MMKE: A Multi-Model Knowledge Extraction System from Unstructured Texts

Qian-Wen Zhang[1], **Zhao Yan**[1] , **Tianyang Zhao**[1 2] *,
Shi-Wei Zhang[1] , **Meng Yao**[1] , **Meng-Liang Rao**[1] , **Yunbo Cao**[1]
[1]Tencent Cloud Xiaowei, Beijing, China
[2]Beihang University, Beijing, China
{cowenzhang, zhaoyan, tanyazhao, zswzhang, mengyao, sekarao, yunbocao}@tencent.com

Abstract

In this work, we present a Multi-Model Knowledge Extraction (MMKE) System which consists of two unstructured text extraction models (RelationSO model and SubjectRO model) based on a multi-task learning framework. Instead of recognizing entity first and then predicting relationships between entity pairs in previous works, MMKE detects subject and corresponding relationships before extracting objects to cope with the diverse object-type problem, overlapping problem and non-predefined relation problem. Our system accepts unstructured text as input, from which it automatically extracts knowledge in the form of (subject, relation, object) triples. More importantly, we incorporate a number of user-friendly extraction functionalities, such as multi-format uploading, one-click extractions, knowledge editing and graphical displays. The demonstration video is available at this link: https://youtu.be/HtOPJrGhSxk.

Introduction

Automated extraction of knowledge (subject, relation, object) from unstructured texts is a fundamental task of information extraction. For instance, with sentence *David Tennant was born in England*, (*David Tennant*, *born_in*, *England*) is considered to be the triplet to be extracted. We propose a **M**ulti-**M**odel **K**nowledge **E**xtraction (MMKE) system that supports multi-format file uploading, one-click extraction, knowledge editing and graphical displays to make the extraction of structured knowledge easier.

Most previous methods divide the extraction task into two parts, first recognize all entities and then predict relationships between each entity pairs (Chan and Roth 2011; Eberts and Ulges 2019; Luan et al. 2019). These methods face three major problems, diverse object-type problem, overlapping problem and non-predefined relation extraction. According to our observation, the types of objects in industrial scenarios tend to be diverse, whereas previous work has focused on entities. It is essential to extend objects to text fragments that can contain not only entities, but also text, numbers, addresses, events, etc. Second, the overlapping problem challenges the early methods of assuming that an entity pair has only one relationship.[1] The facts of relations in a sentence are complex, it may have more than one relational triples, and different triples may share the same entities. Third, existing studies focus more on tasks where relations are predefined, but it is not possible to cover all relation types in a predefined way in real scenarios. Recently, several researchers interested in the task of non-predefined relation extraction have attempted to identify new relations using semi-supervised or unsupervised methods (Lin et al. 2019; Wu et al. 2019). As for our system, we transform the problem into the machine reading comprehension task (Levy et al. 2017; Zhao et al. 2020), and handle the issue of the open growth of non-predefined relations through the question answering mechanism.

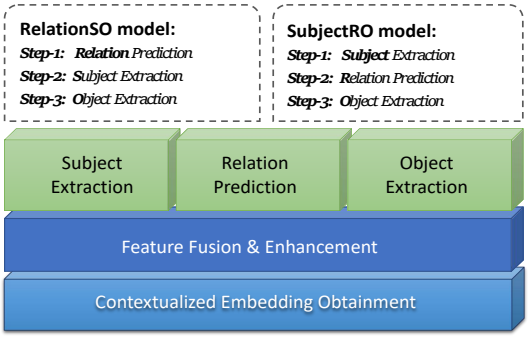
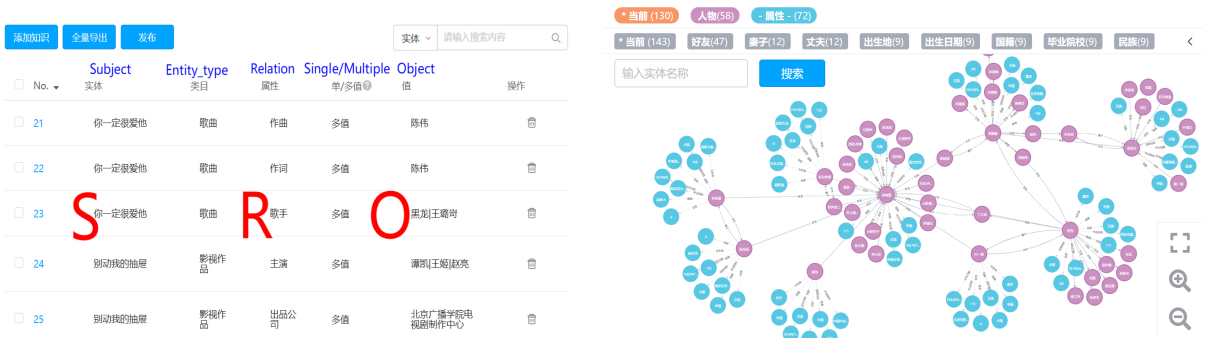


Figure 1: Extraction models framework for MMKE system. Specifically, word representations with contextual information are first generated, then the sentence representation is augmented using feature fusion. Lastly, the top layer matches each sub-task with a model separately.

Different from the way of using API to return results, we are concerned not only with the completeness of the knowledge extraction, but also with the simplicity of using the system. MMKE is a web-based interaction system which is oriented towards general users. Given a text in word or excel format as input, the user can extract the relevant knowledge with one click. We have also integrated editing, indexing, downloading and visualization features to make the system more user-friendly. To summarize, the main contributions of

*Work done during an internship at Tencent.

[1]More information about the overlapping problem can be found in Zeng et al. (2018).



(a) Edit knowledge in the form of triples.

(b) Visualization of knowledge.

Figure 2: Main features of MMKE system.

this work are: 1) An interactive, user-friendly Web system to automatically extract knowledge in the form of triples. 2) A flexible framework with multiple models that splits the task into three sub-tasks, which solves the overlapping problem and non-predefined relation extraction, and extends objects to text fragments to support various values.

Framework

To tackle the three major problems mentioned above, we present two unstructured text extraction models based on a multi-task learning framework as shown in Figure 1, which is the core of MMKE system. The framework contains three sub-tasks, subject extraction, relation prediction and object extraction. Note that both of the two models perform object extraction as a span extraction task in the last step, which has a robuster adaptability to diverse type of object.

The **RelationSO** model, which starts with **Relation** classification and then **S**ubjects and **O**bjects are extracted sequentially with sequence labeling model. Since there may be multiple relations in a sentence, we consider the relation classification as a multi-label learning (Zhang and Zhou 2014) task, which requires predefined relations. The RelationSO model can reduce the cost of traversing relation set, finding those relations that are relevant to the text among a large number of candidate relation set. After specifying relation, the extraction of subjects and objects becomes more explicit. The **SubjectRO** model, which first extracts **Subject** entities, then completes the extraction of **R**elations and **O**bjects in conjunction with a diverse questioning mechanism. This model is mainly based on our previous work (Zhao et al. 2020), which uses natural language questions to enhance the model's understanding of relations and uses question templates to provide the possibility of non-predefined relation extraction.

The benefit of splitting triples into three sub-tasks is that the elements are transmitted superimposed, which maximizes the retention of extracted knowledge. Taking RelationSO as an example, after the first step of obtaining the set of relations, the second step can find the subjects for each relation, and the third step predicts the objects based on each subject-relation pair. This design naturally handles the overlapping problem.

Demonstration

The main functions of our system are to automatically extract knowledge in the form of triples from unstructured text and to assist users in knowledge editing, retrieval and presentation. The first thing a user needs to do before using the system is to determine the relation set. MMKE includes 65 predefined relations as built-in relations for the user to choose from. Users can also define their own relations, and these relations will be treated as cases of non-predefined relation extraction.

Our system allows the user to upload the extracted content in word or excel format according to the system's sample file. After uploading, they can click the start button to extract the unstructured text. MMKE provides the user with a display page of the extraction results, which corresponds the source file to the extraction results. Under this page, the user can click on the knowledge to see where the knowledge is in the original text. As shown in Figure 2a, all the extracted results are stored. Users have the flexibility to edit the extracted knowledge themselves. In addition, our system also offers a download function, which allows the user to download all results in excel format. To display knowledge more visually, our system aggregates related triples in subject units and displays them based on the open source graph database Neo4j (Webber 2012). As shown in Figure 2b, the user can click on a node to get the relations and objects associated with that node. In general, dynamic visualizations can increase user viscosity and arouse their interest in knowledge.

Conclusion

We present an effective Multi-Model Knowledge Extraction (MMKE) system to identify knowledge from unstructured texts, which is the core of Tencent Cloud Xiaowei Knowledge Graph Platform. Considering previous works are limited in handling the diverse object-type problem, overlapping problem and non-predefined relation problem, our system conducts the object identification separately based on the subject and relation extraction. It is of great significance to users for knowledge acquisition and management. In future work, we intend to integrate more extraction algorithms into our system and provide some applications such as question answering.

References

Chan, Y. S.; and Roth, D. 2011. Exploiting syntactico-semantic structures for relation extraction. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 551–560.

Eberts, M.; and Ulges, A. 2019. Span-based Joint Entity and Relation Extraction with Transformer Pre-training. *arXiv preprint arXiv:1909.07755* .

Levy, O.; Seo, M.; Choi, E.; and Zettlemoyer, L. 2017. Zero-Shot Relation Extraction via Reading Comprehension. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*.

Lin, H.; Yan, J.; Qu, M.; and Ren, X. 2019. Learning dual retrieval module for semi-supervised relation extraction. In *The World Wide Web Conference*, 1073–1083.

Luan, Y.; Wadden, D.; He, L.; Shah, A.; Ostendorf, M.; and Hajishirzi, H. 2019. A general framework for information extraction using dynamic span graphs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 3036–3046.

Webber, J. 2012. A programmatic introduction to neo4j. In *Proceedings of the 3rd annual conference on Systems, programming, and applications: software for humanity*, 217–218.

Wu, R.; Yao, Y.; Han, X.; Xie, R.; Liu, Z.; Lin, F.; Lin, L.; and Sun, M. 2019. Open relation extraction: Relational knowledge transfer from supervised data to unsupervised data. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 219–228.

Zeng, X.; Zeng, D.; He, S.; Liu, K.; and Zhao, J. 2018. Extracting relational facts by an end-to-end neural model with copy mechanism. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 506–514.

Zhang, M.-L.; and Zhou, Z.-H. 2014. A review on multi-label learning algorithms. *IEEE transactions on knowledge and data engineering* 26(8): 1819–1837.

Zhao, T.; Yan, Z.; Cao, Y.; and Li, Z. 2020. Asking Effective and Diverse Questions: A Machine Reading Comprehension based Framework for Joint Entity-Relation Extraction. In *Twenty-ninth International Joint Conference on Artificial Intelligence IJCAI-20*.