RESEARCH PAPER

Improving Extraction of Chinese Open Relations Using Pre-trained Language Model and Knowledge Enhancement

Chaojie Wen, Xudong Jia, Tao Chen†

Faculty of Intelligent Manufacturing, Wuyi University, Jiangmen, Guangdong, China

Keywords: Chinese open relation extraction; Pre-trained language model; Knowledge enhancement

Citation: Wen, C., Jia, X., Chen, T.: Improving Extraction of Chinese Open Relations Using Pre-trained Language Model and Knowledge Enhancement. Data Intelligence 5(4), 962-989 (2023).
Received: August 21, 2023; Revised: September 15, 2023; Accepted: October 20, 2023

ABSTRACT

Open Relation Extraction (ORE) is a task of extracting semantic relations from a text document. Current ORE systems have significantly improved their efficiency in obtaining Chinese relations, when compared with conventional systems which heavily depend on feature engineering or syntactic parsing. However, the ORE systems do not use robust neural networks such as pre-trained language models to take advantage of large-scale unstructured data effectively. In respons to this issue, a new system entitled Chinese Open Relation Extraction with Knowledge Enhancement (CORE-KE) is presented in this paper. The CORE-KE system employs a pre-trained language model (with the support of a Bidirectional Long Short-Term Memory (BiLSTM) layer and a Masked Conditional Random Field (Masked CRF) layer) on unstructured data in order to improve Chinese open relation extraction. Entity descriptions in Wikidata and additional knowledge (in terms of triple facts) extracted from Chinese ORE datasets are used to fine-tune the pre-trained language model. In addition, syntactic features are further adopted in the training stage of the CORE-KE system for knowledge enhancement. Experimental results of the CORE-KE system on two large-scale datasets of open Chinese entities and relations demonstrate that the CORE-KE system is superior to other ORE systems. The F1-scores of the CORE-KE system on the two datasets have given a relative improvement of 20.1% and 1.3%, when compared with benchmark ORE systems, respectively. The source code is available at https://github.com/cjwen15/CORE-KE.

† Corresponding author: Tao Chen (E-mail: chentao1999@gmail.com; ORCID: 0000-0002-3634-0854).

1. INTRODUCTION

A well-written text document, no matter whether it is in English or Chinese, often consists of sentences whose entities are linked through semantic relations (for example, "employment" relation between "person" and "company", "has" relation between "product" and "feature", and "is a" relation between two "concepts"). Relation Extraction (RE) is a task of extracting semantic relations from a given text document. As an important step in Information Extraction (IE), relation extraction, after Named Entity Recognition (NER), identifies semantic relations in each sentence of a text document.

Open relation extraction (ORE), different from conventional relation extraction, does not require pre-defined relation types to automatically discover possible relations of interest using text corpus without any human involvement [1]. Given a sentence [在欧洲杯的精彩演出后，鲁尼加盟了曼联 (*After giving a wonderful show in the European Cup, Rooney joined Manchester United*)] (as shown in Figure 1), a triple (鲁尼$_{Rooney}$, 加盟$_{joined}$, 曼联$_{Manchester\ United}$) can be extracted by an ORE system, where the relation 加盟 (*joined*) is not defined in advance in the ORE system.

There are several English oriented ORE systems, such as TextRunner [2], ReVerb [3], and OpenIE6 [4]. These systems, which use morphological features, usually perform well in English corpus but give poor results in Chinese texts [5]. With understanding this incompatibility, a group of researchers recently have paid attention to the studies on Chinese ORE, used external syntactic or semantic knowledge to manually design extraction rules, and extracted open semantic relations from Chinese texts [5, 6]. At the same time, another group of researchers have developed Chinese ORE methods/systems to extract semantic relations through relatively simple supervised neural networks [7, 8]. These Chinese ORE methods/systems however do not use robust neural networks such as Pre-trained Language Models (PLM) to effectively extract open relations from large-scale unlabeled data.

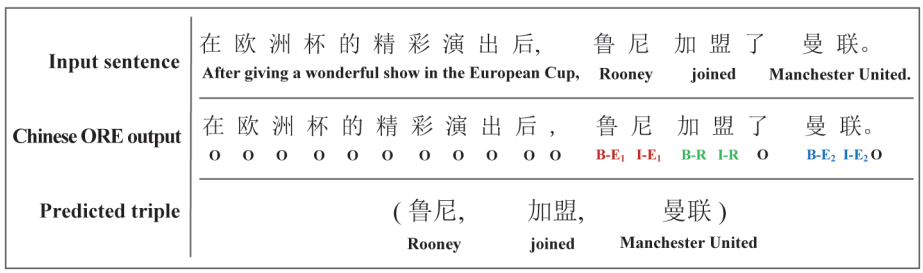

Figure 1. An example of Chinese ORE.

In this paper, a new ORE system, entitled the Chinese Open Relation Extraction with Knowledge Enhancement (CORE-KE) system, is presented to strengthen the Bidirectional LongShort-Term Memory-Masked Conditional Random Field (BiLSTM-Masked CRF) layers on a PLM with external syntactic or semantic knowledge for Chinese open relation extraction. First, Wikidata[①] (a free and open knowledge base

[①] https://www.wikidata.org/wiki/Wikidata:Main_Page

that can be read and edited by both human beings and machines) and the Dependency Semantic Normal Forms (DSNFs) tools developed by Jia et al. [9] are adopted in this study to obtain additional descriptive and triplet knowledge corpus from Chinese ORE datasets. The enhanced knowledge base is further used to pre-train and fine-tune the PLM within the CORE-KE system. Second, the Language Technology Platform (LTP) [10] is used to extract syntactic features and these features are prepared as inputs for the PLM. Third, the BiLSTM layer and the Masked CRF layer [11] are further integrated into the PLM for sequence labeling and Chinese ORE.

Chinese ORE problems (as shown in Figure 1) are delimitated in the CORE-KE system as the problems which can be solved by a sequence labeling task. Given any sentence $s = [c_1, \ldots, c_N]$ with N Chinese characters, the CORE-KE system first assigns each character with a BIO tag and then uses BIO tagging schemes to extract a knowledge triple (E_1, R, E_2) from the system's labeling result. E_1, R and E_2, representing head entity, relation, and tail entity, respectively, each consists of continuous character segment $[c_i, \ldots, c_j]$ where $1 \leq i \leq j \leq N$ of the given sentence s. Using the sequence labeling technique and the PLM model, the CORE-KE system predicts relations from any given Chinese sentences.

The CORE-KE system provides an effective tool in addressing challenges such as fine-grained entity acquisition, long-tail entity mining, and taxonomy evolution in the fields of information retrieval, knowledge graph completion, and intelligent question answering [12, 48]. Entities in a sentence are linked to Wikidata and expanded external knowledge (extracted from Chinese open relation datasets). The linked entities are further combined with their original sentence to form an input into the PLM of the CORE-KE system for model training. In doing so, the CORE-KE system can acquire fine-grained entities from Wikipedia and other knowledge sources. Additionally, the PLM in the CORE-KE system helps train the open relation extraction process with large-scale unstructured texts containing many long-tail words. Furthermore, the CORE-KE system is a robust system which re-trains its ORE process on new data at low costs and can evolve over time.

The main contributions of our work are summarized below:

1) We proposed a Chinese ORE system, CORE-KE, with the support of pre-trained language model and manifold external knowledge. We have also published the code of the system for researchers to reproduce the experiments in the paper. To the best of our knowledge, this is one of the few open-source Chinese ORE systems.
2) Experimental results demonstrate that the CORE-KE system can effectively alleviate fine-grained entity acquisition, which is a challenge in Chinese ORE.
3) Experimental results show that the CORE-KE system performs well in Chinese open relation extraction, giving a relative improvement of 20.1% and 1.3% in F1-score when compared with two state-of-the-art ORE systems on the COER dataset and the SpanSAOKE/SpanSAOKE-NO dataset, respectively.

The rest of this paper is organized as follows: Section 2 describes previous research work related to Chinese open relation extraction. With a good understanding of the research in Chinese ORE, we present

our CORE-KE system in Section 3. Section 4 evaluates the CORE-KE system from the view of its performance in open relation extraction from the COER and SpanSAOKE-NO datasets. Furthermore, this section provides comparisons of the CORE-KE system against several benchmark Chinese ORE systems. The paper is concluded in Section 5 by summarizing the contributions of the research work and outlining the future research directions.

2. RELATED WORK

As an important subtask of information extraction and knowledge acquisition, open relation extraction has attracted many researchers' attention in recent years. Mainstream ORE systems can be divided into 1) unsupervised and rule-based systems, 2) supervised and statistical systems and 3) supervised neural systems.

Unsupervised and rule-based systems mainly apply syntax features to designed syntactic constraints or paradigms in order to extract relationships between entities. Typical systems include TextRunner [2], ReVerb [3], SRLIE [18], ClausIE [19], RelNoun [21], PropS [22], OpenIE4 [23], MinIE [25], Graphene [26], and CALMIE [27]. For example, researchers in the ReVerb system extracted relations in the form of (arg1, relation phrase, arg2) by 1) articulating two simple but powerful constraints (that is, a syntactic constraint and a lexical constraint) and 2) expressing relation phrases via verbs in English sentences. The ReVerb system takes English sentences as inputs, identifies candidate pairs of the noun phrase (NP) arguments (arg1, arg2) from sentences, employs the ReVerb extractor to label each word of any two NP arguments as part of a potential relation phrase, and extracts relations. The ClausIE system explored the linguistic knowledge of English grammars and mapped dependency relations of English sentences to clause constituents. Since the ClausIE system relies on dependency parsing and a small set of domain-independent lexicon, it could lead to over-specification in arguments. To solve this problem, researchers in the MinIE system extracted open relations with semantic annotations, identified and removed specific parts from its relation extraction process, and enhanced the performance of the open relation extraction. It is noted that the MinIE system, constrained by its single annotation type, cannot extract adjectives associated with a noun (for example, assistant director). In summary, unsupervised and rule-based ORE systems can be affected by the implicit error propagation of the tools being used, since they require manually constructed rules and depend on syntactic outputs of NLP tools, which often lead to inefficiency or proneness to errors in the ORE process [49].

The ORE systems supported by supervised and statistical techniques construct open pattern templates over a large training set and then apply these templates to extract open relations. Typical ORE systems include OLLIE [17], Stanford [20], BONIE [24], RSNs [50], and DeepKE [51]. For example, the OLLIE system uses a set of "seed" tuples with high precision from the ReVerb method to bootstrap a large training set and builds open pattern templates. These templates are then applied to individual sentences for ORE. Similarly, the BONIE system also creates "seed" tuples or facts, throws the facts into a training dataset in a bootstrapping process, and develops patterns through dependency parses. The BONIE system constructs numerical tuples by pattern matching and parse-based learning. In summary, supervised and statistical technique-supported ORE systems depend heavily on learning pattern templates. As training datasets vary,

pattern templates may not be well constructed with high quality and diversity. As a result, relations may not be extracted with a good performance.

In recognizing the drawbacks of the above ORE systems, a few researchers have developed another type of ORE systems using deep learning techniques (or neural networks). Syntactic and semantic features of texts can be captured automatically through NLP tasks. These deep learning ORE systems are mainly considered as supervised systems supported by labeling-based, generation-based, and span-based techniques [4].

The labeling-based ORE systems produce a sequence of word labels to mark relationships between words and entities. For example, the SenseOIE system [28] identifies words through syntactic heads of relations and labels each head word for the extraction of semantic relations. It is noted that the labeling-based ORE systems cannot extract relations from sentences where entities and relation are overlapped.

The generation-based ORE systems, including Seq2Seq [29] and IMoJIE [30], use sequence-to-sequence approaches to generate extractions sequentially. For example, the Seq2Seq system treats open relation extraction as a sequence-to-sequence generation problem, where input sequences are sentences, while output sequences are relation tuples with special placeholders. The Seq2Seq system uses an encoder-decoder framework to obtain relation tuples from large-scale unstructured texts. It does not rely on hand-crafted semantic patterns and rules for open relation extraction. Instead, it bootstraps a large volume of high-quality examples (from state-of-the-art Open IE systems) into the model's training process. The IMoJIE system extends the Copy-Attention of the Seq2Seq system and creates additional open relations from extracted tuples. The generation-based systems, in summary, produce new facts or relation triples from previous triples through an iterative process. When false triples are generated from previous triples, a ripple effect may exist within the iterative process. As a result, wrong extraction of open relations may be experienced.

The RnnOIE [31] and SpanOIE [15] systems are examples of span-based ORE systems in which any token subsequence (or span) constitutes a potential entity, while a relation can hold between any pair of spans [32]. The RnnOIE system formulates open relation extraction as a sequence tagging problem and applies the BiLSTM model for training its sequence labeling process. This system addressed several task-specific challenges, including the BIO encoding of predicates with multiple extractions and confidence estimation. The SpanOIE system improves the RnnOIE system by having two modules, with the first module to find predicate spans in a sentence and the second module to combine the found predicate spans and their associated sentences. The SpanOIE system forms combined segments as input and output argument spans for open relation extraction. The span-based ORE systems extract spans to construct unrealistic entity-relations, which may lead to wrong results. For example, given a sentence [他相信湖人队会获得冠军 (*He believes the Lakers will win the champion*)], the extracted relation from a span-based ORE system could be (湖人队$_{the\ Lakers}$, 获得$_{win}$, 冠军$_{the\ champion}$), which is not the true implication of the original sentence.

English has been the primary language in open relation extraction research. Most of the above-mentioned ORE systems use English syntactic features. They cannot be directly applied to Chinese context. With a

good understanding of the limitations of the existing ORE systems, researchers in CORE [5] explored the unsupervised and rule-based methods for Chinese open relation extraction. The CORE system employs word segmentation, part-of-speech (POS) tagging, syntactic parsing, and other NLP techniques, to automatically annotate Chinese sentences. In doing so, input sentences are chunked and entity-relation triples are extracted. The ZORE [6] system is a supervised and statistical method. It first identifies relation candidates from automatically parsed dependency trees and then extracts relations iteratively through a novel double propagation algorithm. It is noted that this system has a logistic regression classifier which assigns features with weights and trains relation triples on the Wiki-500 dataset. In addition, this system gives a confidence score to each extracted relation. It thus reduces false relations significantly in the tuple-generation process and improves the performance of ORE.

Researchers in HNN4ORT [33], PGCORE [8], NPORE [7], DBMCSS [34], and MGD-GNN [13] have used supervised neural networks for Chinese ORE. The HNN4ORT system (a labeling-based system) employs the Ordered Neurons LSTM model [44] to encode syntactic information and capture associations among arguments and relations. The NPORE system implements a graph clique mining algorithm to chunk Chinese noun phrases into modifiers and headwords, generate candidate relation triples, and extract Chinese open semantic relations. The PGCORE system treats relation extraction as a text summary task. It uses a sequence-to-sequence framework and the Pointer-Generator mechanism [38] to solve the ORE problem. The DBMCSS and MGD-GNN systems apply the span-based ORE method in the Chinese context. They extract named entity spans, filter entity pairs, and extract Chinese open relations. The MGD-GNN model constructs a multi-grained dependency graph to incorporate dependency and word boundary information and employs the Graph Neural Network (GNN) to get node representations for predicate and argument predictions.

The CORE-KE system presented in this paper implements a supervised neural ORE approach. It uses syntactic and semantic features, such as POS tagging, dependency parsing, and large-scale external knowledge, to improve the performance of Chinese open relation extraction on unstructured texts.

3. METHODOLOGY

In this section, the Chinese Open Relation Extraction with Knowledge Enhancement (CORE-KE) system powered by a pre-trained language model is presented. Knowledge enhancement in this system refers to the way to understand concepts (or entities in a Chinese sentence) through description information of concepts or other relevant knowledge (including dependency relations). An overview of the CORE-KE system is shown in Figure 2. The CORE-KE system (which implements the WoBERT_plus + BiLSTM + Masked CRF model with knowledge enhancement) has four modules/model:

1) the Descriptive Knowledge Acquisition Module. This module is designed to obtain description information of entities in the training set from Wikidata. The description information is used to pre-train the WoBERT_plus + BiLSTM + Masked CRF model (or the WBM model).

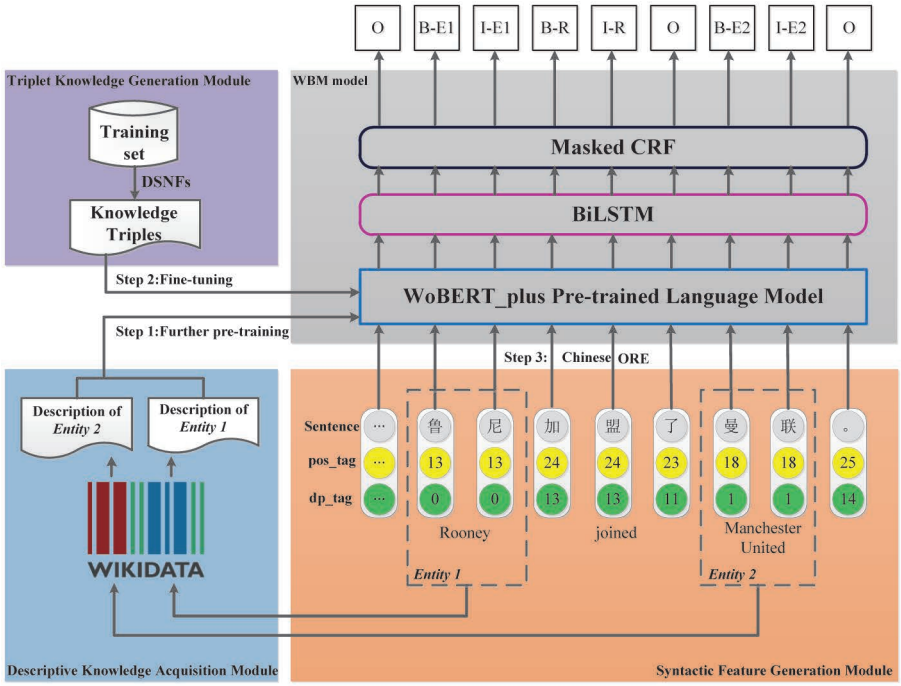

Figure 2. The architecture of the CORE-KE system. In step 1 (the Descriptive Knowledge Acquisition module), the descriptions for entities in a sentence *s* from a Chinese open relation extraction dataset are obtained from Wikidata and concatenated with *s* to further pre-train the WoBERT_plus PLM. In step 2 (the Triplet Knowledge Generation module), extra dependency relation triplets of the sentence *s* are generated by the DSNFs tools. These triplets are used to fine-tune the WoBERT_plus PLM. In step 3 (the Syntactic Feature Generation module), syntactic feature tags for each character in the sentence *s* are generated and concatenated with the sentence *s* to train the WBM model.

2) the Triplet Knowledge Generation Module. This module applies DSNFs tools to generate extra dependency relations for Chinese entities. The dependency relations are further used to fine-tune the WBM model (which is already pretrained by the Descriptive Knowledge Acquisition module).
3) the Syntactic Feature Generation Module. This module introduces the LTP tool to generate POS tag and dependency parsing tag for each character in Chinese sentences.
4) the WBM model. The WBM model is integrated with a PLM. It combines characters (in Chinese sentences) with POS tags and dependency parsing tags and uses the combined sequences as inputs. It produces the characters (attached with their BIO tags) where predicted relations (in the format of triples) are embedded.

3.1 Descriptive Knowledge Acquisition Module

Incorporating individual concepts (or sentences) with various descriptive and supportive pieces of knowledge can improve the ability of understanding concepts [45]. Acquiring and understanding descriptive

knowledge therefore is a vital step to enhance model's cognition on entities [46]. In the CORE-KE system, the Descriptive Knowledge Acquisition module has a set of utilities to help find descriptions for each entity in Chinese sentences and concatenate the found descriptions with their associated entities. Take the sentence *s* [在欧洲杯的精彩演出后，鲁尼加盟了曼联 (*After giving a wonderful show in the European Cup, Rooney joined Manchester United*)] as an example, 鲁尼 (*Rooney*) and 曼联 (*Manchester United*) are two entities. The descriptions of these two entities are extracted from Wikidata. MediaWiki APIs[②] and Wikipedia tools[③] are used in the CORE-KE system to get the entity descriptions. For the entity of 鲁尼 (*Rooney*), the description is 韦恩·马克·鲁尼，已退役的英格兰职业足球运动员，英格兰足坛巨星之一 (*Wayne Mark Rooney, a retired English professional soccer player, is one of the superstars in English soccer*)$_{[Des_A]}$. For the entity of 曼联 (*Manchester United*), the description is 曼彻斯特联足球俱乐部，简称曼联，是一家位于英国曼彻斯特的足球俱乐部，目前比赛于英格兰超级联赛，球队主场为老特拉福德球场 (*Manchester United Soccer Club, abbreviated as Manchester United, is a soccer club located in Manchester, England, currently playing in the English Premier League, the team's home is the Old Trafford stadium*)$_{[Des_B]}$. The descriptions of two entities are further concatenated to the original sentence *s*, that is,

$$Input = S \oplus Des_A \oplus Des_B \tag{1}$$

where *Input* refers to an input instance or sequence to be used to further train the pre-trained language model, $\oplus$ denotes concatenation of any two strings, *s* refers to the original sentence, Des_A and Des_B refer to the description of the first and second entities (entity A and entity B), respectively.

The CORE-KE system also has utilities to clean up entity descriptions. These utilities unify character encodings, change double-byte characters to single-byte characters, and remove noise characters (such as HTML tags and stop words). Additionally, the CORE-KE system sets the description of a Chinese entity to be null if the entity cannot be found in Wikidata.

3.2 Triplet Knowledge Generation Module

The Triplet Knowledge Generation module has implemented the Dependency Semantic Normal Forms (DSNFs) tools to obtain dependency relations for Chinese entities in the training set. Through the dependency semantic normal forms, the CORE-KE system generalizes syntactic and semantic abstractions of relations and structures them with their associated words, POS-tags, dependency path, and labels of dependency path. There are a total of seven dependency semantic normal forms used in the CORE-KE system in support of four relation structures: modified structure, verbal structure, coordination structure, and formulaic structure [9]. With these DSNFs, the CORE-KE system addresses three kinds of unique but ubiquitous Chinese linguistic phenomena, that is, the Nominal Modification-Center (NMC) phenomenon, the Chinese Light Verb construction (CLVC) phenomenon, and the Intransitive Verb (IV) phenomenon.

② https://www.mediawiki.org/wiki/MediaWiki
③ https://github.com/siznax/wptools

The CORE-KE system further maps entity relations into dependency trees and gathers a series of paradigms for relation extractions. The DSNF-based knowledge generation involves the following tasks: 1) pre-process input sentences with word segmentation, POS tagging, and dependency parsing; 2) create candidate entities for each subject sentence by using the LTP tool and the Iterated Heuristic Algorithm; 3) pair candidate entities and classify them into seven DSNFs. Taking advantage of these generated DSNFs, the CORE-KE system extracts dependency relations (in the form of triples) from two large-scale unstructured Chinese ORE datasets for the WBM model.

Given the input sentence *s* in Figure 1 as an example, the following dependency relations [鲁尼 (*Rooney*), 加盟 (*joined*), 曼联 (*Manchester United*)] and [(鲁尼 (*Rooney*), 精彩演出 (*giving a wonderful show*), 欧洲杯 (*European Cup*)] are generated by the CORE-KE system using the DSNFs tools. The first dependency relation [鲁尼(*Rooney*), 加盟 (*joined*), 曼联 (*Manchester United*)] can also be found in the training dataset, however the second dependency relation [(鲁尼 (*Rooney*), 精彩演出 (*giving a wonderful show*), 欧洲杯 (*European Cup*)] does not exist in the training dataset. From the view of the CORE-KE model, the second dependency is an extra one.

3.3 Syntactic Feature Generation Module

Considering that Chinese ORE is highly dependent on the quality of word segmentation and often suffers from the ambiguity of polysemic words [47], the CORE-KE system uses multiple syntactic features to alleviate the ambiguity of polysemy. After obtaining descriptive knowledge and triplet knowledge from the Descriptive Knowledge Acquisition Module and the Triplet Knowledge Generation module, the CORE-KE system generates syntactic feature tags for characters in Chinese sentences. This module describes the process of generating syntactic feature tags, such as POS tags and dependency parsing tags. Figure 3 shows

Range of values for POS tag

0: adjective	10: number	20: preposition
1: other noun-modifier	11: general noun	21: quantity
2: conjunction	12: direction noun	22: pronoun
3: adverb	13: person name	23: auxiliary
4: exclamation	14: organization name	24: verb
5: morpheme	15: location noun	25: punctuation
6: prefix	16: geographical name	26: foreign words
7: idiom	17: temporal noun	27: non-lexeme
8: abbreviation	18: other proper noun	28: descriptive words
9: suffix	19: onomatopoeia	

Range of values for dependency parsing-tag

0: subject - verb	5: attribute	10: left adjunct
1: verb - object	6: adverbial	11: right adjunct
2: indirect - object	7: complement	12: independent structure
3: fronting - object	8: coordinate	13: head
4: double	9: preposition - object	14: punctuation

Figure 3. Syntactic features used in the CORE-KE system.

the types of syntactic features which are tagged in the CORE-KE system. Each character in a Chinese sentence belongs to one of the 29 POS tag types. Numbers ranging from *0* to *28* are used to represent each tag type. For example, the number for the *conjunction* POS tag type is *2*.

There are 15 dependency parsing tag types within the CORE-KE system. Each character in a Chinese sentence can be parsed with one of these dependency parsing tag types. Numbers ranging from *0* to *14* are used to represent each dependency parsing tag type. For example, the number for the *verb-object* dependency parsing tag type is *1*.

Figure 4 shows an example of how these POS and dependency parsing tags are used. The sentence *s* [在欧洲杯的精彩演出后，鲁尼加盟了曼联 (*After giving a wonderful show in the European Cup, Rooney joined Manchester United*)] is incorporated with the *pos_tag* and *dp_tag* lines. The POS tags in the *pos_tag* line and the dependency parsing tags in the *dp_tag* line are generated by the LTP tool. Chinese characters for the same word are assigned with same *pos_tag* and *dp_tag*. Take the Chinese word 鲁尼 (*Rooney*) as an example, the POS tag and the dependency parsing tag are *person name* and *subject-verb*, respectively. The corresponding tag numbers are *13* and *0*, respectively. Additionally, Chinese characters 鲁 and 尼 are also assigned with the POS tag of *13* and the dependency parsing tag of *0*.

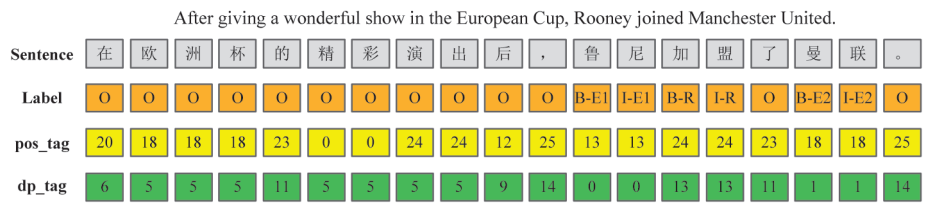

Figure 4. A training example with syntactic features used in the CORE-KE system.

Training sentences are further processed with a series of labels in the CORE-KE system. The BIO tagging scheme is used for sequence labeling. Each Chinese character is labeled as *B-X*, *I-X*, or *O*. *B-X* refers to the beginning of label type *X* (*X* refers to one of the three label categories: head entity (E_1), tail entity (E_2), and relation (*R*) between E_1 and E_2). *I-X* refers to the middle of label type *X*, while *O* indicates that the Chinese character is not a part of an entity or relation. For example, the *Label* line in Figure 4 indicates that 鲁尼 (*Rooney*) is the head entity E_1 in the sentence, 曼联 (*Manchester United*) is the tail entity E_2, and 加盟 (*joined*) is the relation *R* between E_1 and E_2. The sentence *s* is concatenated with its labels, POS tags, and dependency parsing tags to form as a knowledge enhanced Chinese sentence for training in the WBM model.

3.4 WBM Model

The CORE-KE system incorporates descriptive knowledge and triplet knowledge to pre-train the WoBERT_plus pre-trained language model (see Steps 1 and 2 in Figure 2). After these two steps, the CORE-KE system

implements the WBM model for relation extraction. Given a sentence s, a feature vector $V(c_i)$ for every Chinese character $c_i \in S$ in the CORE-KE system is:

$$V(c_i) = emb(c_i) \oplus emb(label(c_i)) \oplus emb(pos(c_i)) \oplus emb(dp(c_i)) \quad (2)$$

where,

- $emb(c_i)$: the embedding of c_i
- $emb(label(c_i))$: the embedding of c_i's ground truth label
- $emb(pos(c_i))$: the embedding of c_i's POS tag
- $emb(dp(c_i))$: the embedding of c_i's dependency parsing tag
- $\oplus$: the concatenation of any two embeddings.

The $emb(c_i)$ and $emb(label(c_i))$ are pre-trained contextual embeddings encoded by the pre-trained language model, which together incorporates information from the character's context in the Chinese sentence s.

The $emb(pos(c_i))$ and $emb(dp(c_i))$, randomly initialized embeddings, incorporate information from POS tags and dependency parsing tags, respectively. These embeddings are encoded according to the following mechanism:

1) Matrices are used to map POS tags and dependency parsing tags. A $V_{pos} \times D$ matrix (S_{pos}) and a $V_{dp} \times D$ matrix (S_{dp}) are generated, where V_{pos} and V_{dp} denote the number of POS tag types and dependency parsing tag types, respectively (which are *29* and *15* in the CORE-KE system). D governs the columns of matrix S_{pos} and matrix S_{dp}. Assume D is set to be 10. Each element in S_{pos} or S_{dp} is a random number ranging from *0* to *10*. According to the number of POS tag and dependency parsing tag of the character, $emb(pos(c_i))$ and $emb(dp(c_i))$ are encoded into the corresponding row of S_{pos} and S_{dp}, respectively. For example, Chinese characters 鲁 and 尼 are both assigned with the POS tag of *13* and the dependency parsing tag of *0*. The $emb(pos(c_i))$ and $emb(dp(c_i))$ of these two characters are encoded, in an embedded space, as the 13^{th} row of S_{pos} and the 0^{th} row of S_{dp}.

Ten columns (D = 10) are selected in the CORE-KE system to consider the constraints of computing resources (storage and computing speed) used for model training. It is tested that the matrices (S_{pos} [29 × 10] and S_{dp} [15 × 10]) are suitable for the WBM model.

2) The sentence s with syntactic features (embedded into the feature vector $V(S)$) is fed into the WoBERT_plus pre-trained language model and the contextualized output embeddings are then computed. The WoBERT_plus[④] model is a variant of the WoBERT model which improves the BERT [35] process dealing with Chinese texts. Typically, the performance of the WoBERT model is similar to that of the BERT model, but it is 1.16, 1.22, and 1.28 times the speed of the BERT model in dealing with Chinese texts of a length

④ https://github.com/ZhuiyiT echnology/WoBERT

of 128, 256, and 512 words, respectively[5]. The WoBERT_plus model improves the RoBERTa-wwm-ext model [36] by training on a much larger corpus for 250,000 steps. Additionally, the WoBERT_plus model increases its vocabulary of the dictionary, when compared with the WoBERT and BERT models. Also, the WoBERT_plus model, similar to the RoBERTa-wwm-ext model, utilizes the whole word masking strategy in various pre-training tasks and thus mitigates the drawbacks of masking partial word piece tokens in the BERT model.

In the CORE-KE system, the output of the WoBERT_plus model is the sequence of character embeddings. This sequence is further fed into the BiLSTM layer to integrate useful information for sequence labeling. The output of the BiLSTM layer is the predicted score for each character. The predicted scores are later referenced to as *logits*. Each *logit* in the sequence is further considered as the emission score for its corresponding character. The emission scores are then used in the Masked CRF layer for computing the scores of possible sequences of BIO tags.

It is noted that each possible sequence of BIO tags is considered as a potential path in the Masked CRF layer. For each possible path, the Masked CRF layer first introduces a transition matrix that models the transition scores from tag i to tag j for any two consecutive characters in the path, where $1 \leq i \leq j \leq N$, N denotes the number of characters in the sentence s.

Assume that a sequence of input characters is $x = \{x_1, \ldots, x_T\}$, a sequence of ground truth BIO tags is $y = \{y_1, \ldots, y_T\}$, and a sequence of logits as $l = \{l_1, \ldots, l_T\}$ for sentence s. Also assume that the number of distinct tags as d, and the set of tag indices as $[d] := \{1, \ldots, d\}$. We then have $y \in [d]$ and $l_i \in \mathbb{R}^d$ where $1 \leq i \leq T$. In addition, the transition matrix is denoted as $A = (a_{ij}) \in \mathbb{R}^{d \times d}$, where a_{ij} is the transition score from tag i to tag j, W as the set of all trainable weights in the encoder (in the WoBERT_plus model and the BiLSTM layer). All transition scores in the transition matrix are randomly initialized. The purpose of the Masked CRF layer is to learn the tagging scheme constraints of the training set during the training process and update transition scores.

By aggregating the emission scores and the transition scores, the Masked CRF layer assigns a score for each possible path. Given the input x, the weights W, and the transition matrix A, the score of a path $p = \{n_1, \ldots, n_T\}$ in the Masked CRF layer can be computed as:

$$s(p, x, W, A) = \sum_{i=1}^{T} l_{i,n_i} + \sum_{i=1}^{T-1} a_{n_i, n_{i+1}} \tag{3}$$

where $l_{i,j}$ denotes the j^{th} entry of l_i.

It is worth noting that the Masked CRF layer, as an improvement of CRF, eliminates the outcomes of illegal paths. An illegal path denotes a path that violates BIO scheme rules. An example rule is that any *I-X* tag must be preceded by a *B-X* tag or another *I-X* tag of the same type (*X* refers to head entity (E_1), tail entity

[5] https://kexue.fm/archives/7758

(E_2) or relation (R)). For example, *O O O B-*E_1 *I-R O* is an illegal path because the transition from *B-*E_1 to *I-R* violates the example rule.

The Masked CRF layer adds a mask to the state transition matrix in advance and sets a minimum value for the score of the position on the illegal path. In doing so, the CORE-KE system does not select illegal paths during the iterative training process. The system can ensure that the predicted path for ORE is a legal path.

The loss function of the Masked CRF layer is described as below:

$$L(W, A) = -\frac{1}{|M|}\sum_{(x,y)\in M}\log\frac{\exp s(x,y)}{\sum_{p\in P/I}\exp s(p,x)} \tag{4}$$

where the dependence of $s(\cdot,\cdot)$ on (W, A) is omitted for the sake of conciseness. Denoting M as the set of all training samples, P as the set of all possible paths, I as the set of illegal paths, P/I as the set of legal paths.

The loss function of the Masked CRF layer is the whole loss function of the CORE-KE system. Our system uses the Adam optimization algorithm (a popular first-order method) to minimize $L(W, A)$. Let (W_{opt}, A_{opt}) be the minimizer of the loss L. The predicted path y_{opt} of a test sample x_{test} is the path having the highest score, that is,

$$y_{opt} = \arg\max_{p\in P/I} s(p, x_{test}, W_{opt}, A_{opt}) \tag{5}$$

In the CORE-KE system, the Viterbi algorithm is used to find the predicted path.

The outputs of the CORE-KE system are the predicted results (in BIO format) which correspond to each character of the input Chinese sentences. In the training process, the loss between the predicted results and the ground truth is back propagated to fine-tune the parameters of the CORE-KE model.

3.5 Training and Inferring Processes

The CORE-KE system follows the data flow as shown in Figure 2 and trains the WBM model using the enhanced knowledge (that is, descriptive knowledge and syntactic features) generated by the Descriptive Knowledge Acquisition Module, the Triplet Knowledge Generation Module, and the Syntactic Feature Generation Module. This section describes the training and inferring processes.

3.5.1 Training process

The CORE-KE system selects each sentence (or sentence s) from the training set and creates a BIO tag for each character in sentence s. Also, the CORE-KE system conducts the following tasks in training the WBM model:

1) Link each entity in *s* to Wikidata and retrieve the description related to each entity (as described in Section 3.1).
2) Further pre-train the WoBERT_plus pre-trained language model (PLM) by using sentence *s* and its description knowledge.
3) Use the DSNFs tools to generate extra dependency relation triplets for sentence *s* (as described in Section 3.2).
4) Use these triplets to fine-tune the WoBERT_plus PLM in Task 1. After fine-tuning, the new PLM model is named as the WoBERT_plus PLM-DT model.
5) Use the LTP tools to generate the syntactic feature tags (such as POS tags and dependency parsing tags) for each character in *s* (as described in Section 3.3).
6) Train the WoBERT_plus PLM-DT model by using sentence *s* and its BIO tags, POS tags, and dependency parsing tags (as shown in Figure 4). After training, the new model is named as the WoBERT_plus PLM-DT + BiLSTM + MaskedCRF model or the WBM model.

3.5.2 Inferring process

The CORE-KE system takes advantage of the test set to verify the inferring process of the WBM model and predict relations for a given sentence in the test set. It involves the following tasks:

1) Select each sentence *s* from the test set and use the LTP tools to generate the syntactic feature tags, such as POS tags and dependency parsing tags, for each character in sentence *s* (as described in Section 3.3).
2) Use sentence *s* and its related POS tags and dependency parsing tags as inputs to the CORE-KE system. The system then predicts the BIO tag for each character of sentence *s* and extracts a relation.

4. EXPERIMENTS

Aiming to evaluate the performance of the CORE-KE system for Chinese open relation extraction, we conducted experiments on two datasets. In this section, we describe the experimental setup and outline the benchmark ORE systems used in comparison with the CORE-KE system in this research. At the end of this section, experimental results are discussed.

4.1 Experimental Setup

The two large-scale datasets used in our experiments were the Chinese Open Entity and Relation Knowledge Base dataset (later referenced to as the COER dataset) and the Span Symbol Aided Open Knowledge Expression dataset (later referenced to as the SpanSAOKE dataset). The COER dataset is a Chinese open entity and relation knowledge base containing 65,778 different named entities and 21,359 different open relation phrases in 218,362 Chinese sentences extracted from several popular Chinese websites, such as Sohu News [http://www.sohu.com], Sina News [http://news.sina.com.cn], and Baidu Baike [https://baike.baidu.com]. The relations in the COER dataset cover many domains, including military, sports, entertainment, economy, etc. We filtered out sentences that contain multiple relations. In doing so,

the COER dataset used in our experiment contained 213,327 triples in the training set and 2,000 triples in the test set, respectively. The COER dataset for the COER-KE system was the same as that for the PGCORE system [8].

The SpanSAOKE dataset is a large-scale sentence-level dataset for Chinese ORE containing 53,869 relation triple facts in 26,496 sentences collected from Baidu Baike. Some of the entities and relations overlap in the SpanSAOKE dataset. For example, in the sentence of [农民收入主要以橡胶为主 (*Farmers' incomes are mainly from rubber*)], 农民收入 (*Farmers' incomes*) and 橡胶 (*rubber*) are two annotated entities, 以橡胶为主 (*from rubber*) is annotated relation in the SpanSAOKE dataset. The tail entity 橡胶 (*rubber*) and the relation 以橡胶为主 (*from rubber*) overlap in this sentence. In our experiments, these overlapping sentences were filtered out and 44,734 triples in 23,856 sentences were retained. The new dataset is later referenced to as the SpanSAOKE without overlapped entity-relation triples (or the SpanSAOKE-NO dataset for short). We randomly divided the new dataset with 35,921 triples for training, 4,373 triples for validation, and 4,440 triples for testing, respectively. The details of the two experimental datasets are shown in Table 1.

Table 1. Statistics of the COER dataset and SpanSAOKE-NO dataset.

Datasets	Split	#Triples
COER	Train	213,327
	Test	2,000
SpanSAOKE-NO	Train	35,921
	Validation	4,373
	Test	4,440

It is noted that the SpanSAOKE-NO dataset is different from the COER dataset. There are many sentences containing multiple relation triples in the SpanSAOKE-NO dataset. The CORE-KE system thus predicts only one result per sentence at a time. Same as the procedures used in MGD-GNN [13], the *gestalt* pattern matching function [37] was used in the CORE-KE system to find the best triple (which matches with the ground truth) from the SpanSAOKE-NO dataset.

In further pre-training and fine-tuning the WoBERT_plus model, we used the learning rate of $5e^{-5}$, the max sequence length of 128, the batch size of 16, and the training epoch of 30. Additionally, we also employed Gaussian Error Linear Units (GELU) as the activation function.

In training the CORE-KE system, we used the max sequence length of 256, the batch size of 16, the dropout rate of 0.5, and the size of 256 as the BiLSTM hidden unit. Additionally, we adopted other parameters the same as those in the fine-tuning process of the WoBERT_plus model.

4.2 Benchmark Systems

The following benchmark systems were used in this research to compare their effectiveness with that of the CORE-KE system:

ZORE [6] is a syntactic ORE system that identifies relation candidates from automatically parsed dependency trees and extracts relations with their semantic patterns iteratively through a novel double propagation algorithm.

UnCORE [14] is a rule-based Chinese ORE system that 1) uses word distance and entity distance constraints to generate candidate relation triples from raw corpus, 2) adopts global and domain ranking methods to discover relation words from candidate relation triples, and 3) employs syntactic rules to filter final relation triples.

DSNFs [9] is an unsupervised Chinese ORE system that establishes its own Dependency Semantic Normal Forms (DSNFs) to map entity relations into dependency trees and considers Chinese unique linguistic characteristics in open relation extraction.

PGCORE [8] is an end-to-end supervised neural Chinese ORE system that applies a Pointer-Generator framework to copy words in input sequences and move them to output sequences via pointers, while retaining its ability to generate new words.

SpanOIE [15] is a span selection based neural open information extraction system which receives competitive results from English corpus. It uses the pre-trained Chinese word embeddings [39] as well as POS tags and dependency labels to apply SpanOIE for Chinese corpus.

CharLSTM [16] applies a vanilla character-based BiLSTM model to encode characters for Chinese open relation extraction.

MGD-GNN [13] is a character-based supervised neural system which constructs a multi-grained dependency graph to incorporate dependency and word boundary information. It employs GNN to get node representations for predicate and argument predictions.

4.3 Measures

The performance of the CORE-KE system on the COER and SpanSAOKE-NO datasets is measured by Precision (P), Recall (R), and micro F1-score (F_1):

$$P = \frac{|C \cap G|}{|C|} \tag{6}$$

$$R = \frac{|C \cap G|}{|G|} \tag{7}$$

$$F_1 = \frac{2 \times P \times R}{P + R} \tag{8}$$

where C refers to the predicted result set of the CORE-KE system, G refers to the golden result set, and $C \cap G$ denotes the triples both in C and G. Both C and G involve all three categories, including head entity, tail entity, and relation.

4.4 Results and Analysis

We conducted numerous experiments by operating the CORE-KE system and the benchmark ORE systems on the COER and SpanSAOKE-NO datasets. The experimental results were obtained and comparisons between the CORE-KE system and the benchmark ORE systems were made.

4.4.1 Overall Comparison

Table 2 shows the experimental results of the CORE-KE system and the benchmark ORE systems on the COER dataset. The best results are highlighted in bold. It is noted that we did not use all the benchmark ORE systems described in Section 4.2, but the four ORE systems (ZORE, UnCORE, DSNFs, and PGCORE) which have been previously reported as the top systems with high ORE performance [8]. The first three of these systems are unsupervised systems.

Table 2. Experimental results of Chinese ORE systems on the COER dataset.

	Systems	Precision	Recall	F1-score
Unsupervised	ZORE	0.838	0.145	0.249
	UnCORE	0.806	0.476	0.599
	DSNFs	0.838	0.587	0.690
Supervised	PGCORE	**0.854**	0.543	0.663
	CORE-KE	0.835	**0.761**	**0.796**

It is observed that all five systems yield similar results in precision (0.83±0.02), the CORE-KE system, however, outperforms the benchmark ORE systems by a large margin in recall and F1-score. In comparison with three statistical and rule-based unsupervised systems (ZORE, UnCORE, and DSNFs), the CORE-KE system ends up with a relative improvement of [(0.761–0.145)/0.145 or 424.8%], [(0.761–0.476)/0.476 or 59.9%], and [(0.761–0.587)/0.587 or 29.6%] in recall, respectively; [(0.796–0.249)/0.249 or 219.7%], [(0.796–0.599)/0.599 or 32.9%], and [(0.796–0.690)/0.690 or 15.4%] in F1-score, respectively. In comparison with the supervised neural ORE system (PGCORE), the CORE-KE system has a relative improvement of [(0.761–0.543)/0.543 or 40.1%] in recall and [(0.796–0.663)/0.663 or 20.1%] in F1-score, respectively. All these results show that the CORE-KE system can significantly improve the performance of Chinese open relation extraction.

Table 3 shows the experimental results of the CORE-KE system and the benchmark ORE systems on the SpanSAOKE/SpanSAOKE-NO dataset. The best results are also highlighted in bold. It is noted that we used only four ORE systems (ZORE, SpanOIE, CharLSTM, and MGD-GNN) since these four systems had been

previously reported as the top systems with high ORE performance [13]. Among these systems, ZORE is an unsupervised system.

Table 3. Experimental results of Chinese ORE systems on the SpanSAOKE/SpanSAOKE-NO dataset.

	Systems	Precision	Recall	F1-score
Unsupervised	ZORE	0.315	0.177	0.227
Supervised	SpanOIE	0.418	0.443	0.430
	CharLSTM	0.404	0.454	0.427
	MGD-GNN	0.450	**0.471**	0.460
	CORE-KE	**0.594**	0.383	**0.466**

It is observed that the CORE-KE system outperforms all the benchmark ORE systems in precision and F1-score. In comparison with the statistical and rule-based system (ZORE), the CORE-KE system gives a relative improvement of [(0.594–0.315)/0.315 or 88.6%] and [(0.466–0.227)/0.227 or 105.3%] in precision and F1-score, respectively. In comparison with the three supervised neural ORE systems (SpanOIE, CharLSTM, and MGD-GNN), the CORE-KE system gives a relative improvement of [(0.594–0.418)/0.418 or 42.1%], [(0.594–0.404)/0.404 or 47.0%], and [(0.594–0.450)/0.450 or 32.0%] in precision and [(0.466–0.430)/0.430 or 8.4%], [(0.466–0.427)/0.427 or 9.1%], and [(0.466–0.460)/0.460 or 1.3%] in F1-score, respectively. All these results indicate the effectiveness of the CORE-KE system. Furthermore, when we compare the CORE-KE system with the above three supervised neural ORE systems, the CORE-KE system ends up with a lower recall. One possible reason might be that the CORE-KE system extracts fewer triples from the SpanSAOKE-NO dataset. The extraction results of the CORE-KE system and the MGD-GNN model are shown in Table 4. It is noted that the CORE-KE system extracted 2,860 triples on the SpanSAOKE-NO dataset, while the MGD-GNN model extracted 5,591 triples on the SpanSAOKE dataset, which is even more than the number of golden triples on this dataset. This may explain why the recall of the CORE-KE system is lower than that of the MGD-GNN model. Table 5 shows the number of head entities, tail entities, and relations extracted by the CORE-KE system. It is noted these three values are inconsistent, the CORE-KE system extracted more head entities than the tail entities and relations. This indicates that at least 723 (3819–3096) triples extracted by the CORE-KE system are incomplete. The actual number is 1580 (4440–2860). These incomplete triples containing only an entity and a relation are not included in the extracted result set. According to the definitions of precision and recall (Equations 6 and 7 in Section 4.3), fewer extracted triples lead to a lower recall and a lower F1 score.

Table 4. The extraction results of the CORE-KE system and the MGD-GNN model on the test set of SpanSAOKE/SpanSAOKE-NO dataset.

	CORE-KE	MGD-GNN
#Golden triples	4,440	5,342
#Extracted triples	2,860	5,591
#Correct triples	1,700	2,516

Table 5. The number of head entities, tail entities and relations extracted by the CORE-KE system on the test set of the SpanSAOKE-NO dataset.

	Golden	Extracted
#head entities	4,440	3,819
#tail entities	4,440	3,128
#relations	4,440	3,096

As a result, the CORE-KE model has limited improvement when compared with the MGD-GNN model. The MGD-GNN incorporates the dependency relations between words and adopts a graph neural network to encode the multi-grained dependency graph (MGD). Different from the MGD-GNN model, the CORE-KE model integrates dependency information in the way of encoding dependency parsing tags in an embedded space and concatenating dependency tags with word embeddings and other information. Compared with the MGD-GNN model, the CORE-KE system pays more attention to capturing the semantic features of the sentence and has a better result in precision.

4.4.2 Comparison with Pre-trained Language Models

Using the COER and SpanSAOKE-NO datasets, we also compared the pre-trained language model used in the CORE-KE system with other pre-trained language models. Table 6 summarizes the experimental results of the pre-trained language models. The best results are highlighted in bold.

Table 6. Experimental results of pre-trained language models on the COER dataset and the SpanSAOKE-NO dataset.

Models/Systems	COER			SpanSAOKE-NO		
	P	R	F_1	P	R	F_1
BERT [35]	0.803	0.740	0.770	0.680	0.296	0.413
RoBERTa [40]	0.806	0.712	0.756	**0.725**	0.277	0.401
RoBERTa-wwm [36]	0.814	0.741	0.776	0.704	0.299	0.419
macBERT [41]	0.807	0.713	0.757	0.700	0.280	0.400
NEZHA [42]	0.819	0.733	0.773	0.625	0.258	0.365
ELECTRA [43]	0.824	0.695	0.754	0.499	0.230	0.316
WoBERT_plus	0.834	0.736	0.782	0.672	0.311	0.425
CORE-KE	**0.835**	**0.761**	**0.796**	0.594	**0.383**	**0.466**

It is noted from Table 6 that the CORE-KE system, when compared with other PLMs, relatively improves the performance of open relation extraction, ranging between 0.1%–4.0% in precision, 2.7%–9.5% in recall, and 1.8%–5.6% in F1-score on the COER dataset, respectively. Additionally, the CORE-KE system enhances the performance of open relation extraction on the SpanSAOKE-NO dataset, ranging between 23.2%–66.5% in recall and 9.6%–47.5% in F1-score, respectively. All these results demonstrate that the CORE-KE system is more effective in Chinese ORE.

4.4.3 Ablation Study

Ablation study was conducted in this research to further assess the performance of the CORE-KE system in Chinese ORE. We had the following five models for the ablation study:

1) Ablation Model #1: the *BiLSTM-Masked CRF* model without the support of any pre-trained language model
2) Ablation Model #2: the *WoBERT_plus + BiLSTM-Masked CRF* model without knowledge enhancement
3) Ablation Model #3: the *CORE-KE* system with the support of external knowledge corpus (descriptive and triplet knowledge) only
4) Ablation Model #4: the *CORE-KE* system with the support of syntactic features only
5) Ablation Model #5: the *CORE-KE* system with the full support of external knowledge corpus (descriptive and triplet knowledge) and syntactic features.

The experimental results of these models on the two datasets are shown in Table 7. The best results are highlighted in bold.

Table 7. Experimental results of the ablation study.

Models	COER			SpanSAOKE-NO		
	P	R	F_1	P	R	F_1
Ablation Model #1	0.771	0.616	0.685	**0.656**	0.124	0.209
Ablation Model #2	0.822	0.752	0.785	0.652	0.336	0.444
Ablation Model #3	0.821	**0.762**	0.792	0.649	0.343	0.449
Ablation Model #4	0.823	0.756	0.788	0.630	0.357	0.456
Ablation Model #5	**0.835**	0.761	**0.796**	0.594	**0.383**	**0.466**

It is noted that the Chinese ORE model with the support of a PLM (as shown with Ablation Model #2) has superior performance on the two datasets, when compared with the Chinese ORE model without the support of a PLM (as shown with Ablation Model #1). The relative improvement of the Chinese ORE model with a PLM are 14.6% and 112.4% in F1-score, respectively. These findings indicate that using a pre-trained language model can boost the performance of Chinese ORE systems.

It is also noted that enhanced knowledge has significant impact on the performance of the CORE-KE system in open relation extraction. When external knowledge corpus was used, Ablation Model #3 relatively improved the F1-score of 0.9% (on the COER dataset) and 1.1% (on the SpanSAOKE-NO dataset), as compared with Ablation Model #2 which did not incorporate any knowledge. When syntactic features were used, Ablation Model #4 had a relative improvement of 0.4% and 2.7% in F1-score, as compared with Ablation Model #2 on the two datasets, respectively. The CORE-KE system, which takes advantage of both external knowledge corpus and syntactic features, outperforms Ablation Model #2 on the two datasets with relative improvements of 1.4% and 5.0% in the F1-score, respectively. All these results demonstrate that the use of external knowledge corpus and syntactic features can improve the performance of Chinese ORE.

From Table 7 we can observe that the PLM (WoBERT_plus) plays a more important role in the CORE-KE system. One possible reason is that the PLM have obtained abundant context-based semantic knowledge on large data sources, which is beneficial to the downstream Natural Language Understanding tasks such as relation extraction. The use of knowledge enhancement in the CORE-KE system creates a new approach to strengthening pre-trained language models in learning entities and relations from Chinese contexts. Syntactic features, such as POS tags and dependency parsing tags for each character in Chinese sentences, help PLMs have a better understanding of each character in terms of syntactic structure.

4.4.4 Case Studies

After the CORE-KE system was trained, we applied it to extract Chinese relations from the COER and SpanSAOKE-NO datasets. The ORE results are shown in Table 8 and Table 9, respectively.

Table 8. ORE results from a Chinese sentence with interferences in the COER dataset.

	Results
Input	《米兰体育报》披露，加里亚尼是前来探营，近距离接触葡萄牙天才维罗索。(La Gazzetta Dello Sport revealed that Galliani came to visit the camp, took a close encounter with the Portuguese genius Veloso.)
Gold (Ground Truth)	Head entity: 加里亚尼 (Galliani) Relation: 接触 (encounter) Tail entity: 维罗索 (Veloso)
DSNFs	Head entity: 米兰体育报 (La Gazzetta Dello Sport) Relation: 接触 (encounter) Tail entity: 维罗索 (Veloso)
CORE-KE	Head entity: 加里亚尼 (Galliani) Relation: 接触 (encounter) Tail entity: 维罗索 (Veloso)

From the sentence [《米兰体育报》披露，加里亚尼是前来探营，近距离接触葡萄牙天才维罗索 (*La Gazzetta Dello Sport revealed that Galliani came to visit the camp, took a close encounter with the Portuguese genius Veloso*)] in Table 8, 加里亚尼 (*Galliani*), 维罗索 (*Veloso*), and 接触 (*encounter*) are annotated as a head entity, a tail entity, and a relation, respectively. There are some interferences in this sentence, such as 《米兰体育报》(*La Gazzetta Dello Sport*), 披露 (*revealed*), 探营 (*visit the camp*), and 葡萄牙 (*Portuguese*). When we run ORE systems (the *CORE-KE* system and the *DSNFs* system), only the *CORE-KE* system can extract 加里亚尼 (*Galliani*) as head entity correctly. The *DSNFs* and *CORE-KE* systems can precisely extract 接触 (*encounter*) as a relation. The tail entity 维罗索 (*Veloso*) can be correctly extracted from two systems. These results indicate that the *CORE-KE* system can eliminate interferences and extract fine-grained entities.

From the sentence [如果想达到好的效果，需要操作者具有精确的时间判断及操控能力 (*To achieve good results, the operator is required to have precise time judgment and manipulation ability*)] in Table 9, 操作

者 (*operator*) is annotated as a head entity, 具有 (*have*) is annotated as a tail entity, and 精确的时间判断及操控能力 (*precise time judgment and manipulation ability*) is annotated as a relation. It is noted that the tail entity in this sentence is very long. The *DSNFs* system cannot extract open relations from this sentence. The *CORE-KE* system can extract the whole triple accurately. These results demonstrate that the *CORE-KE* system can extract very long entities (especially long-tail entities) correctly.

Table 9. ORE results of a Chinese sentence with long entities in the SpanSAOKE-NO dataset.

	Results
Input	如果想达到好的效果，需要操作者具有精确的时间判断及操控能力。(To achieve good results, the operator is required to have precise time judgment and manipulation ability.)
Gold (Ground Truth)	Head entity: 操作者 (Operator) Relation: 具有 (Have) Tail entity: 精确的时间判断及操控能力 (Precise time judgment and manipulation ability)
DSNFs	Head entity: - Relation: - Tail entity: -
CORE-KE	Head entity: 操作者 (Operator) Relation: 具有 (Have) Tail entity: 精确的时间判断及操控能力 (Precise time judgment and manipulation ability)

The CORE-KE system was trained on one NVIDIA 2080Ti GPU for 5 hours on the COER dataset, while the PGCORE system was trained for 17 hours on one NVIDIA 1050Ti GPU [8]. The CORE-KE system is proven to be deployed quickly onto new datasets. As a result, taxonomy evolution can be effectively alleviated.

4.4.5 Error Analysis

In this section, typical errors extracted by the CORE-KE system from the COER and SpanSAOKE-NO datasets are described through examples in Tables 10 and 11, respectively.

It is noted that the ground truth head entity of the example case is [戈麦斯与波斯蒂加 (*Gomez and Postiga*)] in Table 10, the CORE-KE system extracts only part of the correct answer. One reason is that the acquired descriptive knowledge brings the noise to the original training data. For example, the Descriptive Knowledge Acquisition module of the CORE-KE system has acquired the description knowledge of 戈麦斯 (*Gomez*), 波斯蒂加 (*Postiga*), and 葡萄牙队 (*Portugal*). All of these three entities and their descriptions are utilized to further pre-train the WoBERT_plus model. Although in most cases, using granule description knowledge can improve the understanding of a certain concept. However, in this special case with two entities in parallel, the separate description of 戈麦斯 (*Gomez*) and 波斯蒂加 (*Postiga*) indeed creats noise in the relation extraction.

Table 10. Errors extracted by the CORE-KE system from the COER dataset.

	Results
Input	戈麦斯与波斯蒂加则为葡萄牙队打入两球。(Gomez and Postiga scored two goals for Portugal.)
Gold (Ground Truth)	Head entity: 戈麦斯与波斯蒂加 (Gomez and Postiga) Relation: 打入两球 (scored two goals) Tail entity: 葡萄牙队 (Portugal)
CORE-KE	Head entity: 戈麦斯 (Gomez) Relation: 打入两球 (scored two goals) Tail entity: 葡萄牙队 (Portugal)

Table 11. Errors extracted by the CORE-KE system from the SpanSAOKE-NO dataset.

	Results
Input	张女士祖籍黄冈，是香港湖北联谊会会员。(Ms. Zhang, whose ancestral home is Huanggang, is a member of the HongKong-Hubei Friendship Association.)
Gold (Ground Truth #1)	Head entity: 张女士 (Ms. Zhang) Relation: 祖籍 (ancestral home) Tail entity: 黄冈 (Huanggang)
Gold (Ground Truth #2)	Head entity: 张女士 (Ms. Zhang) Relation: 会员 (member) Tail entity: 香港湖北联谊会 (HongKong-Hubei Friendship Association)
CORE-KE	Head entity: 张女士 (Ms. Zhang) Relation: - Tail entity: -

As shown in Table 11, there are two ground truth triples in this example. The CORE-KE system only extracted the common head entity of the two triples. One possible reason is that the CORE-KE system is designed to extract only one triple from a sentence. With the support of knowledge enhancement technology, the CORE-KE system has recognized two potential relations from the example sentence, but is still confused with which one to be used as the final relation.

5. CONCLUSION AND FUTURE WORK

In this paper, a new method entitled Chinese Open Relation Extraction with Knowledge Enhancement (CORE-KE) is presented. The CORE-KE system which implements this method takes advantage of a pre-trained language model (PLM) (with the support of a BiLSTM layer and a Masked CRF layer) on unstructured data and extracts Chinese open relations. To the best of our knowledge, this is the first Chinese ORE method based on a PLM.

Descriptive knowledge from Wikidata and extra triplet knowledge from unstructured Chinese corpus were obtained through two separate modules. Chinese sentences, combined with descriptive and triplet

knowledge, were further used in the CORE-KE system to pre-train and fine-tune the pre-trained language model. In addition, syntactic features were adopted in the training stage of the CORE-KE system.

The experimental results of the CORE-KE system on two large-scale datasets of open Chinese entities and relations demonstrate that the CORE-KE method is superior to other ORE methods. The F1-scores of the CORE-KE method on the two datasets have given a relative improvement of 20.1% and 1.3%, when compared with benchmark ORE methods, respectively. Additionally, the results from the ablation study and the case studies also demonstrate that the CORE-KE method is effective in addressing ORE challenges such as fine-grained entity acquisition, long-tail entity mining, and taxonomy evolution.

There are still some limitations in the CORE-KE method, the future work can be summarized as below:

1) As discussed in the Results and Analysis section, the CORE-KE system, when compared with the benchmark neural ORE systems (that is, SpanOIE, CharLSTM, and MGD-GNN systems), had a lower recall. Further investigation will be undertaken to understand the underlying mechanism for such low performance.
2) Additionally, the CORE-KE method will be improved in the future studies to extract overlapping entities and multiple open relations from Chinese sentences.

ACKNOWLEDGEMENTS

This work was supported by the high-level university construction special project of Guangdong province, China 2019 (No. 5041700175) and the new engineering research and practice project of the Ministry of Education, China (NO. E-RGZN20201036).

AUTHOR CONTRIBUTION STATEMENT

Mr. Chaojie Wen has initiated the proposed system and conducted the experiments. Mr. Chaojie Wen, Dr. Tao Chen, and Dr. Xudong Jia have contributed to the final version of the manuscript. Dr. Tao Chen and Dr. Xudong Jia have provided guidance to the project.

REFERENCES

[1] Pawar, S., Palshikar, G. K., Bhattacharyya, P.: Relation extraction: A survey. arXiv preprint arXiv:1712.05191 (2017)
[2] Etzioni, O., Banko, M., Soderland, S., et al.: Open information extraction from the web. Communications of the ACM 51(12), 68–74 (2008)
[3] Fader, A., Soderland, S., Etzioni, O.: Identifying relations for open information extraction. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pp. 1535–1545 (2011)
[4] Kolluru, K., Adlakha, V., Aggarwal, S., et al.: Openie6: Iterative grid labeling and coordination analysis for open information extraction. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pp. 3748–3761 (2020)

[5] Tseng, Y. H., Lee, L. H., Lin, S. Y., et al.: Chinese open relation extraction for knowledge acquisition. In: Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 12–16 (2014)
[6] Qiu, L., Zhang, Y.: ZORE: A syntax-based system for chinese open relation extraction. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, pp. 1870–1880 (2014)
[7] Wang, C., He, X., Zhou, A.: Open relation extraction for chinese noun phrases. IEEE Transactions on Knowledge and Data Engineering 33(6), 2693–2708 (2019)
[8] Cheng, Z., Wu, X., Xie, X., et al.: Chinese open relation extraction with pointer-generator networks. In: Proceedings of 2020 IEEE 5th International Conference on Data Science in Cyberspace, pp. 307–311 (2020)
[9] Jia, S., E, S., Li, M., et al.: Chinese open relation extraction and knowledge base establishment. ACM Transactions on Asian and Low-Resource Language Information Processing 17(3), 1–22 (2018)
[10] Che, W., Feng, Y., Qin, L., et al.: N-LTP: An open-source neural language technology platform for Chinese. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 42–49 (2021)
[11] Wei, T., Qi, J., He, S., et al.: Masked conditional random fields for sequence labeling. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 2024–2035 (2021)
[12] Zhang, N., Jia, Q., Deng, S., et al.: Alicg: Fine-grained and evolvable conceptual graph construction for semantic search at alibaba. In: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 3895–3905 (2021)
[13] Lyu, Z., Shi, K., Li, X., et al.: Multi-grained dependency graph neural network for Chinese open information extraction. In: Proceedings of Advances in Knowledge Discovery and Data Mining: 25th Pacific-Asia Conference, PAKDD 2021, pp. 155–167 (2021)
[14] Qin, B., Liu, A., Liu, T.: Unsupervised Chinese open entity relation extraction. Journal of computer research and development 52(5), 1029–1035 (2015)
[15] Zhan, J., Zhao, H.: Span model for open information extraction on accurate corpus. In: Proceedings of the AAAI Conference on Artificial Intelligence (Volume 34, No. 05), pp. 9523–9530 (2020)
[16] Lample, G., Ballesteros, M., Subramanian, S., et al.: Neural architectures for named entity recognition. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 260–270 (2016)
[17] Schmitz, M., Soderland, S., Bart, R., et al.: Open language learning for information extraction. In: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 523–534 (2012)
[18] Christensen, J., Soderland, S., Etzioni, O.: An analysis of open information extraction based on semantic role labeling. In: Proceedings of the 6th International Conference on Knowledge Capture, pp. 113–120 (2011)
[19] Del Corro, L., Gemulla, R.: Clausie: clause-based open information extraction. In: Proceedings of the 22nd International Conference on World Wide Web, pp. 355–366 (2013)
[20] Angeli, G., Premkumar, M. J. J., Manning, C. D.: Leveraging linguistic structure for open domain information extraction. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 344–354 (2015)
[21] Pal, H.: Demonyms and compound relational nouns in nominal open IE. In: Proceedings of the 5th Workshop on Automated Knowledge Base Construction, pp. 35–39 (2016)
[22] Stanovsky, G., Ficler, J., Dagan, I., et al.: Getting more out of syntax with props. arXiv preprint arXiv:1603.01648 (2016)

[23] Mausam, M.: Open information extraction systems and downstream applications. In: Proceedings of the 25th International Joint Conference on Artificial Intelligence, pp. 4074–4077 (2016)
[24] Saha, S., Pal, H.: Bootstrapping for numerical open IE. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 317–323 (2017)
[25] Gashteovski, K., Gemulla, R., Corro, L. D.: Minie: minimizing facts in open information extraction. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 2630–2640 (2017)
[26] Cetto, M., Niklaus, C., Freitas, A., et al.: Graphene: Semantically-linked propositions in open information extraction. In: Proceedings of the 27th International Conference on Computational Linguistics, pp. 2300–2311 (2018)
[27] Saha, S.: Open information extraction from conjunctive sentences. In: Proceedings of the 27th International Conference on Computational Linguistics, pp. 2288–2299 (2018)
[28] Roy, A., Park, Y., Lee, T., et al.: Supervising unsupervised open information extraction models. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pp. 728–737 (2019)
[29] Cui, L., Wei, F., Zhou, M.: Neural open information extraction. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp.407–413 (2018)
[30] Kolluru, K., Aggarwal, S., Rathore, V., et al.: Imojie: Iterative memory-based joint open information extraction. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 5871–5886 (2020)
[31] Stanovsky, G., Michael, J., Zettlemoyer, L., et al.: Supervised open information extraction. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pp. 885–895 (2018)
[32] Eberts, M., Ulges, A.: Span-based joint entity and relation extraction with transformer pre-training. In: Proceedings of the 24th European Conference on Artificial Intelligence, pp. 2006–2013 (2020)
[33] Jia, S., Shijia, E., Ding, L., et al.: Hybrid neural tagging model for open relation extraction. Expert Systems with Applications 200, 116951 (2022)
[34] Gan, J., Huang, P., Zhou, J., et al.: Chinese open information extraction based on DBMCSS in the field of national information resources. Open Physics 16(1), 568–573 (2018)
[35] Devlin, J., Chang, M. W., Lee, K., et al.: 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), pp. 4171–4186 (2019)
[36] Cui, Y., Che, W., Liu, T., et al.: Pre-training with whole word masking for chinese bert. IEEE/ACM Transactions on Audio, Speech, and Language Processing 29, 3504–3514 (2021)
[37] Ratcliff, J. W., Metzener, D. E.: Pattern matching: The gestalt approach. Dr Dobbs Journal 13(7), 46 (1988)
[38] See, A., Liu, P. J., Manning, C. D.: Get to the point: Summarization with pointer-generator networks. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1073–1083 (2017)
[39] Li, S., Zhao, Z., Hu, R., et al.: Analogical reasoning on chinese morphological and semantic relations. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 138–143 (2018)
[40] Liu, Y., Ott, M., Goyal, N., et al.: Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)

[41] Cui, Y., Che, W., Liu, T., et al.: Revisiting pre-trained models for Chinese natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings, pp. 657–668 (2020)
[42] Wei, J., Ren, X., Li, X., et al.: Nezha: Neural contextualized representation for chinese language understanding. arXiv preprint arXiv:1909.00204 (2019)
[43] Clark, K., Luong, M. T., Le, Q. V., et al.: Electra: Pre-training text encoders as discriminators rather than generators. In: Proceedings of the International Conference on Learning Representations (2020)
[44] Shen, Y., Tan, S., Sordoni, A., et al.: Ordered neurons: Integrating tree structures into recurrent neural networks. In: Proceedings of the International Conference on Learning Representations (2018)
[45] Li, J., Liu, Z.: Granule description in knowledge granularity and representation. Knowledge-Based Systems 203, 106160 (2020)
[46] Liu, H., Li, W., Li, Y.: A new computational method for acquiring effect knowledge to support product innovation. Knowledge-Based Systems 231, 107410 (2021)
[47] Zhang, J., Hao, K., Tang, X. S., et al.: A multi-feature fusion model for Chinese relation extraction with entity sense. Knowledge-Based Systems 206, 106348 (2020)
[48] Gou, Y., Lei, Y., Liu, L., et al.: A dynamic parameter enhanced network for distant supervised relation extraction. Knowledge-Based Systems 197, 105912 (2020)
[49] Li, Q., Li, L., Wang, W., et al.: A comprehensive exploration of semantic relation extraction via pre-trained CNNs. Knowledge-Based Systems 194, 105488 (2020)
[50] Wu, R., Yao, Y., Han, X., et al: 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, pp. 219–228 (2019)
[51] Zhang, N., Xu, X., Tao, L., et al: DeepKE: A deep learning based knowledge extraction toolkit for knowledge base population. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 98–108 (2022)

AUTHOR BIOGRAPHY

Chaojie Wen was born in Jiangmen, Guangdong, China, in 1996. He received his B.S. degree in computer science and technology and M.S. degree in electronic and communication engineering from Wuyi University in 2019 and 2022, respectively. His research interests include natural language processing, named entity recognition, and relation extraction.
ORCID: 0000-0002-9325-9147

Xudong Jia is a Visiting Scholar of Wuyi University, China. He is also a Professor and the Associate Dean of College of Engineering and Computer Science, California State University, Northridge. He received his B.S. in 1983 and M.S. in 1986 from Beijing Jiaotong University, his M.S. in 1992 from University of Toronto, Canada, and his Ph.D. in 1996 from Georgia Institute of Technology. His research interests include intelligent transportation systems (ITS) standards, geographic information system (GIS) applications in transportation, traffic safety, transportation information systems, travel demand management, and air quality. He is an associate editor of the IEEE Intelligent Transportation Systems Society and the Open Journal of IEEE Intelligent. Transportation Systems.

Tao Chen was born in Shiyan, Hubei, China, in 1981. He received his B. Eng. degree in communication engineering in 2003 from Nanjing Institute of Communication Engineering, China, and his M.Eng. degree in computer application in 2013 from Wuyi University, Guangdong, China, and his Ph.D. degree in computer application from Harbin Institute of Technology, China in 2018. He joined Wuyi University, China, as a lecturer, in 2018. He is the author of one book, 23 articles and 3 patents. His research interests include natural language processing, deep learning, and knowledge acquisition and reasoning.
ORCID:0000-0002-3634-0854