HokuMed in NTCIR-11 MedNLP-2: Automatic extraction of medical complaints from Japanese health records using machine learning and rule-based methods

Magnus Ahltorp
Graduate School of Information Science and Technology
Hokkaido University, Japan
map@kth.se

Hideyuki Tanushi
Dept. of Quality and Patient Safety
Karolinska University Hospital, Sweden
hide-tan@dsv.su.se

Shiho Kitajima
Graduate School of Information Science and Technology
Hokkaido University, Japan
shiho@ist.hokudai.ac.jp

Maria Skeppstedt
Graduate School of Information Science and Technology
Hokkaido University, Japan
mariask@dsv.su.se

Rafal Rzepka
Graduate School of Information Science and Technology
Hokkaido University, Japan
rzepka@ist.hokudai.ac.jp

Kenji Araki
Graduate School of Information Science and Technology
Hokkaido University, Japan
araki@ist.hokudai.ac.jp

ABSTRACT

A conditional random fields model was trained to detect medical complaints in Japanese health record text. Tokenisation was applied by using the dependency parser CaboCha and the conditional random fields model was trained on tokens in a window size of two preceding and three following tokens, as well as on part-of-speech, vocabulary mapping, header name, frequent suffix, orthography and presence of a modality cue.

Modality detection relied on dictionaries of cues for negation, suspicion and family. The scope of negation and suspicion cues was determined by rules relying on the output of CaboCha. For negation and family, cues were gathered by scanning the development corpus for cues, while suspicion cues were obtained by translating English cues.

The best result achieved for recognizing complaints was a precision of 87% and a recall of 77%. For modality detection, positive was detected with a precision of 87% and a recall of 77%, negation with a precision of 76% and a recall of 69%, suspicion with a precision 49% and a recall of 51%, and family with a precision of 78% and a recall of 81%.

Team name

HokuMed (Group ID: SUHUK)

Subtasks

Extraction of Complaint and Diagnosis

Keywords

natural language processing, medical language processing, named entity recognition, modality detection, clinical text, Japanese, conditional random fields

1. INTRODUCTION

Studies of clinical English have dominated clinical NLP research the past years, for instance with the i2b2/VA shared tasks [23, 22] and with the creation of several English corpora [6, 19, 2]. Recently, however, annotated health record corpora in other languages have been created, enabling clinical NLP on non-English languages. The MedNLP-2 shared task is one such example, in which Japanese health record text has been manually annotated and made available to the shared task participants [3].

We participated, as the HokuMed group, in the shared task of recognizing complaints (i.e. entities referring to symptoms and diagnoses) in the text and determining their modality (negation, suspicion and/or concerning a family member). We did not attempt to recognize date and time expressions.

2. METHOD

A development corpus was provided by the shared task organisers, in which about 3300 medical complaints had been manually annotated. Among these complaints, about 1040 were marked as belonging to the class negation, 110 as suspicion, 70 as concerning a family member and exactly 3 as both negated and concerning a family member. As held-out data, the organisers also provided a test corpus, against which the systems developed using the development corpus could be evaluated. This corpus contained about 2140 complaints, among which about 700 were marked as belonging to the class negation, 60 as suspicion, 40 as concerning a family member and exactly 1 as both negated and concerning a family member.

We assessed the number of complaints that had been annotated as non-positive as too small to successfully form a training set for a machine learning approach for modality detection. For this subtask, we therefore instead used cue dictionaries and rules based on the output of a dependency

parser. For the task of recognizing complaints, however, we assessed the number of annotated entity instances to be large enough to be able to successfully apply machine learning. Our final medical complaint extraction system therefore consisted of two parts; a machine learning model was first used for recognizing complaints in the text, and thereafter the rule- and dictionary based subsystem was used for determining the modality of the recognized complaints.

2.1 Entity recognition

Conditional random fields (CRF), introduced by Lafferty et al. [16], was chosen for training a machine learning model for recognizing medical complaints. CRF is a machine learning algorithm suitable for labelling sequential data, e.g. a sequence of tokens among which some denote medical complaints. The CRF implementation CRF++ [15] was used (as a linear chain CRF), which e.g. previously has been used for recognizing entities in clinical text [25, 18, 11, 20].

For segmenting the corpus into the sequential elements required for training the CRF model, white space tokenisation is not an option, since white space is generally not used in Japanese [14, p. 17]. To instead use a segmentation on character level would, however, probably be too granular. Therefore, the Japanese dependency parser CaboCha [5] was used for segmenting the corpus into tokens to use as the sequential elements for training the CRF system. To verify that this did not generally result in tokens that had only partially been annotated as complaints, the result of the tokenisation was matched to the annotated corpus, which showed that the annotation boundary was in the middle of a token in only 5 instances.

The IOB encoding [13, pp. 763–764] of the annotated entities was used, i.e. the CRF model was trained to classify the tokens into three types: *B-complaint* for a token starting an annotated chunk, *I-complaint* for a token inside an annotated chunk and *O* for a token not annotated as a complaint.

When training a CRF model, regularisation is used, which prevents over-fitting [4, p. 10] to the training data. The L1-norm [4, p. 146], governed by the variable C, is one of the methods for regularisation provided by CRF++. Using the L1-norm, with an appropriate value of C, results in features not contributing to the model not being included. This means that a large number of features can be provided to the model, and only the features that are contributing will be included, given that a suitable value for C is used [4, p. 145]. We adopted this approach of training a CRF model with the L1-norm regularisation, and provided the learning algorithm with a large number of features. C was calculated using hill-climbing and a start value of 4, with models trained and evaluated for each value of C, using the F-score of a 10-fold cross-validation. Each fold was evaluated by transforming the data to IOB encoding with one character per line and applying the CoNLL 2000 shared task evaluation script [8].

The following feature types were given as data for training the CRF model (see figure 1):

- The token (at the current position, the 2 previous positions and the 4 following).
- Part-of-speech of the token (at the current position, the 4 previous positions and the 4 following).
- Whether a token can be found in a dictionary of complaint words (at the current position, the 4 previous positions and the 4 following). The dictionaries used are described below.
- The header category in which the token occurs, e.g. 主訴 (main complaint) or 現病歴 (history of present illness) (at the current position).
- The suffix of the token, if it contains a suffix frequent in the complaints dictionaries (at the current position and the 4 following positions).
- The ortographic type of the token, e.g. hiragana, katakana, kanji, romaji, number, symbol (at the current position, the 4 previous positions and the 4 following).
- Whether any cue in a selected subset of the most common modality cues occurs in the same chunk or in the parent chunk, according to the dependency parser (at the current position).

All relevant dictionary resources for medical complaints available to us were used for the dictionary matching:

- The Byomei diagnosis list [17].
- MeSH terms classified under the nodes *Diseases (C)* and *Mental disorders (F03)* [21].
- MedDRA terms, except those classified as *investigations*, *social circumstances* and *surgical and medical procedures* [9].
- English terms from English SNOMED CT [10] belonging to the semantic category *disorder*.
- Translated terms from English SNOMED CT [10] belonging to the semantic category *disorder* that were possible to automatically translate into Japanese using the JMDict dictionary [12].
- Terms denoting complaints that had been automatically extracted from the Japanese patient blog corpus TOBYO using distributional semantics and thereafter manually filtered [1].

The list of suffixes was constructed by first combining the Byomei, MedDRA and MeSH lists into one list of unique terms and extracting one-character and two-character suffixes that occurred at least twice as a suffix in this list. The mapping to these suffixes when constructing the features for CRF prioritised a match to a two-character suffix over a match to a one-character suffix.

2.2 Modality detection

The general idea of the modality detection was to rely on dictionaries of cues for *negation*, *suspicion* and *family*. That is, a complaint was classified as negated if it was affected by a negation cue, as a suspicion if it was affected by a suspicion cue and as related to a family member if it was affected by a family cue.

There are Japanese nouns which can express negation, e.g. 否定 (= negation), 陰性 (= negative), as well as some of these nouns + する (verb, e.g. 否定する) or + だ (nominal adjective, e.g. 陰性だ). The most common method for expressing negation is, however, by inflecting verbs and verbal

Word	POS	Dictionary	Header	Suffix	Orthography	Modality	Result	
嚥下	**名詞, サ変接続**	**B-complaint**	noheader	嚥下	**kanji**	O	B-complaint	
障害	**名詞, 一般**	**I-complaint**	noheader	障害	**kanji**	O	I-complaint	
を	**助詞, 格助詞**	**O**	noheader	O	**hiragana**	O	O	
主	**接頭詞, 名詞接続**	**O**	noheader	O	**kanji**	O	O	
訴	**名詞, 一般**	**O**	**noheader**	**訴**	**kanji**	**O**	O	← current
に	**助詞, 格助詞**	**O**	noheader	**O**	**hiragana**	O	O	
来院	**名詞, サ変接続**	**O**	noheader	**O**	**kanji**	O	O	
し	**動詞, 自立**	**O**	noheader	**O**	**hiragana**	O	O	
た	**助動詞,***	**O**	noheader	**O**	**hiragana**	O	O	
.	記号, 句点	O	noheader	O	symbol	O	O	

Figure 1: CRF features. Used features for the current token are marked with red.

Submission		NER (complaints)	Positive	Family	Negation	Suspicion	Fam.+Neg.	Susp.+Neg.
1	Precision	87.30%	82.14%	96.88%	85.84%	69.70%	14.29%	0.00%
(c=0.69141453)	Recall	79.45%	75.77%	86.11%	74.90%	63.30%	33.33%	0.00%
	F-score	83.19	78.83	91.18	80.00	66.35	20.00	0.00
2	Precision	87.03%	81.87%	96.88%	85.84%	69.47%	14.29%	0.00%
(c=1)	Recall	78.42%	75.05%	86.11%	73.75%	60.55%	33.33%	0.00%
	F-score	82.50	78.31	91.18	79.34	64.71	20.00	0.00
3	Precision	87.63%	82.29%	96.92%	86.12%	70.83%	14.29%	0.00%
(c=0.629,	Recall	78.87%	74.76%	87.50%	74.90%	62.39%	33.33%	0.00%
fewer features)	F-score	83.02	78.34	91.97	80.12	66.34	20.00	0.00

Table 1: Results on the development corpus

adjectives and adding the negation predicate ない or ぬ [14, p. 54]. Limiting used negation cues to those found in the test data would, therefore, be likely to cover a large proportion of frequently used cues. Although there are also grammatical forms for describing level of certainty [14, p. 118], there are a a large number of frequently used non-grammatical cues for expressing suspicion.

Therefore, to find cues for the grammatical constructions expressing negation, as well as other negation cue words, the development corpus was scanned for examples. For suspicion, on the other hand, for which there are potentially a large number of cues, the strategy of translating English suspicion cues was instead adopted. The English cues were obtained from a previous study in which English cues had been collected and translated into Swedish for adapting English modality detection to Swedish [24]. The English versions of these cues were automatically translated into Japanese using the JMDict dictionary [12]. Thereafter, the translated cues were manually filtered by a native Japanese speaker, and only cues that were assessed as valid Japanese cues for expressing suspicion were retained.

As a final step, the constructed modality detection system was used for classifying all annotated complaints, and an error analysis was performed, which resulted in small modifications of the cue lists.

Whether the occurrence of a cue word was assessed as affecting a mentioned complaint, was determined through a set of rules. For negation and suspicion, the implemented rules relied on the output of the dependency parser CaboCha [5]. In general, if a cue was positioned in the same chunk as a complaint or in a parent chunk (i.e. a chunk closer to the root in the dependency graph), it was assessed as affecting the complaint. If the cue was separated from the complaint with a phrase separator, it was, however, assessed as not affecting it. The phrase separators were the following: conjunctive が (ga) following a verb, から (kara), による (ni yoru), ため (tame), に対して (ni taishite), comma unless after a common noun, as well as verbs in the continuative form.

Negation was assigned by checking for cues from three groups and counting the result as negated if cues from an odd number of groups were present. The three groups were:

Submission		NER (complaints)	NER (time & complaints)	Positive	Family	Negation	Suspicion
1	Precision	87.21%	87.21%	87.21%	77.78%	76.18%	48.84%
(c=0.69141453)	Recall	76.92%	65.59%	76.92%	80.77%	68.99%	51.22%
	F-score	81.74	74.87	81.74	79.25	72.41	50.00
2	Precision	87.47%	87.47%	87.47%	80.77%	76.44%	48.84%
(c=1)	Recall	76.12%	64.91%	76.12%	80.77%	68.54%	51.22%
	F-score	81.40	74.52	81.40	80.77	72.27	50.00
3	Precision	88.62%	88.62%	88.62%	77.78%	76.56%	46.67%
(c=0.629,	Recall	75.09%	64.03%	75.09%	80.77%	68.99%	51.22%
fewer features)	F-score	81.30	74.35	81.30	79.25	72.58	48.84

Table 2: Results on the test corpus

negation nouns ((−), 陰性, 否定, 予防, 消失, 清明, 良好, 正常, 不可能, 整, 清), negation predicates (ない, 無い, ぬ, なし), and improvement (改善). For suspicion, the list of suspicion cues mentioned above was used (71 cues). The family modality was assigned if the beginning of the line was present in a list a family cues (母親, 父親, 姉, 兄, 家族, 息子, 母, 娘, 父, 弟, 妹, 両親, 家族歴, 妻, 夫) or if the line was below the header 家族歴. If several cues affected a complaint, the rule was to assign the combined modality (e.g. negation as well as suspicion).

In addition, there was a cue (あり) for positive modality that, if it occurred in the same chunk as the complaint, overruled the negation and suspicion modality assignment. There were also cues that overruled all modalities: 止まらず, and different variants of (+). Finally, the word 意識清明 (lucidity, alert and conscious) was always counted as negated, disregarding any context, since it was always negated in the development corpus.

2.3 Tagging of the test corpus

The C-value that gave the best result for recognizing complaints when using 10-fold cross-validation was used for training a CRF model using the entire development corpus. This model was then applied for recognizing medical complaints in the test corpus. Thereafter, modality detection was performed by applying the final modality cue dictionaries and rules on the detected complaints.

Since the C-value fluctuates depending on the corpus, a second tagging was also performed, using a more "safe" C-value of 1, and a third tagging, using fewer features (only ±1 word was used) and the best C-value when using 10-fold cross validation.

The resulting, automatically tagged, test corpora was then submitted to the organizers for evaluation against the manually annotated version of this corpus.

3. RESULTS

The best F-score on the development corpus, 83.19, was achieved with a C-value of 0.69141453. The precision and recall for this F-score are shown in Table 1, together with the results of the modality detection when using the final modality cue dictionaries and rules on the development corpus.

The results on the test set are shown in Table 2. These were achieved by applying the three models trained on the entire annotated development corpus together with the final modality cue dictionaries and rules.

4. DISCUSSION AND CONCLUSION

The best result achieved on the development set for recognizing complaints was a precision of 87.30% and a recall of 79.45%, not counting time expressions. The best result on the test set was a precision of 87.21% and a recall of 76.92%, again not counting time expressions (since we did not attempt to recognize time expressions, the NER recall including both classes was much lower, 65.59%). On both the development set and the test set, submisson 1 was the most successful. The precision is in line with results from previous studies on English corpora, while the recall is slightly lower (e.g. precision 84% and recall 82% for Wang and Patrick [26] and precision 87% and recall 84% for Jiang et al. [11]).

Our models were optimized on the development set, and the recall were indeed slightly lower when the model was run on the test set. This is to an even larger extent the case for the modality detection, for which the cues and the scope rules have been adapted to the development set, except that the cue lexicon contains suspicion cues not present in the development data, as these cues were obtained by translating English cues.

Choices of cues and rules based on the manual modality annotations in the development set were not always successful on the test set. For example, 意識清明 was always counted as negated in the modality assignment, since it was always negated in the development corpus. It was, however, always positive in the test corpus, which means that the rule perhaps was not a good choice.

The modality assignment results on the development set were: detection of positive with precision 82% of and a recall of 76%, negation with a precision of 85% and a recall of 75%, a detection of suspicion with a precision 70% and a recall of 63%, while family had the best results and was detected with a precision of 97% and a recall of 86%. Inconsistencies in how to classify the cues and parser errors made it difficult to achieve higher results, even when tailoring the rules to the

development corpus. As expected, results were generally lower on the test set; detection of negation had a precision of 76% and a recall of 69%, suspicion had a precision of 49% and a recall of 51%, and family had a precision of 78% and a recall of 81%. Detection of positive increased to a precision of 87% and a recall of 77%.

Previous machine learning based modality detection studies on English have achieved higher results. Clark et al. [7] were, for instance, able to detect the category *present* (corresponding to positive) with a precision of 94% and a recall of 98%, *absent* (corresponding to negative) with a precision of 95% and recall of 92%, and the category *possible* (corresponding to suspicion) with a precision of 77% and recall of 53%. This machine learning model was, however, trained using a larger development set than what was available for this shared task. For the amount of available training data, we therefore believe that the strategy of using a cue and rule based modality detection system is more suitable, while machine learning models are more suitable for the named entity recognition.

5. REFERENCES

[1] M. Ahltorp, M. Skeppstedt, S. Kitajima, R. Rzepka, and K. Araki. Medical vocabulary mining using distributional semantics on Japanese patient blogs. In *Proceedings of SMBM 2014 - The 6th International Symposium on Semantic Mining in Biomedicine*, 2014.

[2] D. Albright, A. Lanfranchi, A. Fredriksen, W. F. Styler, 4th, C. Warner, J. D. Hwang, J. D. Choi, D. Dligach, R. D. Nielsen, J. Martin, W. Ward, M. Palmer, and G. K. Savova. Towards comprehensive syntactic and semantic annotations of the clinical narrative. *J Am Med Inform Assoc*, Jan 2013.

[3] E. Aramaki, M. Morita, Y. Kano, and T. Ohkuma. Overview of the NTCIR-11 MedNLP-2 task. In *Proceedings of NTCIR-11*, 2014.

[4] C. M. Bishop. *Pattern recognition and machine learning*. Springer, New York, NY, 2006.

[5] CaboCha. Cabocha: Yet another Japanese dependency structure analyzer. https://code.google.com/p/cabocha/, 2012.

[6] W. W. Chapman and J. N. Dowling. Inductive creation of an annotation schema for manually indexing clinical conditions from emergency department reports. *Journal of Biomedical Informatics*, 39(2):196–208, 2006.

[7] C. Clark, J. Aberdeen, M. Coarr, D. Tresner-Kirsch, B. Wellner, A. Yeh, and L. Hirschman. Mitre system for clinical assertion status classification. *J Am Med Inform Assoc*, 18(5):563–7, 2011.

[8] Conll. CoNLL-2000. http://www.cnts.ua.ac.be/conll2000/chunking/, Accessed 2011-10-09 2000.

[9] IFPMA. Meddra introductory guide version 14.0. http://www.who.int/medical_devices/innovation/MedDRAintroguide_version14_0_March2011.pdf, 2011.

[10] IHTSDO. International Health Terminology Standards Development Organisation: SNOMED Clinical Terms User Guide, July 2008 International Release. http://www.ihtsdo.org, 2008. Accessed 2011-01-24.

[11] M. Jiang, Y. Chen, M. Liu, S. T. Rosenbloom, S. Mani, J. C. Denny, and H. Xu. A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries. *J Am Med Inform Assoc*, 2011.

[12] JMdict. The jmdict project. http://www.edrdg.org/jmdict/j_jmdict.html, 2013.

[13] D. Jurafsky and J. H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*. Prentice Hall, second edition, Feb. 2008.

[14] M. Kamermans. *An Introduction to Japanese Syntax, Grammar & Language*. Sjgr Publishing, 2010.

[15] T. Kudo. CRF++: Yet Another CRF toolkit. http://crfpp.sourceforge.net/, 2012. Accessed 2012-06-15.

[16] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. 18th International Conf. on Machine Learning*, pages 282–289. Morgan Kaufmann, San Francisco, CA, 2001.

[17] MEDIS-DC. MEDIS Byomei Master v3.11. http://www2.medis.or.jp/stdcd/byomei/index.html, 2014.

[18] J. Patrick and M. Li. High accuracy information extraction of medication information from clinical notes: 2009 i2b2 medication extraction challenge. *J Am Med Inform Assoc*, 17(5):524–527, Sep-Oct 2010.

[19] A. Roberts, R. Gaizauskas, M. Hepple, G. Demetriou, Y. Guo, I. Roberts, and A. Setzer. Building a semantically annotated corpus of clinical texts. *J. of Biomedical Informatics*, 42:950–966, October 2009.

[20] M. Skeppstedt, M. Kvist, G. H. Nilsson, and H. Dalianis. Automatic recognition of disorders, findings, pharmaceuticals and body structures from clinical text: an annotation and machine learning study. *J Biomed Inform*, 49:148–58, Jun 2014.

[21] U.S. National Library of Medicine. MeSH (Medical Subject Headings). http://www.ncbi.nlm.nih.gov/mesh, 2014.

[22] Ö. Uzuner, I. Solti, F. Xia, and E. Cadag. Community annotation experiment for ground truth generation for the i2b2 medication challenge. *J Am Med Inform Assoc*, 17(5):519–523, 2010.

[23] Ö. Uzuner, B. R. South, S. Shen, and S. L. DuVall. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *J Am Med Inform Assoc*, 18(5):552–556, 2011.

[24] S. Velupillai, M. Skeppstedt, M. Kvist, D. Mowery, B. E. Chapman, H. Dalianis, and W. W. Chapman. Cue-based assertion classification for Swedish clinical text – developing a lexicon for pycontextswe. *Artif Intell Med*, Jan 2014.

[25] Y. Wang. Annotating and recognising named entities in clinical notes. In *Proceedings of the ACL-IJCNLP Student Research Workshop*, pages 18–26, Singapore, 2009.

[26] Y. Wang and J. Patrick. Cascading classifiers for named entity recognition in clinical notes. In *Proceedings of the Workshop on Biomedical Information Extraction*, pages 42–49, 2009.