IJCSE International Journal of Computer Sciences and Engineering Open Access
Survey Paper Volume-5, Issue-3 E-ISSN: 2347-2693

A Survey on Relation Classification from Unstructured Medical Text

S. Gupta[1*], A.K. Manjhvar[2]

[1*]Department of Computer Science and Engineering, Madhav Institute of Technology and Science, Gwalior
[2]Department of Computer Science and Engineering, Madhav Institute of Technology and Science, Gwalior

[*]*Corresponding Author: saumaya.may.09@gmail.com, Mob.: 7415367993*

Available online at: www.ijcseonline.org

Received:05/Mar/2017 Revised: 11/Mar/2017 Accepted: 21/Mar/2017 Published: 31/Mar/2017

***Abstract*— Medical documents are rich in information and such information can be useful in building many health applications. Since information in medical documents is often unstructured and in nonstandard natural language so it is difficult to collect and present this information in a structured way. Structured information can be present as named-entity in the text, relationship between clinical entities, summary of the text, etc. To get the specific information from the text, many rule based and machine learning techniques are widely used. In this paper, we present several existing techniques for relation classification from unstructured medical text. We focus on rule based approaches, feature based relation classification approaches and convolutional neural network based approach in context of relation classification from unstructured text. We will also discuss semi supervised approaches for the cases where tagged data set is not much available to train the classifier.**

Keywords—Data Mining, Relation Classification, Natural Language Processing

I. INTRODUCTION

Information extraction is the process of mining useful information from raw data using machine learning and natural language processing approaches. Information extraction includes extraction of named entities, relationship between named entities, collecting temporal information from text, and many such insights of data. In this process of information extraction, several challenges occur when applied to different domain and data sources. For example, information extraction from Facebook post, social media tweets and comments is challenging due to non-standard use of language. Named entities includes names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc [1]. In medical text, named entities represent medical terms e.g. medicine, treatment. Relation classification is a process to find whether a pair of medical entities are related or not [2]. A relation is defined in the form of a tuplet = (e_1, e_2, ..., e_n) where the e_i are entities in a predefined relation r within text document D. Most relation classification systems focus on extracting binary relations. Examples of binary relations include treats (crocin, fever). In this paper we will discuss about existing techniques which are useful in relation classification between named entity present in the raw medical text.

In this paper, we begin with rule based approaches which performs relation classification based on hand build patterns in section II. Section III contain the related work of supervised approaches which formulate the relation extraction task as a binary classification problem. Further, we discuss feature based, convolution neural network based approaches, semi supervised machine learning methods and kernel based methods of supervised relation classification. The major advantage of kernel methods is they offer efficient solutions that allow us to explore a large (often exponential, or in some cases, infinite) feature space in polynomial computational time, without the need to explicitly represent the features. More recently, semi-supervised and bootstrapping approaches have gained special attention. Section IV contain various measures to evaluate result on the classification approaches and data set available for relation classification. And in the last section V, we compare various approaches and their impact on relation classification task.

II. RULE BASED APPROACHES

In rule based approach, hand build patterns are identified by domain experts. These hand build patterns are designed to identify the relationship between named entities present in the text. This whole task is executed by analyzing set of sample examples and draw 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.

A human can predict there is a hyponym relation between *red algae* and *Gelidium* after a reading of the text. This is possible by observing the connecting words 'such as' between these two words. For identifying relation such as hyponyms this approach can be used. Many such rule based system exist in medical domain to capture relations between medical entities [3]. SemRep is one of them. It uses many such patterns for relation identification task [4]. For example,

X TREATS Y & Z OCCOURS_IN Y → X TREATS Y

This indicates, if in a sentence X is a medicine and it treats a problem Z. And Z is also occurring in some problem Y, then X can treat problem Y.

III. SUPERVISED MACHINE LEARNING APPROACHES

The supervised machine learning algorithms learn from the pre-tagged corpus with the help of a set of features. These features are carefully designed by domain experts. The idea behind the supervised learning is to model the relation classification task as a binary or multiclass classification problem. In this model a classifier is trained with different techniques for relation classification on pre-tagged data. A binary classification problem can be described with the example: Given a sentence S = w_1, w_2, e_1, w_3, ... e_2, …, w_n. Where e_1 and e_2 are medical entities present in the sentence S. A binary classifier predicts weather a relation R exist between e_1 and e_2 or not based on the features extracted from the sentence S. The examples for which relation exists are tagged as true instances and if relation does not exist then tagged as a false instance.

Classifiers like Support Vector Machines (SVMs) or any other classifier are used to classify relations [5]. These classifiers are trained using a feature set collected after textual analysis (like POS tagging, dependency parsing, etc) of the tagged sentences from the data. Classifiers can also take rich structural representations like parse trees as input while training and testing. There exist other methods, for which we can design a neural network which may represent the words using words vectors and positional features and will automatically learn the most important features [1] [6].

A. Feature Based Approach

Feature based approach depends on classification models for specifying the category of relation exist between entities. Classifier classifies the relation between entities based on relevant features vector which are extracted from text. Kambhatla has shown that contextual features can be used to identify semantic relations between two medical entities in a specific sentence and represents features as a feature vector [7]. Features for relation identification can be in a different domain such as lexical, syntactic, dependency related or word embedding properties [8].

- Lexical Features includes lexical and context-based features, for example, words between medical entities, words nearby each medical entity in the sentence. These features play a main role in relation classification. Other such features include bag of word approach. Consider the following sentence

  *she continues to have **fever** which is very well controlled with **crocin.***

  In this example, first entity is fever and second entity is crocin. There is a relation between both as crocin *treats* fever. To capture this relation, lexical feature for the sentence can be number of words between medical entities (5), words between entities (which is very well controlled with), word before first entity (have) and word after second entity (empty). Figure (1) shows the basic architecture of the feature based approach.

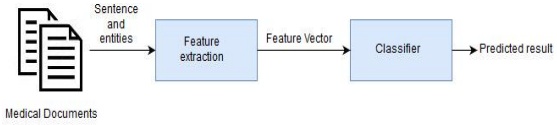


Figure 1. A Feature based relation classification model.

- Syntax tree Features includes syntax level information about mentioned pair of entities. For example, such feature can be part of speech tags and chunk head of the candidate entities. In our sentence, part speech tags for first entity and second entity are considered as features in feature vector. Chunk head in our example will be another feature which is NP for *fever*.

- Dependency tree has more information about the relation between the medical entities mentioned in the sentence. Dependency tree presents the grammatical relationship between the words in the sentence. For example, dependency tree for our sentence is shown in (2).

Figure 2. An example of dependency tree.

Features from dependency tree can be words in shortest path between entities, shortest path labels and shortest path length. In our example, feature value for shortest path between entities is controlled, shortest path labels are "rcmod prep_with" and shortest path length is 1.

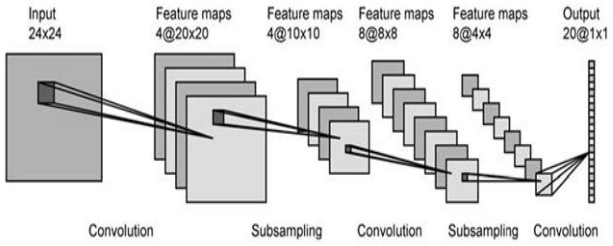
- Entity features includes entity level information which have a huge impact in relation classification task. These features captures other medical between mention pair of entities and number of such other medical entities.

- Word embedding features includes distance between word embedding of the entities and cosine similarity between medical entities. These features can be easily obtained by already available word embedding.

B. Convolutional Neural Network Based Approach

Feature based approaches are widely used in relation classification but feature engineering is a very complicated task. It's not easy to find useful features for relation classification task and it is very much depending on the domain. But in convolutional neural network methods, system learns important features from the text based on the positive and negative tagged data set. These systems also minimize the dependence on external modules and resources [9].

The input to a convolutional neural network (CNN) will be words represented by word embedding and positional features based on the relative distance from the mentioned entities. The layers in CNN gives a correlation between features in the initial layers and learns long distance features in the subsequent layers. As shown in (3), each CNN layer performs a convolution operation, which takes care of the local convolution of input, and a max pooling layer, which cuts the input dimension without losing the dominant features and one nonlinear layer at the end. The nonlinear layer transforms input into a linearly separable space. Convolutional network shows promising results in the relation classification task [9].



.

Figure 3. A Convolutional Neural Network architecture.

C. *Semi Supervised approaches*

In supervised methods, a lot of tagged data is required for training the classifier. If we don't have enough tagged data to train the classifier for relation classification then, a good result is not expected. The alternative approach to overcome this problem is bootstrapping technique. In this method, we have some seed instances, which is manually tagged data used for the first phase of training called the seed instances. We train the classifier with seed instances and test the classifier on remaining data, by this we get more train examples by adding the test results to the training set [10]. Thus, the training set grows up to a sufficient amount. This approach is called as a semi-supervised model.

D. *Kernel Methods*

The kernels used for relation-extraction (or relation-detection) are based on string-kernels described in Text classification using string kernels [11]. String kernels have been discussed in the context of text classification. However, an understanding of the workings of string-kernels is essential for interpreting the kernels used for relation classification. Given two strings a and b, the string-kernel computes their similarity based on the number of sub-sequences that are common to both of them. More the number of sub-sequences common, greater the similarity between the two strings. Each string can be mapped to a higher dimensional space where each dimension corresponds to the presence (weighted) or absence (0) of a particular sub-sequence. For example, a string a= **cat** can be expressed in a higher dimensional space of sub-sequences as follows:

$$\phi(a = cat) = [\phi a(x)..\phi c(x)..\phi t(x)..\phi at(x)..\phi ca(x)..\phi a(x)..\phi cat(x)]$$

$$= [\ \lambda..\ \lambda\ ..\ \lambda\ ..\ \lambda^2\ ..\ \lambda^2\ ..\ \lambda^2\ ..\ \lambda^3\ ..\ ]$$

Where $\lambda \in (0, 1]$ is a decay factor such that longer and non-contiguous sub-sequences are penalized. In fact, Φ_{ct}(cat) is penalized more (λ^3) than φ_{ca} (cat) (λ^2) and φ_{at}(cat) (λ^2) since ct occurs non-contiguously in **cat**.

Dataset	Domain	Language(s)	Year	Task(s)
MUC	Military and news Reports	English	1987 - 1997	Named Entity Extraction(NER)
ACE	Newswier, broadcast news, Speech transcript	English, arabic, chinese	1999-present	NER, Relation extraction, event detection
Medline	Medical Publication	English	1950-present	Protein-Protein Interaction and Gene Binding

Table 1. Data set available for relation classification evaluation.

Table (1) shows available data set for Relation classification.

IV. RESULTS AND DISCUSSION

We have seen various existing methods for relation classification and each one shows different results. To calculate accuracy of a method few parametric measures are done on the predicted result on the test data set. First classifier is trained with train data set and then following parameters are computed to evaluate accuracy of the method:

- True Positive (TP) – number of correctly predicted positive instances. That is, number of test cases for which classifier predicts true relationship for true relationship test cases.
- True Negative (TN) – number of correctly predicted negative instances. That is, number of test cases for which classifier predicts negative relationship for negative relationship test cases.
- False Positive (FP) – number of incorrect predicted instances as positive. That is, number of test cases for which classifier predicts true relationship for negative relationship test cases.
- False Negative (FN) – number of incorrect predicted instances as false. That is, number of test cases for which classifier predicts false relationship for true relationship test cases.
- Precision – It measures probability of correctly identified true test cases out of all predicted true instances.

$$\text{Precision} = \frac{TP}{TP + FP}$$

- Recall – It measures probability of probability of correctly predicted true test cases out of all actual true instances.

$$\text{Recall} = \frac{TP}{TP + FN}$$

- F-measure – It is the harmonic mean of Precision and Recall. Higher the F score shows well balanced classification system.

$$F = 2.\frac{precision.recall}{precision + recall}$$

Different relation classification models show different result on different data set. Results also various on the nature of relationship.

V. CONCLUSION and Future Scope

So far, we have seen all the aspects of the entity-relation classification problem starting with the algorithms, discussing the evaluation criteria finally culminating with a discussion of some important applications. Among the supervised approaches, dependency path methods stand out as the best both in terms of computational complexity and performance. Surprisingly the tree kernel has not been comparatively evaluated with other kernels which leaves room for speculation. It is clear that kernel methods outperform feature-based approaches for supervised relation extraction. Semi-supervised approaches seem to be well suited for open domain relation extraction systems since they can easily scale with the database size and can extend to new relations easily. Supervised approaches on the other hand can do well when the domain is more restricted like the case of bio-text mining. The problem of N-ary relation extraction is often factored in sets of binary relation-extraction problems which can be sub-optimal. It would be interesting to investigate approaches that handle higher order relations efficiently without factorizing them.

REFERENCES

[1] Collobert, Ronan, "*Natural language processing (almost) from scratch.*" Journal of Machine Learning Research, Vol. (12), pp.2493-2537, 2011.

[2] Bach N, Badaskar S. *"A review of relation extraction"*. Literature review for Language and Statistics II. 2007.

[3] Hearst, Marti A. "*Automatic acquisition of hyponyms from large text corpora.*" Proceedings of the 14th conference on Computational linguistics, Association for Computational Linguistics, Vol. (2), pp.539-545, 1992.

[4] Rindflesch, Thomas C., et al. "*Medical facts to support inferencing in natural language processing.*" AMIA. 2005.

[5] Hong, Gumwon. "*Relation extraction using support vector machine.*" In International Conference on Natural Language Processing, pp. 366-377, 2005.

[6] Nguyen, Thien Huu, and Ralph Grishman. *"Relation extraction: Perspective from convolutional neural networks."* In Proceedings of NAACL-HLT, pp. 39-48, 2015.

[7] Kambhatla, Nanda. "*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, pp. 22-23, 2004.

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

[9] Nguyen, Thien Huu, and Ralph Grishman. *"Relation extraction: Perspective from convolutional neural networks."* In Proceedings of NAACL-HLT, pp. 39-4,. 2015.

[10] Carlson, Andrew, et al. "*Toward an Architecture for Never-Ending Language Learning*." AAAI. Vol. (5), 2010.

[11] Lodhi, Huma, Craig Saunders, John Shawe-Taylor, Nello Cristianini, and Chris Watkins. "*Text classification using string kernels."* Journal of Machine Learning Research, Vol. (2), pp 419-444, 2002.

Authors Profile

Ms. Saumaya Gupta pursed Bachelor of Engineering from Poornima Institute of Engineering and Technology, Jaipur, India in 2011. She is currently pursuing Master of Technology in Computer Science and Engineering from Madhav Institute of Technology and Science, gwalior.

Mr Amit Kumar Manjhvar pursed Bachelor of Engineering from branch Computer Engineering and Master of Technology in Software System. He is currently working as Assistant Professor in Department of Computer Science Engineering, Madhav Institute of Technology and Science, Gwalior.

.

