BARY at the NTCIR-11 MedNLP-2 Task for complaints and diagnosis recognition

Yusuke Matsubara
Social ICT Research Center,
University of Tokyo, Japan
7-3-1 Hongo, Bunkyo-ku, Tokyo
matsubara@sict.i.u-tokyo.ac.jp

Mizuki Morita
Social ICT Research Center,
University of Tokyo, Japan
7-3-1 Hongo, Bunkyo-ku, Tokyo
mizuki@sict.i.u-tokyo.ac.jp

Hasida Kôiti
Social ICT Research Center,
University of Tokyo, Japan
7-3-1 Hongo, Bunkyo-ku, Tokyo
hasida.koiti@i.u-tokyo.ac.jp

ABSTRACT

This paper describes a machine-learning based approach to recognizing diagnosed disease names and corresponding temporal expressions. Using CRFs (conditional random fields) to learn and predict tags, the systems described in this paper are characterized by a character-level formulation and heuristic features extracted from medical terminologies. Experimental results on the NTCIR-11 MedNLP-2 datasets suggest that the approach effectively exploit terminological resources and combine them with other NLP (natural language processing) resources including morphological analyzers.

Categories and Subject Descriptors

I.2.7 [**Artificial Intelligence**]: Natural Language Processing – *Language parsing and understanding*

General Terms

Experimentation, Human Factors, Languages

Keywords

Computational linguistics, clinical reports, machine learning, medical informatics

Team Name

BARY

Subtasks

Task 1 (Extraction task)

1. INTRODUCTION

Recognizing disease names in diagnoses with modalities (assertion, negation, etc.) and relevant time frames (e.g., whether observed since yesterday or three years ago) is an important problem of medical natural language processing. We have participated the NTCIR-11 MedNLP-2 Task [1] and explored effective ways to exploit presently available linguistic and terminological resources to tackle with the problem, under a standard approach of sequence labeling problems solved in supervised machine learning. Our team ("BARY") has participated the extraction subtask, where participants are given with medical reports annotated with disease names, their modalities and temporal expressions (for training), and non-annotated reports (for testing), written in Japanese. The overview paper of the MedNLP-2 task [1] gives a comprehensive description of the task and comparisons among the participants' approaches.

Our focus in this participation is three-fold: we aim to investigate how well a character-level formulation work, how effective a two-stage model is for the task, and how to induce a feature set suited for the domain and the scale of the given task.

We use a character-level formulation, instead of morpheme-level (or word-level) formulation in order to flexibly combine information from segmentation given by morphological analyzers and that of domain-specific terminological resources. In particular, our approach is supposed to mitigate negative effect caused by errors in word segmentation given by the morphological analyzer.

2. MATERIALS AND METHOD

Extraction of disease name, its modality attributes, and temporal expression in free-form text were approached as a sequence labeling problem. Each sentence in the input corpus was split into characters and IOB2 tags for each character were predicted with supervised machine learning. Both complaint and diagnosis (<c>) and temporal (<t>) tags were simultaneously predicted.

2.1 External language resources

For extracting disease names and symptoms with higher precision, Hyojun Byomei Master ver. 3.13[1] and Shojo Shoken Master <Shintai Shoken Hen> (PHYXAM-beta) ver 20140306[2] were used as glossaries of disease names and physical findings, respectively. Both were developed by and downloaded from the website of Medical Information System Development Center (MEDIS-DC). [3]

2.2 Machine learning and features

We model the task as two variations of machine learning formulations.

[1] 標準病名マスター (*Hyojun Byomei Master)* http://www2.medis.or.jp/stdcd/byomei/index.html

[2] 症状所見マスター【身体所見編】(*Shojo Shoken Master [Shintai-shoken-hen]*), http://www2.medis.or.jp/master/syoken/

[3] Medical Information System Development Center, http://www.medis.or.jp

1. Sequences of tag-modality pairs as labels (single-stage model)

2. Sequences of tags as 1st-stage labels, and sequences of modality markers as 2nd-stage labels (two-stage model)

For both models, we encode complaint and diagnosis (<c>) tags and temporal (<t>) tags into an IOB2 encoding as shown in Table 1 and train character-based CRFs (conditional random fields), and let them predict those tags against new data. The single-stage model yields one CRF instance, while the two-stage model yields two. To train their parameters associated with features (described below in detail), we use CRF++[4], a widely use open-source implementation

Our feature set includes traditional features such as surface strings, morphological analysis results and medical terminologies, as well as more task-specific features we created such as regular expression patterns to capture temporal expressions and prefixes and suffixes of disease names. Following traditional methods of natural language processing, features are automatically expanded using a (-N,M) character window, where N and M are the numbers of preceding/succeeding characters respectively, whose values are to be determined by the parameter tuning described in the next section. We use essentially the same feature set for the two models; the only difference is that the two-stage model utilizes 1st-stage labels (machine-predicted tags of disease and temporal expressions) as additional features.

An overview of the feature set we use is as follows. Note that although we use character-based models, some features work on multiple characters, giving sequences of feature values at once, as shown in Table 1.

Character surface and **character type** features are derived purely from the input string. The former is Unicode-normalized characters themselves and the latter categorizes characters into "Latin alphabet", "roman numerals", "plus sign", "parentheses", etc. Certain characters were chosen to have a type of their own, based on our observation in the training data that notations such as "(+)" are highly indicative of existence of disease diagnoses.

Morphological analysis features encode part-of-speech tags and word boundaries given by MeCab[5], the morphological analyzer we use.

Medical terminology features detect whether the enclosing substring is an entry of a subset of the MEDIS-DC dictionaries. The subset is discussed in Section 2.2. When matched, IOB-encoded feature values will be added. This works independently with the morphological analysis. (i.e., the input is matched against entries based on a simple string match.)

Temporal expression pattern Based on our observation in the training data that temporal expressions tend to be expressed regularly and have less variations, we created and used 10 regular expressions to capture most frequent ones. They include "\d{4}(|)年(から|)" (4 digits, full-width space (optional), "year", "since" (optional) in sequence) and "数日(前|後|)(から|)" ("several days", "before/later" (optional), "since" (optional)).

[4] T. Kudo, "CRF++: Yet Another CRF Tool Kit", https://code.google.com/p/crfpp/

[5] Kudo, T., "MeCab: Japanese morphological analyzer", https://code.google.com/p/mecab/

Family expression pattern features similarly capture mentions of family members with a set of regular expression patterns we created based on our observations. The patterns consist of expressions normally used to refer to family members in Japanese such as "父" *chichi* ("father") and "祖母" *sobo* ("grandmother").

Medical terminology affix features match the input string with a set of prefixes and suffixes derived from the MEDIS disease dictionary described above. We added this based on our observation in the training data that experts mention diseases using parts of formal disease names registered in the dictionary; some suffixes coincide with a class of diseases, and others are highly indicative of a disease name. This is in particular the case when a disease name is a long compound. We extracted 2289 maximal prefixes and suffixes of disease names in the MEDIS dictionary that occur 10 times or more and are longer than 1 character, removing redundant (enclosed by a longer affix) ones. The extracted affixes include: 好酸球 *kosankyu* ("acidocyte" or "acidophilic") and 水疱 *suiho* ("bladder" or "bullous").

We refer to temporal expression pattern, family expression pattern and medical terminology affix features as heuristic features in this paper.

Table 1: A sample corpus in our feature representation (with a subset of features). "B-T", "I-T","B-C","I-C" denote a beginning/inside/outside of temporal/disease expressions. F1 is a column for character type features, F2 is for medical terminology, and F3 is for temporal expression pattern.

Input token	Tag	F1	F2	F3
2 *ni* ("two")	B-T	NUMBER	O	I
日 *nichi* ("days")	I-T	CJK	O	I
前 *mae* ("ago")	I-T	CJK	O	I
か *ka* ("since")	I-T	HIRAGANA	O	I
ら *ra*	I-T	HIRAGANA	O	I
発 *hatsu* ("fever")	B-C	CJK	I	O
熱 *netsu*	I-C	CJK	I	O
。(EOS)	O	KUTEN	O	O

2.3 Parameter tuning

We use 5-fold cross-validation to select features from variations of the ones described above, and to select the most promising CRF hyperparameters. We split the training data into 5 sections manually to ensure the volume and topics are reasonably balanced. The cross validation lead us to decide:

- To use UniDic[6] 2.1.2 instead of IPAdic 2.7.0, and to not use MEDIS-derived resources as the dictionary used in MeCab
- To use Hyojun Byomei (標準病名, "Standardized disease names") dictioanry and Shintai Shoken (身体所見, "physical diagnosis") dictionary from the MEDIS-DC dictionaries and not use Rinsho Kensa (臨床検査, "clinical examination") dictionary
- To use feature 2-grams and 3-grams within a (-N,M)=(-2,2) window
- To set the regularization parameter C=1.0 and frequency cut-off threshold F=1

Due to time constraints, the above combination was not chosen from all possible combinations of options, but only the promising ones we tried.

2.4 Submitted systems

We have submitted three systems described below to show the differences of using the two-stage model and of using task-specific features we created.

BARY-1 is based on the single-stage model described in Section 2.2. and incorporates all machine-learning features except for the three task-specific heuristic features: temporal expression pattern, family expression pattern and medical terminology affix features.

BARY-2 is based on the two-stage model, and incorporate the same set of features as BARY1 does.

BARY-3 is based on the two-stage model, and incorporate all features including the three heuristic features described in Section 2.1.

3. RESULTS AND DISCUSSIONS

We have evaluated our system using the training set and the test set of annotated clinical findings reports, provided at the NTCIR MedNLP-2 Task. The training data consists of 102 reports, and the test data consists of 51.

As part of preliminary experiments, we have evaluated the different types of features used in the training of BARY-1, our baseline system. Table 2 shows how precision, recall, and (non-weighted) F-measure change as we add features. The results indicate that all feature types considered here contribute mostly incrementally, despite improvements being marginal in some cases. It should be noted that the character-type features give a major gain. We consider this to be a result of the characteristics of Japanese orthography in which a change in character types (e.g. from *kanji* to *hiragana*) in a sentence often coincides with a word boundary. It is also worth noting that the medical terminology features contribute positively to recall, indicating that terminology resources could help capturing unseen or less-frequent terms that are likely to be missed due to the data sparseness .

Table 3 shows official evaluations of our three systems on the MedNLP-2 gold-standard test set. A major trend is that two-stage models (BARY-2 and 3) outperform the single-stage one (BARY-1), albeit by a small margin (<1 point in F-measure). Another apparent trend is that affix features of BARY-3 contributed to recognizing assertive mentions of disease names ("C-positive" in Table 3) better. Because positive disease-name mentions cover a large potion of the all annotation entities, this resulted in BARY-3 outperforming the other two systems overall. On the other hand, the temporal expression and family expression patterns we introduced in BARY-3 have had only negligible contributions for their corresponding tags ("C-family" and T).

[6] Centre for Corpus Development, National Institute for Japanese Language and Linguistics (NINJAL), http://www.ninjal.ac.jp/corpus_center/unidic/

Table 2: Contributions of features in the BARY-1 system, evaluated with 5-fold cross-validation using the 102 reports of the MedNLP-2 training set. (best scores in bold.)

	Precision	Recall	F-measure
(1): Character surface	ALL: 73.94 C: 73.45 T: 75.80	ALL: 58.11 C: 56.45 T: 65.83	ALL: 65.04 C: 63.81 T: 70.36
(2): (1) + Character type	ALL: 88.00 C: 88.44 T: 86.01	ALL: 79.41 C: 79.57 T: 78.52	ALL: 83.46 C: 83.72 T: 82.02
(3): (2) + Morphological analysis	ALL: **89.40** C: **90.28** T: 85.70	ALL: 79.81 C: 79.90 T: 79.42	ALL: 84.28 C: **84.70** T: 82.34
(4): (3) + Medical terminology	ALL: 89.34 C: 90.08 T: **86.17**	ALL: **80.00** C: **80.00** T: **80.03**	ALL: **84.35** C: 84.67 T: **82.87**

Table 3: Performances on the 51 reports of the MedNLP-2 test sets of the three systems, trained with the 102 annotated reports of the training set. (best scores in bold.)

		Precision	Recall	F
BARY-1	ALL	89.44	77.41	82.99
	C-positive	89.50	76.64	82.57
	C-negative	**75.00**	57.69	65.22
	C-suspicion	70.67	66.07	68.29
	C-family	60.87	34.15	43.75
	T	89.09	**81.84**	**85.31**
BARY-2	ALL	89.33	78.84	83.76
	C-positive	89.34	**78.46**	83.55
	C-negative	**75.00**	57.69	65.22
	C-suspicion	72.14	**68.09**	**70.06**
	C-family	**65.22**	**36.59**	**46.88**
	T	**89.25**	81.03	84.94
BARY-3	ALL	**89.66**	**78.92**	**83.95**
	C-positive	**89.77**	**78.46**	**83.74**
	C-negative	71.90	**67.87**	**69.83**
	C-suspicion	**71.90**	67.87	69.83
	C-family	**65.22**	**36.59**	**46.88**
	T	89.05	81.57	85.15

3.1 Error analysis

As an example to show effectiveness of the dictionaries we used, we identified an instance 原発性アルドステロン血症 ("primary hyperaldosteronism") that our system was able to recognize correctly in the document 051 in the test corpus. This term did not occur in the test corpus, proving that the successful recognition was due to the Hyojun Byomei Master dictionary which incorporates the term.

We also identified an instance that an entry from a dictionary was considered to introduce noise. 心配 (*shinpai*, "anxiety"), which is a term found in the Hyojun Byomei Master, was wrongly recognized as a disease in the document 051, where the word was considered to be used in its general meaning.

好酸球 (*kosankyu*, "eosinophil granulocyte") was extracted from the Hyojun Byomei Master as an affix and matched against the document 051 in the test data. We consider this term indicative of diseases such as 急性好酸球性肺炎 (*kyusei-kosankyu-haien*, "eosinophilic pneumonia").

Similarly, 異常陰影 (*ijou-inei*, "abnormal findings on diagnostic imaging") was shown to be effective to help recognizing predicative expressions that denote different classes of abnormal imagery findings. Since only noun-noun compound terms for the class of findings are registered in the dictionaries we used, adding the suffix as a clue was considered beneficial.

心臓 (*shinzo*, "heart") and 血管 (*kekkan*, "blood vessel") were considered harmful when introduced as medical terminology affixes, because they are commonly used both expressions mentioning diseases and those do not. As an example, 心臓血管外科転科 (*shinzo-kekkan-geka-tenka*, "transferred to department of cardiovascular surgery") was wrongly recognized as a disease.

アメリカリウマチ学会 ("American College of Rheumatology") was mistakenly recognized as a disease in the document 043, probably due to the fact that it contains リウマチ (*riumachi*, "rheumatism"), which has been extracted as a suffix. Bary1 and Bary2 were not suffered from this because while リウマチ does not stand as a full disease name in the dictionaries while it is a component of many.

Temporal expressions were relatively easier to recognize due to their obvious orthographic and local patterns. A typical error in recognizing temporal expressions include false positives of dates and years mentioned in non-disease context. An example is 2008年3月 (“March 2008”) being incorrectly tagged by the system, which was part of a reference to a publication (日胸: 67巻3号, 2008年3月 (“Nikkyo: vol. 67, no. 3, March 2008”)). (document 013)

We also observe failed examples that require longer context than used in our systems to correctly recognize. A typical subcategory concerns modality distributed via coordinating structures; 経過中に頭痛や上気道症状、尿路症状はなく (“headache, upper respiratory symptoms, and urinary tract symptoms not being found”) in the document 063 require one to distribute the negation to the three symptoms.

4. CONCLUSION

The character-level approach using the two stages of CRFs, used with traditional features and heuristic feautres we have introduced in this paper, have been shown to be effective to the task of recognizing disease names and their modalities. Among the heuristic features, affix features using maximal prefixes and suffixes are an promising way to exploit existing terminological resources containing disease names.

5. REFERENCES

[1] Aramaki, E., Morita, M., Kano, Y., and Ohkuma, T. 2014. Overview of the NTCIR-11 MedNLP-2 Task. In *Proceedings of the 11th NTCIR Workshop Meeting on Evaluation of Information Access Technologies.*

[2] Lafferty, J., McCallum, A., and Pereira, F. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the 18th International Conference on Machine Learning*, 282-289