Volume 5, Issue 11, November 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

An Overview of Entity Relation Extraction Techniques

Ashwini Zadgaonkar
Department of Computer Science & Engineering
RCOEM, India

Abstract*— Natural language processing requires deep understanding of semantic relationships between entities. This paper presents comprehensive review of various aspects of the Entity Relation Extraction task. Here, an attempt is made to cover in detail some of the important supervised and Semi-supervised classification approaches to the relation extraction task along with critical analyses. One of the most important aspects of Entity relation is higher-order relations. So there is a need of inventing approaches capable of handling higher order relations efficiently. Finally, in this paper some relation extraction applications like Question answering and Bio-text mining are also discuss.

***Keywords*— NLP, Supervised learning, Entity relation, Information Extraction, Binary Relation**

I. INTRODUCTION

In today's Era of WWW, vast amount of unstructured electronic text is available on the Web in various forms like newswires, blogs, email communications, governmental documents, chat logs, etc. For information Processing applications this large volume of information is analysed based on the desired interest .The major drawbacks observed over here are a) Large volume b) Heterogeneous nature of information. Effective solution to overcome these problems is to turn unstructured text into structured one by annotating semantic information of interest. Manual analysis of such large volume of data is practically impossible and new techniques are needed to automate this process. Relations between the entities can play an important role in text analysis. For task automation, machine needs to learn how to recognize a piece of text having a semantic property of interest in order to make a correct annotation. Extracting semantic relations between entities in natural language text is a crucial step towards natural language understanding. The problem of obtaining structured information from the text is dealt with Information Extraction (IE), a field of NLP. So this paper focuses on various techniques of recognizing relations between entities in unstructured text.

A relation is generally defined in the form of a tuple t = (e1 , e2 , ..., en) where the ei are entities in a predefined relation r within document D. Though Most of the relation extraction systems focus on binary relations, higher-order relations are equally important. In the sentence "*At codons 12, the occurrence of point mutations from G to T were observed*" exists a 4-ary biomedical relation. The biomedical relationship between a type of variation, its location, and the corresponding state change from an initial-state to an altered-state can be extracted as point mutation (codon, 12, G, T).

In this review, we will begin by discussing knowledge based methods along with their limitations followed by supervised approach which formulate the relation extraction task as a binary classification problem. Recently, Semi-supervised and Bootstrapping approach has got special recognition in the field. There exist higher-order relations extraction systems also. (McDonald et al., 2005). The novelty of (McDonald et al., 2005) such systems is to factorize complex relations into binary relations which are represented as a graph, and an algorithm to reconstruct complex relations by making tuples from selected maximal cliques in the graph.

II. KNOWLEDGE BASED METHODS

The Knowledge based Relation extraction methods are preferred for domain-specific tasks where texts are similar and a closed set of relations needs to be identified. Systems which use these methods generally rely on pattern-matching rules manually crafted for each domain (Riloff and Jones 1999; Pasca 2004). Still there are some exceptions where relations are domain-independent .Hearst (1992) describes the usage of lexico-syntactic patterns for extraction of hyponymy relations in an open domain. These patterns capture such hyponymy relations as between "author" and "Shakespeare", "wound" and "injury", "England" and "European country". However, the author notes that this method does not work well for all relations as same patterns cannot uniquely identify the given set of relation.

To summarize we can say that, knowledge-based methods are not easily portable to other domains and involve too much manual labour but they can be used effectively if the main aim is to get results quickly in well-defined domains and document collections.

III. SUPERVISED METHODS

Supervised methods for Relation Extraction rely on machine learning by using a training set of tagged domain-specific examples. Such systems automatically learn extractors for relations by adopting machine-learning techniques. The major disadvantage of this approach is that the development of a suitably tagged corpus can take a lot of time and efforts. But these systems can be portable to a different domain provided training data for that domain is available.

Some supervised systems use bootstrapping to make construction of the training data easier. These methods are referred as "*weakly- supervised information extraction*". Though bootstrapping appears quite promising , error propagation becomes a serious issue for this approach. Mistakes in extraction at the initial stages generate more mistakes at later stages and decrease the accuracy of the process. Another problem that can occur is that of semantic drift. This happens when multiple senses of the word are not taken into consideration and therefore each iteration results in a diversion move from the original meaning.

In general, supervised extraction methods have some limitations.

1. These methods are difficult to extend to new entity-relation types for want of labelled data.
2. Extensions to higher order entity relations are difficult.
3. They are relatively computationally burdensome and do not scale well with increasing amounts of input data.
4. M ost of the methods require pre-processed input data in the form of parse tree, dependency parse trees etc. Thus, the pre-processing stage is error prone and can affect the performance of the system

IV. SELF SUPERVISED SYSTEMS

Self-supervised systems attempts to make the process of information extraction completely unsupervised. The KnowItAll Web IE system (Etzioni et al. 2005), is an example of a self-supervised system. The Intelligence in Wikipedia (IWP) project (Weld et al. 2008) is another example of a self-supervised system. Self Supervised approach can be

A. Open Information Extraction

Etzioni et al. (2008) introduced the notion of Open Information Extraction, which is opposed to Traditional Relation Extraction. Open information extraction is "a novel extraction paradigm that tackles an unbounded number of relations". This method does not presuppose a predefined set of relations and is targeted at all relations that can be extracted. The Open Relation extraction approach is relatively a new one, so there is only a small amount of projects using it. A set of relation-independent lexico-syntactic patterns are used to build a relation-independent extraction model. It was found that 95 % of all relations in English can be described by only 8 general patterns, e.g. "E1 Verb E2 ". The input of such a system is a corpus and some relation-independent heuristics where relation names are not known in advance. These systems are able to extract instances of the four most frequently observed relation types: Verb, Noun+Prep, Verb+Prep and Infinitive". They are subjected to a number of limitations, which are common to all RE systems:

i) it extracts only explicitly expressed relations that are primarily word-based.

ii) relations should occur between entity names within the same sentence.

B. Distant Learning

Mintz et al. (2009) introduces a new term "*distant supervision*". The author use a large semantic database containing 7,300 relations between 9 million named entities. For each pair of entities that appears in relation, they identify all sentences containing those entities in a large unlabeled corpus. At the next step textual features to train a relation classifier are extracted. Even though the 67,6 % of precision achieved ,this method has a wide scope for improvement, it has inspired many researchers to further investigate in this direction. Currently there are number of papers trying to enhance "distant learning" in several directions. Some researchers target the heuristics that are used to map the relations in the databases to the texts. for example, (Takamatsu et al. 2012) argue that improving matching helps to make data less noisy and therefore enhances the quality of relation extraction .

V. BEYOND BINARY RELATIONS

All Relation extraction systems focus primarily on binary relations. Semi-supervised systems such as TextRunner claim that their system can deal with n-ary relations but not very clear about the algorithmic changes that are required. (McDonald et al., 2005) proposed a framework for extracting complex relations (tuples) between entities in the text. Their algorithm is based on extracting 4-ary relations from biomedical abstract text. An instance in a relation is a list of entities (e_1 , e_2 , ..., e_n) where e_i is entity type. For example, we are interested in the ternary relation (organizer, conference, location) that relates an organizer to a conference at a particular location. For a sentence "ACL-2010 will be hosted by CMU in Pittsburgh", the system should extract (CMU, ACL-2010, Pittsburgh).

Given a sentence, to find out N-ary relation,it is required to list all possible tuples. Using all these tuples to train a binary classifier , it distinguish valid instances from invalid ones. for example a relation type with *n* entity elements, each element has *m* possible way then there are $O(m^n)$ possible complex relation candidates. Instead of trying to classify all possible relation instances, the key ideas of (McDonald et al., 2005) are

1. Start by recognizing binary relation instances that appear to be arguments of the relation of interest.
2. Extracted binary relations can be treated as the edges of graph with entity mention as nodes.
3. Reconstruct complex relations by making tuples from selected maximal cliques in the graph.

There are two major advantages of factoring complex relation into binary relations. First it allows for the use of almost any binary relation classifier which has been well studied and is often accurate. Second, the number of possible binary relations is much smaller than the number of possible complex relations.

VI. APPLICATIONS

The World Wide Web is a big storehouse of unstructured information. Structuring this information need to identify relational structure between them. For example, it would be possible to extract the entire family-tree of a prominent personality using a resource like Wikipedia. Relations describe the semantic relationships among the entities present in the text which is useful for a better understanding of natural language. In this section we will overview describe two important applications of relation extraction namely: Automatic Question- Answering and Bio-text mining.

A. Question Answering (QA)

If a query to a search engine is *"When was Gandhi born ?"*, then the expected answer would be *"Gandhi was born in 1869"*. The template of the answer is *<PERSON> born-in <YEAR>* which is the relational triple born in(PERSON, YEAR) where PERSON and YEAR are the entities. To extract the relational triples, a large database (ex: web) can be queried using a small initial question-answer set (ex: *"Gandhi 1869"*). The best matching (or most confident) patterns are then used to extract answer templates which in turn can be used to extract new entities from the database. The new entities are again used to extract newer answer templates and so on till convergence. This bootstrapping based method for QA is described in (Ravichandran & Hovy,2002).

B. Mining Biotext

Relation extraction methods are useful in discovering protein-protein interactions, and gene-binding conditions. Patterns like "Protein X binds with Protein Y" are often found in biomedical texts where the protein names are entities which are held together by the "bind" relation. Such protein-protein interactions are useful for applications like drug discovery. Other relations of interest are, a protein's location inside an organism. Such ternary relationships are extracted using linear kernels computed over features in (Liu et al., 2007). Cancer researchers can use inferences like *"Gene X with mutation Y leads to malignancy Z"* in order to isolate cancerous genes. These information patterns can be pieced together by extracting ternary relations between genes, mutations and malignancy conditions in a large corpus of biotext.

VII. CONCLUSION

Relation extraction is very important in NLP and can be beneficial for: semantic search, machine reading, question answering, knowledge harvesting, paraphrasing, building the- sauri etc. (Nakashole et al. 2012b, 2013). The field is becoming more and more interdisciplinary and methods from data mining and Pattern recognition domains are frequently used to assist in the task of relation extraction (Cergani and Miettinen 2013; Riedel et al. 2013; Nakashole et al. 2012a).After reviewing all the aspects of the entity-relation extraction problem, it can be said that Supervised approach works well with the specific domain in terms of computational complexity and performance. 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 be extended to new relations easily. Supervised approaches on the other hand can do well when the domain is more restricted like the case of biotext mining.

Extracting N-ary relations often play an important role in the relation Extraction task.. To better understand the semantic relations between entities in text N-ary relations could be more useful. So there is a need of investigating approaches capable of handling higher order relations efficiently without factorizing them. Research in terms of relation extraction has still room for improvement, however, it targets a very difficult problem where language ambiguity is a significant obstacle. The majority of research in the field is done for English language, therefore targeting local languages and exploring multilingual information extraction can be the future direction of Relation Extraction task

REFERENCES

[1] Etzioni, O., Cafarella, M., Downey, D., Popescu, A.M., Shaked, T., Soderland, S., Weld, D.S., Yates, A.: Unsupervised named-entity extraction from the web: an experimental study. Artif. Intell. 165, 91–134 (2005). (Elsevier Science Publishers Ltd., Essex,UK)

[2] Etzioni, O., Banko, M., Soderland, S., Weld, D.S.: Open information extraction from the web. Commun. ACM 51, 68–74 (2008) (2002) The IEEE website. [Online]. Available: http://www.ieee.org/

[3] Grishman, R.: Information extraction: techniques and challenges. In: Pazienza, M.T (ed.) SCIE 1997. LNCS, vol. 1299, pp. 10–27. Springer, Heidelberg (1997)

[4] Hovy, E., Kozareva, Z., Rilff, E.: Toward completeness in concept extraction and clas- sification. In: EMNLP '09: Proceedings of the 2009 Conference on Empirical Meth- ods in Natural Language Processing, pp. 948–957. Association for Computational Linguistics, Morristown (2009)

[5] Hearst, M.A.: Automatic acquisition of hyponyms from large text corpora. In: Proceedings of the 14th Conference on Computational Linguistics, pp. 539–545. Association for Computational Linguistics, Morristown (1992)

[6] Liu, Y., Shi, Z., & Sarkar, A. (2007). Exploiting rich syntactic information for relationship extraction from biomedical articles. Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers (pp. 97–100). Rochester, New York: Association for Computational LinguisticsMcDonald, R., Pereira, F., Kulick, S., Winters, S., Jin, Y., & White, P. (2005). Simple algorithms for complex relation extraction with applications to biomedical ie. ACL '05: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (pp. 491–498). Ann Arbor, Michigan

[7] Pasca, M.: Acquisition of categorized named entities for web search. In: Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management, CIKM '04, pp. 137–145. ACM, New York (2004)

[8] Riloff, E., Jones, R: Learning dictionaries for information extraction by multi-level bootstrapping. In: Proceedings of the Sixteenth National Conference on Artificial Intelligence and the Eleventh Innovative Applications of Artificial Intelligence Con- ference, Menlo Park, CA, USA, AAAI '99/IAAI '99, pp. 474–479. American Asso- ciation for Artificial Intelligence (1999)

[9] Weld, D.S., Wu, F., Adar, E., Amershi, S., Fogarty, J., Hoffmann, R., Patel, K., Skinner, M.: Intelligence in Wikipedia. In: Proceedings of the 23rd AAAI Conference, Chicago, USA (2008)

[10] Zhao, S., & Grishman, R. (2005). Extracting relations with integrated information using kernel methods. Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (pp. 419–426).