ImNER Indonesian Medical Named Entity Recognition

Conference Paper · August 2014

DOI: 10.13140/2.1.3431.6163

3 authors:

Wiwin Suwarningsih
Indonesian Institute of Sciences
17 PUBLICATIONS **0** CITATIONS
SEE PROFILE

Iping Supriana
Bandung Institute of Technology
243 PUBLICATIONS **145** CITATIONS
SEE PROFILE

Ayu Purwarianti
Bandung Institute of Technology
122 PUBLICATIONS **140** CITATIONS
SEE PROFILE

2014 2nd International Conference on Technology, Informatics, Management, Engineering & Environment
Bandung, Indonesia, August 19-21, 2014

ImNER Indonesian Medical Named Entity Recognition

Wiwin Suwarningsih[1,2], Iping Supriana[3], Ayu Purwarianti[4]
[1,3,4]*School of Electrical Engineering and Informatics, Institut Teknologi Bandung*
Bandung, Indonesia
[1]wiwin.suwarningsih@students.itb.ac.id, [3]iping@informatika.org, [4]ayu@informatika.org
[2]Researcher on Research Center for Informatics
Indonesian Institute of Science
wiwin.suwarningsih@lipi.go.id

***Abstract*—We propose a medical named entity recognition for medical question answering system with Indonesian language. The aim is to provide a good medical named entity grammar by only using the available language resource. Our strategy here is to build the features most often used for the recognition and classification of medical named entities. We organize them along two different axes: word-level and list features, document and corpus features. For the reason we built our own features to Indonesian medical named entities and used it as the feature of the available with SVM Software. By using 3000 sentences, the highest accuracy score achieved is about 90%.**

Keywords*—Medical named entity, Word-level features, Document and corpus features, SVM engine.

I. INTRODUCTION

Medical named entity recognition is an important and difficult task in computational linguistics such as medical question answering system. The discovery of new and potentially meaningful relationships between named entities in medical literature can take great advantage from the application of multirelational data mining approaches in text mining [7]. In recent years, conditional random fields (CRFs) have shown good performance in named entity recognition tasks. However, a direct application of it to medical named entity recognition incurs a very high training cost [1].

There are many biomedical/medical/clinical named entity recognition (NER) research with various approach and method. Many of them utilized some classifiers such as CRFs, SVMs, and biomedical corpus for experiment such as GENIA.

Among NER research, Cai et al [5], Chan et al [1], Yang et al [6], and Liao et al [11], have investigated biomedical/medical NER with classifiers such as CRFs.

Cai et al [5], proposed a novel selection method for tri-training learning, using three classifiers: CRFs, SVMs and ME. In tri-training process, Cai's select new labeled samples based on the selection model maximizing training utility, and compute the agreement according to the agreement scoring function. Chan et al [1], evaluate two alternatives to training a CRFs with a traditional single-phase maximum likelihood training method. For the cascaded method, Chan propose to include a "margin" in the model that leads to better recognition results. Yang et al [6], represents a two-phase approach based on semi-CRFs and novel feature sets. Semi-CRFs put the label to a segment not a single word which is more natural than the other machine learning methods. Yang's approach divides the whole biomedical NER into two sub-tasks: term boundary detection and semantic labeling. Liao et al [11], employ a skip-chain conditional random fields (CRFs) model for BioNER (Biomedical named entity recognition). This model completely considers to the long-range dependencies about biomedical information.

Gong et al [4], Cai et al [5], and Liao et al [11], have employed biomedical NER approaches that exploit GENIA corpus for experiment and the experiment result achieves precision, recall and F-score of the best.

Gong et al [4], presents a hybrid approach to recognize biomedical entity, which includes POS (Part-of-Speech) tagging, rules-based and dictionary-based approach using biomedical ontology. By using the GENIA 3.02 corpus for aiding biologist tagging biomedical entity in the biomedical literature and obtain a recall of 66%, a precision of 78% and an F-score 71.5%. Cai et al [5], using the GENIA corpus for experiment, show that Cai's proposed tri-training learning approach can more effectively and stably exploit unlabeled data. Liao et al [11], using the GENIA corpus for testing, their approach obtains significant improvement over other methods, which achieves 72.8% for precision, 73.6% for recall and 73.2% for F-score.

An other research with propose a new novelty for clicical NER such as Dehghan et al [2], explores a variety of common challenges faced by clinical named entity recognition and classification methods as well as current approaches to handling them. Gu et al [3], propose a novel approach aimed at minimizing the annotation requirement. The Gu's idea is to use a dictionary which is essentially a list of entity names compiled by domain experts and sometimes more readily available than domain experts themselves.

Colmenar et al [8], presents a Named Entity Recognition (NER) system based on Hidden Markov Models. The system design is language independent, and the target language and scope of the NER is determined by the training corpus. Fatiha [9], develop a tool for extracting named entities from free French reports. A rule-based approach is applied to clinical reports corpus of infectious diseases to extract semantic content in the form of named entities and properties. Two objectives can be achieved through this work; (1) the patient data anonymization and (2) the structured database fill in dedicated for future medical applications.

Here, we present the features most often used for the recognition and classification of Indonesian medical named entities. We organize them along two different axes: Word-level and List features, and Document and corpus features. For the reason, we built our own features to Indonesian medical named entities and used it as the feature of the available with SVM Software.

The rest of the paper is organized as follows : Sec 2. Related work; Sec 3. Describes about medical named entity recognition for medical question answering; Sec 4. Present the Indonesian medical features; Sec 5. describes our experimental data and its result with the SVM engine.

II. Related Work

There are two main approaches for named entity recognition : manual and automatic. Named entity recognition using manual such as Dehghan et al [2], identification and classification of mentions of relevant clinical concepts is a crucial preprocessing step in designing and developing clinical decision support systems. While this task has gained significant attention in recent years, there are still a number of issues that need further investigation. Apice et al [7], investigate the application of such an approach to address the task of identifying informative syntactic structures, which are frequent in biomedical abstract corpora. Initially, named entities are annotated in text corpora according to some biomedical dictionary (e.g. MeSH taxonomy).

The second system is automatic named entity recognition Liu et al [10] to recognize and extract the exact Japanese sight seeing domain named entities. It is a basic step for the following processing: question analysis and keyword extraction information retrieval. As well as, through doing the named entity recognition, we consider that it can mine exact information from text document to respond to user. Colmenar et al[8], implements a different approach of that statistical theory, showing that each component may complement the results of the other one. Unlike most of the previous works, two labels are returned when the components provide different results. This redundancy is an advantage when human supervision is mandatory at the end of the process such as in intelligence environments.

An other related work to adapting SVMs to named entity recognition, the multi-class problem and the unbalanced class distribution problem become very serious in terms of training cost and performance by Lee et al [14]. Zhang and Elhadad [15], propose an unsupervised approach to extracting named entities from biomedical text. A noun phrase chunker followed by a filter based on inverse document frequency extracts candidate entities from free text. Classification of candidate entities into categories of interest is carried out by leveraging principles from distributional semantics.

III. Medical Named Entity Recognition for Medical Question Answering

Most question answering system, NER is typically used as an aid to filter out strings that do not contain the answer. The NER is therefore used to single out the entity types appearing in a text fragment. If a piece of text does not have any entity with a type compatible with the type of the expected answer, the text is discarded or heavily penalised. With this in mind, the desiderata of a NER are related with the range of entities to detect and with the recall of the system [12].

A question answering system typically uses both a taxonomy of expected answers and the taxonomy of named entities produced by its NER to identify which
named entities are relevant to a question [10]. In this paper we combine between NER includes the identification and classification of certain proper nouns (like location, facility, diagnosis, definition organization, person, data, and others) in a text and medical NE features. Features are descriptors or characteristic attributes of words designed for algorithmic consumption. Feature vector representation is an abstraction over text where typically each word is represented by one or many boolean, numeric and nominal values [13].

Similar with medical question answering system, we devide analysis for medical NER into two component such as medical range entity; recall and precision medical NER. Following subsections will discuss each component in detail.

A. *Medical range of Entity*

Different domains (such as medical or clinical or biomedical) require different types of answers. Typically, the medical question classification component determines the type of medical question and the type of the expected medical answer. In this paper we construct medical range of entity about medical taxonomy based on identification and classification of proper nouns. Indonesian medical taxonomy shown in table I.

TABLE I. Indonesian Medical Taxonomy

Classification	Medical range of entity
Location	Kota, Provinsi, Kabupaten, desa, Bandung, Jakarta, Didalam, Keluar
Facility	Rawat inap, rawat jalan, VIP, perawatan dokter spesialis, fisioterapi, laboratorium, USG, EKG
Diagnosis	Penyakit, Obat, Terapi, Perawatan, Pengobatan, Tindakan pencegahan, antisipasi komplikasi, dampak over dosis obat, makanan pendukung, ASI eksklusif, pengendalian ASI, perawatan luka, penanganan tersedak, luka bakar, luka gigitan binatang berbisa
Definition	Nama penyakit, Jenis penyakit, Penggunaan obat, Risiko terapi, Risiko Obat, kemoterapi, darah tinggi, kekurangan darah, anemia, penyakit lupus, lekemia,
Person	Pasien, perawat, dokter, bidan, terapis.

A problem that arises here is the medical named entity types usually are not matched against the types the medical question requires. Consequently, even though a medical question classifier could determine a very specific type of medical answer, this type needs to be mapped to the types provided by the medical NER features.

B. *Attribute extraction*

Our first challenge is to named entity recognize and extract identifying as well as sensitive information from the unstructured data on the heterogeneous data. We use a statistical learning approach for extracting identifying and sensitive attributes in medical data.

A key to our classifier based approach is the selection of the feature set based on Indonesian taxonomy. Once the text is parsed, a set of features is generated for each token or term in the medical text. The features of a token contain the token itself such as previous word, next word, things, and number, etc.

The use of local features allow our medical NER to be more portable and work across many different types of data. Once the medical feature data is generated, we feed them to a classifier for training. We note that our extraction component so far only extracts atomic attributes and we use simple heuristics to associate these attributes to an entity.

We will discuss in section 3 about two classification approaches we studied.

C. *Recall and Precision for medical NER*

Given that the medical NER is used to filter out candidate medical answers, it is important that only wrong answers are removed, while all correct answers stay in the set of possible answers. Therefore, recall in a medical NER in medical question answering is to be preferred above precision. A NER developed for a generic NE recognition task is fine-tuned for a good balance between recall and precision[13].

In order to increase recall, it is therefore theoretically to allow to return multiple labels and then let further modules of the medical QA system do the final filtering to detect the exact answer. This is the hypothesis that we want to test a medical NER that assigns single labels and a variation of the medical NER features.

IV. INDONESIAN MEDICAL NE FEATURES

Indonesian medical named entities organize along two different axes such as word-level features, list features, document and corpus features. Following subsections will discuss each component in detail.

A. *Indonesian Medical Word-level and list features*

Indonesian medical word-level features are related to the character makeup of words. They specifically describe word case, punctuation, numerical value and special characters. Table II lists subcategories of word-level features.

TABLE II. INDONESIAN MEDICAL WORD-LEVEL FEATURES

Features	Examples
Morphology	- Prefix (e.g menahun, mengganggu, penyakit, penderita, pengobatan,.) - Suffix (e.g perawatan, gangguan jantung,) - Stem (e.g. obat, terapi, sakit, gejala, cegah, sesak, nafas)
Part-of-speech	- verb (e.g. konsumsi obat, serangan asma, operasi) - noun (e.g. kalsium, fosfat, jantung, hati, tulang, paru-paru, ginjal) - Foreign word (e.g. farmakoterapi, sodium, hypercalciuria, urin, hypocitraturia, hyperoxaluria)
Function	- Alphanumeric (e.g. satu, dua puluh, seratus, seribu, enam puluh) - non-alpha(e.g. *, $, #, @) - n-gram (e.g. bigram : _M,ME, ED, DI, IS,S_; trigram : _ME, MED, EDI, DIS, ID_, S_ _)
Case	- Starts with a capital letter (e.g. Imunisasi, Balita, Tekanan darah tinggi) - Word is all uppercased (e.g. ASI, RSUD, BKIA)
General list	- Stop words (e.g. hanya, pun, itu, ini, beberapa, sebagian) - Capitalized nouns (e.g. January, Monday)

B. *Indonesian Medical Document and Corpus features*

Document features are defined over both Indonesian medical document content and document structure. We list in this section features that go beyond the single word and multi-word expression and include meta-information about indonesian medical documents and corpus statistics.

TABLE III. INDONESIAN MEDICAL DOCUMENT AND CORPUS FEATURES

Features	Example
Multiple occurrences	- Other entities in the context - Anaphora, coreference
Local syntax	- Enumeration (e.g. dokter, obat, dosis, perawatan) - Apposition (e.g. minggu yang lalu, dokter pendamping, terapis pengganti)
Corpus frequency	- Word and phrase frequency (e.g. gejala penyakit, diagnosa penanganan penyakit, penanganan penyakit) - Multiword unit permanency (e.g. dosis obat, perulangan terapi, pengetesan sampling darah, pengujian diagnosa)

V. EXPERIMENTS

A. *Experimental Data*

In our knowledge, we built our own Indonesian medical named entity. We collected Indonesian medical articles from two popular Indonesian sites (http://health.detik.com/ and detikhealth.com and *health.kompas.com/konsultasi*) for data years 2013 (after eliminating the similar medical entity, it gave us 1000 medical sentences for each classification and features).

B. *Experimental Result*

In the experiment, we used an SVM algorithm are available in the WEKA. For the baseline, we used Indonesian medical taxonomy.

We compared the highest baseline result with our proposed features (see table IV. For the result) : the Indonesian medical word level-list feature (WL) and the Indonesian medical document and corpus feature (DC). We used 5-fold cross validation for the accuracy calculation.

TABLE IV. ACCURACY SCORE OF INDONESIA MEDICAL NER

Method	Accuracy Score
Baseline	85%
Baseline + WL	84%
WL + DC	89%
Baseline + WL + DC	90%

From table IV, we can see that using several medical world gives higher score than all feature in the sentences.

TABLE V. CONFUSION MATRIX FOR BASELINE + WL FEATURE

	Loc	Fac	Diagn	Def	Person
Loc	23	4	237	0	5
Fac	34	56	11	0	1
Diagn	1	1	0	67	1
Def	0	0	0	0	19
Person	0	0	0	0	4

As shown in table V, the highest misclassification for the Baseline + WL system is the "Diagnosis" category falls into "Facility" category. This is mostly for "penyakit" named entity, such as "Jenis penyakit yang saya derita membahayakan atau tidak".

TABLE VI. CONFUSION MATRIX FOR BASELINE + WL +DC FEATURE

	Loc	**Fac**	**Diagn**	**Def**	**Person**
Loc	12	4	37	123	5
Fac	34	56	11	0	1
Diagn	1	1	0	67	1
Def	0	45	1	0	19
Person	23	0	0	0	4

Based on table VI, the highest misclassification for the Baseline + WL + DC system is the "Definistion" category falls into "Facility" category. This is mostly for "risiko terapi" named entity, such as "Kemoterapi adalah cara memelihara ketika telah melakukan suatu operasi, resiko yang akan muncul biasanya berupa mual dan pusing". Compared to the baseline + WL system result, the full feature gives better accuracy score almost for all classes medical taxonomy.

Our initial experimental results show that our medical NER system effectively detects a variety of identifying attributes with high precision, and provides flexible identification options that anonymizes the data. It sightseeing domain named entity recognition we have got excellent precision and recalling rates. It shows that our method is effective and can be used in a practical medical question answering system

VI. CONCLUSION

We presented Indonesian medical name entity recognition as well as attribute extraction for anonymizing heterogeneous health information including both structured and unstructured data. Our experiment showed that the Indonesian medical named entities using SVM was organize for two different fitures and SVM was able to achieve good accuracy score.

ACKNOWLEDGMENTS

We thank anonymous reviewers for their helpful comments.

REFERENCE

1. Shing-Kit Chan, Wai Lam Efficient, Methods for Biomedical Named Entity Recognition, *In Proceedings of the 7th IEEE International Conference on Bioinformatics and Bioengineering*, 2007. BIBE 2007. Page: 729 – 735.
2. Dehghan, A, Keane, J.A, Nenadic, G., Challenges in Clinical Named Entity Recognition for Decision Support, *IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2013.* Page : 947 – 951.
3. Baohua Gu ; Dahl, V. ; Popowich, F., Recognizing Biomedical Named Entities in the Absence of Human Annotated Corpora , *International Conference on Natural Language Processing and Knowledge Engineering, 2007.* IEEE Conference Publications 2007, Page(s): 74 – 81
4. Le-Jun Gong ; Yi Yuan ; You-Bing Wei ; Xiao Sun, A Hybrid Approach for Biomedical Entity Name Recognition, *2nd International Conference* on *Biomedical Engineering and Informatics, 2009. BMEI '09.* IEEE Conference Publications 2009 , Page(s): 1 - 5
5. YueHong Cai ; Xianyi Cheng, Biomedical Named Entity Recognition with Tri-Training Learning, *2nd International Conference on Biomedical Engineering and Informatics, 2009. BMEI '09.* IEEE Conference Publications 2009 , Page(s): 1 - 5
6. Li Yang ; Yanhong Zhou, Two-phase biomedical named entity recognition based on semi-CRFs, *IEEE Fifth International Conference on Bio-Inspired Computing: Theories and Applications (BIC-TA), 2010,* Page(s): 1061 - 1065
7. Appice, A. ; Ceci, M. ; Loglisci, C., Discovering Informative Syntactic Relationships between Named Entities in Biomedical Literature, *Second International Conference on Advances in Databases Knowledge and Data Applications (DBKDA), 2010* . IEEE Conference Publications : 2010 , Page(s): 120 - 125
8. Colmenar, J.M. ; Abanades, M.A. ; Poza, F. ; Martin, D. ; Cuesta, Alfredo ; Herran, A. ; Hidalgo, J.I, On a generalized name entity recognizer based on Hidden Markov Models, *11th International Conference on In Intelligent Systems Design and Applications (ISDA), 2011,* IEEE Conference Publications: 2011 , Page(s): 952 - 958
9. Fatiha, B. ; Bouziane, B. ; Baghdad, A**,** MedIX: A named entity extraction tool from patient clinical reports**,** *International Conference on Communications, Computing and Control Applications (CCCA), 2011*, IEEE Conference Publications 2011 , Page(s): 1 - 6
10. Ye Liu ; Ren, F**,** .Japanese named entity recognition for question answering system**,** *In International Conference on Cloud Computing and Intelligence Systems (CCIS), 2011 IEEE* , Page(s): 402 – 406.

11. Zhihua Liao ; Hongguang Wu, Biomedical Named Entity Recognition Based on Skip-Chain CRFS**,** *International Conference on Industrial Control and Electronics Engineering (ICICEE), 2012,* IEEE Conference Publications : 2012 , Page(s): 1495 – 1498.
12. Diego Moll, Menno van Zaanen and Daniel Smith, Named Entity Recognition for Question Answering, *In Proceedings of the 2006 Australasian Language Technology Workshop (ALTW2006)*, pages 51–58.
13. Bick, Eckhard. A Named Entity Recognizer for Danish. In *Proc. Conference on Language Resources and Evaluation,* 2004
14. Ki-Joong Lee, Young-Sook Hwang, Seonho Kim, Hae-Chang Rim, Biomedical named entity recognition using two-phase model based on SVMs, *Journal of Biomedical Informatics 37,* 2004, pp: 436–447.
15. Shaodian Zhang, Noémie Elhadad, Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts, *Journal of Biomedical Informatics 46,* 2013, pp: 1088–1098.