Patient Information Extraction in Noisy Tele-health Texts

Mi-Young Kim, Ying Xu, Osmar Zaiane and Randy Goebel
Department of Computing Science
University of Alberta
Edmonton AB Canada
{miyoung2, yx2, zaiane, rgoebel}@ualberta.ca

Abstract— We explore methods for effectively extracting information from clinical narratives, which are captured in a public health consulting phone service called HealthLink. The currently available data consists of dialogues constructed by nurses while consulting patients on the phone. Since the data are interviews transcribed by nurses during phone conversations, they include a significant volume and variety of noise: First is explicit noise, which includes spelling errors, unfinished sentences, omission of sentence delimiters, variants of terms, etc. Second is implicit noise, which includes non-patient's information and negation of patient's information. To filter explicit noise, we propose our biomedical term detection/normalization method: it resolves misspelling, term variations, and arbitrary abbreviation of terms by nurses. In detecting temporal terms and other types of named entities (which show patients' personal information such as age, and sex), we propose a bootstrapping-based pattern learning to detect all kinds of arbitrary variations of the named entities. To address implicit noise, we propose a dependency path-based filtering method. The result of our de-noising is the extraction of normalized patient information. The experimental results show that we achieve reasonable performance with our noise reduction methods.

I. INTRODUCTION

Extraction of clinical information such as medications, symptoms, diseases, and patient's personal information from clinical text is an important task of clinical natural language processing. Our data comes from captured Tele-Health dialogues in Alberta, Canada from a publicly accessible system that is called HealthLink. With HealthLink, the public can access health advice and information by calling a phone line and discussing real time with a registered nurse who simultaneously transcribes the conversation in text. As an integral part of Electronic Health Records (EHR), clinical notes pose special challenges for analyzing EHRs due to their unstructured nature and substantial noise, since they are written by health practitioners in real time, while talking with patients. The noise can be divided into two types: First is noise which is shown explicitly, such as spelling errors, abbreviations, acronyms, unfinished sentences, term variants, and omission of sentence delimiters. Second is implicit noise which means noise is only revealed by a variety of inference methods. Written information which is not about a patient, and the negation information which is not true, are examples of implicit noise.

To identify explicit noise, we embed a misspelling correction module in our unsupervised language-model based biomedical term detection method. For temporality and other types of named entities, we set up seed patterns and run our own bootstrapping method: it detects variants of the seed patterns in the data using Damerau-Levenshtein distance [13]. To identify implicit noise, we use more detailed natural language processing method employing syntactic analysis, and filter out untrustworthy information.

The contents of this paper are as follows. In Section 2, we explain how to identify explicit noise. In Section 3, we describe our own method for identifying implicit noise by classifying named entities into facts, non-facts, and concerns for a patient. Section 4 shows some experimental results, and related work is described in Section 5. Section 6 concludes with a summary and future work.

II. REMOVING EXPLICIT NOISE

A. Language model-based biomedical named entity recognition

Here we describe our method to recognize biomedical terms such as symptom, disease, drug, virus, etc. Because nurses write arbitrarily different forms for the same term, we also perform normalization of terms. For literature mining in medical records, the medical ontology known as Unified Medical Language System (UMLS) [4] enables physicians to classify signs, symptoms, and diseases using accepted medical concepts. Our hypothesis is that, combined with an integrated information retrieval method, the UMLS is a powerful and appropriate tool to use as the basis for automatically mapping biomedical names with variant forms into one concept. More specifically, we only keep concepts belonging to the following semantic types as biomedical concepts: {disease or syndrome, finding, sign or symptom, virus, pharmacologic}.

A central architectural aspect of our processing of medical documents is based on treating sentences of those documents as queries and UMLS entries as documents. In this information retrieval (IR) model, we infer a language model for each UMLS concept entry, and rank each related entry according to how likely it generates the input sentence based on its language model. We would like to estimate $\hat{p}(Q \mid M_d)$, the probability of the query Q given the language model of document d as follows.

$$\hat{p}(Q \mid M_d) = \prod_{w \in Q} \hat{p}(w \mid M_d) \times \prod_{w \notin Q} 1.0 - \hat{p}(w \mid M_d).$$

The first term is the probability of generating words in the query and the second term is the probability of not generating other terms. The specific probabilities for $\hat{p}(Q \mid M_d)$ are defined as follows:

$$\hat{P}(w \mid M_d) = \begin{cases} \hat{p}_{ml}(w,d)^{(1.0-\hat{R}_{w,d})} \times \hat{p}_{avg}(w)^{\hat{R}_{w,d}} & \text{if } wf_{(w,d)} > 0 \\ \dfrac{c_w t}{cs} & \text{otherwise} \end{cases},$$

$$\hat{R}_{w,d} = \left(\frac{1.0}{(1.0+\bar{f}_w)}\right) \times \left(\frac{\bar{f}_w}{(1.0+\bar{f}_w)}\right)^{wf_{(w,d)}},$$

$$\hat{p}_{ml}(w,d) = \frac{wf_{(w,d)}}{dl_d},$$

$$\hat{p}_{avg}(w) = \frac{\left(\sum_{d(w \in d)} \hat{p}_{ml}(w,d)\right)}{df_w}.$$

For more details on each probability, see Ponte and Croft [6].

We use Damerau-Levenshtein distance to identify explicit noise such as misspelling, and arbitrary abbreviations. When we compute term frequency, we include the term variants of which the Damerau-Levenshtein distance is less than a threshold. By using the Damerau-Levenshtein distance measure, we compute term frequency *wf(w,d)* as follows:

$$wf(w,d) = \sum_{t \in DL_w} \text{count(t,d)} \times (1 - \frac{DL\_dist(t,w)}{length(t)}),$$

where DL_w is a group of a variant t for word w where $DL\_dist(t,w) <=$ threshold. $DL\_dist(t,w)$ is a damerau-levenshtein distance between t, and w.

We assume there is only one concept ID corresponding to a medical term. But since it is typical that more than one concept is retrieved for each medical term mentioned in a sentence, we need to cluster the concepts according to their shared words. We then apply a Hierarchical Agglomerative Clustering (HAC) algorithm, which is the most commonly used method for document clustering [7], and which does not require a prespecified number of clusters.

Once the clustering has been completed, we select a concept that shows the highest rank in each cluster and is within a threshold. We select the threshold dynamically based on the ranking score distribution, specifically choosing the point at which there is a significant drop in ranking scores which means the ratio of score[i]/score[i+1] is biggest.

B. Bootstrapping-based other named entity recognition

Temporality, location, and other named entities such as age, and sex also have various surface forms including misspelling, and arbitrary abbreviations. Given these considerations, we address the following question: How can the named entities having arbitrary different surface forms be automatically learned from the data with minimal effort using lexical and part-of-speech patterns?

Many successful methods have used an unsupervised iterative bootstrapping framework [11]. This kind of bootstrapping is often considered to be minimally supervised, as it is initialized with a small set of seed terms of the target category to extract. These seeds are used to identify patterns that can match the target category, which in turn can extract new patterns [12].

Starting from the original seed, each new pattern produced by the Damerau-Levenshtein distance algorithm can be considered an input seed for another instance of the algorithm. This procedure can be iterated over all the new patterns. We assume that the new patterns produced by the expansion of the original seed are the most semantically similar. Therefore, after one iteration, we stop. The stop criterion reduces the number of computations and guarantees a semantic similarity between the original seed and the new patterns. The final output of the bootstrapping process is the union without duplicates of all the new patterns that are evaluated as correct by the stop criterion.

The more specific description of our method is in the following: There are many types of temporal words. Since manually constructing seed patterns cannot cover all types, we use WordNet to retrieve all words related to time. We collect all the words of which semantic category is <noun.time> from WordNet, and then annotate the words in our data as "temporal noun (TEM)" if they are included in the <noun.time> category. We regard each temporal noun as a seed. We input each seed *s* to the bootstrapping algorithm and get the output of the seed variants *s'* if Damerau-Levenshtein distance(s, s') is the same with or less than a threshold. We set the threshold value as 2, except for the short words, where the length is the same as or less than 4, we set the threshold value as 1.

We have to check if the obtained new patterns are semantically same with the original seed. To do that, we use WordNet dictionary and part of speech tagging. If an obtained pattern word occurs in the WordNet dictionary with a different meaning or has a different part-of-speech (POS) tag from that of the seed word, then we consider the new pattern has a different meaning, and we filter them from the obtained pattern set. We use the POS-tag results of the Stanford parser (http://nlp.stanford.edu/software/lex-parser.shtml).

We extract temporal NP, sex, and age using the regular expressions of Figure 1. For the age, we use the pattern 'age'+num' and 'num-year-old', and for the sex, we choose the more frequently appearing category between 'man' and 'woman' in the data. We can get the noun's sex information from the definition of the noun ('man', 'woman', 'female', 'male', etc.) in WordNet, and we also have a small dictionary which includes commonly occurring terms such as 'he', 'she', 'him', 'her', 'his', 'son', 'daughter', 'husband', and 'wife'.

1. Temporality
($NUM)* ($temporal_noun)$^{+}$ ($NUM)*

($determiner) ($temporal_noun)$^{+}$

[$when|$time] ($any_word)* (NUM)

2. Age
[$age] ($any_word)* ($NUM)
($NUM) [$year] [$old]

3. Sex
Choose the more frequent category between 'woman' and 'man' category words

Figure 1. Regular expressions

III. REMOVING IMPLICIT NOISE

We want to extract only facts from patient information by removing implicit noise. We need to do two tasks: First is to extract information only for patients, not for their classmates, friends, or family members. Second is to detect facts from the patients' information. Details are given in the following subsections.

A. Extracting information only for patients

We need to know the subject of each named entity to remove named entities not associated with a particular patient. Based on the syntactic analysis, we determine the subject of a named entity, and filter it if the subject is not a patient.

We perform syntactic analysis using the Stanford parser, and detect the subject of each named entity. The subject should be a person. To improve the syntactic analysis results, we replace all misspelled words with the corrected results according to Section 2. We make the following two assumptions.

1. The most frequent subject indicates the patient.

2. If the parser does not explicitly indicate the subject of a named entity, then the nearest person noun is its subject.

B. Removing negations of patient information

We assess the factuality degree of events (whether they correspond to facts, counter-facts, or only concerns) based on heuristic patterns. We detect polarity based on the negation words such as "no", "deny", "not", "impossible", and "refuse". The boundary that the polarity is applied to is determined based on simple syntactic analysis.

Presence of a medical term in a clinical note does not necessarily imply its presence in a patient. The negation annotator looks at the surrounding text of each medical term annotation and filters term mentions found in negated contexts based on simple heuristics such as presence of negation related words noted above. We collect negation words, and then prune negations based on the negation words and detect the boundaries of negation based on parsing. Negation words can either be adjectives or verbs. The two cases show different kinds of syntactic graphs, and we need different rules for each case.

Table 1 Performance of our system

	Precision	Recall	F-measure
Biomedical term detection	0.6684	0.9345	0.7794
Temporality	0.8839	0.9010	0.8924
Age	0.9286	0.8529	0.8891
Sex	0.9519	0.8895	0.9197
Factuality Assessment	0.9386	0.8736	0.9049

Table 2. Comparison of our biomedical term detection system with others

	Precision	Recall	F-measure
Our Biomedical term detection	0.6684	0.9345	0.7794
MetaMap [1]	0.7882	0.4867	0.6017

If the POS-tag of a negation word is an adjective, its main syntactic function is to modify the following noun. Therefore, the boundary of negation includes its governor, which is the following noun, and all the children/descendant nodes of the governor.

If the negation word is a verb, its main function is to be a governor of its modifiers. In this case, we just include its children/descendant nodes within a negation boundary.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

We evaluated our system's performance on our HealthLink data, which consists of 3,328 sentences for 200 patients. We constructed gold standard dataset by manually annotating seven named entities in the data (biomedical, age, sex, temporality, travel, negation, and concern). We obtained the following experimental results.

1. Our proposed method in biomedical term detection achieved 77.94% in F-measure (see Table 1).

2. Our method significantly outperformed MetaMap [1] by 17% (see Table 2).

We measure the performance of our system based on precision and recall as follows: Precision = (the number of correctly detected terms)/(the number of all detected terms), and Recall = (the number of correctly detected terms)/(the total number of existing terms in the data). F-measure is the harmonic mean of precision and recall, and computed by (2*precision*recall)/(precision+recall).

As shown in Table 1, among those systems using the same evaluation data, our system outperformed all previous reported systems with a precision of 66.84%, recall of 93.45%, and F-measure of 77.94%. We compare our method with MetaMap [1] (http://metamap.nlm.nih.gov/), a tool developed at the National Library of Medicine, for mapping raw English text to standardized medical concepts in the UMLS metathesaurus. The precision, recall, and F-measure represent proportions of populations. In trying to determine the difference in performance of two systems, we employ the z-test on two proportions. The z-test with alpha=0.05 showed that our method significantly outperformed the previous method.

V. RELATED WORK

Biomedical term detection has been studied extensively in recent years, including the mapping of text phrases to UMLS concepts [1]. In previous work, an approach of A. Jimeno [5] is based on the identification of weighted words that compose terms denoting ontology concepts. They integrate two new aspects in their scoring method: the proximity between words in text and the amount of information carried by each individual word, and they do not use any threshold methods to choose relevant concepts among the ranked concepts. They also do not consider the noise in the data.

Some other machine learning approaches have also been investigated. H.W. Chun et al. [3] used a maximum entropy-based method to filter candidate disease names found by dictionary-based methods. M. Bundschus et al. [2] tried cascaded conditional random fields (CRF) using various features based on contexts, dictionary and orthogonal form to detect disease terms and the functional relations between them, but they need annotated data for training.

Among the few systems in the medical domain that treat time expressions, the study by Denny et al. [8] is most relevant to our work. They proposed timing and status descriptors for colonoscopy testing data. While they used the KnowledgeMap concept identifier to extract colonoscopy concepts, they developed a rule-based method with regular expressions to extract time descriptors and then normalized them. But they rely on meticulous manual rule writing. For pattern learning, many previous studies have suggested bootstrap-based pattern learning [9-10] in a variety of applications.

VI. CONCLUSION

We propose a method and system for patient information extraction from noisy tele-health records. When we retrieve the patient-related valuable information from the noisy data, there are two kinds of noise that we have to remove: First is explicit noise which includes spelling errors, unfinished sentences, omission of sentence mark, etc. Second is implicit noise, which includes non-patient's information and patient's untrustworthy information. To remove explicit noise, we propose our biomedical term detection/normalization method which deals with misspelling, imperfectness, and arbitrary abbreviation by nurses. In detecting temporal named entity and other types of named entities which shows patients' personal information such as age, and sex, we propose bootstrapping-based pattern learning to detect all kinds of arbitrary variations of the named entities. To identify and remove implicit noise, we propose a dependency path-based filtering method. Finally, we obtain normalized patient information. For the biomedical term detection, we use our own unsupervised method using a simple language model and the Damerau-Levenshtein distance. We also presented a temporality detection system that provides a practical and extensible state-of-the-art system for extracting time expressions. In addition, we introduce our regular expression patterns to detect other types of named entities. Our system is useful for experts to mine patient information in the city and analyze trends of patients' concerns/symptoms, because all the retrieved terms are cleaned by detecting noise.

ACKNOWLEDGEMENTS

This research was supported by the Alberta Innovates Centre for Machine Learning (AICML) and the iCORE division of Alberta Innovates Technology Futures.

REFERENCES

[1] A.R Aronson, "Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program". Proc. AMIA Symp, pp.17-21, 2001

[2] M. Bundschus, M. Dejori, M. Stetter, V. Tresp, H.P. Kriegel, "Extraction of semantic biomedical relations from text using conditional random fields", BMC Bioinformatics, Apr 23;9:207, 2008

[3] H.W. Chun, Y. Tsuruoka, J.D. Kim, R. Shiba, N. Nagata, T. Hishiki, and J. Tsujii, "Extraction of gene-disease relations from Medline using domain dictionaries and machine learning", Pac Symp Biocomput, pp. 4-15, 2006

[4] M. Dai, "An Efficient Solution for Mapping Free Text to Ontology Terms". AMIA Summit on Translational Bioinformatics. San Francisco, CA, 2008

[5] A. Jimeno, E. Jimenez-Ruiz, V. Lee, S. Gaudan, R. Berlanga, and D. Rebholz-Schuhmann, "Assessment of disease named entity recognition on a corpus of annotated sentences", BMC bioinformatics, 9 Suppl 3():S3, 2008

[6] Jay M. Ponte and W. Bruce Croft, "A Language Modeling Approach to Information Retrieval", Proc. of ACM SIGIR conference on Research and development in information retrieval, pp.206-214, 1998

[7] P. Willet, "Recent trends in hierarchical document clustering: a critical review". Information Processing and Management, Vol.24, pp.577-597, 1988.

[8] J.C. Denny, J.F. Peterson, N.N. Choma, H. Xu, R.A. Miller, L. Bastarache, N.B. Peterson, "Extracting timing and status descriptors for colonoscopy testing from electronic medical records", J Am Med Inform Assoc, 17(4):383-8, 2010

[9] T. Hao, Bootstrap-based equivalent pattern learning for collaborative question answering. LNCS 7182, pp.318-329, 2012

[10] N. Nakashole, M. Theobald, and G. Weikum, Find Your Advisor: Robust Knowledge Gathering from the Web, WebDB, 2010

[11] Ellen Riloff and Jessica Shepherd. A corpus-based approach for building semantic lexicons. In Proceedings of the Second Conference on Empirical Methods in Natural Language Processing, pages 117–124, Providence, RI, USA. 1997

[12] T. McIntosh, Unsupervised discovery of negative categories in lexicon bootstrapping, EMNLP, pp.356-365, 2010

[13] F. J. Damerau, A technique for computer detection and correction of spelling errors, Communications of the ACM, Vol.7 Issue 3, pp.171-176, 1964