DuIE: A Large-scale Chinese Dataset for Information Extraction

Shuangjie Li, Wei He, Yabing Shi, Wenbin Jiang, Haijin Liang, Ye jiang, Yang Zhang, Yajuan Lyu, Yong Zhu

Baidu Inc., Beijing 100193, China
{lishuangjie,hewei23,shiyanbing01,jiangwenbin,lianghaijin,
jiangye,zhangyang08,lvyajuan,zhuyong}@baidu.com

Abstract. Information extraction is an important foundation for knowledge graph construction, as well as many natural language understanding applications. Similar to many other artificial intelligence tasks, high quality annotated datasets are essential to train a high-performance information extraction system. Existing datasets, however, are mostly built for English. To promote research in Chinese information extraction and evaluate the performance of related systems, we build a large-scale high-quality dataset, named DuIE, and make it publicly available. We design an efficient coarse-to-fine procedure including candidate generation and crowdsourcing annotation, in order to achieve high data quality at a large data scale. DuIE contains 210,000 sentences and 450,000 instances covering 49 types of commonly used relations, reflecting the real-world scenario. We also hosted an open competition based on DuIE, which attracted 1,896 participants. The competition results demonstrated the potential of this dataset in promoting information extraction research.

Keywords: Information extraction, Dataset, Performance evaluation.

1 Introduction

Information extraction (IE) aims to extract structured information from unstructured or semi-structured text. Representative structured information includes entities, their attributes and relations, carrying important semantic information conveyed by the text. IE enables machines to understand the semantics of text and acts as the foundation of many important applications, such as knowledge graph construction, semantic information retrieval and intelligent question answering etc. Many efforts focus on the task of IE and achieve significant progress, especially with deep learning techniques [1~8].

Similar to most artificial intelligence applications, high-performance IE systems require supervised learning and adequate annotated datasets. However, existing datasets for IE are mainly built for English. To the best of our knowledge, there is no large-scale dataset for Chinese IE. In fact, even the existing English datasets are subjected to limited scale or poor quality. For example, NYT dataset [9] is automatically constructed without manual annotation, and suffers from the poor data quality problem. SemEval-

2010 dataset [11] and FewRel dataset [12] achieve relatively higher quality by introducing manual annotations, but their data scales are still not sufficient.

To better evaluate the performance of Chinese IE techniques, we build a large-scale good quality dataset, DuIE, and make it publicly available for research use. We design an effective coarse-to-fine procedure including candidate generation and crowdsourcing annotation in order to achieve large data scale and high data quality.

To the best of our knowledge, DuIE is the first large-scale, high-quality dataset for Chinese IE. Specifically, it contains 450,000 instances, with 49 commonly used relation types, 340,000 unique Subject-Predicate-Object (SPO) triples and 210,000 sentences. The text in DuIE covers a variety of domains in real-world applications, such as news, entertainment, user-generated contents. The annotations contain single-valued and multi-valued triples, reflecting the real-world scenario. Table1 gives an example of annotated sentences in DuIE.

Table 1. A sample data in DuIE dataset

Sentence	SPO list
《**给最开心的人**》是**梁咏琪**于2004年12月15日发行的音乐专辑《**娱乐大家**》中的歌曲。 *To the Happiest People* is a song in *Gigi Leung*'s music album *The Great Entertainer,* which was released on December 15, 2004.	S：给最开心的人，P：歌手，O：梁咏琪 S: *To the Happiest People,* P: singer, O: *Gigi Leung* S:给最开心的人，P:所属专辑，O:娱乐大家 S: *To the Happiest People,* P: fromAlbum, O: *The Great Entertainer*

We hosted an open competition based on DuIE dataset as a part of 2019 Language and Intelligence Challenge[1], which is jointly organized by China Computer Federation (CCF), Chinese Information Processing Society of China (CIPS) and Baidu Inc. As one of the three tasks in this challenge, the IE task attracted 1,836 teams from around the world. During the competition, 324 teams submitted 3,367 results in total. The performance of these results shows the effectiveness of DuIE on the evaluation of the IE techniques.

The rest of this paper is organized as follows. We first briefly describe the preparation of data and the schema for dataset construction. After that, we describe in details the coarse-to-fine dataset construction procedure, including candidate generation and crowdsourcing annotation. Then, we give the statistical analysis of the dataset and the competition on it. Finally, we conclude the paper and discuss future directions.

2 Construction of DuIE

This section describes the procedure of constructing DuIE dataset. In general, we design an effective coarse-to-fine method combining automatic distant supervision and human annotation, which is the key to achieve high data quality in large data scale.

[1] http://lic2019.ccf.org.cn/

As shown in Figure 1, our construction procedure is composed of the following three steps: (1) preparing all kinds of required data, including the schema, related SPO triples and a large-scale real-world corpus. (2) generating candidates by distant supervision methods on both SPO level and schema level to ensure high recall and precision. (3) using crowdsourcing to label the correct triples among all candidates according to sentence contexts.

Fig. 1. Procedures of DuIE construction

2.1 Data Preparation

We design a schema to guide the dataset construction. A schema is a set of triple templates, each of which is composed of a head entity type, a relation and a tail entity type:

$$\text{Schema} = \{(Subject\ type, Predicate, Object\ type)\}$$

By analyzing Baidu information retrieval and recommendation logs, we include 49 most frequently used predicate types. Table 2 shows some examples of our schema.

According to the schema, we select related subject-predicate-object triples from the structured info-boxes of Baidu Baike[2]. In details, the predicate in a triple should be semantically equivalent to a predicate in the schema, and the subject/object should be an instance of corresponding subject/object type respectively, as specified in the schema. These triples are used to annotate large amounts of raw sentences in order to produce IE instances. The raw sentences are extracted from Baidu Baike and Baidu News Feeds[3], covering major domains in real-world information requirement, including entity descriptions, entertainment news, user-generated articles and so on.

2.2 Candidate Generation

We use two types of distant supervision methods, namely SPO-level distant supervision and schema-level distant supervision, to ensure candidate quality.

SPO-level Distant Supervision. SPO-level distant supervision is a popular distant supervision method, which was widely used in existing dataset construction work. It is based on the closed-world assumption, i.e., entity information in the knowledge base is complete. In other words, if there was a relation between two entities, the triple found in the knowledge base and sentences that mention these two entities should express that relation. According to this assumption, we obtained all candidate instances in the form of $(e_1, p_1, e_2, sentence_1)$ if (e_1, p_1, e_2) and $sentence_1$ are in triple and text candidates

[2] https://baike.baidu.com/
[3] https://baijiahao.baidu.com

we got in the previous step separately, and both entity e_1 and entity e_2 appeared in sentence $sentence_1$.

Table 2. Schema examples in DuIE dataset

Subject type	Predicate	Object type	SPO example
人物 (Person)	毕业院校 (almaMater)	学校 (Educational institutions)	S: 杜道生, P: 学校, O: 北京大学 S: *Du Daosheng,* P: almaMater, O: *Peking University*
影视作品 (Film and TV works)	导演 (directedBy)	人物 (Person)	S: 逆光之恋, P: 导演, O: 任海曜 S: *The Backlight of Love,* P: directedBy, O: *Kenne Yam*
图书作品 (Book)	作者 (author)	人物 (Person)	S: 呐喊, P: 作者, O: 鲁迅 S: *Call to Arms,* P: author, O: *Lu Xun*

Schema-level Distant Supervision. In our method, schema-level distant supervision was utilized to compensate for the data incompletion problem of SPO-level distant supervision. Although information extraction datasets could be built by the SPO-level distant supervision method without any human intervention, the quality of such datasets is often limited. One crucial reason is that the closed-world assumption is not always hold. In reality, no knowledge base could include the entire set of knowledge in the world. Therefore, some correct triples mentioned a sentence could be missed in the previous step.

In order to compensate for the data incompletion problem, we proposed a schema-level distant supervision method. Firstly, for each candidate sentence, named entities with our target types were labeled by Named Entity Recognition (NER) algorithms. Secondly, entity pairs are recalled if their types matched one of the triple patterns specified in the schema. For example, in the sentence given in table 1, (*To the Happiest People, fromAlbum*, *The Great Entertainer*) would be recalled as a candidate triple in that sentence, if we know that *To the Happiest People* is a song and *The Great Entertainer* is an album, which matches the target subject and object types of predicate *fromAlbum*, even though this triple is missing in the knowledge base.

2.3 Crowdsourcing Annotation

Finally, to filter out noise instances and improve dataset accuracy, we invited some annotators to judge whether or not every candidate instance was correct on a crowdsourcing platform. For the convenience and efficiency of human annotating, we presented instances in a special question pattern. Given one instance to be labeled as $(sentence, S, P, O)$, we converted it to a judgment question:

Is this correct? <*P*> of <*S*> (*Subject type*) is <*O*> (*Object type*) according to the *sentence*.

An annotation candidate example is shown in Figure 2. Annotators had to judge whether the annotation questions were correct according to the following three criteria:

(1) Clues should be found from and only from the sentence provided. There is no need to consider whether the triple is true in the real world. (2) Subjects and objects should match the given types, which are pre-defined in the schema. (3) Predicates do not need to appear explicitly in the sentence.

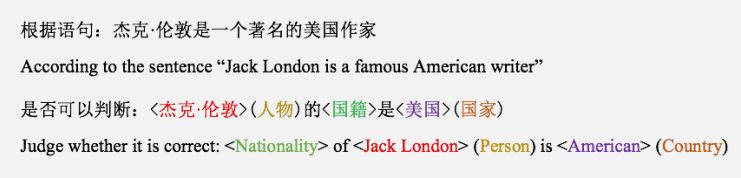

Fig. 2. An example of annotated data

While annotating test dataset, to ensure the labeling quality, each instance was first assigned to two annotators. Instances with consistent answers from the both annotators would be sent to a third annotator. During the entire annotation step, about 10 crowdsourced users were involved to work on about 640,000 candidate instances. Finally, we gathered all correct instances as the final dataset.

3 Data Statistics

Based on the above construction procedure, we build the largest Chinese information extraction dataset, DuIE, which contains 458,184 instances with 49 different predicate types, 239,663 entities, 347,250 triples, and 214,739 real-world Chinese sentences, as shown in Table 3. The average length of all sentences is 54.58, and there are 8,490 unique tokens in total. In final dataset, 78% instances are from SPO-level method, while 22% instances are from schema-level method. This shows the effectiveness of our two-level distant supervision approaches.

Table 3. Statistics of DuIE dataset in detail

#Instance	#Entity	#Relation Types	#Triple	#Sentence
458,184	239,663	49	347,250	214,739

Table 4. Comparison of DuIE Dataset with existing IE dataset

Dataset	#Relation types	#Instance
NYT-10	57	143,391
SemEval-2010 Task 8	10	10,717
FewRel	100	70,000
DuIE	49	458,184

Table 4 provides a comparison of our DuIE dataset to the existing popular IE datasets, including NYT-10, SemEval 2010 Task 8 dataset, and FewRel. It shows that DuIE is significantly larger than existing IE datasets.

DuIE dataset is split into three parts, a training set, a development set and a testing set, as show in Table 5, and there is no overlap among sentences among these three sets. Currently, the training set and development set are available to download[4].

Table 5. Statistics of DuIE training, dev, testing sets

Dataset	Training Set	Dev Set	Testing Set
#Sentence	173,108	21,639	19,992
#Instance	364,218	45,577	48,389

We further analyze the data distribution in several aspects. As we can see from Figure 3, 63% sentences are from Baidu Baike encyclopedia corpus, while 37% sentences are from Baidu Feed news. In Figure 4, distribution on different entity types is given. The most common types in DuIE are Person, Film and TV works, Song, and Book, which are is consistent with the set of top entity type found in Baidu search logs.

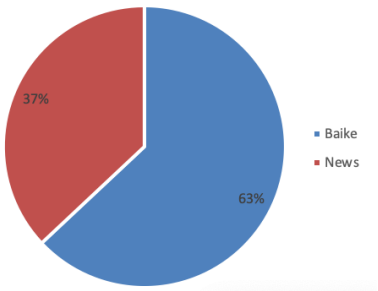

Fig. 3. Distribution on text source types

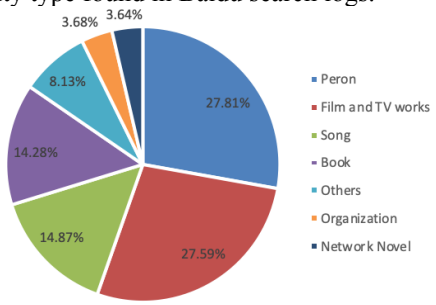

Fig. 4. Distribution on entity types

4 Evaluation on Information Extraction Task

This section provides details on the information extraction competition using the DuIE dataset, including competition task description, evaluation results and detailed analysis.

4.1 Competition Task

We hosted an IE task in 2019 Language and Intelligence Challenge[3], whose objective is to extract all correct triples according to the given sentences and a pre-defined schema. Specifically, a triple predicted by a participant system is considered as correct when its relation and two corresponding entities are matched with the triple annotated on the testing set. Considering that some entities are mentioned in sentences using aliases, we use a dictionary of alias in Baidu Knowledge Graph in the evaluation. Standard *Precision*, *Recall* and *F1* scores are used as metrics to evaluate the performance of participating systems. The final results are ranked according to the F1 value. During the

[4] http://ai.baidu.com/broad/download

period of competition, IE task have attracted 1,836 teams from both academia and industry, and 324 teams submitted 3,367 results in total.

Table 6. Evaluation results of the top 10 systems

System No.	Precision	Recall	F1	System No.	Precision	Recall	F1
S1	89.8%	88.9%	89.3%	**S6**	89.5%	87.0%	88.2%
S2	89.6%	88.9%	89.2%	**S7**	89.4%	86.8%	88.1%
S3	89.8%	88.5%	89.1%	**S8**	88.3%	86.9%	87.6%
S4	89.5%	88.6%	89.0%	**S9**	86.2%	87.9%	87.1%
S5	89.2%	88.2%	88.7%	**S10**	89.3%	84.8%	87.0%

4.2 Evaluation results

The overall competition results are published in the competition website[5]. Table 6 shows the list of top participant systems with their performance metric, ordered by their F1 values. We found that some techniques widely adopted by our participants, such as pre-trained models like BERT [13], lexical features, ensemble techniques, rule-based post-processing. In addition, some teams use parameter sharing, self-attention mechanism and manual-designed features to further improve performances.

4.3 Result Analysis

For comprehensive understanding for our dataset and related IE technologies, we performed a detailed analysis on the performance results of top participants.

Overall error analysis. We sampled incorrect triples in the top ten systems and manually labeled their error types. The top error types are shown in Table 7. "relation error", which means that an incorrect relation is extracted for an entity pair, is the most common error type and accounts for 38% of all errors. This indicates that the extraction model still has room for improvement in identifying relations between entity pairs.

The second common error type is "non-relation error", which accounted for 22% of all errors. This error type means that there is no semantic relation between the extracted subject and object in the sentence. This often happens when there are multiple entities of the same type in the given sentence. We further break down this category by source text types. An interesting observation is that "non-relation error" happens much more frequently on News text (30%) than on Baike text (17%). This indicates that it is more challenging to identify relations on more complex text styles.

The "entity boundary error" means that target entities could be found but the boundary recognition is not accurate enough which accounts for 21% of the total errors. In addition, 11% of the errors are due to the fact that entities in the triples do not conform to the types provided in schema constraint. This indicates that participants do not make

[5] http://lic2019.ccf.org.cn/

full use of labels of entity types when training models or extracting triples. There are also 8% other dispersed errors such as inference knowledge error which means that the SPO cannot be extracted without the background knowledge.

Table 7. Major error types

Error type	Error description	Example	Ratio
Relation error	The relation between subject and object is wrong.	《人龙传说》是 1999 年香港电视广播有限公司出品的古装神话剧，由罗永贤监制，陈浩民、袁洁莹主演。 Error: S:人龙传说, P: 导演, O: 罗永贤	38%
Non-relation error	There is no semantic relation between extracted subject and object in the sentence.	余思经典长篇代表作《细雨湿流光》作为青春成长的经典长篇代表作，曾得到《中国式离婚》《新结婚时代》编剧王海鸽和青春文学代表人物饶雪漫的携手力推。 Error: S: 细雨湿流光, P: 编剧, O: 王海鸽	22%
Entity boundary error	Entity recognition is incomplete or redundant.	唐寅《溪山渔隐图》（下为全卷 引首“渔隐”为乾隆御笔） 唐寅（1470-1524），字伯虎，后改字子畏，号六如居士、桃花庵主等，明代画家、书法家、诗人 Error: S: 唐寅, P: 号, O: 桃花庵主等	21%
Entity type error	Entity types do not conform to schema constraint.	《电子电路与电子技术入门》是新电气编辑部所著，科学出版社出版的图书。 Error: S: 电子电路与电子技术入门, P: 作者, O: 新电气编辑部	11%
Other errors	Other dispersed errors	杨渺，男，汉族，1970 年 9 月出生于甘肃临夏，中国甘肃国际技术合作公司副总经理。 Error: S: 杨渺, P: 国籍, O:中国	8%

Effects of source text types. Table 8 shows the average performance metrics of the top 5 systems and the top 10 systems on Baike and news texts respectively. The results show that compared with news text, the average F1 value of top 10 extraction systems on Baike text is 11.9% higher. One possible reason is that, Baike texts are usually edited by domain experts in a rather fixed format, while news texts is more complex in style and often involved with diversified linguistic patterns. Therefore, information extraction on news texts is much more difficult.

Table 8. Performance results in different sources of text

System	Baike			News		
	Precision	Recall	F1	Precision	Recall	F1
Avg-top5	92.6%	92.3%	92.4%	82.4%	80.1%	81.2%
Avg-top10	92.2%	91.5%	91.9%	81.5%	78.6%	80.0%

Single-valued v.s. multi-valued triples. We evaluated the recall of single-valued and multi-valued triples. Multi-valued triple means that a single S-P pair corresponds to multiple O values or a single P-O pair corresponds to multiple S values in the given sentence. The performance results of top five average and top 10 average systems in

multi-valued and single-valued triples are shown in Table 9 respectively. It can be seen that in top 10 systems, the average recall of the single-valued triples is 6.4% higher than that of the multi-valued triples, which indicates that it is more challenging to extract all the multi-valued triples.

Table 9. Evaluation of results in multi-valued and single-valued sentences

System	Multi-valued triples recall	Single-valued triples recall
Avg-top5	84.5%	90.3%
Avg-top10	83.1%	89.5%

We sampled some unrecalled multi-valued triples and found there were two types. As shown in Table 10, the first type is that multiple entities are adjacent or simply concatenated by a separator, while the second type is that multiple entities are not adjacent in text. It can be seen that sentence characteristics of multi-valued triples are significant, and researches need to be carried on how to model such cases in the future.

Table 10. Two types of unrecalled multi-valued triples

Multi-valued type	Instance examples
Multiple entities are adjacent	《**新六指琴魔**》是由**海润影视制作有限公司**、**耳东影业（北京）有限公司**、**华视网聚**、**横店影视制作有限公司**联合出品，香港导演王晶担任总监制、总导演。 S：新六指琴魔， P：出品公司 Multi-valued O：海润影视制作有限公司，耳东影业（北京）有限公司，华视网聚，横店影视制作有限公司
Multiple entities are not adjacent to the text	但在学习过程中她却爱上了主持这个行当，从1997年的《**五星奖**》，1998 年《**正大综艺**》，2001年《**猜猜谁会来**》，一直到的《**家庭演播室**》、《**娱乐星天地**》，**吉雪萍**主持的节目留给了人们很深刻的印象，她的名字在上海可谓家喻户晓。 Multi-valued S：五星奖，正大综艺，猜猜谁会来，家庭演播室，娱乐星天地 P：主持人 O：吉雪萍

5 Conclusion

In this paper, we present DuIE dataset, the largest high-quality Chinese information extraction dataset， which was built in a coarse-to-fine procedure combining of distant supervision and crowdsourcing annotation. To validate the dataset, we conduct a technical evaluation and analyze the errors in top systems. We found that the most common errors in information extraction systems are relation error and entity error, and current models still have rooms for improvement in these areas. For texts from different source types, the error distributions are quite different. In addition, further research is needed for small sample sizes and multi-valued triples. DuIE could help in evaluating and advancing information extraction techniques in future research.

References

1. Zeng D, Liu K, Lai S, et al. Relation classification via convolutional deep neural network[J]. (2014).
2. Jiang X, Wang Q, Li P, et al. Relation extraction with multi-instance multi-label convolutional neural networks[C]//Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. pp. 1471-1480 (2016).
3. Zeng X, He S, Liu K, et al. Large scaled relation extraction with reinforcement learning[C]//Thirty-Second AAAI Conference on Artificial Intelligence. (2018).
4. Miwa M, Bansal M. End-to-end relation extraction using lstms on sequences and tree structures[J]. arXiv preprint arXiv: 1601.00770 (2016).
5. Dai, D., Xiao, X., Lyu, Y., Dou, S., She, Q., Wang, H.: Joint Extraction of Entities and Overlapping Relations using Position-Attentive Sequence Labeling. In: AAAI (2019)
6. Takanobu R, Zhang T, Liu J, et al. A Hierarchical Framework for Relation Extraction with Reinforcement Learning[J]. arXiv preprint arXiv:1811.03925 (2018).
7. Zheng, S., Wang, F., Bao, H., Hao, Y., Zhou, P., Xu, B.: Joint extraction of entities and relations based on a novel tagging scheme. In: ACL (2017)
8. Zeng X, Zeng D, He S, et al. Extracting relational facts by an end-to-end neural model with copy mechanism[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 148-163 (2018).
9. Riedel S, Yao L, McCallum A. Modeling relations and their mentions without labeled text[C]//Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, Berlin, Heidelberg, pp. 148-163 (2010).
10. Hoffmann R, Zhang C, Ling X, et al. Knowledge-based weak supervision for information extraction of overlapping relations[C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. Association for Computational Linguistics, pp. 541-550 (2011).
11. Hendrickx I, Kim S N, Kozareva Z, et al. Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals[C]//Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions. Association for Computational Linguistics, pp. 94–99 (2009).
12. Han X, Zhu H, Yu P, et al. Fewrel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation[J]. arXiv preprint arXiv:1810.10147 (2018).
13. Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).