ThinkMiners: Disorder Recognition using Conditional Random Fields and Distributional Semantics

Ankur Parikh **Avinesh PVS** **Joy Mustafi** **Lalit Agarwalla** **Ashish Mungi**
IBM India Pvt Ltd, IBM Software Group, Watson
{anparikh,pavinesh,jmustafi,lalit.agarwalla,r1amungi}@in.ibm.com

Abstract

In 2014, SemEval organized multiple challenges on natural language processing and information retrieval. One of the task was analysis of the clinical text. This challenge is further divided into two tasks. The task A of the challenge was to extract disorder mention spans in the clinical text and the task B was to map each of the disorder mentions to a unique Unified Medical Language System Concept Unique Identifier. We participated in the task A and developed a clinical disorder recognition system. The proposed system consists of a Conditional Random Fields based approach to recognize disorder entities. The SemEval challenge organizers manually annotated disorder entities in 298 clinical notes, of which 199 notes were used for training and 99 for development. On the test data, our system achieved the F-measure of 0.844 for entity recognition in relaxed and 0.689 in strict evaluation.

Keywords: medical language processing, clinical concept extraction, conditional random fields.

1 Introduction

Mining concepts from the electronic medical records such as clinical reports, discharge summaries as well as large number of doctor's notes has become an utmost important task for automatic analysis in the medical domain. Identification and mapping of the concepts like symptoms, disorders, surgical procedures, body sites to a normalized standards are usually the first steps towards understanding natural language text in the medical records.

In this paper, we describe a machine learning based disorder recognition system for the Task 7A of 2014 SemEval challenge. In Section 2 we give a background of the existing solutions to tackle the problem. Section 3 covers our approach in detail, followed by evaluation and conclusion in Section 4 and Section 5 respectively.

2 Background

In recent times, many systems have been developed to extract clinical concepts from various types of clinical notes. The earlier natural language processing (NLP) systems were mainly built heavily using domain knowledge i.e. medical dictionaries. These systems include MetaMap (Aronson and Lang, 2010), HiTEX (Zeng et al., 2006), KnowledgeMap (Denny et al., 2003), MedLEE (Friedman et al., 1994), SymText (Koehler, 1994) and Mplus (Christensen et al., 2002). In the past couple of years, researchers have been exploring the use of machine learning algorithms in the clinical concept detection. To promote the research in this field many organizations such as ShARe/CLEF, SemEval have organized a few clinical NLP challenges. In CLEF 2013 (Pradhan et al., 2013), the challenge was to recognize medication-related concepts. Both rule-based (Fan et al., 2013; Ramanan et al., 2013; Wang and Akella, 2013) and machine learning based methods as well as hybrid methods (Xia et al., 2013; Osborne et al., 2013; Hervas et al., 2013) were developed. In this shared-task sequential labeling algorithms (i.e., Conditional Random Fields (CRF)) (Gung, 2013; Patrick et al., 2013; Bodnari et al., 2013; Zuccon et al., 2013) and machine learning methods (i.e., Support Vector Machine (SVM)) (Cogley et al., 2013) have been demonstrated to achieve promising performance when provided with a large annotated corpus for

Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 652–656,
Dublin, Ireland, August 23-24, 2014.

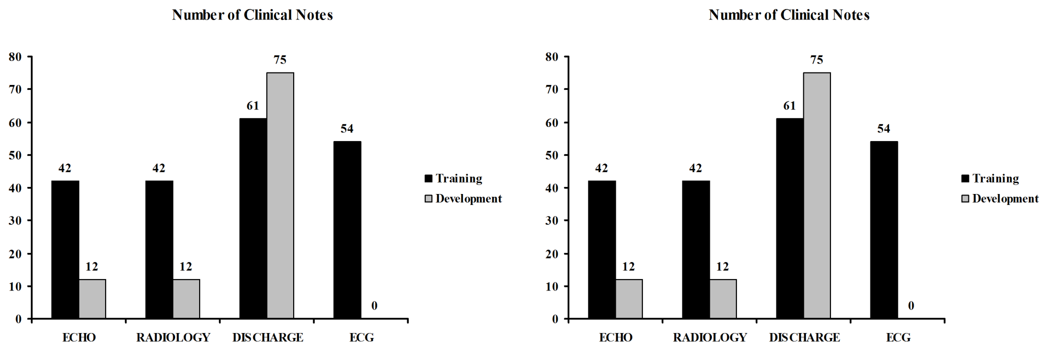

Figure 1: Dataset distribution

training.

3 Approach

Entity recognition has been tried in various domains like news articles, Wikipedia, sports articles, financial reports and clinical texts. In clinical text, entities can vary from medical procedures, disorders, body site indicators etc. Clinical text also presents with a peculiar concept of disjoint disorders/entities. This phenomenon is common in clinical domain compared to others and further complicates entity extraction from clinical notes.

3.1 Data

The data consisted of around 298 notes from different clinical types including radiology reports, discharge summaries, ECG and ECHO reports. For each note, disorder entities were annotated based on a pre-defined guidelines. The data set was further divided into two, with 199 notes in the training set and 99 notes in the development set. The training set contains 5811 disorders where as the development contained 5340 disorders. Figure 1 shows the distribution of the training and development set respectively.

3.2 Data Preprocessing

In the pre-processing step we tokenized, lemmatized and tagged the text with part of speech using the Apache cTAKES[1] (Savova et al., 2010). Further, section and source meta data extraction is done for the text in the documents.

In Named Entity Recognition (NER), when solved using machine learning, the text is typically converted to BIO format (Beginning, Inside and Outside the entity). BIO representation means the words in the text are assigned one of the following tags B - begin, I - inside and O - outside of the entity i.e. in this case a disorder. So now the task of NER is a sequence labeling problem to assign the labels to the tokens. Especially in the medical domain, the challenge is more complicated due to the presence of disjoint disorders (<10%), which could not be solved using the traditional BIO-notation. BIO approach works well with entities which are consecutive. So, we took an enhanced approach (Tang et al., 2013a) where the consecutive disorders are assigned traditional BIO tags and for disjoint disorders we create two tag sets a) D{B,I} : for disjoint entity words which are not shared by multiple concepts; and b) H{B,I}: for disjoint entity words which belong to more than one disjoint concept.

The following examples show the annotations of consecutive as well as disjoint disorders.

1: “The **left atrium** is moderately **dilated**.”
“The/O left/DB atrium/DI is/O moderately/O dilated/DB ./O”

2: “The **left** & **right atrium** are moderately **dilated**.”
“The/O left/DB &/O right/DB atrium/HB are/O moderately/O dilated/HB ./O”

3.3 Sequence Labeling

We have used Conditional Random Fields (CRF), a popular approach to solve sequence labeling tasks. CRF++[2] was used as an implementation of CRF for our purpose.

[1] https://ctakes.apache.org/

[2] http://crfpp.googlecode.com/svn/trunk/doc/index.html

Feature set used for the learning algorithm:

- **Word level features**: words [-2,2], suffix and prefix.
- **Syntactic features**: parts-of-speech(POS).
- **Discourse features**: source & section. Sentence containing disorder mentions usually have similar syntactic patterns based on sections (ex: 'Past Medical History') and source type (ex: discharge summary, radiology report). To capture this, source and section meta data have been provided as a feature.
- **Distributional semantics**: We used a contextual similarity based approach from the popular concept called NC-value (Frantzi et al., 2000).

  We followed the following steps to encapsulate the distributional semantics into the learning model:

  - For all the disorders in the training data we created two sets of contextual words namely context before (CB$_a train$) and context after (CA$_a train$). These words belong to open class (Noun, Verb, Adjective, Adverb) allocated for each section (S$_j$).
  - Weights are calculated for the contextual words.

    Weight(b$_{train}$) = $\frac{freq(disorders,b)}{freq(disorders)}$

  - For each word in the test data we created a similar sets of contextual words(CB$_a$, CA$_a$) as above.
  - Two scores are calculated for each token based on the product of frequency of the contextual word per section S$_j$ with weight calculated of that word in the training set.

    For each section (S$_j$):

    $$NC{-}value_B(a) = \sum_{b \epsilon CB_a, S_j} f_a(b_{test}) * weight(b_{train}) \quad (1)$$

    $$NC{-}value_A(a) = \sum_{b \epsilon CA_a, S_j} f_a(b_{test}) * weight(b_{train}) \quad (2)$$

    where
    a is the candidate term,
    CB$_a$ is the set of context words of "a" in a window of [-2,0],
    CA$_a$ is the set of context words of "a" in a window of [0,2],
    S$_j$ is a section like "Past Medical History", "Lab Reports" etc.
    b is a word from CB$_a$ or CA$_a$,
    f$_a$(b$_{test}$) is the frequency of b as a term context word of "a" in the test set,
    weight(b$_{train}$) is the weight of b as term context word of a disorder in the training set,
    NC-value$_B$(a) is the distributional semantic score of contextual words ***before*** the candidate term,
    NC-value$_A$(a) is the distributional semantic score of contextual words ***after*** the candidate term.
  - Further a similarity class is calculated based on a set of thresholds on the NC-value namely High-Sim, Med-Sim, Low-Sim and assigned to the tokens.

Most of the features were similar to that of the previous approaches (Tang et al., 2013a; Tang et al., 2012; Tang et al., 2013; Jiang et al., 2011) with an addition of an innovative distributional semantics based features (Nc-value$_B$, NC-value$_A$), which we have tried and tested for concept mining in clinical text.

4 Evaluation

The evaluation was done in two categories a) strict evaluation: exact match, which requires the starting and ending of the concept to be the same as the gold standard data b) relaxed evaluation: here the concepts don't match exactly with the start and end of the concept but may overlap.

In the strict and relaxed evaluation, the best F-measure among our system was 0.689, 0.844 without the distributional semantics where as best Precision was 0.907, 0.749 with the distributional semantics as a feature. Table 1. shows the detailed result.

5 Conclusion

Extraction of the concepts from the medical text is the fundamental task in the process of analysing patient data. In this paper we have tried a CRF based approach to mine the disorder terms from the clinical free text. We have tried various word

SemEval-2014 Shared Task 7A	**Strict**			**Relaxed**		
	Precision	Recall	F-measure	Precision	Recall	F-measure
Disorder Recognition without Distributional Semantics Feature	0.734	0.65	0.689	0.892	0.802	0.844
Disorder Recognition with Distributional Semantics Feature	0.749	0.617	0.677	0.907	0.758	0.826

Table 1: Results of the system on test set

level, syntactic, discourse and distributional semantic based features as adapted to the medical domain.

We have observed an increase (+1.5%) in precision but a drastic fall (-4.4%) in recall while using the distributional semantic feature. Ideally this feature has to improve the results because it takes contextual features into consideration. In our opinion inappropriate scaling of the feature values might have caused the drop. Further we would like to investigate the use of large unlabeled data, dependency tree based context and more experiments have to be carried out like threshold setting, feature value scaling to show better results. Also due to license issues we could not use UMLS dictionary. From our survey we figured out that 2-3% of improvement has been observed when the concepts from the dictionary are used.

References

B. Tang, H. Cao, Y. Wu, M. Jiang, and H. Xu. 2013. *Recognizing clinical entities in hospital discharge summaries using Structural Support Vector Machines with word representation features*. BMC Med Inform Decis Mak, vol. 13 Suppl 1, p. S1.

M. Jiang, Y. Chen, M. Liu, S. T. Rosenbloom, S. Mani, J. C. Denny, and H. Xu. 2011. *A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries*. J Am Med Inform Assoc, vol. 18, no. 5, pp. 601606.

B. Tang, Y. Wu, M. Jiang, Y. Chen, J. C. Denny, and H. Xu. 2013. *A hybrid system for temporal information extraction from clinical text.* J Am Med Inform Assoc.

B. Tang, H. Cao, Y. Wu, M. Jiang, and H. Xu. 2012. *Clinical entity recognition using structural support vector machines with rich features*. in Proceedings of the ACM sixth international workshop on Data and text mining in biomedical informatics, New York, NY, USA, pp. 1320.

C. Friedman, P. O. Alderson, J. H. Austin, J. J. Cimino, and S. B. Johnson. 1994. *A general natural-language text processor for clinical radiology*. J Am Med Inform Assoc, vol. 1, no. 2, pp. 161174.

S. B. Koehler. 1994. *SymText: a natural language understanding system for encoding free text medical data*. University of Utah.

L. M. Christensen, P. J. Haug, and M. Fiszman. 2002. *MPLUS: a probabilistic medical language understanding system*. in Proceedings of the ACL-02 workshop on Natural language processing in the biomedical domain - Volume 3, Stroudsburg, PA, USA, pp. 2936.

J. C. Denny, P. R. Irani, F. H. Wehbe, J. D. Smithers, and A. Spickard. 2003. *The KnowledgeMap Project: Development of a Concept-Based Medical School Curriculum Database*. AMIA Annu Symp Proc, vol. 2003, pp. 195199.

G. K. Savova, J. J. Masanz, P. V. Ogren, J. Zheng, S. Sohn, K. C. Kipper Schuler, and C. G. Chute. 2010. *Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications*. J Am Med Inform Assoc, vol. 17, no. 5, pp. 507513.

Q. T. Zeng, S. Goryachev, S. Weiss, M. Sordo, S. N. Murphy, and R. Lazarus. 2006. *Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system*. BMC Med Inform Decis Mak, vol. 6, p. 30.

A. R. Aronson and F. M. Lang. 2010. *An overview of MetaMap: historical perspective and recent advances*. J Am Med Inform Assoc, vol. 17, no. 3, pp. 229236.

. Uzuner, I. Solti, and E. Cadag. 2010. *Extracting medication information from clinical text*. J Am Med Inform Assoc, vol. 17, no. 5, pp. 514518.

Katerina Frantzi, Sophia Ananiadou, and Hideki Mima 2000. *Automatic recognition of multi-word terms:. the C-value/NC-value method.* International Journal on Digital Libraries 3(2):115–130.

James Cogley, Nicola Stokes and Joe Carthy. 2013. *Medical Disorder Recognition with Structural Support Vector Machines*. Online Working Notes of the CLEF 2013 Evaluation Labs and Workshop, 23 - 26 September, Valencia - Spain.

Robert Leaman, Ritu Khare and Zhiyong Lu. 2013. *NCBI at 2013 ShARe/CLEF eHealth Shared Task: Disorder Normalization in Clinical Notes with Dnorm*. Online Working Notes of the CLEF 2013 Evaluation Labs and Workshop, 23 - 26 September, Valencia - Spain.

James Gung. 2013. *Using Relations for Identification and Normalization of Disorders: Team CLEAR in the ShARe/CLEF 2013 eHealth Evaluation Lab*. Online Working Notes of the CLEF 2013 Evaluation Labs and Workshop, 23 - 26 September, Valencia - Spain.

Hongfang Liu, Kavishwar Wagholikar, Siddhartha Jonnalagadda and Sunghwan Sohn. 2013. *Integrated cTAKES for Concept Mention Detection and Normalization*. Online Working Notes of the CLEF 2013 Evaluation Labs and Workshop, 23 - 26 September, Valencia - Spain.

Jon D. Patrick, Leila Safari and Ying Ou. 2013. *ShARe/CLEF eHealth 2013 Named Entity Recognition and Normalization of Disorders Challenge*. Online Working Notes of the CLEF 2013 Evaluation Labs and Workshop, 23 - 26 September, Valencia - Spain.

Andreea Bodnari, Louise Deleger, Thomas Lavergne, Aurelie Neveol and Pierre Zweigenbaum. 2013. *A Supervised Named-Entity Extraction System for Medical Text*. Online Working Notes of the CLEF 2013 Evaluation Labs and Workshop, 23 - 26 September, Valencia - Spain.

Guido Zuccon, Alexander Holloway, Bevan Koopman and Anthony Nguyen. 2013. *Identify Disorders in Health Records using Conditional Random Fields and Metamap AEHRC at ShARe/CLEF 2013 eHealth Evaluation Lab Task 1*. Online Working Notes of the CLEF 2013 Evaluation Labs and Workshop, 23 - 26 September, Valencia - Spain.

Jung-wei Fan, Navdeep Sood and Yang Huang. 2013. *Disorder Concept Identification from Clinical Notes An Experience with the ShARe/CLEF 2013 Challenge*. Online Working Notes of the CLEF 2013 Evaluation Labs and Workshop, 23 - 26 September, Valencia - Spain.

S. V. Ramanan, Shereen Broido and P. Senthil Nathan. 2013. *Performance of a multi-class biomedical tagger on clinical records*. Online Working Notes of the CLEF 2013 Evaluation Labs and Workshop, 23 - 26 September, Valencia - Spain.

Chunye Wang and Ramakrishna Akella. 2013. *Performance of a multi-class biomedical tagger on clinical records*. Online Working Notes of the CLEF 2013 Evaluation Labs and Workshop, 23 - 26 September, Valencia - Spain.

Yunqing Xia, Xiaoshi Zhong, Peng Liu, Cheng Tan, Sen Na, Qinan Hu and Yaohai Huang. 2013. *Combining MetaMap and cTAKES in Disorder Recognition: THCIB at CLEF eHealth Lab 2013 Task 1*. Online Working Notes of the CLEF 2013 Evaluation Labs and Workshop, 23 - 26 September, Valencia - Spain.

John David Osborne, Binod Gyawali and Thamar Solorio. 2013. *Evaluation of YTEX and MetaMap for clinical concept recognition*. Online Working Notes of the CLEF 2013 Evaluation Labs and Workshop, 23 - 26 September, Valencia - Spain.

Lucia Hervas, Victor Martinez, Irene Sanchez and Alberto Diaz. 2013. *UCM at CLEF eHealth 2013 Shared Task1*. Online Working Notes of the CLEF 2013 Evaluation Labs and Workshop, 23 - 26 September, Valencia - Spain.

Sameer Pradhan, Noemie Elhadad, Brett R. South, David Martinez, Lee Christensen, Amy Vogel, Hanna Suominen, Wendy W. Chapman and Guergana Savova. 2013. *Task 1: ShARe/CLEF eHealth Evaluation Lab 2013*. Online Working Notes of the CLEF 2013 Evaluation Labs and Workshop, 23 - 26 September, Valencia - Spain.