Semi-Open Attribute Extraction from Chinese Functional Description Text

Li Zhang ZHANGL16@UCI.EDU
University of California Irvine

Yanzeng Li* LIYANZENG@STU.PKU.EDU.CN
Ruoyu Zhang RY_ZHANG@PKU.EDU.CN
Peking University
Wangxuan Institute of Computer Technology, Peking University

Wenjie Li LIWENJIEHN@PKU.EDU.CN
Peking University
ChongQing Research Institute of Big Data, Peking University

Editors: Vineeth N Balasubramanian and Ivor Tsang

Abstract

Attribute extraction is a task to identify the attribute and the corresponding attribute value from unstructured text, which is important for extensive applications like web information retrieval and the recommended system. The traditional relation extraction-based methods or joint extraction-based systems are often perform attribute classify based on subject and attribute-value pairs, and extract the attribute triples in the scope of ontology schema categories, which is in the assumption of the close-world and cannot satisfy the diversity of attributes.

In this work, we propose a semi-open information extraction system for attribute extraction in a multi-component framework. With the proposed semi-open attribute extraction system (SOAE), more attribute-value pairs can be discovered by extracting literal triples without the limitation of pre-defined ontology. An additional co-trained ontology-based attribute extraction model is appended as a component following the assumption of the partial-closed world (PCWA), remission the performance degradation of SOAE caused by missing of the literal predicate in raw text and contribute to extract richer attribute triples and construct more dense knowledge graph. For evaluating the performance of the attribute extraction system, we construct a Chinese functional description text dataset CNShipNet and conduct experiments on it. The experimental results demonstrate that our proposed approach outperforms several state-of-the-art baselines with a large margin.

Keywords: Attribute Extraction, Information Extraction, Deep Learning, Semi-Open Information Extraction

1. Introduction

With the information explosion on Internet, how to conduct the massive unstructured text in a structured way is a great challenge for providing services like online search engines

* Corresponding author

or Knowledge Base Question Answering systems. Thus, the Natural Language Processing (NLP) community has organized numerous research on Information Extraction (IE), which is to extract structured knowledge from unstructured text (Zhang et al., 2019; Martinez-Rodriguez et al., 2018; Zheng et al., 2018), and developed semantic web technology to build and store structured knowledge as Knowledge Graph (KG) (Vileniškis and Butkienė, 2020).

KG typically organizes and expresses knowledge in the form of triple (*subject*, *predicate*, *object*) to describe things and semantic relationships (Zou et al., 2014). The most common triple in KG is composed of (*head_entity*, *relation*, *tail_entity*), which can be extracted by IE methods of Named Entity Recognition (NER) (Nadeau and Sekine, 2007), Relation Extraction (RE) (Wang et al., 2016; Zhang et al., 2018) and Joint Extraction (JE) (Wei et al., 2020; Yu et al., 2020). There has been a plethora of research in these areas.

Practically, considering the application of information retrieval, entity attribute is more useful and informative than atomic entity (Ghani et al., 2006; Ravi and Paşca, 2008). Hence there are plenty of works on the database to study how to store and manage EAV (Entity-Attribute-Value) data model (Marenco et al., 1999; Nadkarni et al., 1999; Paul and Latiful Hoque, 2011). Specifically, an attribute could be also presented as a triple (*entity*, *attribute*, *value*), which can be constructed via Attribute Extraction (AE).

Traditionally, AE is usually recognized as a classification problem and performed with NER/RE simultaneously (Shi et al., 2019), which is inflexible and limited to the design of ontology schema. Under the open-world assumption (OWA), the ontology schema cannot cover all attributes, so the AE system also needs to have the ability to discover new attributes. Various Open Information Extraction (OpenIE) systems have explored attribute extraction in the open-world scenario (Zheng et al., 2018; Zhang et al., 2019; Yu et al., 2021). However, there are still some problems in the past OpenIE and Semi-OpenIE systems: the long-range dependencies are hard to be captured, the dependence of complete literals of arguments and predicates limited accuracy of OpenIE and Semi-OpenIE in the production environment. Therefore, we model the AE system following the partial-closed world assumption (PCWA), which is an intermediate ground between OWA and CWA (close-world assumption).

In this paper, we formulate the attribute extraction task as a multi-component process to satisfy PCWA:

- **Semi-Open Attribute Extraction (SOAE)** for jointly extracting textual attribute name and attribute value which present in the text for discovering abundant attribute information in the functional description text.
- **Ontology-based Attribute Extraction (OBAE)** for extracting attribute-value pair into schema-based attribute category and completing the missed attribute name field in SOAE.

In this AE system, SOAE is the main component, which aims to extract the literal attribute name and attribute value when the core entity is explicitly described in a functional description text. [1] OBAE is a fallback component, which is used to extract ontology-based attribute triples when the attribute name is not mentioned in the text. The extraction results of SOAE and OBAE would be merged directly to obtain the final attribute triple set.

1. Our code and data are available on https://github.com/lsvih/SOAE

To evaluate our approach, we construct and release a new dataset CNShipNet[2] which contains around 5,000 triplet annotations from functional description text of ships. Experimental results show that our method outperforms several baselines for extracting attribute triples.

The main contributions of this work can be summarized as follows: We present a multi-component framework for attribute extraction from unstructured functional description text. In which, an SOAE model is introduced for extracting textual attribute and attribute value jointly, a OBAE model is introduced for extracting attribute and attribute value based on ontology schema. To study the performance of SOAE, OBAE, and AC, we construct and publish a new Chinese attribute extraction dataset CNShipNet. The experiments empirically show our approach achieves significant improvements over RE-based and JE-based baselines.

2. Preliminaries

2.1. Problem Definition

Functional description text refers to the texts that describe the functionality or performance characteristics of a certain entity (Liu et al., 2019). Given a set of unstructured functional description text and a designed ontology-based attribute set (e.g. Section 4.1), our objective is to extract the attribute names and corresponding attribute values in the raw text, as well as the attribute values for the ontology attribute categories. In short, for solving the AE problem, it is necessary to satisfy PCWA, including partial CWA and OWA. Therefore, we adopt two extraction paradigms: SOAE and OBAE to satisfy the aforementioned assumptions respectively.

Semi-Open Attribute Extraction In functional description text, the most common scenario is that the core entity (*subject*) has been determined. Therefore, how to utilize the determined subject to extract textual attributes and attribute values from unstructured text is a semi-open IE problem.

Definition 1. Let $S = \{s_i | 1 \leq i \leq n\}$ denotes a set of *subject*, $D = \{d_i | 1 \leq i \leq n\}$ presents a collection of functional description text, each d_i describes the corresponding subject s_i. Text d_i is consists of a sequence of token $d_i = \{t_{i,1}, t_{i,2}, ..., t_{i,l}\}$, the l is the length of sequence d_i. Extract all attribute triple $\{< s_i, t_{i,[u:u+w]}, t_{i,[p:p+q]} >, ...\}$, in which $t_{i,[u:u+w]}$ and $t_{i,[p:p+q]}$ are extracted textual token from d_i, w and q denote the length of attribute name and attribute value respectively.

Ontology-based Attribute Extraction Given an attribute category from the designed ontology schema and unstructured functional description text, extracting the attribute value according to the subject is a closed-world IE problem.

Definition 2. Given a subject set S, a functional description text set D and an attribute category set $A = \{a_i | 1 \leq i \leq m\}$ defined by ontology schema, extract all attribute triple $\{< s_i, a_j, v_{i,k} >, ...\}$, the $v_{i,k}$ is attribute value described in d_i.

Problems in Real-World Scenarios In SOAE and OBAE, we have summarized several features which are essential in attribute extraction based on real-world scenarios:

2. The processed dataset is publicly available at: https://github.com/lsvih/SOAE

- **Overlapping** (OVERLAP): Overlapped or nested attribute values with shared attribute name.
- **Subject-agnostic** (NON-SUBJ): The subject is missing in some long-distance dependency or coreference situations.
- **Textual-predicate** (T-PRED): According to the OWA, the system needs to be able to extract the out-of-ontology attributes (textual predicates) from raw text.
- **Non-fact predicate** (NON-PRED): The ability of systems to extract attribute in ontology that are not explicitly mentioned in the raw text.

These problems will obviously affect the accuracy of the extraction of attributes, therefore the AE system that has more of the above features would achieve better performance.

2.2. Pre-trained Language Model

Recently, pre-trained language models (PLM) (Peters et al., 2018; Devlin et al., 2019; Radford et al., 2018; Yang et al., 2019) have demonstrated the effectiveness on a variety of NLP tasks such as NER, sentence pair matching (SPM), and machine reading comprehension (MRC). An explanation for the strong performance of the PLM is that PLMs could leverage large-scale unlabeled corpus and obtain prior knowledge of the language in the pre-training stage, which provides necessary information for fine-tuning stage and improves the performance of downstream tasks.

For example, a well-known PLM, BERT (Devlin et al., 2019) introduced Next Sentence Predict (NSP) pre-training task to estimate whether the two sentences are in continuous context. Specifically, the NSP is conducted input sequence in the form of

$$[\texttt{CLS}]\{t_1, t_2, ..., t_n\}_1[\texttt{SEP}]\{t'_1, t'_2, ..., t'_m\}_2[\texttt{SEP}]$$

in which $\{(\cdot)\}_i$ denotes the sequence of tokens in the i-th sentence. Benefit from the design of TransformerBlock (Vaswani et al., 2017), tokens in the two sentences could perform information interaction by multi-head attention operator effectively. The downstream fine-tuning tasks like SPM and MRC, which are following this paradigm to contrast two sentences, are gaining the maximizing promotion of performance.

2.3. Joint Extraction

Joint Extraction (JE) aims to detect *subject-object* pairs with the corresponding *predicate* in a single model, bridges the gap between extract-then-classify and unified labeling approaches. The experimental results of prior works prove that a joint learning framework could bring a remarkable improvement compared to the assembly of several NER and RE models. Besides, benefit from the interaction between different tasks, JE could get some additional features, such as extracting overlapping relationships, etc.

JE system usually constructs a two-stage model: extracting entity or subject first, and then classify the relation (Wei et al., 2020; Yu et al., 2020). Recently, TPlinker (Wang et al., 2020b) proposed a single-stage JE method via transforming the JE task into a token pair linking problem. TPLinker obtained better performance by bridging the gap between training and inference, which is leading by inconsistent two-stage processes. Specifically, TPLinker designed the paradigm of JE extraction as:

Given a sentence, two positions p_1, p_2 and a relation r. To estimate:

- can p_1 and p_2 determine an entity?
- whether two entities respectively start with p_1 and p_2 have relation r?
- whether two entities respectively end with p_1 and p_2 have relation r?

After answering the above three questions, all the entity spans and their relations are extracted naturally. In other words, the JE task transforms into a classification problem of the token pair, which is composed of the tokens' position at p_1 and p_2. We abstract the main idea of TPLinker and adopt it in the OBAE component, which is detailedly described in Section 3.3.

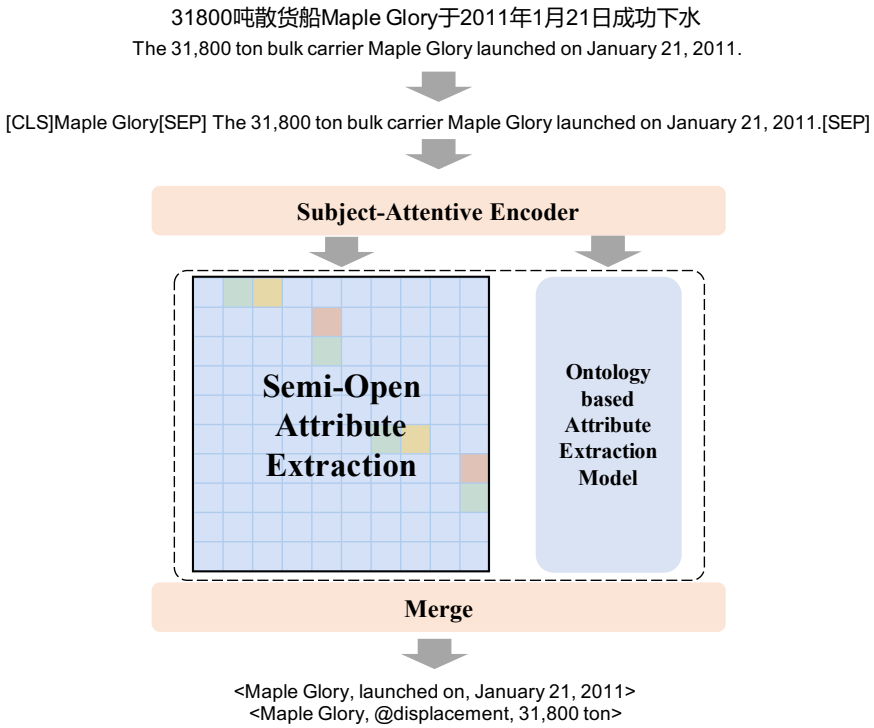

Figure 1: Overview of the architecture, components and processes of our system. The attribute name starting with the @ symbol indicates that is an ontology-based attribute category.

3. Methodology

Figure 1 demonstrates the overall architecture of our approach.

3.1. Subject-Attentive Encoder

In this work, we adopt BERT (Devlin et al., 2019) as the basic encoder. As described in Section 2.2, the PLM could obtain prior knowledge of the language and semantic through specific model architecture and well-designed pre-training tasks.

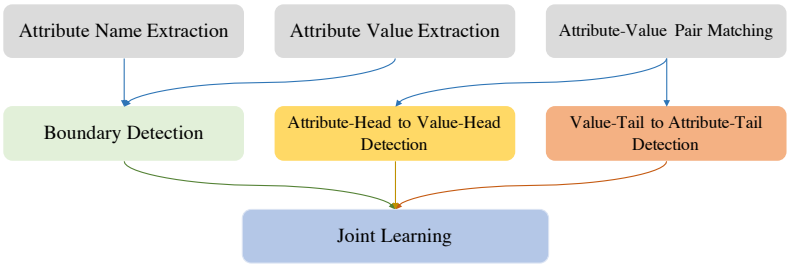

Figure 2: Decomposed Subtasks of Semi-Open Attribute Extraction.

Considering our problem statement, integrating *subject* information into text representation could guide the extraction of the corresponding *predicate* and *object*. Inspired by MRC-NER (Li et al., 2020), we organize the input sequence of encoder as a query-sentence formulation as follows:

$$\text{TEXT}_{subject} = [\text{CLS}]\{s_1, s_2, ..., s_n\}^s[\text{SEP}]\{t_1, t_2, ..., t_m\}[\text{SEP}] \quad (1)$$

$$\hat{\mathbf{H}} = \text{BERT}(\text{TEXT}_{subject}) \quad (2)$$

the $\{(\cdot)\}^s$ represents the *subject*, n and m denotes the lengths of *subject* and sentence respectively. BERT receive the combined token sequence as Equation 2 and then output the encoded text representation $\hat{\mathbf{H}} \in \mathbb{R}^{(n+m+3)\times d}$, in which d denotes the encoding dimension, $(n+m+3)$ is the sequence length of combined input sequence.

3.2. Semi-Open Attribute Extraction

Semi-Open Attribute Extraction (SOAE) aims to extract attribute name and attribute value pair from text according to the *subject*. The SOAE has been decomposed into three subtasks:

- Attribute extraction, to recognize the name of attribute value like NER. In SOAE, the attribute name should present as a mention in the raw text. The attribute name which not appear in the raw text would be ignored in the SOAE stage and fall back to OBAE, which is detailed in Section 3.3.
- Attribute value extraction, to extract the value of attributes. Generally, the attribute value is numeric *object* or a specific segment of description text.
- Attribute-value pair relationship extraction, to match the extracted attribute names and attribute values and construct triple, which is similar to the definition of RE task.

Through the analysis of these subtasks, we combined the attribute extraction and attribute value extraction task into a boundary detection task, which aims to divide the textual segments of attribute name and attribute value jointly. Inspired by TPLinker (Wang et al., 2020b), the attribute-value pair matching task is disassembled into two path detection tasks: attribute head to attribute value head (ANH-AVH) path detection and attribute value tail to attribute tail (AVT-ANT) path detection. With the determination of the ANH-AVH and AVT-ANT path, the attribute name and attribute value are naturally linked and extracted. By extracting attribute triples in this way, the problem of missing extraction

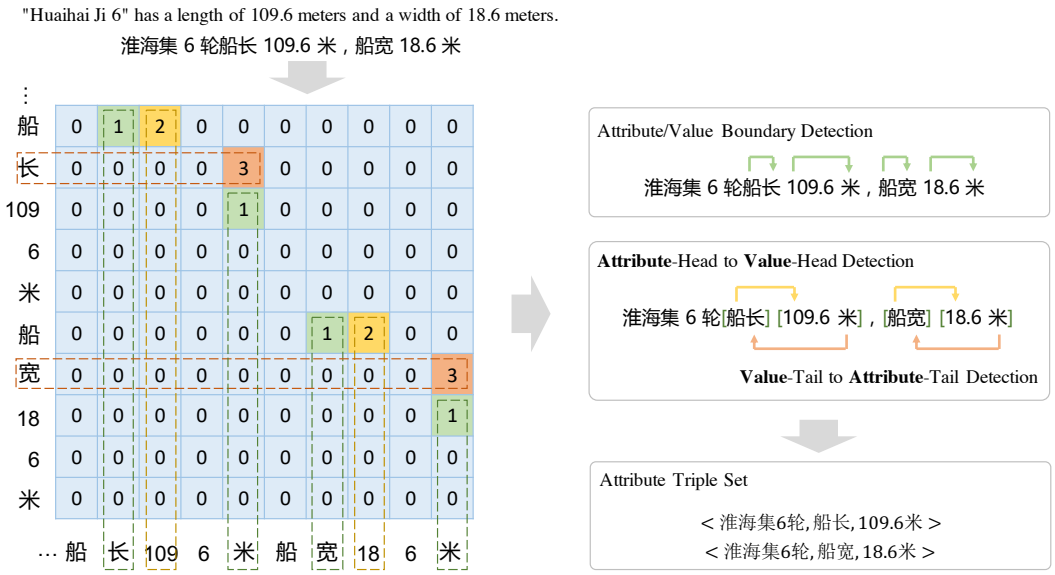

Figure 3: Left: the tagging scheme of semi-open attribute extraction, different colors represent the different subtasks. The symbol "1,2,3" in table represents label [AN | H-AN | T],[AN | H-AV | H] and [AV | T-AN | T] respectively. Right: the exploded view of subtasks corresponding to the left table, and the decoded results.

caused by overlap can be effectively avoided, which occurs in AE frequently. Figure 2 shows how the SOAE is decomposed into subtasks and the re-association of subtasks.

To reduce cascade errors, we model and train these three subtasks jointly. As illustrated in Figure 3, boundary detection task, ANH-AVH, and AVT-ANT path detection task are converted to a single token-pairs tagging task. Specifically, we define tagging label for token-pairs: $[\alpha|\beta\text{-}\alpha|\beta]$, in which $\alpha \in \{\texttt{AV}, \texttt{AN}, \texttt{O}\}$, $\beta \in \{\texttt{H}, \texttt{T}, \emptyset\}$. α is symbol for identifying the type of token, β is symbol for clearing weather a token is head or tail of a attribute or attribute value. An example is provided in Figure 3, there is a attribute name [船长](length), thus the token pair (船,长) has been tagged as [AN | H-AN | T] to recognize these two tokens forming an attribute name (AN). Similarly, the token pair (船,109) and (米,长) are tagged as [AN | H-AV | H] and [AV | T-AN | T], implying a attribute-value pair (船长,109.6米).

By utilizing such unified labeling of token pairs, the loss function is defined as:

$$P(y_{i,j}) = \texttt{softmax}(\mathbf{W}[\mathbf{h}_i, \mathbf{h}_j] + \mathbf{b}) \tag{3}$$

$$J_{SOAE} = -\frac{1}{l} \sum_{i=1, j \geq i}^{l} \sum_{(* \neq \emptyset)} \log P\left(y^*_{i,j} = \hat{y}^*\right) \tag{4}$$

in which $[\mathbf{h}_i, \mathbf{h}_j]$ is the combined representation of token pair (t_i, t_j), $\hat{y}$ is the golden tag, l is the length of sequence, the condition $j \geq i$ limit the token pair constructed in an orderly manner for pruning, $(* \neq \emptyset)$ stipulates that only the label of token pairs include

in {[AN | H-AN | T], [AN | H-AV | H], [AV | T-AN | T]} would be calculated gradient, **W** and **b** are trainable parameters.

3.3. Ontology-based Attribute Extraction

As Section 2.1 described, an OBAE component is introduced to perform attribute extraction based on ontology categories, and cover the non-fact predicate problem that SOAE cannot extract attributes due to the absence of predicate in raw text.

To solve the aforementioned problem, a JE-based model (Wang et al., 2020b) is adopted to build this OBAE component. The basic approach and statement of this JE-based OBAE component have described in Section 2.3.

This work is toward semi-open attribute extraction. Facing the dataset of semi-open IE, the *predicate*s in each sample are exists as literal mention. In general, transplanting the semi-open IE dataset to the close world model will cause a serious long-tailed classification problem and cause the model performance degradation, as most of the literal attributes only appear a few times. Fortunately, this work is mainly to explore the semi-open attribute extraction system. The OBAE model is only employed as an accessory to supplement the attributes of the ontology category, and the attribute organized in the ontology category is generally sufficient for training.

In addition, since the *subject* is determinate in the semi-open IE paradigm, similarly, it is no longer necessary to predict the *subject* when process joint learning in the JE-based model, which reduces the difficulty of training. The loss function of the JE-based OBAE model can be denoted as:

$$J_{OBAE} = J'_{predicate} + J'_{object} \tag{5}$$

Formally, only the predicate classification loss and the object tagging loss should be calculated in training.

4. Experiments

4.1. Dataset

We collected our data from an open-access Chinese ship information website[3], it is a Chinese ship news website that reports global newly built ships. For efficiency, we only extracted the paragraphs that mention the brief introductions of ships with usually 1-3 long sentences as our text input, which is the most informative part of functional description text. These sentences might have attributes such as ship height, ship weight, ship speed, etc.

Dataset Schema We defined the attribute categories shown in Table 1 for the dataset.

Data Annotations For each sample, we manually annotated the ship name with an annotator tool (Li et al., 2021) then find all attributes that belong to the ship specifically. The dataset is constructed as a triplet set, each triplet is composed of a ship name, an attribute name, and the corresponding attribute value.

If the ship name does not have a literal value in the text, we annotated it as a placeholder; if the attribute name is not appeared in the sample meanwhile attribute value exists, we

3. www.cnshipnet.com

Table 1: Pre-defined attribute schema of CNShipNet dataset.

Attribute Name	Description
船长(length)	The maximum horizontal distance from the very front of the bow to the very end of the stern parallel to the design waterline.
船宽(width)	At the widest point of a ship, the horizontal distance measured from the outer edge of the ribs on one side to the outer edge of the ribs on the other side is called the ship's width.
船深(depth)	The distance from keel line to deck perpendicular to base plane.
吃水(draught)	The depth of a ship below the surface of the water.
航速(speed)	The distance covered by a ship in unit time.
载重(tonnage)	Maximum carrying capacity of a ship.
下水(launch date)	The date of a ship enters the water for the first time.
试航(trial voyage date)	The date of a ship makes a trial voyage in the designated area.

annotated the attribute as an attribute class based on the pre-defined schema. Finally, we tagged around 5,000 entity-attribute-value triples for this dataset.

4.2. Experimental Setting

Evaluation Metrics Following previous studies on IE (Baldini Soares et al., 2019; Zhou et al., 2016; Wang et al., 2020b), we evaluate micro average F1 scores on dev set and test set respectively. Each experiment is run five times with a random seed, and the average score is reported.

Experimental Configurations We built our model with PyTorch, the implementation of BERT-based models and pre-trained PLM weights are provided by huggingface (Wolf et al., 2020). In our experiments, the epoch and batch size is set to 50, 32 respectively. The optimizer is BERTAdam with an initial learning rate of $5e$-5. The input text is truncated at the position of 256. CoreNLP (Manning et al., 2014) is employed to obtain the dependency trees for GCN and AGGCN model. Early-stopping is adopted to choose the model with the best F1 score on dev set. All the hyper-parameters are tuned on the dev set.

Baseline Models We compared our AE system with a series of strong baselines, including heuristic method, RE-based method and JE-based system: **Heuristic method:** Since the attributes and attribute values in the CNShipNet dataset have an obvious pattern, we use syntactic and lexicon features and employ regular expressions to design a heuristic method refer to Zhao et al. (2010). **CNN:** a simple convolutional network to learn the sentence representation for classification, which contains word embedding and position embedding (Zeng et al., 2014). **Att-GRU:** an Attention-Based Bidirectional Recurrent Neural Network (Att-RNN) based model, with the ability to capture the most important semantic information in a sentence automatically (Zhou et al., 2016). **C-GCN:** a graph model, conduct contextualized GCN over the dependency tree to capture the dependency structure (Zhang et al., 2018) [4]. **AGGCN**: an upgraded C-GCN, incorporates dependency trees into the model and utilizes attention mechanism for pruning (Guo et al., 2019). **BERT:** a simple baseline,

4. https://github.com/zjunlp/deepke

Table 2: F1 score of different IE systems over proposed CNShipNet dataset. Function List columns present the ability of these models to solve practical problems in real-world scenarios, features and abbreviations are described in Section 2.1.

System	CNShipNet		Function List			
	dev	test	OVERLAP	NON-SUBJ	T-PRED	NON-PRED
Heuristic method	37.882	42.295		✓	✓	
CNN(Zeng et al., 2014)	68.326	67.513				✓
Att-GRU(Zhou et al., 2016)	75.930	70.572				✓
C-GCN(Zhang et al., 2018)	79.563	79.180				✓
AGGCN(Guo et al., 2019)	79.019	80.671				✓
BERT(Devlin et al., 2019)	78.866	79.250				✓
MTB(Baldini Soares et al., 2019)	81.996	83.924				✓
ETL-Span(Yu et al., 2020)	80.935	79.257	✓			✓
CasRel$_{LSTM}$(Wei et al., 2020)	79.102	80.324	✓			✓
CasRel$_{BERT}$(Wei et al., 2020)	83.090	83.846	✓			✓
OBAE$_{LSTM}$(Wang et al., 2020b)	81.496	80.177	✓			✓
OBAE$_{BERT}$(Wang et al., 2020b)	84.655	85.280	✓			✓
Ours(SOAE)	87.614	88.927	✓	✓	✓	
Ours(SOAE+OBAE)	**89.137**	**91.455**	✓	✓	✓	✓

the of *subject* and attribute value are sent to the BERT (Devlin et al., 2019) incorporating positional encoding, and classify their integrated representation for determining attribute category. **MTB** uses special tokens to mark the location of the subject and object in the sentence, then concatenates the contextual word representations of special token to predict the relationship (Baldini Soares et al., 2019). **ETL-Span**[5]**:** Yu et al. (2020) proposed a novel span-based tagging scheme, which could be solved by a hierarchical boundary tagger conveniently to model the internal dependencies of triples jointly. **CasRel**[6]**:** CasRel (Wei et al., 2020) is a JE model, proposed a novel hierarchical binary tagging framework that recognizes head-entities and all possible tail-entities-relation pair in two steps, achieve state-of-the-art on the public benchmark.

4.3. Experimental Results

As shown in Table 2, we compare our systems with the baseline systems. Generally, our method consistently outperforms all baselines on dev and test set, which demonstrates the effectiveness of the proposed approach.

In detail, comparing to classification-based RE models (CNN, Att-GRU, C-GCN, AGGCN, BERT, and MTB), our model achieves significant improvements on the test set of CNShipNet. The largest margin of performance is 23.942% (compared with CNN) and the smallest performance gap reached 7.531% (compared with MTB). The main problem

5. https://github.com/yubowen-ph/JointER
6. https://github.com/longlongman/CasRel-pytorch-reimplement

of the obvious performance gap is the close-world relation extraction model is cannot discover diversified attributes well. While modeling the attribute extraction as a classification problem, it is hard for models to deal with the long tail problem with the lack of data.

Comparing with JE systems, which extract attribute and attribute-value jointly and can handle the problem of the overlapping argument, our system also shows stable improvements. Our system respectively achieves 12.198%, 7.609%, and 6.175% average improvement compared with ELT-Span, BERT-based CasRel, BERT-based TPLinker. We consider that it is because our model is capable of identifying more attribute triple with textual attribute names.

4.4. Ablation Study

Table 3: An ablation study of our AE system, evaluate on CNShipNet dev set.

Objective	**F1**
Ours(SOAE + OBAE)	89.137
- w/o OBAE	87.614
- w/o Subject-Attentive Encoder	84.793
- w/o OBAE and Subject-Attentive Encoder	84.162

We conduct ablation experiments to explore the effectiveness of each component in our system. The ablation experiments are organized on the dev set of CNShipNet, we respectively remove one particular component at a time to evaluate the impact on the performance. The results of the ablation study are demonstrated in Table 3. We can observe that: (1) After removing OBAE, the performance dropped by 1.523%, and the model lost the ability to extract attribute triples with missing predicates. (2) Without the subject-attentive encoder (replaced it with the original BERT encoder), the result was reduced by 4.344%, which indicates that the subject plays a critical role in the SOAE although the subject is not the target of SOAE. Specifically, the subject-attentive encoder could guide the model filter attribute-value pairs that are not related to the core entity.

5. Related Works

Relation Extraction Relation extraction is an important sub-task in NLP and IE, which aims to extract semantic relations between pairs of co-occurring entities in text. The most common setting is constructed under CWA, where a set of predefined relation types are provided. Therefore, RE has been often modeled as a classification problem (Jiang, 2012).

In recent years, neural networks have dominated the RE task with their powerful discriminative feature learning ability. Zeng et al. (2014) introduced a convolutional neural network to model over the whole sentence with some linguistic tags and improved the classification accuracy. Zhou et al. (2016) proposed attention-based bidirectional long short-term memory (LSTM) networks for RE. The attention mechanism could automatically focus on the words which have a decisive effect on classification. Baldini Soares et al. (2019) proposed a BERT based model where they adopted marker technique (i.e. special token pairs `[SBJ]`, `[\SBJ]`, `[OBJ]`, `[\OBJ]`) to enclose the subject and object entities, and then used the vector of the start token of pair entities for relation classification.

Joint Extraction The traditional pipelined IE manner is usually composed by NER and RE. Those pipelined IE systems usually encounter error propagation and unable to capture the interaction between entities and relations. To overcome these shortages, researchers proposed Joint Extraction (JE) which targets to jointly extract the entities and relations.

CoType (Ren et al., 2017) formulates the JE problem as a global embedding problem, learns the representations of entity mentions, relation mentions and type labels jointly, and promote the extraction performance. CasRel (Wei et al., 2020) provides a novel cascade binary tagging framework that models relations as functions that map subjects to objects in a sentence to extract head-entities and relation-tail-entity pairs in two-stage. Similar to the CasRel, ETL (Yu et al., 2020) is a two-stage model and can extract triple jointly, benefit from hierarchical boundary tagger, ETL has the ability to extract discontinuous predicates and objects. TPLinker (Wang et al., 2020b) is a state-of-the-art method in JE, which employs a token pair linking model with a novel handshaking tagging scheme and performs binary classification to detect start and end positions of entities. The detail about TPLinker is described in Section 2.3.

Attribute Extraction Similar to RE and JE, Attribute Extraction (AE) is also an important subtask of knowledge graph construction. Early works on attribute value extraction use rule-based extraction techniques (Vandic et al., 2012; More, 2016) which bring in the dictionary or rule-based feature to identify attributes and attribute values. Recently, researchers tend to model AE as an IE and solved with RE and JE incidentally (Baldini Soares et al., 2019; Wei et al., 2020; Yu et al., 2020). Besides, sequence tagging is also welcome to model the attribute extraction problem. OpenTag (Zheng et al., 2018) proposed an end-to-end Attention-BiLSTM-CRF-based sequence tagging model to extract attribute value.

As described in Section 2.1, classification-based AE often faces the problem that cannot discover new attributes in the open world. Open Information Extraction (OpenIE) systems are rise to discover attributes in a more open perspective (Etzioni et al., 2008; Zhang et al., 2019). ReNoun (Yahya et al., 2014) propose an OpenIE system that bootstrapping a seed set from training data, and extracts noun phrase attribute. Martinez-Rodriguez et al. (2018) developed an OpenIE system that utilizes NLP approaches to find named entities including attributes, and obtain binary relations from open knowledge sources to construct the knowledge graph. Besides, Wang et al. (2020a) adopted a novel question answering formulation to extract attribute value.

6. Conclusion

In this paper, we propose a semi-open attribute extraction system to extract abundant ontology-free entity-attribute-value triple from raw text. To detect the fact of attribute and attribute value, we recompose subtasks of SOAE and adopt an end-to-end matrix tagging scheme to convert the SOAE problem to a head-tail-path detection task. An OBAE sub-component is introduced as a supplement to SOAE, used to extract ontology-category-based attribute-value pairs, and to solve the problem of the non-fact predicate in OpenIE. By the modeling of SOAE and the combination of an OBAE sub-component, this SOAE system provides the feature of extracting both explicit and implicit attribute triples jointly and satisfies the assumption of the partial-closed world (PCWA). We conduct experiments on a benchmark dataset CNShipNet, which is constructed and annotated for evaluating the

comprehensive abilities of AE systems, and the experimental results prove the effectiveness of our approach. In the future, we will further explore the design of this SOAE system, combining OBAE and SOAE into a unified end-to-end model.

References

Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. Matching the blanks: Distributional similarity for relation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2895–2905, Florence, Italy, 2019.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. 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)*, pages 4171–4186, Minneapolis, Minnesota, 2019.

Oren Etzioni, Michele Banko, Stephen Soderland, and Daniel S Weld. Open information extraction from the web. *Communications of the ACM*, 51(12):68–74, 2008.

Rayid Ghani, Katharina Probst, Yan Liu, Marko Krema, and Andrew Fano. Text mining for product attribute extraction. *ACM SIGKDD Explorations Newsletter*, 8(1):41–48, 2006.

Zhijiang Guo, Yan Zhang, and Wei Lu. Attention guided graph convolutional networks for relation extraction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 241–251, Florence, Italy, 2019.

Jing Jiang. Information extraction from text. In *Mining text data*, pages 11–41. Springer, 2012.

Xiaoya Li, Jingrong Feng, Yuxian Meng, Qinghong Han, Fei Wu, and Jiwei Li. A unified MRC framework for named entity recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5849–5859, Online, 2020.

Yanzeng Li, Bowen Yu, Li Quangang, and Tingwen Liu. FITAnnotator: A flexible and intelligent text annotation system. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*, pages 35–41, Online, June 2021.

Qu Liu, Zhenyu Zhang, Yanzeng Li, Tingwen Liu, Diying Li, and Jinqiao Shi. Icnet: Incorporating indicator words and contexts to identify functional description information. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2019. doi: 10.1109/IJCNN.2019.8852428.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland, 2014.

Luis Marenco, Prakash Nadkarni, Emmanouil Skoufos, Gordon Shepherd, and Perry Miller. Neuronal database integration: the senselab eav data model. In *Proceedings of the AMIA Symposium*, page 102. American Medical Informatics Association, 1999.

Jose L Martinez-Rodriguez, Ivan López-Arévalo, and Ana B Rios-Alvarado. Openie-based approach for knowledge graph construction from text. *Expert Systems with Applications*, 113:339–355, 2018.

Ajinkya More. Attribute extraction from product titles in ecommerce. *arXiv preprint arXiv:1608.04670*, 2016.

David Nadeau and Satoshi Sekine. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26, 2007.

Prakash M Nadkarni, Luis Marenco, Roland Chen, Emmanouil Skoufos, Gordon Shepherd, and Perry Miller. Organization of heterogeneous scientific data using the eav/cr representation. *Journal of the American Medical Informatics Association*, 6(6):478–493, 1999.

Razan Paul and Abu Sayed Md Latiful Hoque. Optimized entity attribute value model: A search efficient re-presentation of high dimensional and sparse data. *Interdisciplinary Bio Central*, 3(3):9–1, 2011.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, 2018.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding with unsupervised learning. 2018.

Sujith Ravi and Marius Paşca. Using structured text for large-scale attribute extraction. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 1183–1192, 2008.

Xiang Ren, Zeqiu Wu, Wenqi He, Meng Qu, Clare R. Voss, Heng Ji, Tarek F. Abdelzaher, and Jiawei Han. Cotype: Joint extraction of typed entities and relations with knowledge bases. In *Proceedings of the 26th International Conference on World Wide Web, WWW 2017, Perth, Australia, April 3-7, 2017*, pages 1015–1024. ACM, 2017.

Xue Shi, Yingping Yi, Ying Xiong, Buzhou Tang, Qingcai Chen, Xiaolong Wang, Zongcheng Ji, Yaoyun Zhang, and Hua Xu. Extracting entities with attributes in clinical text via joint deep learning. *Journal of the American Medical Informatics Association*, 26(12): 1584–1591, 2019.

Damir Vandic, Jan-Willem Van Dam, and Flavius Frasincar. Faceted product search powered by the semantic web. *Decision Support Systems*, 53(3):425–437, 2012.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008, 2017.

Tomas Vileiniškis and Rita Butkienė. Leveraging predicate-argument structures for knowledge extraction and searchable representation using rdf. *International Journal of Knowledge Engineering*, 6(1):30–34, 2020.

Linlin Wang, Zhu Cao, Gerard de Melo, and Zhiyuan Liu. Relation classification via multi-level attention CNNs. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1298–1307, Berlin, Germany, 2016.

Qifan Wang, Li Yang, Bhargav Kanagal, Sumit Sanghai, D. Sivakumar, Bin Shu, Zac Yu, and Jon Elsas. Learning to extract attribute value from product via question answering: A multi-task approach. In *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020*, pages 47–55. ACM, 2020a.

Yucheng Wang, Bowen Yu, Yueyang Zhang, Tingwen Liu, Hongsong Zhu, and Limin Sun. TPLinker: Single-stage joint extraction of entities and relations through token pair linking. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1572–1582, Barcelona, Spain (Online), 2020b. International Committee on Computational Linguistics.

Zhepei Wei, Jianlin Su, Yue Wang, Yuan Tian, and Yi Chang. A novel cascade binary tagging framework for relational triple extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1476–1488, Online, 2020.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, 2020.

Mohamed Yahya, Steven Whang, Rahul Gupta, and Alon Halevy. ReNoun: Fact extraction for nominal attributes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 325–335, Doha, Qatar, 2014.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 5754–5764, 2019.

Bowen Yu, Zhenyu Zhang, Xiaobo Shu, Yubin Wang, Tingwen Liu, Bin Wang, and Sujian Li. Joint extraction of entities and relations based on a novel decomposition strategy. In *Proc. of ECAI*, 2020.

Bowen Yu, Zhenyu Zhang, Jiawei Sheng, Tingwen Liu, Yubin Wang, Yucheng Wang, and Bin Wang. Semi-open information extraction. In *Proceedings of the Web Conference 2021*, WWW '21, pages 1661–1672. ACM, 2021.

Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. Relation classification via convolutional deep neural network. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2335–2344, Dublin, Ireland, 2014. Dublin City University and Association for Computational Linguistics.

Dongxu Zhang, Subhabrata Mukherjee, Colin Lockard, Luna Dong, and Andrew McCallum. OpenKI: Integrating Open Information Extraction and Knowledge Bases with Relation Inference. 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)*, pages 762–772, Minneapolis, Minnesota, 2019.

Yuhao Zhang, Peng Qi, and Christopher D. Manning. Graph convolution over pruned dependency trees improves relation extraction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2205–2215, Brussels, Belgium, 2018.

Yanyan Zhao, Bing Qin, Shen Hu, and Ting Liu. Generalizing syntactic structures for product attribute candidate extraction. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 377–380, Los Angeles, California, 2010.

Guineng Zheng, Subhabrata Mukherjee, Xin Luna Dong, and Feifei Li. Opentag: Open attribute value extraction from product profiles. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, August 19-23, 2018*, pages 1049–1058. ACM, 2018.

Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 207–212, Berlin, Germany, 2016.

Lei Zou, Ruizhe Huang, Haixun Wang, Jeffrey Xu Yu, Wenqiang He, and Dongyan Zhao. Natural language question answering over rdf: A graph data driven approach. In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, SIGMOD '14, pages 313–324, New York, NY, USA, 2014. ACM.