Uncovering Algorithmic Approaches in Open Information Extraction: A Literature Review

Injy Sarhan[1,2] & Marco Spruit[2]

[1] Computer Engineering, College of Engineering, Arab Academy for Science, Technology and Maritime Transport (AAST), Abukir, Alexandria 1029, Egypt
[2] Information and Computing Sciences, Utrecht University, Princetonplein 5, 3584 CC Utrecht, The Netherlands
{i.a.a.sarhan,m.r.spruit}@uu.nl

Abstract

The demand for fast and efficient methods to extract pivotal information encourages researchers towards IE tasks. *Open Information Extraction* (OIE) is the process of extracting relation tuples from text, targets to ease the process of identifying domain-independent relations in texts .

Example

"Barack Obama born August 4, 1961 in Hawaii served as the 44th President of the United States."

OIE relation triplets format: **(argument 1, relation, argument 2):**

(Barack Obama, BornIn-Loc, Hawaii)
(Barack Obama, BornIn-Year, August 4, 1961)
(Barack Obama, Served-as, President of the United States)

Challenges

1.Uninformative Extractions:
Incorrect handling of relational phrases that results in leaving out crucial information.

2.Incoherent Extractions:
Purposeless extractions that are derived from opaque relation phrases that the extractor fails to correctly identify

	Machine-Learning Classifiers	Hand-Crafted Rules
Shallow Syntactic Analysis	**TextRunner [1]:** • 1st OIE system. • Naïve Bayes model. • Two-stage Technique that learns relation mapping rules and domain relations. **WOE^{POS} [2]:** • Employs a CRF extractor • Utilizes Wikipedia to train data for their extractors. • Automatic assembly of training examples from Wikipedia info box.	**ReVerb [4]:** • Utilizes syntactic and lexical constraints. • Confidence score is then allocated to extractions. **R2A2 [5]:** • Merges ReVerB with an argument enhance argument extraction. • Identifies arguments by utilizing patterns. **ExtrHech [6]:** • Applies syntactic constraints as regular expressions. • Multi-lingual (Spanish and English) **LSOE [7]:** • Implements lexical-syntactic patterns to POS-tagged texts to extract relation triples. • Uses Generic patterns and Rule-based patterns.
Depen-dency Parsing	**WOE^{Parse} [2]:** • Dependency path patterns acquired from from Wikipedia extraction. • Parses handles complicated distance relationships. **OLLIE [3]:** • Bootstraps training set from seed tuples to learn pattern templates. • Pattern templates determine the argument and the relation phrase. • Learning component ensures that all the important information had been captured. • Confidence function is trained to rule out non-factual extractions.	**KraKen [8]:** • Extracts N-ary facts. • Examines fact completeness and correctness. **DepOE [9]:** • Multi-lingual system (Portuguese, Spanish, English) • Extract verb-based triples from Wikipedia. **ClausIE [10]:** • Identify clauses in an input sentence. • Determines category of each clause to be consistent with the grammatical function of surrounding text. **CSD-IE [11]:** • Tree expressing the semantics is derived. • Tree constituents are combined to form the contexts creating the phrase.

Future Trends and Conclusion

Neural Networks

OIE paradigm that implements an encoder-decoder framework [12].
- Employs recurrent neural network
- Utilizes three-layer LSTM.

Future Research:
- Multilingual & N-ary extractions.
- Analysis supports Hand-Crafted approaches and the novel Neural Network approaches.

References

[1] Banko, M., Cafarella, M. J., Soderland, S., Broadhead, M., & Etzioni, O.: Open information extraction from the web. In IJCAI (Vol. 7, pp. 2670-2676) (2007).
[2] Wu, F. and Weld, D, S.: Open information extraction using Wikipedia. In Proceedings of the 48th annual meeting of the association for computational linguistics. Association for Computational Linguistics, (2010).
[3] Schmitz, M., Bart, R., Soderland, S., & Etzioni, O: Open language learning for information extraction. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Association for Computational Linguistics, (2012).
[4] Fader, A., Soderland, S., & Etzioni, O.: Identifying relations for open information extraction." Proceedings of the conference on empirical methods in natural language processing. Association for Computational Linguistics, (2011).
[5] Etzioni, O., Cafarella, M., Downey, D., Popescu, A. M., Shaked, T., Soderland, S., ... & Yates, A.: Unsupervised named-entity extraction from the web: An experimental study. In Artificial intelligence (165.1: 91-134) (2005).
[6] Zhila, A., & Gelbukh, A.: Comparison of open information extraction for English and Spanish. In 19th Annual Int. Conference Dialog pp. 714-722 (2013).
[7] Xavier, C., de Lima, V., & Souza, M: Open information extraction based on lexical-syntactic patterns. In Braz. Conf. Intelligent Systems, pp. 189– 194, (2013).
[8] Akbik A, Löser A.: Kraken: N-ary facts in open information extraction. In Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction, pp. 52– 56, (2012).
[9] Gamallo, P., Garcia, M., & Fernández-Lanza, S.: Dependency-based open information extraction. In Proceedings of the joint workshop on unsupervised and semi-supervised learning in NLP (pp. 10-18). Association for Computational Linguistics. (2012).
[10] Del Corro, L., & Gemulla, R.: Clausie: clause-based open information extraction. In Proc. of the 22nd int. conference on WWW pp. 355-366. ACM, (2013).
[11] Bast, H., & Haussmann, E.: Open information extraction via contextual sentence decomposition. In Semantic Computing (ICSC), IEEE Seventh International Conference on. pp. 355-366. IEEE, (2013).
[12] Cui, Lei, Wei, F. and Zhou, M.: Neural Open Information Extraction. arXiv preprint arXiv:1805.04270 (2018).