IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Volume-6, Issue-1 E-ISSN: 2347-2693

Natural Language Understanding Using Open Information Extraction Technique

Ashwini V. Zadgaonkar

Shri Ramdeobaba College Of Engineering And Management , RTMNU, Nagpur, India

*Corresponding Author: zashwini@rediffmail.com, Tel.: +00-0712-25341861

Available online at: www.ijcseonline.org

Received: 24/Dec/2017, Revised: 04/Jan/2018, Accepted: 19/Jan/2018, Published: 31/Jan/2018

***Abstract*—** Natural language understanding (NLU) task deals with use of computer software to understand human text or speech in the form of sentences. IE is the integral component of this task. IE extracts information about desired entities from diverse resources and stored it in machine readable format for future processing. IE systems developed so far uses either supervised or unsupervised approach for information extraction. **Distant supervision**, **Open information extraction** and **Joint prediction** are few more techniques which claims to improve IE system performance. This paper is an attempt to give comparative analysis of these advanced approached and the need of combination of these techniques for further enhancement. To conclude, few application areas were identified like machine reading which can be benefited from this combined approach.

***Keywords*—** Information Extraction, Open Information Extraction, Distant Supervision, Joint Prediction

I. INTRODUCTION

IE systems extract semantic relations between entities to form relevant information segments of text documents. Generally Extracted information is presented in relation tuple form. e.g. given the sentence "*Barack Obama is the President of the United States*", extracted relation tuple is ***PresidentOf (Barack Obama, the United States)***. IE systems uses pattern-matching techniques for extraction so performance of these systems is very much dependent on domain specific knowledge possess by the system. However traditional IE systems failed on scalability and portability parameters across domains.

In supervised IE systems [2] sentences in a corpus are hand-labelled for the presence of entities and relations. Supervised IE uses variety of lexical, syntactic and semantic features to label the relation hold between a given pair of entities. Problems associated with Supervised approach can be i) *cost associated with producing labelled training data* and 2)*biased classifiers*. On the other hand Unsupervised IE systems ,extracts word strings between entities in large amounts of text and apply clustering on these word strings to produce relation-strings [3]. Unsupervised IE uses large amount of text/speech to extract large numbers of unspecified relations. It uses a very small number of seed instances or patterns to do bootstrap learning [1]. These seeds are used with large corpus to extract new set of patterns for more instances in iterative fashion. However these systems suffer from *low precision* and *semantic drift problems*. To overcomes problems associated with both supervised and unsupervised IE, there is a need to move towards high precision IE systems which are highly scalable on large corpora.

Information extraction domain demands for systematic evaluation of different approaches suggested so far for the task and comment on its utility on the basis of its merits and demerits. Such comprehensive analysis will act as a guiding source for all the researchers who wnat explore this domain further for their problem of interest whicg is objective behind this paper.

The remainder of this paper is organized as follows. Section II gives overview of Open information extraction techniques for Information Retrieval. Section III discuss about Distant learning paradigm. Section IV discuss about recently suggested Joint prediction model for Information Retrieval. Section V concludes with a comparative analysis of these different approaches and possible directions of future research work.

II. OPEN INFORMATION EXTRACTION

Open information extraction [1] is "*a novel extraction paradigm that tackles an unbounded number of relations*". It facilitates domain independent relation discovery from text corpus. Also this approach is well scaled to the diversity and volume of the Web corpus. The input to

OIE is text corpus and output is extracted relations in three tuple form.

	IE	OIE
Input	Sentences + Labelled relations	Sentences
Relation	Need to be specified in advance	Free discovery
Extractor	Domain specific relations	Domain independent relations

Table 1. Comparison between IE and OIE

First generation OIE systems were designed to express a relation based on lexicalized features like Part-of-Speech or shallow tags. These OIE systems used their relation independent model of self-training to learn relations and entities in the corpora. **TextRunner [1]** OIE system used Naive Bayes model with POS and Chunking features to trained tuples. TEXTRUNNER's input is a corpus and output is a set of efficiently indexed extractions.WOE [4] generated relation-specific training examples by matching Infobox attribute values to corresponding sentences to learn an unlexicalized extractor. StatSnowball [5] proposed statistical extraction framework to perform both traditional relation extraction and Open IE. StatSnowball used discriminative Markov logic networks (MLNs) by learning their weights in a maximum likelihood estimate sense. Empirical results of StatSnowball showd a significantly higher recall and high precision.

However first generation OIE systems sufferd from incoherent and uninformative relation extraction problems. To overcome these problems second generation OIE systems were introduced which focuses on thorough linguistic analysis. ReVerb [6] is the second generation OIE system based on verb phrase-based relations. This system built a set of syntactic and lexical constraints to identify relations based on verb phrases. One of the limitation of Reverb was that it ignored the relation context by considering verb only which might lead to false or incomplete relations. OLLIE [10] is an extended ReVerb system stands for Open Language Learning which performed deep analysis on the identified verb-phrase relation. OLLIE followed ReVerb to identify potential relations based on verb-mediated relations. The system applied bootstrapping to learn other relation patterns using its similarity relations found by ReVerb. In each pattern the system used dependency path to connect a relation and its corresponding arguments for extracting relations mediated by noun, adjective and others. After identifying the general patterns, the system applied them to the corpus to obtained new tuples. Therefore, OLLIE extracted a higher number of relations from the same corpus compared to ReVerb.

A more recent Open IE system named as ClausIE [8] used clause structures to extract relations and their arguments from natural language text. ClausIE used dependency parsing and a set of rules for domain-independent lexica to detect clauses without any requirement for training data. ClausIE exploited grammar clause structure of English for detecting clauses and its constituents in sentence. As a result, ClausIE obtained high-precision relation extraction which can be flexibly customized to new application domain.

III. DISTANT SUPERVISION

Distant supervision is an extension of the approach used for exploiting WordNet to extract hypernym (is-a) relations between entities. It used Freebase , a large semantic database to provide distant supervision for relation extraction. Distant supervision [9] used Freebase to identify all sentences containing entities mentioned in freebase. This approach does not required labelled corpora which makes it domain dependence and managed scalability factor. For each pair of entities that appeared in some Freebase relation, they found all sentences containing those entities in a large unlabeled corpus and extracted textual features to trained a relation classifier. This model of distant supervision was able to extract 10,000 instances of 102 relations at a precision of 67.6%. [10] identified two main problems of distant supervision approach : (1) some training examples obtained through heuristic were not valid (2) the same pair of entities had several relations. Distant supervision can be further improved by considering Multi-instance Multi-label relations. Bayesian framework can also be utilized which can captured label dependency and learnt incorrect and incomplete labels. Multi-Instance Multi-label (MIML) approach [13] identified the problem of multiple relations between two entities. They Combined Distant supervision with Multi-Instance overlapping relations (where two same instances may be in two different relations) to extract relations .

IV. JOINT PREDICTION MODEL

Finkel [14] explored the idea of joint modeling using Bayesian networks. The approach was tested on two tasks: semantic role labelling and recognizing textual entailment. The end-to-end performance of natural language processing systems is often hampered by use of a greedy pipeline architecture, which causes errors to propagate and compound at each stage. To overcome this ,a novel architecture was presented to model these pipelines as Bayesian networks, with each low level task corresponding to a variable in the network, and performed approximate inference at each stage to find the best labeling. This approach gained benefits of sampling the entire distribution over labels at each stage in the pipeline. Roth and Yih [12] employed the idea of combining two stages on Information

Extraction: named entity recognition and relation extraction. Singh [15] proposed a single, joint graphical model that represented the various dependencies between the IR tasks (entity tagging, relation extraction, and co reference). Their joint modelling approach helped to avoid cascading errors. The joint model obtained 12 % error reduction over the other proposed models.

V. CONCLUSION

This paper discusses three recent IE trends namely Open information extraction, Distance supervision and Joint prediction. OIE techniques rely on lexical, syntactic and semantic features which makes it domain independent to a large extent. Distance supervision approach uses large semantic databases like freebase for the task. Joint prediction techniques focuses on implementing joint inference model to generate the inference form each stage of information extraction.

Our objective behind analysing these approaches is to identify the most suitable approach for extracting info from unstructured text. Each of the approaches discussed here have its own shortcomings which leads to a way for further exploration. If we select any one technique for designing IE system then performance can not be guranteed. So we concluded that IE systems should follow combined architecture which add the advantages of each technique to improve overall system performance. Relation extraction is a major step of IE and layered approach will ensure that no relation left unexplored from unstructured text specially for unlabelled set of relations.

Future research directions in Open Information Extraction domain focuses on integrating scalable methods for resolving multiple mentions of entities in a corpus (Multi instance multi label relations). Distant supervision methods can use Combination of syntactic and lexical features to provides better performance.

Another goal of the IE systems demands for representing extracted relation tuples into a graph-based structure popularly known as Knowledge graphs so that it can be efficiently utilized by NLP applications like machine reading or Semantic search.

Future challenges can be summarized as

- Developing systems which scales up information extraction to world wide web.
- Designing efficient structures for representing extracted information of new kinds.
- Automatic accumulation of large collections of facts to support know- ledge-based AI systems.

REFERENCES

[1] Etzioni, O., Banko, M., Soderland, S., Weld, D.S.: *Open information extraction from the web*. Commun. ACM 51, 68–74 ,2008

[2] 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.

[3] Michele Banko, Michael J. Cafarella, Stephen Soderland, Matthew Broadhead, and Oren Etzioni. : *Open Information Extraction from the Web*. In IJCAI.volume 7, pages 2670–2676, 2007

[4] 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

[5] J. Zhu, Z. Nie, X. Liu, B. Zhang, J.R. Wen, *Stat Snowball: a statistical approach to extracting entity relationships*. In Proceedings of WWW 2009.

[6] A. Fader, S. Soderland, O. Etzioni, *Identifying Relations for Open Information Extraction*. In Proceedings of EMNLP, 2011

[7] Mausam, Schmitz, M., Bart, R., Soderland, S. *Open Language Learning for Information Extraction.* In *Proceedings of EMNLP*, 2012.

[8] L.D. Corro, R. Gemulla, *ClausIE: Clause-Based Open Information Extraction.* In Proceedings of WWW, 2013

[9] Mintz, M., Bills, S., Snow, R., Jurafsky, D.: *Distant supervision for relation extraction without labeled data*. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2, ACL '09, pp. 1003–1011. Association for Computational Linguistics, Stroudsburg,2009

[10] Surdeanu, M., Tibshirani, J., Nallapati, R., Manning, C.D.: *Multi-instance multi-label learning for relation extraction*. In: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, EMNLP-CoNLL '12, pp. 455–465. Association for Computational Linguistics, Stroudsburg,2012

[11] Finkel, J.R., Manning, C.D., Ng, A.Y.: *Solving the problem of cascading errors: approximate bayesian inference for linguistic annotation pipelines*. In: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, EMNLP '06, pp. 618–626. Association for Computational Linguistics, Stroudsburg ,2006

[12] Roth, D., Yih, W.: *Global inference for entity and relation identification via a linear programming formulation*. In: Getoor, L., Taskar, B. (eds.) Introduction to Statistical Relational Learning. MIT Press, Cambridge, 2007

[13] Hoffmann, R., Zhang, C., Ling, X., Zettlemoyer, L., Weld, D.S.: Knowledge-based weak supervision for information extraction of overlapping relations. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1, HLT '11, pp. 541–550. Association for Computational Linguistics, Stroudsburg, 2011

[14] Finkel, J.R., Manning, C.D., Ng, A.Y.: Solving the problem of cascading errors: approximate bayesian inference for linguistic annotation pipelines. In: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, EMNLP '06, pp. 618–626. Association for Computational Linguistics, Stroudsburg (2006)

[15] Roth, D., Yih, W.: Global inference for entity and relation identification via a linear programming formulation. In: Getoor, L., Taskar, B. (eds.) Introduction to Statistical Relational Learning. MIT Press, Cambridge (2007)

Authors Profile

Ashwini Zadgaonkar received Bachelor of Engineering Degree in Computer Technology from PCE, Nagpur University and Masters of Technology in Computer science and Engineering from RCOEM, Nagpur University ,India in 2012 and is currently working as an Assistant Professor in Computer Science and Engineering Department of Shri Ramdeobaba College of Engineering and Management, Nagpur.

