Relation Extraction

Nandakumar, Pushpak Bhattacharyya
Indian Institute of Technology, Bombay
{nandu, pb}@cse.iitb.ac.in

2-Jul-2016

Abstract : Many applications in information extraction, natural language understanding, information retrieval require an understanding of the semantic relations between entities. We present a comprehensive review of various aspects of the entity relation extraction task. We also cover relation extraction task in the medical domain and various challenges and useful resources available in the medical domain.

1 Introduction

There is a large amount of unstructured data available in everywhere in various forms in different domains. It will be very useful if we can extract the information from these documents. Relation extraction is one of the major task for extracting structured data.

In this survey we discuss five major class of approaches of relation extraction. Starts from the very simple rule-based approach with help of hand built features then will move to more complex supervised such as feature-based convolutional net based approaches and also cover semi-supervised systems, which helps in situations where less training data is available.

2 Hand-built patterns

Hand built rule-based systems have a naive approach towards relation extraction task. By analyzing set of sample examples and written a possible set of rules which obey it.

for example,

Agar is a substance prepared from a mixture of *red algae*, such as *Gelidium*, for laboratory or industrial use.

While reading this sentence human can predict, there is a hyponym relation between *red algae* and *Gelidium*. We predict this by observing the connecting words *'such as'* between these two words. For identifying relation such as hyponyms [Hearst [10]] we can use this technique. We can add more and more these kinds of rules (Y such as X, such Y as X, such Y as X, X, and other Y, Y including X, Y especially X) to capture different possibilities. Using all kinds of possible rules hyponym extraction has a performance of 66% accuracy Hearst [10].

The main issue with this approach is, it is too hard to write all set of possible rules. We have to update the set of rules once we find an exception to the existing rules. And more importantly, we have to redo this whole work for the other kind of relation extraction. For example, if we have to extract meronyms extraction, then the possible set of rules will be totally different. The performance of this kind of system will depend on how many rules we covered and will depend on the type of relation we have to extract.

Pattern based relation extraction is also applicable in the medical domain. Rule based system, for extraction of treatment relations, Abacha and Zweigenbaum [1] between a treatment (e.g. medication) and a problem (e.g. disease) obtain 75.72% precision and 60.46% recall.

3 Supervised methods

Supervised systems are the state-of-the-art system for many natural language processing tasks. The supervised system will learn from the already tagged corpus with the help of features (carefully designed by experts in the specific domain) of or learn by the systems. The key idea for the supervised learning is to model the relation extraction task as a classification problem (Binary, or multi-class) and train the classifier with different techniques for prediction of new relations. We take help from various available machine learning algorithm for design the classification problem.

For the sake of simplicity and clarity, we restrict our discussion to binary relations between two entities. Given a sentence $S = w_1, w_2, .., e_1.., w_j, ..e_2, .., w_n$,

where e_1 and e_2 are the entities, a mapping function f (.) can be given as

$$f_R(T(S)) = \begin{cases} \text{True}, & \text{if } e_1 \text{ and } e_2 \text{ are related according to relation } R \\ \text{False}, & \text{Otherwise.} \end{cases}$$

Where $T(S)$ are the features extracted from the sentence S. The mapping function $f(.)$ defines whether there exist a relation between these entities. If a labeled set of positive and negative relation examples are available for training, the function $f(.)$ can be constructed as a discriminative

classifiers like Support Vector Machines (SVMs) Hong [11] or any other classifier. These classifiers can be trained using a set of features selected after performing textual analysis (like POS tagging, dependency parsing, etc) of the labeled sentences, which we discuss in section 3.1. On the other hand, input to the classifiers can also take the form of rich structural representations like parse trees. Depending on the nature of inputs to the classifier training, known as kernel methods. Apart from these methods, we can design a neural network which will represent the words using words vectors and positional features and will learn the most dominant features using neural networks and classifier will run based on these features, will cover in section 3.2.

3.1 Feature Based Approach

Feature-based methods rely on classification models for automatically specifying the category of relation exist between entities, based on relevant features vector belongs Chan and Roth [4]. For that, surrounding contextual features Kambhatla [12] are used to identify semantic relations between the two entities in a specific sentence and represent them as a feature vector. Features will define the relation existence can be in a different domain such as lexical, syntactic, dependency related or word embedding Gormley et al. [8]. Some features are good indicators of entity relations while others are not and it is important to select only those features which are relevant to the task, and thus the feature engineering is a key factor in this approach.

3.1.1 Lexical Features

Lexical and context-based features, like words between entities, nearby words of each entity within a window size etc can be added as a feature for relation extraction problem. Apart from these we can add bag-of-words

3.1.2 Syntactic Tree Features

The syntactic tree represents the grammatical structure of a sentence. We parse the sentences using the syntactic parser and extract the labels of parse tree constituent that exactly covers the mention, and also labels of all constituents that cover the mention. These will include the POS tags, Chunk head of entity, the syntactic path between the entities, head word of entity etc.

3.1.3 Dependency Features

The words and part-of-speech and chunk labels of the words on which the mentions are dependent in the dependency tree derived from the syntactic parse tree. We can identify the related entities clearly from the dependency tree. This insight will help to add dependency features in the relation extraction task Culotta and Sorensen [6]. Dependency feature will consist of dependency path in the dependency tree, labels, length of dependency tree etc can be added as features for this task.

3.1.4 Entity Features

We are identifying the relation between entities. So the type of relation between the entities will change according to the entity type we are considering. The entity types will help to classify the type of relation exist between them.

3.1.5 Word Embedding

The lexical feature will give a good representation of sentence with entities we are focusing. An alternative approach to capturing lexical information relies on continuous word embedding Mikolov et al. [14] as representative of words. Word Embedding features have improved many tasks, including Named Entity Recognition, Chunking, Dependency parsing, semantic role labeling, and relation extraction. Embedding can capture lexical information, but alone they are insufficient in state-of-the-art systems, they are used alongside features of the broader linguistic context Gormley et al. [8].

3.2 Convolutional Neural Network Based Approach

Till recent past relation extraction systems have made extensive use of features generated by various natural language processing modules (includes Syntactic and dependency parse information, lexical information, entity details etc). Errors in these features will propagate to next level and lead to errors of relation detection and classification. Here we depart from these traditional approaches with complicated feature engineering by introducing a convolutional neural network for relation extraction that automatically learns features from sentences and minimizes the dependence on external modules and resources Nguyen and Grishman [15].

The input to a convolutional neural network will be words represented by word embedding Collobert et al. [5] and positional features based on the relative distance from the mentioning entities. So there will not be any dependency to other natural language processing modules here. The convolutional layers will give a local correlation between features in the starting layers and will learn long distance features in the later layers. Each convolutional layer will consist of a convolution operation, which takes care of the local convolution of input, and a max pooling layer, which will cut the input dimension without losing the dominant features and one nonlinear layer at the end. The nonlinear layer will transform input into a linearly separable space. Convolutional network shows promising results in the relation extraction in relation classification tasks.

The basic structure of CNN approach consists of three basic layers. One convolution layer, (includes, convolution, max pooling, and nonlinear layer), one non-linear transformation layer and a softmax layer for classification.

There are lots of hyperparameters which we need to tune for the best performance of the system, make this approach a difficult task. Word Embedding dimension, a number of units in the hidden layer, number of hidden layer. Convolution filter size all we need to tune for the best performance.

4 Semi Supervised

The main problem with supervised methods are, we need lots of tagged data for learning the classifier. If we don't have enough annotated data to train and lots and lots of unannotated text for relation extraction then, we will not get a good result. The solution for this approach is bootstrapping technique. In this approach, we have some seed instances, which is manually tagged data used for the first phase of training called the seed instances. We train with seed instances and learn the classifier, and test with the classifier, and get more train examples by adding the test results to the training set. Thus, the training set will grow up to a sufficient amount. This approach can be called as a semi-supervised model.

DIPRE (Duality of patterns and Relations) is such a system developed by brin in 1998, Brin [3] for an extracting author-book relationship from web text. An improved version of DIPRE, which will cut the noise by the newly added seed tuple, called snowball Agichtein and Gravano [2]. The workflow of the system as follows.

Algorithm

1. Search the whole text and extract the matching sentences, which consist the seed relation pairs.

2. Generate pattern from each matching sentence.

   Pattern

   $$< w_l - left >< entity_1 >< w_m - middle >< entity_2 >< w_r - right >$$

   where,

   $entity_1$: First Entity.

   $middle$: words between the entities in the sentence.

$entity_2$: Second Entity.

$Right$: Words after the second entity in a window of size k(k=2).

W_l : Weight vector for left terms.

W_m : Weight vector for middle terms.

W_r :Weight vector for right terms.

(Weight vector will calculate based on frequency of terms in the context)

3. Once create the pattern for each possible sentence, Undergo single pass clustering algorithm using match score, and form clusters of similar patterns. Each cluster is represented by the centroid of weight vectors (w'_l, w'_r, w'_s).

4. Extract new possible tuples from the text for which match the pattern and the similarity score is greater than the threshold value (Tsim). The similarity score is calculated as the dot product of weight vectors.

$$\begin{aligned} Similarity(t_s, t_p) &= w_{ls}.w_{lp} + w_{ms}.w_{mp} + w_{rs}.w_{rp} \ (\ ifthetagsarematching \ ) \\ &= 0 \ (\ otherwise \ ) \end{aligned} \tag{1}$$

5. Add a new set of tuple which have good confidence value and continue from step 2

   Confidence value calculated based number or of patterns which derive the tuple.

5 Relation Extraction in Medical Domain

There are many supervised and semi-supervised techniques exist for solving relation extraction problem. In medical domain also the impact of supervised approach is very significant. SVM based relation extraction systems show state of the art relation extraction system Roberts et al.

[16] using various syntactic, dependency, lexical (Grouin et al. [9], Frunza and Inkpen [7]) and domain knowledge Zhang et al. [17] from the existing systems like UMLS (Unified Medical Language System)[1].

Support vector based approaches are common among the most effective relation extraction systems. Since the medical data consist of a large number of concept pairs with no relations, some of the relation extraction systems do the task in two stages. First, separate related pairs of entities from all possible pair of entities and in the second stage, classify the type of the relation exist between them. Most of the systems used n-grams with specific semantics, hand-built linguistic patterns and made use of simplified representations of text. The state of the art system in relation extraction system, Roberts et al. [16] secured an F-score of 73.7 for the i2b2 data set.

6 Challenges

Medical documents such as prescription sheets, discharge documents etc. includes lots of challenge for the relation extraction task.

- **Nonstandard Structure** : Medical documents from different hospitals didn't follow any standard structure. It varies with different hospital and doctors.

- **Non Grammatical Language** : Medical document sentences are not well formed in the documents. Most of the sentence are very short and just mention the medicine name and usage. In some documents, doctors list all medicine one by one connected by multiple punctuations.

- **History Mention** : Most patient records consist of a patient history of problems and treatment he underwent. It is real hard task to identify the related entities from the history

[1] http://www.nlm.nih.gov/research/umls/

mentions, since some of old treatments are related to current treatment and some are not.

- **Negation** : Negation is a well known standard problem in the field of information extraction. It has a dominant role in the relation extraction task also. In some documents it is mentioned that the treatment we not done or not under treatment X etc. Such instances we should handle properly.

7 Resources

In the clinical domain, annotated corpora are not only expensive but also often unavailable for research due to patient privacy and confidentiality requirements. In 2010, i2b2[2] partnered with VA Salt Lake City Health Care System in manually annotating patient reports from three institutions and created a challenge in which the research community could participate in a head-to-head comparison of their systems. I2b2 (Informatics for Integrating Biology and the Bedside) is developing a scalable informatics framework that will enable clinical researchers to use existing clinical data for discovery research and combined with IRB-approved genomic data, facilitate the design of targeted therapies for individual patients with diseases having genetic origins. This platform currently includes wide international adoption by the CTSA network, academic health centers, and industry.

Medline (Medical Literature Analysis and Retrieval System Online)13 is a large repository of citations, abstracts, and publications from the biosciences and biomedical domain. PubMed is a freely available access point to the Medline repository. Full papers are also provided wherever available. Medline is a valuable resource for medical domain natural language data. Citations and abstracts serve as title keyword-pool whereas content of papers provides millions of documents of medical text. Medical vocabularies can be built based on this corpus. Medline is widely used in Biomedical Natural Language Processing research. Subsets of the Medline

[2] https://www.i2b2.org

dataset have been annotated at the sentence and token level to identify names of clinical entities, biomolecules, gene and protein names etc.

8 Conclusion

In this survey, we discuss different approaches existing in relation extraction task and its importance in natural language processing field. Till the recent past, tree kernel outperforms than the feature-based approaches in supervised approaches. But later neural net based approaches using word embedding and feature embedding with Wordnet and NER Gormley et al. [8] getting a state-of-the-art result on SemEval2010 (83 %) and ACE 2005 (74 %) data set. Recent works using the convolutional network with synonym coding Liu et al. [13] showing state of art result on ACE 2005 (83.8%) dataset and getting comparable result Gormley et al. [8] (82.8%) of the relation classification task. In medical domain, feature based state of art system Roberts et al. [16] obtained F-Score of 73.7 for relation extraction and classification task, in 2010i2b2/VA relation data set, leaving about a quarter of the relations in the corpus incorrectly classified. The difficulty of classifying these relations comes from the lack of explicit contextual information that describes the relations and/or the complexity of the language used in presenting the relations. While deeper syntactic analysis may help with the complex language, in the absence of context, domain knowledge may provide a better performance for relation extraction task.

References and Notes

1. Asma Ben Abacha and Pierre Zweigenbaum. Automatic extraction of semantic relations between medical entities: a rule based approach. *J. Biomedical Semantics*, 2(S-5):S4, 2011.

2. Eugene Agichtein and Luis Gravano. Snowball: Extracting relations from large plain-text

collections. In *Proceedings of the fifth ACM conference on Digital libraries*, pages 85–94. ACM, 2000.

3. Sergey Brin. Extracting patterns and relations from the world wide web. In *The World Wide Web and Databases*, pages 172–183. Springer, 1999.

4. Yee Seng Chan and Dan Roth. Exploiting syntactico-semantic structures for relation extraction. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 551–560. Association for Computational Linguistics, 2011.

5. Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *The Journal of Machine Learning Research*, 12:2493–2537, 2011.

6. Aron Culotta and Jeffrey Sorensen. Dependency tree kernels for relation extraction. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 423. Association for Computational Linguistics, 2004.

7. Oana Frunza and Diana Inkpen. Extracting relations between diseases, treatments, and tests from clinical data. In *Canadian Conference on Artificial Intelligence*, pages 140–145. Springer, 2011.

8. Matthew R Gormley, Mo Yu, and Mark Dredze. Improved relation extraction with feature-rich compositional embedding models. *arXiv preprint arXiv:1505.02419*, 2015.

9. Cyril Grouin, Asma Ben Abacha, Delphine Bernhard, Bruno Cartoni, Louise Deleger, Brigitte Grau, Anne-Laure Ligozat, Anne-Lyse Minard, Sophie Rosset, and Pierre Zweigenbaum. Caramba: concept, assertion, and relation annotation using machine-learning based approaches. In *i2b2 Medication Extraction Challenge Workshop*, 2010.

10. Marti A Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics-Volume 2*, pages 539–545. Association for Computational Linguistics, 1992.

11. Gumwon Hong. Relation extraction using support vector machine. In *Natural Language Processing–IJCNLP 2005*, pages 366–377. Springer, 2005.

12. Nanda Kambhatla. Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations. In *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*, page 22. Association for Computational Linguistics, 2004.

13. ChunYang Liu, WenBo Sun, WenHan Chao, and WanXiang Che. Convolution neural network for relation extraction. In *Advanced Data Mining and Applications*, pages 231–242. Springer, 2013.

14. Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

15. Thien Huu Nguyen and Ralph Grishman. Relation extraction: Perspective from convolutional neural networks. 2015.

16. Kirk Roberts, Bryan Rink, and Sanda Harabagiu. Extraction of medical concepts, assertions, and relations from discharge summaries for the fourth i2b2/va shared task. In *Proceedings of the 2010 i2b2/VA Workshop on Challenges in Natural Language Processing for Clinical Data. Boston, MA, USA: i2b2*, 2010.

17. Yaoyun Zhang, Ergin Soysal, Sungrim Moon, Jingqi Wang, Cui Tao, and Hua Xu. Integrating multiple on-line knowledge bases for disease-lab test relation extraction. *AMIA Summits on Translational Science Proceedings*, 2015:204, 2015.