A NLP system of DCUMT in NTCIR-11 MedNLP-2: RNN for ICD/Time Entity Recognition and ICD Classification Tasks

Tsuyoshi Okita
Dublin City University
tokita@computing.dcu.ie

Qun Liu
Dublin City University
qliu@computing.dcu.ie

ABSTRACT

This paper describes the medical NLP system developed at Dublin City University for participation in the Second Medical NLP Shared Task (MedNLP 2) in NTCIR-11 [1]. This shared task is a Japanese task. Our system detects International Classification of Diseases (ICD) and time entities and classifies ICD entities. We participated in the task 1 which detects the ICD and time entities, and the task 2 which classifies the detected ICD entities among the ICD codes. Our system uses deep learning to learn and classify those entities. Our result was F1 score of 67.8 for the ICD entity recognition task (task 1), 77.4 for the time entity recognition task (task 1), and 54.0 for the ICD classification task (task 2 for gold standard).

Team Name

DCUMT

Subtasks

Task 1 and Task 2.

Keywords

named-entity recognition, recurrent neural network, sequential classification

1. INTRODUCTION

This paper describes the medical NLP system developed at Dublin City University for participation in the Second Medical NLP Shared Task (MedNLP 2) in NTCIR-11 [1]. This system handles International Classification of Diseases (ICD) and time entity recognition tasks in Japanese. We participated in task 1 which detects the ICD/time entities, and task 2 which classifies the detected ICD entities among the ICD codes.

The remainder of this paper is organized as follows. Section 2 describes the overview of our systems. Our experimental results are presented in Section3. We conclude in Section 4.

2. OUR METHODS

Tag Scheme.

The tag scheme we used is the UO-BILOU scheme that we propose. Ratinov [8] discussed that the BILOU scheme is superior to the BIO scheme.

- the BILOU scheme: the BILOU scheme suggests to learn classifiers that identify (1) the Beginning, (2) the Inside and (3) the Last tokens of multi-token chunks, (4) the Outside of the text segments, and (5) the inside of Unit-length chunks [8].
- the BIO scheme suggests to learn classifiers that identify (1) the Beginning, (2) the Inside and (3) the Outside of the text segments.

These schemes are often used for a single classifier system such as conditional random field (CRF). As is mention in the next paragraph our system is an interactive two classifier system, it would need slight extension of this to deploy the decision in two stages. Our classifier makes two decisions: the first one is to decide whether the given sequence of token is terminology or not, the second one is to decide whether the given one unit of token is terminology or not. Noted that the input sequence of the first and the second classifiers are different which are decided interactively. In sum, the UO-BILOU scheme that we propose can be written in the following way.

- the UO-BILOU scheme: the UO-BILOU scheme suggests that the first classifiers learns the UO tags and the second one learns the BILOU tags where the classifiers are invoked according as the sequence. The first classifier handles UO tags where the ICD entity is a unit-length chunks where (1) the inside of Unit-length chunks, and (2) the Outside of unit-length chunks. The second classifier handles BILOU tags where the classifier identifies (1) the Beginning, (2) the Inside of multi-token chunks, (3) the Last tokens of multi-token chunks, (4) the Outside of the text segments, and (5) the inside of Unit-length chunks.

Algorithm.

The ICD/time entity recognizer and the ICD classifier are shown in Figure 1. Although ICD/time entity recognizer and ICD classifier commonly use Recurrent Neural Networks (RNNs), the architectures of these are slightly different. In both figures the lower rectangles include the subsystems that learns the ICD/time entities.

For the ICD/time entity recognition task, we deploy an interactive two classifier system. Two classifiers make distinctive decisions: the first one is to decide whether the given sequence of tokens is an ICD/time entity or not, and the second one is to decide whether the given unit of token is within or outside of the ICD entities.

The algorithm goes as follows.

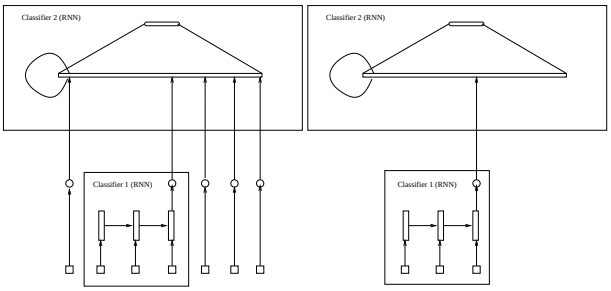

Figure 1: The left figure shows the architectures of ICD/time entity recognizer while the right figure shows that of ICD classifier.

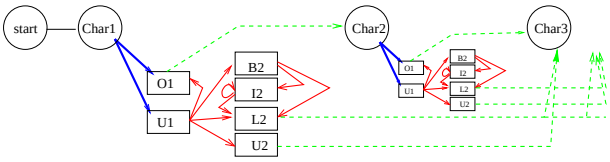

Figure 2: Figure shows an interactive two classifier system. For the given input of sequence (character-based or word-based), the first classifier is to decide whether the given sequence of token is ICD/time entities or not (O1 and U1 correspond to this), and the second one is to decide whether the given one unit of token is within or outside of the ICD entities (B2, I2, L2, O2 (=O1), and U2 correspond to this).

1. Upon the arrival of (first) character, the first classifier makes a binary decision whether this is inside/outside of an ICD/time entity.
2. If it is inside of an entity, the second classifier is invoked to determine the boundary whose next character is outside of ICD/time entities when ICD/time entity contains a long characters. If it is outside, go to Step 1.
3. If the second classifier detects the end of an ICD/time entities and if the sequence still continues, it replies the next character which is outside of an ICD/time entries and go to Step 1, otherwise it finishes.

We separate two classifiers since such separation may capture two distinctive difficulties which are common to ICD and time entities [6]. First, they have rich compositional structures which can be easily confused with non-ICD/temporal phrases (e.g. the word "May" can be a month name or a verb). Other example is a long ICD entity such as B型肝炎ウイルスキャリア, which can be considered as the combination of various words (Noted that it may not appropriate to use the terminology *compositional* here though). It is recently recognized that recurrent/recursive neural network can efficiently capture such compositional structures [7, 3].

Second, ICD and time entities can carry different meaning in different linguistic contexts (e.g. the word "Friday" refers to different dates: "We met on Friday" and "We will meet on Friday"). This difficulty is natively different from whether the classifier is good at handling compositional structures or not.

Due to these two reasons, we do not employ single classifier system such as CRF-based system with the BILOU tag, but to separate a classifier into two where one classifier is good at handling rich compositional structures and the other one is good at handling the binary decision whether the token is ICD/time entity or not.

2.1 Sub-systems

Compositional Sub-system for ICD Entities.

The second classifier is trained with the ICD/time entity with a preceding/succeeding character (or word) if these exist. The ICD/time entities which are provided in byomei master do not contain the information about a preceding/succeeding character. It is noted that one medical disease has often multiple corresponding ICD entities as is shown in Figure 6. It is also noted that we do not have a big corpus other than a small train set provided by MedNLP2 organizers.

We use a recurrent neural network (RNN) [4, 7], more specifically RNN encoder-decoder like setup (Two figures in the lower rectangles in Figure 1). RNN encodes a variable-length sequence into a fixed-length vector representation. The RNN reads each symbol of an input sequence x sequentially. Let a variable-length sequence $x(= (x_1, \ldots, x_T))$, f be a non-linear activation function, and $h_{(t)}$ be a hidden state of the RNN. At each time step t the hidden state $h_{(t)}$ of the RNN is updated by (1):

$$h_{(t)} = f(h_{(t-1)}, x_t) \quad (1)$$

Suppose we use the 1-of-K coding (or the hot representation) and a softmax activation function, the output for all possible symbols $j(= 1, \ldots, K)$ at each time step t, which is $p(x_t|x_{t-1}, \ldots, x_1)$, can be written as in (2)

$$p(x_{t,j}|x_{t-1}, \ldots, x_1) = \frac{\exp(w_j h_{(t)})}{\sum_{j'=1}^{K} \exp(w_{j'} h_{(t)})} \quad (2)$$

The probability of the sequence x can be written as in (3):

$$p(x) = \prod_{t=1}^{T} p(x_t|x_{t-1}, \ldots, x_1) \quad (3)$$

Upon the reading of the end of the sequence the hidden state of the RNN can be considered as the updated summary c, i.e. updated word embedding, of the sequence x. Using stochastic gradient-descent, we maximize the negative log-likelihood.

Compositional Sub-system for Time Entities.

We used the same subsystem for ICD entities. Figure 1 shows example of time entities in training set.

2.2 Overall Systems

Upon the new replaced ICD/time entity with the word embedding, this classifier makes the binary decision based on such word embedding. Noted that since we have an annotated training set both of these classifiers can be trained in supervised manner.

We use the deep algorithm here as well. Our motivation lies in the fact that one characteristic of the deep learning is in the distributed representations where non-mutually exclusive features/attributes create a combinatorially large set of

ここ１年間, その夜から, 今回入院半年前から, 今朝から, 入院同日から, 入院後, 入院時, 来院時, 救急外来受診時, 術後経過, 退院後, 午前９時０５分, 同年６月１２日, 同日, 平成１２年, 当日中, 昨夕, 昨年, 昨日, 昨晩, 昭和６２年, 最近１年, 現在, 第３病日より, 約１時間５５分後, 約３ヵ月前から, 経過約１０年, 翌年１月４日, １カ月前, ７月終旬, ２月初めから, １０月２９日夕刻まで, ２月１４日夕, ６月２日～７日, １０日後より, １１／２２, １９９６年, ２００８年2月２１日, ２００８／１１／１０, ２０２７年8月6日夕食時, ２０２８年１０月１６日AM７時１０分頃, ２：３０, ４０ｙｅａｒ, 5日間連続, ６－１０　週, Ｄａｙ１５, Ｄａｙ３－５, Ｈ２８年頃～

Table 1: Figure shows a variety of time entities taken from the training set.

distinguishable configurations [2]. Learning a set of features that are not mutually exclusive can be exponentially more statistically efficient having nearest-neighbor-like models.

The left figure in Figure 1 shows the ICD/Time entity detection system while the right figure shows the ICD classification system. The input of the former system is a sequence while that of the latter system is only the ICD entity. The supervised sequence labeling task by RNN relates to the basic characteristics of RNN (Two figures in the upper rectangles in Figure 1).

Although it is not required to embed two subsystems into the overall classifier at the same time in this context we can embed ICD/time entity classifier (the second classifier) into the overall classifier (the first classifier) by the Fisher information matrix $I_M = \mathbf{E}[g(\theta^0, x), g(\theta^0, x)']$ [5] to embed two sub-systems of recurrent neural networks.

Normalization.

We apply the following normalization.

- The lexical features are normalized in terms of dates and numbers. For example, 1980 becomes $*DDDD*$ and $212-325-4751$ becomes $*DDD*-*DDD*-*DDDD*$.

3. EXPERIMENTAL RESULTS

Corpus used for the experiment is described in Table 2. Table 3 shows the experimental results for tasks 1 and 2. The depth of RNN in the subsystems was variable according as the length of the ICD entities while the depth of RNN in the overall system was set to 5.

type	test	training	ours
ICD	763(1329)	1099(2075)	658(1554)
temporal	245(300)	442(578)	241(325)

Table 2: Statistics from the gold standard data.

4. DISCUSSION

4.1 Character-Based Analysis

Intuitively, the cluster of Kanji characters looks like an ICD entity. According to Table 4 which shows the constituency of ICD entities, around 71 percents are indeed Kanji only compositions.

task 1			
NE only(Accuracy)	92.10		
NE only	76.78	60.72	67.81
NE+positive	71.10	60.36	65.29
NE+family	91.89	80.95	86.08
NE+ Negation	78.69	39.46	52.56
NE+Suspicion	51.52	30.91	38.64
Time	72.41	83.20	77.43
task 2			
NE+ICD	N/A		
GoldStandard+ICD	54.00		

Table 3: Result of detection of ICD/Temporal expressions (task1) and ICD classification (task2).

71.2%	Kanji
9.7 %	Kanji-katakana
7.4%	Kanji-katakana-hiragana
5.7 %	Kanji-hiragana
3.4 %	Kanji-figure
0.9 %	Kanji-katakana-figure
0.7 %	Kanji-katakana-hiragana-figure
0.4 %	Kanji-hiragana-figure
0.3 %	katakana-hiragana
0.3 %	katakana
0.1 %	hiragana

Table 4: Table shows examples of compositions of BCD expressions.

We can observe several characteristics from Table 5. First, despite that the whole characters of ICD entities may include 71.2% of Kanji characters, the Kanji characters are only 25% of whole characters in medical descriptions. Second, from the third and the fourth rows, 85.5% of the last chars of ICD entities are Kanji and 61.8% of the next chars are hiragana. Third, similarly from the fifth and the sixth rows, 48.0% of the previous words of ICD entities are hiragana and 87.8% of the top chars of ICD entities are Kanji characters.

4.2 Word-Based Analysis

For example, an ICD expression such as B型肝炎ウイルスキャリア has rich compositional structures which consists of B型, 肝炎, ウイルス, and キャリア. Each of these words have distinctive meaning of representation. One interesting observation is that, if we consider 下行結腸癌１型, which can be decomposed into 下行, 結腸, 癌, and １型, we noticed that we may cluster some similar meaning representation

char type	ICD bagOf Char	last char of ICD	ICD→ next word	prev word →ICD	top char of ICD
Kanji	25%	85.5%	17.5%	33.4%	87.8%
alphabet	12%	8.3%	13.3%	16.3%	5.4%
katakana	3%	3.8%	0.4%	0.1%	2.6%
hiragana	5%	1.5%	61.8%	48.0%	2.5%
symbol	39%	0.4%	4.4%	2.0%	1.2%
figure	14%	0.3%	2.5%	4%	0.4%

Table 5: Table shows examples of compositions of ICD expressions.

such as B型 and 1型 (since these refer the types), and 肝炎 and 癌 (since these refer to the name of disease). It is worthy to consider tentative categorization of meaning representation of ICD expressions (Noted these are informal tentative categorization by human beings): (1) body parts (e.g. 消化管, 仙骨部, 蜂巣, 抹消, 眼瞼結膜, 右目, 胃, 左室, 心筋, 僧帽弁, 全身, 脳, 左目眼底), (2) name of the disease (e.g. 肝炎, 肺炎, 梗塞, 潰瘍, 褥瘡, ヘルニア, ポリープ), (3) diagnosis (e.g. 便, 血餅, 胸水, 喀痰, 嘔吐, 出血, 冷汗, 貧血, 斜視, 失調, 幻覚, 妄想, 嘔気), (4) name (e.g. 残遺型, V3～V6, B型,I度), (5) upper/lower/both sides (e.g. 上部, 両側, 統合), (6) increase/high (e.g. 低下, 付着, 慢性, 低値, 発, 高, 減少, 巨大, 陰影, 低, 肥大, 高値, 異常, 急性, 逆流, 不振, 倦怠, めまい, 浸潤), (7) biological name (e.g. 炎症細胞, キャリア, 細菌, ウイルス, 菌), (8) color(黒色, 赤), (9) measurement (血糖, 体重, 呼吸, CPK, Q波, 心電図, 心音), (10) substance such as tobacco and alcohol (酒, タバコ, アルコール, ツルゴール, タール, 石灰), (11) postposition (e.g. 症, 様, 化, 傾向, 病, 性, 感, 苦, 食欲, 炎, 覚, 自覚, 不整), and (12) others (e.g. atypical epithelium GroupIII (adenoma), AV block).

It is easily predicted that this word-based method using meaning representation in which we can judge whether the words coming from some clusters may have better performance than the character-based method.

A162	肺結核, consoLidation
B181	B 型慢性肝炎, B 型慢性肝疾患, B 型肝硬変
B24_	HIV 感染, HIV 感染症 (AIDS), HIV 感染 (AIDS), HIV 感染症,AIDS
B59_	ニューモシスチス肺炎, ニューモシスチス PCR, PCP
C170	十二指腸癌, 腫瘍, GroupV (moderately-poorly diff. adenocarcinoma of the drodenum), GroupV, moderately-poorly diff. adenocarcinoma of the drodenum
C186	下行結腸癌 1 型 (腺癌), 下行結腸癌 1 型, 腺癌
C189	大腸癌, 腫瘍, 高分化管状腺癌 trb1, adenocarcinoma
C220	単発癌, 肝細胞癌, 腫瘍
C342	腫瘍, 腫瘍内部, 腫瘍性病変
C349	原発, 原発巣, 原発性肺癌 (扁平上皮癌), 原発性肺癌 (腺癌), 未治療 IIIB/IV 期非小細胞肺癌, 肺腺癌, 非小細胞癌, 非小細胞肺癌, non-small cell carcinoma, 原発性肺癌, 扁平上皮癌, 腺癌
R91_	びまんせいスリガラス陰影, スリガラス, スリガラス影, 不正形腫瘤, 両側スリガラス陰影, 右中肺異常陰影, 小結節, 棍棒状陰影, 気管支壁肥厚, 浸潤影, 異常影, 異常陰影, 石灰化結節, 石灰化陰影, 粒状影, 粒状陰影, 索状影, 結節, 結節影, 網状影, 網状陰影, 胸膜嵌入像, 胸部レントゲン上異常, 胸部異常影, 胸部異常陰影, 胸部 X 線陰影, 軽度網状影, 過膨張, 間質影, 陰影, honey comb, spicula, spiculation
R943	不整, 右室負荷所見, 心病変, 心音不整, P 波, f 波認めなず, V3～V6
Z720	タバコ, 喫煙, 喫煙歴
Z721	アルコール, 大量飲酒, 常習飲酒家, 機会飲酒, 毎日飲酒, 飲酒, 飲酒歴

Table 6: Table shows example of ICD names in training corpus together with the ICD codes. One difficulty of this problem is that a varieties of ICD names correspond to a single disease.

5. CONCLUSION

This paper describes the medical NLP system developed at Dublin City University for participation in the Second Medical NLP Shared Task (MedNLP 2) in NTCIR-11 [1]. This shared task is a Japanese task. Our system detects International Classification of Diseases (ICD) and temporal entities and classifies ICD entities. We participated in the task 1 which detects the ICD and temporal entities, and the task 2 which classifies the detected ICD entities among the ICD codes. Our system uses deep learning to learn and classify those entities. Our result was F1 score of 67.8 for the ICD entity recognition task (task 1), 77.4 for the time entity recognition task (task 1), and 54.0 for the ICD classification task (task 2 for gold standard).

6. ACKNOWLEDGMENTS

This research is supported by the Science Foundation Ireland (Grant 12/CE/I2267) as part of the CNGL at Dublin City University. We would also like to thank the Irish Centre for High-End Computing.

7. REFERENCES

[1] E. Aramaki, M. Morita, Y. Kano, and T. Ohkuma. Overview of the ntcir-11 mednlp-2 task. *NTCIR*, 2014.

[2] Y. Bengio, I. Goodfellow, and A. Courville. Deep learning. *MIT Press (preparation version 22/10/2014)*, 2014.

[3] K. Cho, B. van Merrienboer, C. Gulcehre, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *In Conference on Empirical Methods in Natural Language Processing (EMNLP) 2014*, 2014.

[4] A. Graves. Supervised sequence labelling with recurrent neural networks. *Studies in Computational Intelligence*, 2012.

[5] T. Jaakkola and D. Haussler. Exploiting generative models in discriminative classifiers. *In Advances in Neural Information Processing Systems 11*, page 487—493, 1998.

[6] K. Lee, Y. Artzi, J. Dodge, and L. Zettlemoyer. Context-dependent semantic parsing for time expressions. *In Proceedings of the 2014 Conference of Association for Computational Linguistics*, 2014.

[7] R. Pascanu, C. Gulcehre, K. Cho, and Y. Bengio. How to construct deep recurrent networks. *In Proceedings of the International Conference on Learning Representations (ICLR) 2014 Conference Track*, 2014.

[8] L. Ratinov and D. Roth. Design challenges and misconceptions in named entity recognition. *In Proceedings of the CoNLL 09*, 2009.