Research on information extraction technology applied for knowledge graphs

Wei Zhou

Renmin University of China,No. 59 Zhongguancun Street, Haidian District Beijing, 100872, P.R. China

908768554@qq.com

Abstract. Information extraction is an important part of natural language processing and is an important basis for building question and answer systems and knowledge graphs. A growing number of new technologies are being applied to information extraction with the development of deep learning techniques. As a first step, this paper introduces information extraction techniques and their main tasks, then describes the development history of information extraction techniques, and introduces the practice and application of different types of information extraction techniques in knowledge graph construction, including entity-extraction, relationship extraction and attribute extraction. Finally, some problems and research directions faced by information extraction techniques are discussed.

Keywords: Knowledge Graph, Information Extraction, Entity Extraction, Relationship Extraction.

1. Introduction

The big data era has brought about the emergence of enormous amounts of data, and as a result, there is an increasing need to identify relevant information and derive insight from it. In response to this demand, knowledge graph technology has emerged and is increasingly important in realizing knowledge interconnection.

Because the building of knowledge graphs begins with a systematic description of concepts, entities, and their relationships in the objective world the correctness of information extraction of concepts, entities, and relationships is critical to the construction process.Information loss, redundancy, and overlap are often the most significant challenges to the construction of knowledge graphs. Information extraction, as the first step in knowledge graph construction, is critical to obtaining candidate knowledge units [1-2]. The completeness and accuracy of information extraction directly and explicitly affect the quality and efficiency of the subsequent knowledge graph construction steps and the quality of the final knowledge graph.

Through the method of bibliometrics, this study makes a visual analysis of the key technologies, frontier breakthroughs, and research hotspots of information extraction technology.

Through the method of literature review, this paper systematically introduces the historical background and development of information extraction. There are three sub-tasks that it can be divided into based on the technical characteristics: entity extraction, relationship extraction, and attribute extraction. Among them, each subtask is divided into specific domain-oriented and open domain-oriented according to its application field, and text-oriented and web-oriented according to its data source.

Firstly, this paper discusses the significance of information extraction research based on the framework of knowledge atlas; Then it examines the history of information extraction through the eyes of three international assessment conferences: MUC, ACE, and ICDM. Following that, the major methods of information extraction are explained in detail, including entity extraction, relationship extraction, and attribute extraction; finally, the information extraction research trend is examined.

2. Previous works on information extraction

In general, information extraction takes the text above and extracts particular information using machine learning, natural language processing (NIP), and other techniques. It then stores that information in a structured database so that users may search and use it.

Early information extraction research began in the mid-1960s, represented by two long-term projects, the linguistic string of New York University and frump of Yale University. However, it wasn't until the late 1980s that the Message Understanding Conference (MUC) was held that the study and use of information extraction progressively entered a profitable period [3]. After MUC, the National Institute of Standards and Technology (NIST) hosted the automated content extraction (ACE) assessment conference, which was a significant worldwide gathering for information extraction research from 1999 to 2008.Compared with MUC, ACE evaluation is not specific to a specific field or scenario. It adopts a set of an evaluation systems based on false positives (yes in the standard answer but not in the system output) and false positives (no in the standard answer but not in the system output). It also evaluates the cross-document processing ability of the system.

There are several types of information extraction, including entity extraction, relationship extraction, attribute extraction, and sub-tasks. There are mainly two types of entity recognition methods: specific field and open field. Domain-specific entity recognition methods mainly include some classical models, such as the maximum entropy classification model、Hidden Markov model and Conditional Random Field model, etc. In the open domain-oriented information extraction, the source of information is no longer a specific knowledge field, but a whole network of information and a large amount of Web corpus [4]. For example, KnowItAll system deals with large-scale and heterogeneous Web corpora, such as Twitter, Wikipedia, etc [5]. Due to the limitations of traditional statistical models that require a large number of corpus annotations and manual construction of a large number of features, some new methods have emerged, such as using semi-supervised algorithms, remotely supervised algorithms, self-learning method based on massive data redundancy to solve the problem of open entity extraction [6-8]. The open domain-oriented entity extraction method is often applied to the novel question-answering system based on common sense [9].

Information extraction methods are divided into natural text-oriented information extraction, Web-oriented text extraction, and social network-oriented information extraction according to the different sources of processing information. In entity recognition extraction, rule-based and statistics-based entity recognition methods are usually used to process natural language texts, which have strong pertinence and high accuracy, and can usually obtain good recognition results under manual annotation, For example, document uses a rule-based method to realize the company name as the processing object [10] The document combines the KNN classifier with the linear conditional random field (CRF) model to realize named entity recognition from short informal Twitter articles. The method based on deep learning does not need to define relevant features manually [11]. It can be applied to process natural texts in a single field. For example, literature [12] takes scientific articles as the processing object and uses a neural marking model to extract keyword phrases from scientific research articles. It can also be applied to process web data. For example, literature [13] proposes a semi-supervised system for entity recognition and distributed representation of Twitter.

In addition to natural language text and Web text, social network data is also a rich data source. Social network nodes are massive and feature a diverse set of relationships. Literature [14] proposed using a sequential joint clustering algorithm based on an unsupervised method to extract a variety of relationships in social networks containing multiple nodes.

3. Discussion

Early named entity recognition often used rule-based methods. Generally, linguistic experts first selected various features that can represent a certain type of entity according to the characteristics of the entity type to be recognized, such as the surname of a person's name, the title of a position, etc., built a limited rule template, and then completed the extraction of named entities using pattern matching [15]. Most of these systems rely on the domain knowledge of linguistic experts, which is not only time-consuming and labor-intensive but also inevitable.

With the development of machine learning, machine learning based on statistics is also continuously applied to information extraction In this method, various features of each word in the text (such as lexical features, part of speech tagging, word meaning features, etc.) are expressed as a feature vector, and then large-scale training corpus is trained through different model methods. Finally, entity recognition is carried out through the trained model. The common models are: Hmm (hidden Markov model), Me (Maxmium Entropy), SVM (support vector machine) and CRF (conditional random fields) and so on [16-19].

In recent years, with the introduction of word embedding, the application of deep learning methods in natural language processing has reached a climax. Wod2vec is the representative of word vectors. Its basic idea is to use vectors with unified dimensions to represent each word in the model [20]. This not only solves the problem of data sparsity caused by high-dimensional vector space but also integrates more semantic features into it. At the same time, heterogeneous texts can be represented by unified dimensional vector features.

Liu et al. first used CNN (Convolutional Neural Networks) to extract features automatically. It encoded sentences with word vectors and lexical features and then completed the classification with convolution, full connection, and softmax layers [21]. It improved the F1 value by 9% on the ACE 2005 dataset compared with the kernel-based method. Zeng et al. use pre-training word vectors and location features, as well as a wide maximum pooling layer behind the CNN layer [22]. Nguyen and Grishnian completely abandon the lexical features [23], allowing CNN to learn automatically, and use multi-window convolution to obtain different scales of n-gram information, achieving better results through end-to-end neural networks.

Compared with traditional machine learning methods, CNN-based methods have achieved good results, but CNN has a weak ability to extract time series features. And RNN (Recurrent Neural Network) model is suitable for extracting time series features. Zhang et al. used BRNN (Bidirectional RNN) for relationship extraction for the first time [24]. BRNN is equivalent to integrating forward and backward RNNs, which input words in sentences into two RNNs according to the forward and reverse directions respectively, and then overlay the implicit layers of the two RNNs.

Cai et al. proposed a deep learning relationship extraction module based on Shortest Dependency Path (SDP) in 2016: the bidirectional recursive convolution neural network model (BRCNN) [25]. The main idea of this paper is to model the SDP of the network syntax between two entities, encode the global information of the SDP using a dual-pass LSTM (Long Short-Term Memory), and capture the local characteristics of two words of each dependency link using CNN, to enhance the ability of entities to classify the direction of the relationship between them.

Miwa et al. first applied the neural network method to the joint model of named entity recognition and entity relationship extraction in 2016 [26]. The model is based on LSTM-RNN and executes end-to-end. The model consists of Three Representation layers. The bottom layer is the word embedding layer to complete information encoding. There are two bidirectional LSTM-RNN in the word embedding layer. One is based on word sequence for entity recognition tasks, and the other is based on dependency tree structure for relationship extraction. The two parts share encoding information and stack to form an overall model. As part of the input of the latter structure, the former output and the hidden layer make entity recognition and extraction interact.

Katiyar et al. combined Attention, an attention mechanism, with BiLSTM for named entity recognition and relationship extraction in 2017 [27]. The model draws on the model of Miwa et al. and improves the disadvantage of the original model depending on the intersectional sequence, dependency tree, etc [25]. The model has an input layer with one embedded word representation, two output layers, an entity for output identification, and a relationship classification using the attention model.

In 2018, Devlin et al. proposed the BERT (Bidirectional Encoder Representations from Transformers) model. BERT belongs to the pre-training language model [28]. The so-called pre-training model is to pre-train the model with a large number of customized text so that the model can acquire general language knowledge, and then perform Fine-tuning phase training according to downstream tasks so that the model parameters can be fine-tuned according to specific task requirements and domain knowledge.

In recent years, the emergence of pre-training models such as GPT and BERT has made the Q&A reading comprehension task a good downstream task of theirs [28-29]. Simply reconstructing the original network structure and fine-tuning can get good results. Wang et al. improved their performance on the SQuAD dataset by using multi-paragraph prediction based on the original BERT [30-31]. Alberti and others improved on BERT and SQuAD and applied them to a more difficult question and answer data set, NQ [32-33].

At present, information extraction methods based on deep-learning have made good progress, but there are still many aspects worth further study. First, the deep-learning model is good at processing single-sentence semantic information, but in practice, many entity relationships are expressed by multiple statements together, which requires the model to comprehensively understand, memorize and infer multiple statements in the document, and extract document-level relationships. Secondly, the current research on information extraction is mostly focused on the preset set of extraction tasks, but future applications will be open-domain-oriented information extraction. Therefore, it is necessary to continuously explore how to automatically discover new entity relationships and their facts in the open domain. Finally, current research is often limited to single-language text information, and human beings can process multiple information synthetically when receiving information. Therefore, it is necessary to explore how to extract relationships using multilingual text, sound and video information synthetically.

4. Conclusion

This paper first introduces the concept of information extraction based on the concept of knowledge map and the construction of technical framework. Then it briefly introduces the history of information extraction through three international evaluation conferences and three development stages (rule-based stage, statistical learning stage, deep-learning stage). Subsequently, the latest development and a series of cases of key information extraction technologies combined with CNN, RNN, LSTM, BERT and other deep-learning algorithms are introduced in detail. Finally, some problems and research trends that need to be solved for future information extraction are analyzed.

References

[1] Etzioni, O., Fader, A., Christensen, J., et al. (2011) Open Information Extraction: The Second Generation.

[2] Wu, X.D., Wu, J., Fu, X.Y., Li, J.C., Zhou, P. and Jiang, X. (2019) Automatic Knowledge Graph Construction: A Report on the 2019 ICDM/ICBK Contest. 2019 IEEE International Conference on Data Mining (ICDM), Beijing, China, 8-11 November 2019, 1540-1545. https://doi.org/10.1109/ICDM.2019.00204

[3] Ralph Grishman, Beth Sundheim: Message Understanding Conference - 6: A Brief History. In: Proceedings of the 16th International Conference on Computational Linguistics (COLING), I, Copenhagen, 1996, 466–471.

[4] Zhao, J., Liu, K., Zhou, G.Y., et al. (2011) Open Information Extraction. Journal of Chinese Information Processing, 25, 98-110.

[5] Etzioni, O., Cafarella, M., Downey, D., et al. (2005) Unsupervised Named-Entity Extraction from the Web: An Experimental Study. Artificial Intelligence, 165, 91-134. https://doi.org/10.1016/j.artint.2005.03.001

[6] Shi, B., Zhang, Z., Sun, L., et al. (2014) A Probabilistic Co-Bootstrapping Method for Entity Set Expansion. Proceedings of the 25th International Conference on Computational Linguistics: Technical Papers, Dublin, august 2014, 2280-2290.

[7] Agichtein, E. and Gravano, L. (2000) Snowball: Extracting Relations from Large Plain-Text Collections. Proceedings of the 5th ACM Conference on Digital Libraries, San Antonio, June 2010, 85-94.

[8] Fader, A., Soderland, S. and Etzioni, O. (2011) Identifying Relations for Open Information Extraction. Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, John McIntyre Conference Centre,Edinburgh, 27-31 July 2011, 1535-1545.

[9] RatnaParkhi, A. (1997) A Simple Introduction to Maximum Entropy Models for Natural Language Processing. Institute for Research in Cognitive Science, Technical Reports, University of Pennsylvania, Pennsylvania, 97-108.

[10] Rau, L.F. (1991) Extracting Company Names from Text. Proceedings of the 7th IEEE Conference on Artificial Intelligence Applications Piscataway, Miami Beach, 24-28 February 1991, 29-32.

[11] Zhu, J., Nei, Z.Q., Liu, X.J., et al. (2009) StatSnowball: A Statistical Approach to Extracting Entity Relationships. Proceedings of the 18th International Conference on World Wide Web, Madrid, 20-24 April 2009, 101-110.

[12] Yi, L., Mari, O. and Hannaneh, H. (2017) Scientific Information Extraction with Semi-Supervised Neural Tagging. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, September 2017 2641-2651.

[13] Godin, F., Vandersmissen, B., Neve, W.D., et al. (2015) Multimedia Lab @ ACL W-NUT NER Shared Task: NamedEntity Recognition for Twitter Microposts Using Distributed Word Representations. Proceedings of the Workshop on Noisy User-Generated Text, Beijing, July 2015, 146-153.

[14] Bollegala, D.T., Matsuo, Y. and Ishizuka, M. (2010) Relational Duality: Unsupervised Extraction of Semantic Relations between Entities on the Web. Proceedings of the 19th International Conference on World Wide Web, WWW 2010, Raleigh, 26-30 April 2010, 151-160.

[15] SODERLANDS. Learning information extraction rules for semi - structured and Free Text[J]. Machine Learning, 1999,34(1-3):233 - 272.

[16] ZHOU G D, SU J. Named entity recognition USing an HMM— based chunk tagger[C]//Proceedings of 40th Annual Meeting of the Association for Computatoional Linguistics. Philadelphia, PA, USA, 2002 ：473-480.

[17] BORTHWICK A. A maximum entropy approach to named entity recognition[D]. New York: New York University, 1999.

[18] CRISTANINI N, SHAWE - TAYLOR J. An introduction to support vector machines[M]. Cambridge: Cambridge University Press, 2000.

[19] LAFFERTY J, MCALLUM A, PEREIRA F. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data [C] //Proceedings of the Eighteenth International Conference on Machine Learning, 2001 ：282 -289

[20] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient Estimation of Word Representations in Vector Space[J]. arXiv preprint arXiv: 1301.3781, 2013

[21] LIUC Y, SUNWB, CHAO W H, et al. Convolution Neural Network for Relation Extraction L C]//International Conference on Advanced Data Mining and Applications. Springer, Berlin, Heidelberg, 2013 ： 231 - 242.

[22] ZENG D, LIU K, LAI S, et al. Relation classification via convolutional deep neural network [C]//Proceedings of the 25th International Conference on Computational Linguistics, 2014：2335 -2344.

[23] NGUYEN TH, GRISHMAN R. Combining neural networks and log - linear models to improve relation extraction [J. arXiv preprint arXiv： 1511.05926,2015.

[24] ZH DONGXU, DONG W. Relation Classification via Recurrent Neural Network [J], arXiv preprintarXiv ： 1508.01006,2015 ： 121 -128.

[25] CAI R, ZHANG X, WANG H. Bidirectional Recurrent Convolutional Neural Network for Relation Classification 1 C J//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin, Germany ,2016 ： 756 -765.

[26] MIWA M, BANSAL M. End - to - end relation extraction using LST- Ms on sequences and tree structures [C] //Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin, Germany, 2016 ： 1105 -1116.

[27] KATIYAR A, CARDIE C. Going out on a limb: Joint Extraction of Entity Mentions and Relations without Dependency Trees [C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver, Canada, 2017 ： 917 - 928.

[28] DEVLIN J, CHANG M W, LEK K, et al. BERT： Pre - training of Deep Bidirectional Transformers for Language Understanding] C」 // Proceedings of the 2019 Confence of the North American Chapter of the Association for Computational Linguistics: Human Lnguage Technologies. 2019 ： 4171 -4186.

[29] Young, T., Hazarika, D., et al. Recent trends in deep learning based natural language processing[J]. IEEE Computational Intelligence Magazine. 2018, 13(3), 55-75.

[30] Wang, Zhiguo, et al. Multi-passage BERT: A Globally Normalized BERT Model for Open-domain Question Answering[C]// Proceedings of the 2019. Conference on Empirical Methods in Natural Language Processing and the 9th International. Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019.

[31] Rajpurkar, Pranav, et al. Squad: 100,000+ questions for machine comprehension of text[J]. arXiv preprint arXiv:1606.05250, 2016.

[32] Alberti, Chris, Kenton Lee, et al. A bert baseline for the natural questions[J]. arXiv preprintarXiv:1901.08634, 2019.

[33] Kwiatkowski, Tom, et al. Natural questions: a benchmark for question answering research[J]. Transactions of the Association for Computational Linguistics 7. 2019: 453-466.