. 2010. Rectified linear units improve restricted boltzmann machines. In *International Conference on International Conference on Machine Learning*, 807–814.

Ning, W.; Yu, M.; and Zhang, R. 2016. A hierarchical method to automatically encode chinese diagnoses through semantic similarity estimation. *BMC medical informatics and decision making* 16(1):30.

Pakhomov, S. V. S.; Buntrock, J. D.; and Chute, C. G. 2006. Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques. *Journal of the American Medical Informatics Association* 13(5):516–525.

Palangi, H.; Deng, L.; Shen, Y.; Gao, J.; He, X.; Chen, J.; Song, X.; and Ward, R. 2015. Deep sentence embedding using long short-term memory networks: Analysis and application to information retrieval. *IEEE/ACM Transactions on Audio Speech & Language Processing* 24(4):694–707.

Pang, L.; Lan, Y.; Guo, J.; Xu, J.; Wan, S.; and Cheng, X. 2016. Text matching as image recognition. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Robertson, S. E.; Walker, S.; Jones, S.; Hancock-Beaulieu, M. M.; Gatford, M.; et al. 1995. Okapi at trec-3. *Nist Special Publication Sp* 109:109.

Shen, Y.; He, X.; Gao, J.; Deng, L.; Mesnil, G.; and goire. 2014. A latent semantic model with convolutional-pooling structure for information retrieval. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, 101–110.

Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. 2014. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research* 15(1):1929–1958.

Wan, S.; Lan, Y.; Guo, J.; Xu, J.; Pang, L.; and Cheng, X. 2015. A deep architecture for semantic matching with multiple positional sentence representations. *Computer Science*.

Wan, S.; Lan, Y.; Xu, J.; Guo, J.; Pang, L.; and Cheng, X. 2016. Match-srnn: Modeling the recursive matching structure with spatial rnn. *arXiv preprint arXiv:1604.04378*.

Xu, C.; Tao, D.; and Xu, C. 2013. A survey on multi-view learning. *arXiv preprint arXiv:1304.5634*.

Xue, X.; Jeon, J.; and Croft, W. B. 2008. Retrieval models for question and answer archives. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, 475–482. ACM.

Zhang, X.; Zhao, J.; and LeCun, Y. 2015. Character-level convolutional networks for text classification. In *Advances in neural information processing systems*, 649–657.