A Comparison of Rule-Based and Machine Learning Methods for Medical Information Extraction

Osamu Imaichi, Toshihiko Yanase, and Yoshiki Niwa
Hitachi, Ltd., Central Research Laboratory
1-280, Higashi-Koigakubo
Kokubunji-shi, Tokyo 185-8601
{osamu.imaichi.xc,toshihiko.yanase.gm,yoshiki.niwa.tx}@hitachi.com

Abstract

This year's MedNLP (Morita and Kano, et al., 2013) has two tasks: de-identification and complaint and diagnosis. We tested both machine learning based methods and an ad-hoc rule-based method for the two tasks. For the de-identification task, the rule-based method achieved slightly higher results, while for the complaint and diagnosis task, the machine learning based method had much higher recalls and overall scores. These results suggest that these methods should be applied selectively depending on the nature of the information to be extracted, that is to say, whether it can be easily patternized or not.

1 Introduction

Machine learning based and rule-based methods are the two major approaches for extracting useful information from natural language texts. To clarify the pros and cons of these two approaches, we applied both approaches to this year's MedNLP tasks: de-identification and complaint and diagnosis.

For the de-identification task, ages and times, for example, are seemingly a type of information that can be patternized quite easily. In such cases, an ad-hoc rule-based method is expected to perform relatively well. In contrast, the complaint and diagnosis task would seem to have much more difficulty patternizing information, so a machine learning approach is expected to provide an effective methodology for tackling these problems.

2 Machine Learning Approach

In this section, we explain how the machine learning based approach works.

2.1 Sequential Labeling by using CRF

We formalized the information extraction task as a sequential labeling problem. A conditional random field (CRF) (Lafferty and McCallum, et al., 2001) was used as the learning algorithm. We used CRFsuite[1], which is an implementation of first-order linear chain CRF.

The CRF-based sequential labeling proceeds as follows. First, we applied a Japanese morphological parser (MeCab[2]) to documents and segmented the sentences into tokens with part-of-speech and reading. Then, the relationship between tokens was estimated using CaboCha[3], which is a common implementation of the Japanese dependency parser (Kudo and Matsumoto, 2002). Finally, we extracted the features of the tokens and created models using CRFsuite.

2.2 Basic Features

We used the following features to capture the characteristics of the token: surface, part-of-speech, and dictionary matching. The surface and part-of-speech of the target token were converted into numerical expressions in what is known as one-hot representation: the feature vector has the same length as the size of the vocabulary, and only one dimension is on. The dictionary feature is a binary expression that returns one if a word is in the dictionary and zero if not.

[1] http://www.chokkan.org/software/crfsuite/
[2] http://mecab.googlecode.com/svn/trunk/mecab/doc/
[3] http://code.google.com/p/cabocha/

International Joint Conference on Natural Language Processing Workshop on Natural Language Processing for Medical and Healthcare Fields, pages 38–42,
Nagoya, Japan, 14-18 October 2013.

We prepared ten kinds of dictionaries featuring age expressions, organ names, Japanese era names, family names, time expressions, names of hospital departments, disease names from the Japanese Wikipedia, Chinese characters related to diseases, suspicious expressions, and negative expressions. These dictionaries were created on the basis of the rules explained in Section 3.

To capture the local context of a target token, we combined features of several neighbor tokens. First, we merged the features of five adjacent tokens. Let w_i be the i-th token of the sentence. We concatenated the features of w_{i-2}, w_{i-1}, wi, w_{i+1}, and w_{i+2} and created $w_{[i-2:i+2]}$ to express the i-th node. Second, we concatenated the features of $w_{[i-2:i+2]}$ and w_i^{src} (w_i^{tgt}) to denote source (target) token of w_i.

2.3 Unsupervised Feature Learning

In addition to the basic features, we used clustering-based word features (Turian and Ratinov, et al., 2010) to estimate clusters of words that appear only in test data. These clusters can be learned from unlabeled data by using Brown's algorithm (Brown and deSouza, et al., 1992), which clusters words to maximize the mutual information of bigrams. Brown clustering is a hierarchical clustering algorithm, which means we can choose the granularity of clustering after the learning process has been finished.

We examined two kinds of Brown features: those created from training and test data related to the MedNLP Task (1,000 categories) and those created from the Japanese Wikipedia (100 categories). We decreased the number of categories of the latter because clustering Wikipedia is computationally expensive. The computational time of Brown clustering is $O(VK^2)$, where V denotes the size of vocabularies and K denotes the number of categories.

3 Rule-based Method

In this section, we explain the rule-based method.

3.1 De-identification task

- <a>: age
  - The basic pattern is "d1[歳才台代] (SAI (years old), SAI (years old), DAI (10's, 20's, ..., etc.), DAI (10's, 20's, ..., etc.))", where d1 is a positive integer, and [ABC] refers to A, B, or C.
  - If an age region is followed by specific modifiers "時|頃|[こご]ろ|代|[前後]半|以[上下] (JI (when), KORO (about), DAI, ZENHAN (anterior half), KOUHAN (posterior half), IJOU (over), IKA (under))", that region is expanded to the end of the modifier. A disjunctive expression "aaa|bbb|ccc" means aaa, bbb, or ccc.
  - If an age region is followed by one of interval-markers "から|より|まで|～ (KARA (from), YORI (from), MADE (to))", that region is expanded to the end of the marker.
  - If one age region is followed by another age region directly or with only hyphen-type characters (-ー－―～) between them, the two regions are joined to one.
    - eg. <a>27 歳 (27 SAI (27 years old))～<a>47 歳 (47 SAI (47 years old))→<a>27 歳～47 歳
- <t>: time
  - The basic pattern of time tags is "d1 年 d2 月 d3 日 d4 時 d5 分 d6 秒 (d1 NEN (year) d2 GATSU (month) d3 NICHI (day) d4 JI (hour) d5 FUN (minute) d6 BYO (second))", where d1 to d6 are non-negative integers. Any partial pattern starting from d1 or d2 or d3 is also eligible.
  - The special numerical pattern d1/d2 (1900 <= d1 <= 2099, 1 <= d2 <= 12) is interpreted as year = d1 and month = d2. In addition, the special numerical pattern "d1/d2 [に|から|より|まで|～] (NI (at), KARA (from), YORI (from), MADE (to))" (1 <= d1 <= 12, 1 <= d2 <= 31) is interpreted as month = d1 and day = d2.
  - Exceptional patterns are: "[同当即翌前][日年月]|翌朝|翌未明|その後 (same year, this year, next morning, ... etc.)".
  - While a time region is preceded by a prefix-type modifier, or followed by a postfix-type modifier, the region is expanded to the beginning or to the tail of the modifiers.
    - Prefix type modifiers:
      - [翌昨同当本][年月日] (last year, last month, same day, ... etc.)
      - AM/PM type prefix: 午後 (GOGO (PM))| 午前

(GOZEN (AM) | AM | am | PM | pm
- ●Ambiguity type prefix: 約 | およそ | ほぼ | 概ね (YAKU (about), OYOSO (about), HOBO (about), OOMUNE (about))

✧ Postfix type modifiers:
- ●[上中下]旬|初め|午[前後]|深夜|早朝|昼|朝方?|夕[方刻]?|[春夏秋冬] (late at night, early in the morning, ...etc)
- ●Ambiguity type: 頃 | ころ | ごろ | 前後 | 程 | 以[降後前]

✧ Intervals (from ~~~ to ~~~)
- ●<t>...<t> (から|より|まで|～) → <t>... (から|より|まで|～) </t>
- ●<t>aaaaa</t><t>bbbbb</t> → <t>aaaaa bbbbb</t>
- ●<t>aaaaa</t> [[- ー － — ～] <t>bbbbb</t> → <t>aaaaa [- ー－—～] bbbbb </t>

- <h>: hospital
  - ➢ First hospital tags were added by using the below hospital words dictionary composed of seven words, and temporary division tags were added by using the division-word dictionary of 27 words.
    - ✧ Hospital words: 当院|近医|同院|病院|クリニック|総合病院|大学病院 (TOUIN (my/our hospital), KINNI (near hospital), DOUIN (same hospital), BYOUIN (hospital), KURINIKKU (clinic), SOUGOUBYOUIN (general hospital), DAIGAKUBYOUIN (university hospital)
    - ✧ Division words: 外科|眼科|循環器内科|皮膚科|内科 ... etc. (GEKA (surgery), GANKA (ophthalmology), JUNKANKINAIKA (cardiovascular internal medicine), HIFUKA (dermatology), NAIKA (internal medicine) (27 words)
  - ➢ While a hospital region is preceded by any number of division regions, the hospital region is extended to the beginning of the division regions.
    - ✧ 内 科 皮 膚 科 <h>病院</h> → <h>内科皮膚科病院</h>
  - ➢ If a hospital region is preceded by a sequence of name characters (■), the region is expanded to the beginning of the name sequence.
    - ✧ ■ ■ ■<h>皮膚科病院</h> → <h>■ ■ ■皮膚科病院</h>
  - ➢ If a division region is preceded by a sequence of name characters, the region is expanded to the beginning of the name sequence, and the tag is changed to a hospital tag.
    - ✧ ■ ■ ■ 内科 →<h>■ ■ ■内科</h>
  - ➢ As a special case, if a name character sequence is followed by "[をに]?(紹介|緊急)(受診|入院) (SHOKAI (refer), KINKYU (emergency), JUSIN (consult), NYUIN (stay in hospital))", the name character sequence is taken as a hospital region.
    - ✧ ■ ■ ■[をに]?(紹介|緊急)(受診|入院) → <h> ■ ■ ■ </h>[をに]?(紹介|緊急)(受診|入院)
- <p>: person name
  - ➢ This tag was skipped.
- <x>: sex
  - ➢ The sex tags were added only by a simple pattern: "男性|女性 (DANSEI (male), JOSEI (female))".

3.2 Complaint and diagnosis task

- All <c> tags of the training data were extracted and a dictionary of complaints was made containing 1,068 words
- The <c> tags were added to the test data by the longest match method using the dictionary. In case of a single character word (咳 and 痰), a tag is added only if both the preceding character and the following character are not Kanji characters.
- If a <c> tag region is followed by the cancelling expressions below, the <c> tag is cancelled.
  - ➢ postfix type cancelling expressions: [歴剤量時室率]|検査|教育|反応|導入|胞診|精査|を?施行|培養|細胞|成分|取り?扱|ガイ[ダド]|分類基準|[^予防]*予?防|[^療]*療法|=[0-9] (history, inspection, prevention, ... etc.)

- <family> tags are added by using the following family-words:
  - 祖父母|兄弟?|姉妹?|[叔祖][父母][父母]親?|息子|娘|弟|妹 (SOHUBO (grandparent), KYOUDAI (brother), SHIMAI (sister), CHICHIOYA (father), HAHAOYA (mother), MUSUKO (son), MUSUME (daughter), OTOUTO (younger brother), IMOUTO (younger sister))
- Exception: some of following words are not tagged.
  - 親指|母指|娘細胞 (OYAYUBI (thumb), BOSI (thumb), MUSUMESAIBOU (daughter cell))
- If a <c> tag is preceded by a <family> tag in the same sentence, then "family" modality is added to the <c> tag.
  - <family>祖母</family> ... <c> aaaaa </c> ... <c>bbbbb</c> → <c modality=family> aaaaaa </c> ... <c modality=family> bbbbb </c>
- <negation> tags added to negation words like "ない (NAI (not))" or "ぬ (NU (not)", using Japanese morphological analysis.
- Also negation expressions like "否定的|否定され|(-) (HITEITEKI (negative), HITEISARE (denied))" are tagged with <negation> tag.
- <suspicion>, <recognition> and <improvement> tags are also tagged by pattern matching.
  - suspicion: 疑[いうっ]|疑わ[しせれ]|うたが[いうっ]|うたがわ[しせれ]|可能性|危険性|否定でき<negation>|考慮され|考え|思われ (UTAGAU (to suspect), KANOUSEI (possibility), KIKENSEI (dangerous), KOURYOSARE (considering), KANGAE (think), OMOWARE (appear))
  - recognition: 認め|診断|出現|訴え|みとめ (MITOME (recognize), SHINDAN (diagnosis), SHUTSUGEN (appearance), UTTAE (complain), MITOME (recognize))
  - improvement: 改善|消失|解消|離脱|軽快 (KAIZEN (improve), SHOUSHITU (disappear), KAISHOU (reverse), RIDATSU (separation), KEIKAI (resolve))
- If an <improvement tag or a <suspicion> tag is directly followed by a <negation> tag, then both tags are cancelled.
  - <improvement>改善</improvement>せ<negation>ず</negation> → 改善せず
  - <suspicion>疑われ</suspicion><negation>ず</negation> → 疑われず
- If a <recognition> tag is directly followed by a <negation> tag, then the recognition tag is cancelled and the negation tag is extended to the beginning of the recognition tag.
  - <recognition>認め</recognition> <negation>ず</negation> → <negation>認めず</negation>
- If a <c> tag is followed by a <negation> tag or <improvement> tag in the same sequence, and if the in-between part (M) does not contain any recognition/suspicion tags, then
  - if no other <c> tag exists in the in-between part M, "negation" modality is added to the <c> tag.
  - if other <c> tags exist in M, and if the in-between parts of <c> tags are composed of the following connecting expressions, then the negation modality is added to the <c> tag.
    - あるいは|または|および|及び?|乃至は?|ないしは?|その他の?|など|や|と|等 (ARUIWA (or), MATAWA (or), OYOBI (or), NAISHIWA (or), SONOHOKANO (other), NADO (and others), YA (or), TO (and), NADO (and others))
- If a <c> tag is followed by a <suspicion> tag, then "suspicion" modality is added under a similar condition as above.

4 Result

4.1 De-identification task

The results of the de-identification task are as follows.

	P	R	F	A
Rule	89.59	91.67	90.62	99.58
ML1	92.42	84.72	88.41	99.49
ML2	91.50	84.72	87.98	99.46

The Rule column shows the results of the rule-based method, and the ML1 and ML2 columns show the results of the machine learning methods. The ML1 is the result with Brown clustering using training and test data of the MedNLP Task. In addition to this, the ML2 is the result using Japanese Wikipedia for Brown clustering.

4.2 Complaint and diagnosis task

The results of complaint and diagnosis task are as follows.

	P	R	F	A
Rule	72.47	58.12	64.50	93.40
ML1	88.98	74.24	80.94	96.08
ML2	88.55	75.32	81.40	96.06

5 Conclusion

For the de-identification task, the rule-based method achieved slightly higher results, while for the complaint and diagnosis task, the machine learning based method had much higher recalls and overall scores. These results suggest that we should use these methods selectively depending on the nature of the information to be extracted, that is to say, whether it can be easily patternized or not.

References

Morita, M., Kano, Y., Ohkuma, T., Miyabe, M., and Aramaki, E. 2013. Overview of the NTCIR-10 MedNLP Task, In *Proceedings of the 10th NTCIR Workshop Meeting on Evaluation of Information Access Technologies.*

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.

Kudo, T., and Matsumoto, Y. 2002. Japanese Dependency Analysis using Cascaded Chunking, In *Proceedings of the 6th Conference on Natural Language Learning* (COLING 2002 Post-Conference Workshop), 63-69.

Turian, J., Ratinov, L., and Bengio, Y. 2010. Word representations: A simple and general method for semi-supervised learning, In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 384-394.

Brown, P. F., deSouza, P. V., Mercer, R. L., Pietra, V. J. D., and Lai, J. C. 1992. Class-based n-gram models of natural language, *Computational Linguistics*, 18:467-479.