Cross-lingual Inference with A Chinese Entailment Graph

Anonymous ACL submission

Abstract

Predicate entailment detection is a crucial task for question-answering from text, where previous work has explored unsupervised learning of entailment graphs from typed open relation triples. In this paper, we present the first pipeline for building Chinese entailment graphs, which involves a novel high-recall open relation extraction (ORE) method and the first Chinese fine-grained entity typing dataset under the FIGER type ontology. Through experiments on the Levy-Holt dataset and a boolean QA task, we verify the strength of our Chinese entailment graph, and reveal the cross-lingual complementarity: on the parallel Levy-Holt dataset, an ensemble of Chinese and English entailment graphs beats both monolinguals, and raises unsupervised SOTA by 4.7 AUC points[1].

1 Introduction

Predicate entailment detection is important for many tasks of natural language understanding (NLU), including reading comprehension and semantic parsing. Suppose we wish to answer a question by finding a relation **V** between entities **A** and **B**. Often, **V** cannot be found directly from the reference passage or database, but another relation **U** can be found between **A** and **B**, where **U** entails **V** (for instance, suppose **U** is *buy*, **V** is *own*). If we can identify this with predicate entailment detection, we can then answer the question.

To detect predicate entailments, previous work has explored unsupervised learning of typed entailment graphs (Szpektor and Dagan, 2008; Berant et al., 2011, 2015; Hosseini et al., 2018, 2019, 2021). Entailment graphs are directed graphs, where each node represents the predicate of a relation, and an edge from node **U** to node **V** denotes "**U** entails **V**". Entailment graphs are built based on the Distributional Inclusion Hypothesis (DIH) (Dagan et al., 1999; Geffet and Dagan, 2005; Herbelot and Ganesalingam, 2013; Kartsaklis and Sadrzadeh, 2016). Predicates are disambiguated according to their arguments' types, predicates taking the same types of arguments go into one subgraph.

While previous work on entailment graphs has mostly been limited to English, building entailment graphs in other languages is interesting and challenging. The importance is two-fold: for that language, a native entailment graph would facilitate NLU in it; for cross-lingual inference, entailment graphs in different languages host exploitable complementary information. In particular, we argue that by jointly consulting strong entailment graphs in multiple languages, improvements can be gained for inference in **all** participant languages.

In this paper, we choose Chinese as our target language to build entailment graphs, as it is distant enough from English to exhibit rich complementarity, while relatively high-resource. The main challenge for building Chinese entailment graphs, is to extract reliable **typed relation triples** from raw corpora as strong input. This involves open relation extraction (ORE) and fine-grained entity typing (FET), which we discuss below.

ORE extracts predicate-argument triples from sentences, where previous work has used rule-based methods over syntactic parsers either directly (Fader et al., 2011; Etzioni et al., 2011; Angeli et al., 2015), or for distant supervision (Cui et al., 2018; Stanovsky et al., 2018; Kolluru et al., 2020). The challenge in ORE can be largely attributed to the poor definition of "open relations". The situation worsens in Chinese, as the parts of speech are more ambiguous and many linguistic indicators of relations are poorly represented. Previous work on Chinese ORE (Qiu and Zhang, 2014; Jia et al., 2018) has defined narrow sets of open relations, failing to identify many relational constructions. Conversely, we propose a novel dependency-based ORE method, which we claim provides comprehensive coverage of relational constructions.

[1]Our codes and data-sets will be available on Github.

FET assigns types to the arguments of extracted relations, so that word-senses of predicates can be disambiguated. The challenge in Chinese FET lies mainly in the lack of datasets over a suitable type ontology: too coarse a type set would be insufficient for disambiguation, too granular a type set would result in disastrous sparsity in the entailment graph. Following Hosseini et al. (2018), we use the popular FIGER type set (Ling and Weld, 2012), and build CFIGER, the first FIGER-labelled Chinese FET dataset via label mapping. Entity typing models built on this dataset show satisfactory accuracy and are helpful for predicate disambiguation.

We evaluate our Chinese entailment graph on the Levy-Holt entailment dataset (Levy and Dagan, 2016; Holt, 2019) (via translation) and a natively-Chinese boolean QA task following McKenna et al. (2021). Results show that our Chinese entailment graph outperforms baselines by large margins, and is comparable to the English graph. We verify our cross-lingual complementarity argument on Levy-Holt dataset: by ensembling English and Chinese graphs, we show a clear advantage over both monolingual graphs, and set a new SOTA.

Our contributions are as follows: 1) we present a novel Chinese ORE method sensitive to a much wider range of relations than previous SOTA, and a Chinese FET dataset, the first under the FIGER type ontology; 2) we construct the first Chinese entailment graph, comparable to its English counterpart; 3) we reveal the cross-lingual complementarity of entailment graphs with an ensemble.

2 Background and Related Work

Predicate entailment detection has been an area of active research. Lin (1998); Weeds and Weir (2003); Szpektor and Dagan (2008) proposed various count-based entailment scores; Berant et al. (2011) proposed to "globalize" typed entailment graphs by closing them with transitivity constraints; Hosseini et al. (2018) proposed a more scalable global learning approach with soft transitivity constraints; Hosseini et al. (2019, 2021) further refined the entailment scores with link prediction.

Our work is closely related to Hosseini et al. (2018), with key adaptations for Chinese in ORE and FET. Their ORE method is based on a CCG parser (Reddy et al., 2014), while ours is based on a dependency parser (Zhang et al., 2020); their FET is done by linking entities to Wikipedia entries, while we use neural entity typing for the task.

Dependency parses are less informative than CCG parses, and require heavier adaptation. However, Chinese dependency parsers are currently more reliable than CCG parsers (Tse and Curran, 2012). Previous Chinese ORE methods (Qiu and Zhang, 2014; Jia et al., 2018) are based on dependency parsers, but they omit many common constructions essential to ORE. In §3, we present the most comprehensive Chinese ORE method so far.

Linking-based entity-typing can be more accurate than neural methods, since the type labels are exact as long as linking is correct. However, current Chinese entity linking methods require either translation (Pan et al., 2019) or search logs (Fu et al., 2020). Both hurt linking accuracy, and the latter grows prohibitively expensive with scale. On the other hand, since the seminal work of Ling and Weld (2012), neural fine-grained entity typing has developed rapidly (Yogatama et al., 2015; Shimaoka et al., 2017; Chen et al., 2020), with a common interest in FIGER type set. For Chinese, Lee et al. (2020) built an ultra-fine-grained entity typing dataset, based on which we build our CFIGER dataset via label mapping.

Weber and Steedman (2019) aligned English and German entailment graphs, and showed that the English graph can help with German entailment detection. Yet it was uncertain whether this effect comes from genuine complementarity or the mere fact that the English graph is stronger. We take one step further, and show that complementarity can be exploited in both directions: for English, the higher resource language, entailment detection can also benefit from the ensemble to reach new heights.

As a related resource, Ganitkevitch et al. (2013) created a multi-lingual database for symmetric paraphrases; in contrast, entailment graphs are directional. More recently, Schmitt and Schütze (2021) proposed to fine-tune language models on predicate entailment datasets via prompt learning. In contrast to our entailment graphs, their approach is supervised, which carries the danger of overfitting to dataset artifacts (Gururangan et al., 2018).

Another related strand of research, e.g. SNLI (Bowman et al., 2015), is concerned with the more general NLI task, including hypernymy detection and logic reasoning like $A \wedge B \rightarrow B$, but rarely covers the cases requiring external knowledge of predicate entailment. Conversely, entailment graphs aim to serve as a robust resource for directional predicate entailments induced from textual corpora.

3 Chinese Open Relation Extraction

We build our ORE method based on DDParser (Zhang et al., 2020), a SOTA Chinese dependency parser. We mine relation triples from its output by identifying patterns in the dependency paths.

Depending on the semantics of the head verb, instances of a dependency pattern can range from being highly felicitous to marginally acceptable as a relation. Motivated by our downstream task of entailment graph construction, we go for higher recall and take them in based on the **Relation Frequency Assumption**: less felicitous relations occur less frequently, and are less likely to take part in entailments when they do occur, thus they are negligible.

Due to the lack of a commonly accepted benchmark or criterion for "relations", we did not perform an intrinsic evaluation for our Chinese ORE method; its significant benefit to our EG_{Zh} graph (§6.3, §7) should suffice to demonstrate its strength.

3.1 Parsing for Chinese ORE

The task of open relation extraction on top of LM-driven dependency parsers, is really the task of binding the relations in surface forms to the underlying relation structures. Though trivial at first sight, the definition of these underlying and essentially semantic relations demands detailed analysis.

Jia et al. (2018) is the latest to propose an ORE method on dependency paths. They defined a set of rules to extract relations patterns, which they call dependency semantic normal forms (DSNFs)[2].

However, their set of DSNFs is inexhaustive and somewhat inaccurate. We show below that many linguistic features of Chinese demand a more principled account, more constructions need to be considered as relations, some to be ruled out. These observations are made from a multi-source news corpus, which we use to build entailment graphs (§5)[3]. Below, we highlight 5 additional constructions we identify, explained with examples[4].

A. PP Modifiers as "De" Structures One key feature of Chinese is its prevalent use of *"De"* structures in the place of prepositional phrases, where *"De"* can be seen as roughly equivalent to the possessive clitic ***'s***. For instance, in *"咽炎(pharyngitis) 成为(becomes) 发热(fever) 的(De) 原因(cause); Pharyngitis becomes the cause of fever"*, the root clause in Chinese is (Pharyngitis, becomes, cause), but we *additionally* extract the underlying relation **(pharyngitis, becomes·X·De·cause, fever)**, where the true object *"fever"* is a **nominal** attribute of the direct object *"cause"*, and the true predicate subsumes the direct object[5].

The same also applies to the subject, though somewhat more restricted. For sentences like *"苹果(Apple) 的(De) 创始人(founder) 是(is) 乔布斯(Jobs); The founder of Apple is Jobs"*, we additionally extract the relation **(Apple, founder·is, Jobs)**, where the true subject *"Apple"* is a **nominal** attribute of the direct subject *"founder"*, and the true predicate subsumes the direct subject[6].

B. Bounded Dependencies In Chinese, bounded dependencies, especially control structures, are expressed with a covert infinitival marker, equivalent to English "to". We capture the following phenomena in addition to direct relations:

- Sequences of VPs: for sentences like *"我(I) 去(go-to) 诊所(clinic) 打(take) 疫苗(vaccine); I go to the clinic to take the vaccine"*, the two verb phrases "去(go-to) 诊所(clinic)" and "打(take) 疫苗(vaccine)" are directly concatenated, with no overt connection words. Here we additionally extract the relation **(I, take, vaccine)** by copying the subject of the head verb to subsequent verbs.
- Subject-control verbs: for the famous example *"我(I) 想(want) 试图(try) 开始(begin) 写(write) 一个(a) 剧本(play); I want to try to begin to write a play"*, again the verbs are directly concatenated; this time, all verbs but the first one behaves as infinitival complements to their direct antecedents. In such cases, we extract sequences of relations like **(I, want, try)**, **(I, want·try, begin)**, **(I, want·try·begin, write)**, **(I, want·begin·try·write, a play)**.

Notably, the above relations are different from Jia

[2]We refer readers to Appendix A for a brief summary.

[3]A referee has commented that when refining our ORE method, we might have inadvertently or unconsciously fine-tuned the system for the evaluation tasks. However, as entailment graph construction is fully unsupervised, this source corpus is independent of the evaluations in §6 and §7. Particularly, the Levy-Holt dataset used in §6 consists of short sentences, which is a vastly different genre, involving much simpler structures, with a single relation per sentence and few subordinating constructions discussed above (see Appendix L for supporting statistics); the QA dataset used in §7 is built from a separate news corpus, strictly excluding overlaps with those used to develop the parser and the entailment graphs.

[4]We refer readers to Appendix J for diagram illustrations.

[5]Here and below, examples are paired with English metaphrases, and when necessary, paraphrases; relation triples are presented as English metaphrases (inflections ignored) .

[6]These relations are more felicitous with frequent predicate argument combinations, and less so for the infrequent ones. As in line with the Relation Frequency Assumption, less felicitous relations are also less statistically significant.

et al. (2018)'s conjunctions in Table 4: the event sequences here involve subordination (control) rather than coordination, thus need a separate account.

C. Relative Clauses Relative Clauses also take the form of modification structures in Chinese, for which additional relations should also be extracted. For example, in *"他(he) 解决(solve) 了(-ed) 困扰(puzzle) 大家(everyone) 的(De) 问题(problem); He solved the problem that puzzled everyone"*, we extract not only the direct relation **(he, solve, problem)**, but also the relation embedded in the modification structure **(problem, puzzle, everyone)**.

D. Nominal Compounds Relations can be extracted from nominal compounds, where an NP has two consecutive "ATT" modifiers: in *"德国(Germany) 总理(Chancellor) 默克尔(Merkel); German Chancellor Merkel"*, "Germany" modifies "Chancellor", and "Chancellor" modifies "Merkel". Jia et al. (2018) extracted relations like **(Germany, Chancellor, Merkel)** for these NPs.

However, they overlooked the fact that prepositional compounds in Chinese with omitted "De" take exactly the same form (see construction **A**). For example, in NPs with nested PP modifiers like *"手续(formalities) 办理(handle) 时效(timeliness); Timeliness of the handling of formalities"* , we have the same structure, but it certainly does not mean *"the handling of formalities is timeliness"*!

We take a step back and put restrictions on such constructions: only when all 3 words in the NP are nominals (but not pronouns), the third word is the head, the second is a 'PERSON' or 'TITLE', and the first is a 'PERSON', then it is a relation, like **(Merkel, is·X·De·Chancellor, Germany)**. Otherwise, such NPs rarely host felicitous relations.

E. Copula with Covert Objects The copula is sometimes followed by modifiers ending with *"De"*. Examples are *"玉米(Corn) 是(is) 从(from) 美国(US) 引进(introduce) 的(De); Corn is introduced from US"*, *"设备(device) 是(is) 木头(wood) 做(make) 的(De); The device is made of wood"*.

In these cases, there exists an object following the indicator *"的(De)"*, but the object is an empty ***pro*** considered inferable from context. In the absence of the true object, the *VOB* label is given to *"的(De)"*, leading to direct relations like **(Corn, is, De)**. However, the true predicates are rather *"is introduced from"* or *"is made of"*. To fix this, we **replace** the direct relations with ones like **(Corn, is·from·X·introduce·De·*pro*, America)**, reminiscent of the constructions **A**.

3.2 Our ORE Method

With the above constructions taken into account, we build our ORE method on top of DDParser. For part-of-speech labels, we use the POS-tagger in Stanford CoreNLP (Manning et al., 2014). We detect negations by looking for negation keywords in the adjunct modifiers of predicates: for predicates with an odd number of negation matches, we insert a negation indicator to them, treating them as separate predicates from the non-negated ones.

4 Chinese Fine-Grained Entity Typing

As shown in previous work (Berant et al., 2011; Hosseini et al., 2018), the types of a predicate's arguments are helpful for disambiguating a predicate in context. To this end, we need a fine-grained entity typing model to classify the arguments into sufficiently discriminative yet populous types.

Lee et al. (2020) presented CFET dataset, an ultra-fine-grained entity typing dataset in Chinese. They labelled entities in sentence-level context, into around 6,000 free-form types and 10 general types. Unfortunately, their free-form types are too fragmented for predicate disambiguation, and their general types are too ambiguous.

We turn to FIGER (Ling and Weld, 2012), a commonly used type set: we re-annotate the CFET dataset with FIGER types through label mapping. Given that there are around 6,000 ultra-fine-grained types and only 112 FIGER types (49 in the first layer), we can reasonably assume that each ultra-fine-grained type can be unambiguously mapped to a single FIGER type. For instance, the ultra-fine-grained type "湖 (lake)" is unambiguously mapped to the FIGER label "location / body_of_water".

Based on this assumption, we manually create a mapping between the two, and re-annotate CFET dataset with the mapping. We call the re-annotated dataset **CFIGER**, as it is the first in Chinese with FIGER labels. As with CFET, this dataset consists of 4.8K crowd-annotated data (equally divided into crowd-train, crowd-dev and crowd-test) and 1.9M distantly supervised data from Wikipedia[7].

For training set we combine the crowd-train and Wikipedia subsets; for dev and test sets we use crowd-dev and crowd-test respectively. We train two baseline models: *CFET*, the baseline model with CFET dataset; *HierType* (Chen et al., 2020), a SOTA English entity typing model.

[7]For detailed statistics, please refer to Appendix B.

Macro F1 (%)	dev	test
CFET with CFET dataset	-	24.9
CFET with CFIGER dataset	75.7	75.7
HierType with FIGER dataset	-	82.6
HierType with CFIGER dataset	74.8	74.5

Table 1: F1 scores of baseline models for CFIGER dataset, compared with the results on the datasets where they were proposed. Macro-F1 scores are reported because it is available in both baselines.

Results are shown in Table 1: the F1 score of *HierType* model is slightly lower on CFIGER dataset than on FIGER dataset in English; contrarily, thanks to fewer type labels, the F1 score of *CFET* baseline increases on CFIGER, bringing it on par with the more sophisticated *HierType* model. This means our CFIGER dataset is valid for Chinese fine-grained entity typing, and may contribute to a benchmark for cross-lingual entity typing.

For downstream applications, we nevertheless employ the *HierType* model, as empirically it generalizes better to our news corpora. As shown in later sections, the resulting FET model can substantially help with predicate disambiguation.

5 The Chinese Entailment Graph

We construct the Chinese entailment graph from the Webhose corpus[8], a multi-source news corpus crawled from 133 news websites in October 2016. Similarly to the NewsSpike corpus used in Hosseini et al. (2018), the Webhose corpus contains multi-source non-fiction articles from a short period of time. This means it is also rich in reliable and diverse relation triples over a focused set of events, ideal for building entailment graphs.

For the 313K valid articles in Webhose, we get their CoreNLP POS tags and feed them into our ORE method in §3, to extract the open relation triples. Then, with *HierType* model (Chen et al., 2020) in §4, we type all arguments of the extracted relations; we type each predicate with its subject-object type pair, such as *person-event* or *food-law*.

We finally employ the entailment graph construction method in Hosseini et al. (2018), taking in only binary relations. The detailed statistics of our Chinese entailment graph are shown in Table 2: compared with EG_{En}, our graph is built on just over half the number of articles, yet we have extracted 70% the number of relation triples, and built a graph

	EG_{Zh}	EG_{En}
# of articles taken	313,718	546,713
# of triples used	7,621,994	10,978,438
# of predicates	363,349	326,331
# of type pairs where:		
subgraph exists	942	355
$\lvert$subgraph$\rvert$ > 100	442	115
$\lvert$subgraph$\rvert$ > 1,000	149	27
$\lvert$subgraph$\rvert$ > 10,000	26	7

Table 2: Stats of our Chinese entailment graph (EG_{Zh}) compared with the English graph in Hosseini et al. (2018) (EG_{En}). $|\cdot|$ denotes the number of predicates.

involving even more predicates. In general, our EG_{Zh} is of comparable size to EG_{En}. We encourage interested readers to check Appendix D for details of our graph-building process and a quick introduction to Hosseini et al. (2018).

6 Evaluation by Entailment Detection

6.1 Benchmark and Baselines

To evaluate the quality of our Chinese entailment graph, we first perform an intrinsic evaluation on the predicate entailment detection task. Our experiments are based on the popular Levy-Holt dataset (Levy and Dagan, 2016; Holt, 2019), with the same dev/test configuration as Hosseini et al. (2018). We convert the Levy-Holt dataset to Chinese through machine translation, then do evaluation on the translated premise-hypothesis pairs.

In Levy-Holt dataset, the task is: to take as input a pair of relation triples about the same arguments, one premise and one hypothesis, and judge whether the premise entails the hypothesis. To convert Levy-Holt dataset into Chinese, we concatenate each relation triple into a pseudo-sentence, use Google Translate to translate the pseudo-sentences into Chinese, then parse them back to Chinese relation triples with our ORE method in §3. If multiple relations are returned, we retrieve the most representative ones, by considering only those relations whose predicate covers the HEAD word.[9]

To type the Chinese relation triples, we again use *HierType* model to collect their subject-object type-pairs. The premise and hypothesis need to take the same types of arguments, so we take the intersection of their possible type-pairs[10]. We search the entailment subgraphs of these type-pairs, for entailment edges from the premise to the hypothesis,

[8] https://webhose.io/free-datasets/chinese-news-articles/

[9] See Appendix C for more details.

[10] Unless the intersection is empty, then we take the union.

and return the entailment scores associated with these edges. When edges are found from multiple subgraphs, we take their maximum score[11].

We compare our Chinese entailment graph with a few strong baselines:

BERT: We take the translated pseudo-sentence pairs, and compute the cosine similarity between their pretrained BERT representations at [CLS] token. This is a strong baseline but symmetric;
Jia: We build entailment graph in the same way as §5, but with the baseline ORE method by Jia et al. (2018); accordingly, Jia et al. (2018) method is also used in parsing the translated Levy-Holt;
DDPORE: Similar to *Jia* baseline, but with the baseline ORE method from DDParser (2020).

6.2 Cross-lingual Ensembles

In order to examine the complementarity between our Chinese entailment graph (EG_{Zh}) and the English graph (EG_{En}) (2018), we ensemble the predictions from the two graphs, $pred_{en}$ and $pred_{zh}$[12]. We experiment with four ensemble strategies: lexicographic orders from English to Chinese and Chinese to English, max pooling and average pooling:

$$
\begin{aligned}
pred_{en\_zh} &= pred_{en} + \gamma * \Theta(pred_{en}) * pred_{zh} \\
pred_{zh\_en} &= \gamma * pred_{zh} + \Theta(pred_{zh}) * pred_{en} \\
pred_{max} &= MAX(pred_{en}, \gamma * pred_{zh}) \\
pred_{avg} &= AVG(pred_{en}, \gamma * pred_{zh})
\end{aligned}
$$

where $\Theta(\cdot)$ is the boolean function $IsZero$, γ is the relative weight of Chinese and English graphs. γ is a hyperparameter tuned on Levy-Holt dev set, searched between 0.0 and 1.0 with step size 0.1.

For instance, suppose our premise is *"he, shopped in, the store"*, and our hypothesis is *"he, went to, the store"*, then our Chinese relations, by translation, would be "他, 在·X·购物, 商店" and "他, 前往, 商店" respectively. Suppose we find in the English graph an edge from *"shop in"* to *"go to"*, scored $pred\_en = 0.6$, and we find in the Chinese graph an edge from "在·X·购物" to "前往", scored $pred\_zh = 0.7$. Then we would have $pred_{en\_zh} = 0.6$, $pred_{zh\_en} = 0.7$, $pred_{max} = 0.7$, $pred_{avg} = 0.65$.

In addition to ensembling with EG_{En}, we also ensembled our entailment graph with the SOTA English graph EG_{En}++ (2021). We call the latter ones **Ensemble++** here and below.

[11]We provide a human evaluation on the quality of the resulting Chinese Levy-Holt dataset in Appendix I.
[12]"zh" is the abbreviation for Chinese by convention.

AUC (%)	dev	test
BERT ⋆	5.5	3.2
Jia (2018) ⋆	0.9	2.4
DDPORE (2020) ⋆	9.8	5.9
EG_{Zh} ⋆	**15.7**	**9.4**
EG_{En} (2018) ⋄	20.7	16.5
EG_{En}++ (2021) ⋄	23.3	19.5
Ensemble En_Zh ⋄	**28.3** ($\gamma: 0.8$)	**21.2**
Ensemble Zh_En ⋄	**27.4** ($\gamma: 0.9$)	**21.5**
Ensemble MAX ⋄	**29.9** ($\gamma: 0.8$)	**22.1**
Ensemble AVG ⋄	**30.0** ($\gamma: 1.0$)	**22.1** †
Ensemble++ AVG ⋄	**31.2** ($\gamma: 0.3$)	**24.2** †
EG_{Zh} *-type* ⋆	11.1	7.0
DataConcat En ⋄	20.6	17.8
DataConcat Zh ⋆	19.0	14.2
DataConcat Esb ⋄	31.8	25.0
BackTrans Esb ⋄	23.0	17.5

Table 3: Area Under Curve values on Levy-Holt dataset, for Chinese entailment graph (EG_{Zh}), its baselines, ensembles with English graphs, and ablation studies. EG_{En} is the English graph from (Hosseini et al., 2018); EG_{En}++ is the English graph from (Hosseini et al., 2021). Entries with ⋆ uses Chinese lemma baseline; entries with ⋄ uses English lemma baseline; entries with † are the best ensemble strategies by dev set results.

6.3 Results and Discussions

To measure the performance of our constructed Chinese entailment graphs, we follow previous work in reporting the Precision-Recall (P-R) Curves plotted for successively lower confidence thresholds, and their Area Under Curves (AUC), for the range with $> 50\%$ precision.

A language-specific lemma baseline sets the left boundary of recall, by exact match over the lemmatized premise / hypothesis. For our Chinese entailment graph (EG_{Zh}) and its baselines, the boundary is set by Chinese lemma baseline. For the ensembles, in order to get commensurable AUC values with previous work instead of being over-optimistic, we use the English lemma baseline.

As shown in Table 3, on the Chinese Levy-Holt dataset, our EG_{Zh} graph substantially outperforms the BERT pre-trained baseline. EG_{Zh} is also far ahead of entailment graphs with baseline ORE methods, proving the superiority of our Chinese ORE method against previous SOTA.

EG_{Zh} and EG_{En} are built with the same algorithm (Hosseini et al., 2018), and evaluated on parallel datasets. Learnt from 57% the data, EG_{Zh} achieves an AUC exactly 57% of its English counterpart. Note that the Chinese entailment graph is under-

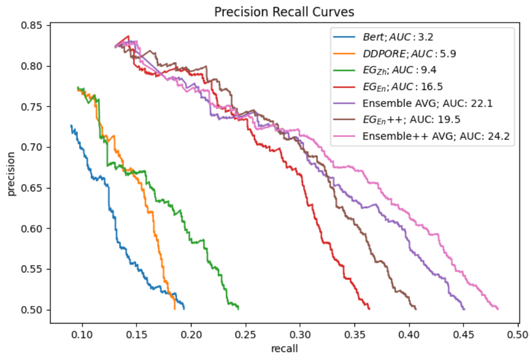


Figure 1: P-R Curves on Levy-Holt test set for EG$_{Zh}$, ensembles and baselines; *Jia(2018)* baseline is far behind others, thus omitted for the clarity of the figure.

estimated with the use of translated dataset: out of the 12,921 relation pairs in Levy-Holt test set, only 9,337 of them are parsed into valid Chinese binary relations. This means, for Chinese entailment graphs, the upper bound for recall is not 100%, but rather 72.3%, as is the upper bound for AUC. Besides, the translationese language style in Chinese Levy-Holt also poses a gap in word-choice to the natively-built entailment graph, resulting in more mismatches. Considering the above extra noise, the performance of EG$_{Zh}$ means our pipeline is utilizing the source corpus very well.

The ensemble between Chinese and English entailment graphs sets a new SOTA for unsupervised predicate entailment detection. With all 4 ensemble strategies, improvement is gained upon both monolingual graphs; with **Ensemble AVG**, the best on dev-set, the margin of test set improvement is more than 5 points. Moreover, when ensembling with EG$_{En}$++, we get a test-set AUC of 24.2 points (**Ensemble++ AVG**), raising SOTA by 4.7 points.

In Table 3, we additionally present three **ablation studies** to verify the solidarity of our approach.

In the first ablation study, *EG$_{Zh}$ -type*, we take away entity typing and train an untyped entailment graph. In this setting, we lose 2.4 AUC points. This means our entity typing method is indeed helpful for disambiguating predicates in entailment graphs.

In the second ablation study, the *DataConcat* settings, we disentangle cross-lingual complementarity from the effect of extra data. We machine-translate NewsSpike corpus into Chinese, Webhose into English. We build an English graph "DataConcat En" using *NewsSpike + translated-Webhose*, and a Chinese graph "DataConcat Zh" using *Webhose + translated-NewsSpike*. Results show that while both graphs improve with data from the other side, they are still far behind our **Ensemble** settings above. Further, ensembling the two DataConcat graphs delivers an AUC of 25.0 points, 7.2 higher than DataConcat En, an even wider margin than our main setting. This suggests, the success of cross-lingual ensemble **cannot** be reproduced by sticking all the data together for a monolingual graph.

In the third case study, *BackTrans Esb*, we disentangle cross-lingual complementarity from the effect of machine-translation. Machine translation can be noisy, but it can also map synonyms in the source language to the same words in the target language. To single out this effect, we translate the Chinese Levy-Holt dataset back into English, and perform an ensemble between predictions on the original and the back-translated Levy-Holt. The gain in this case is only marginal, suggesting that cross-lingual complementarity is the reason for our success, while the synonym effect is not.

To further analyse the improvements with our ensembles, we conduct a case study over the difference in predictions between our ensemble and the English monolingual, thresholded at 65% precision. We find that the majority of the improvements can be attributed to the additional evidence of entailment; we refer readers to Appendix E for details.

In conclusion, from the entailment detection experiment, we have learnt that: 1) our Chinese entailment graph is strong in the monolingual setting, with contributions from the ORE method and entity typing; 2) a cross-lingual complementarity is clearly shown between Chinese and English entailment graphs, where the effect of ensembles is most significant in the moderate precision range (see Figure 1). We expect that ensembling strong entailment graphs in more languages would lead to further improvements.

7 Evaluation by Question Answering

In addition to the entailment detection evaluation, we further demonstrate the strength of our Chinese entailment graph in application with an extrinsic question answering task, natively in Chinese. Since the existing Chinese QA datasets (Cui et al., 2019; Zheng et al., 2019; Sun et al., 2019) barely concern predicate entailments, we evaluate using a more adversarial boolean QA task following McKenna et al. (2021), inspired by Poon and Domingos (2009).

This task is designed as a boolean variant of *machine reading at scale* (Chen et al., 2017): given a proposition and a pool of context articles, a model attempts to judge whether the queried proposition is true, based on the context pool. Our QA dataset is built off CLUE (Xu et al., 2020), a huge news corpus similar to Webhose[13]. The assumption is, frequently-mentioned predicates between frequently-mentioned arguments are high-quality events to be used as positives; absent predicates between frequently-mentioned arguments are probably not true, and can be used as negatives.

Articles in CLUE corpus are parsed into relation triples as in §3, and partitioned into 3-day time spans to get uniquely identifiable events. Those relation triples whose predicates appear over 30 times in the corpus, and whose argument-pairs appear in over 15 articles in their partitions, are selected as **high-quality positives**. For each positive, we generate negatives designed to be challenging for machines: following McKenna et al. (2021), we replace the positive predicates with their hyponyms / troponyms in Chinese WordNet (Wang and Bond, 2013). Since Chinese predicates are often multi-token and discontinuous, we look for substitutions of spans in predicates rather than entire predicates. If a substituted predicate is absent between this argument-pair in this partition, but exists elsewhere in the corpus with other argument-pairs, we consider it an **adversarial negative**: meaningful, but neither mentioned nor entailed in this context.

We use a sample of 40,000 positives along with their 70,583 adversarial negatives (a similar proportion to Levy-Holt) as our final QA dataset, split evenly into dev / test sets. We have omitted many details in dataset elicitation for the sake of brevity; we share this dataset as part of our release, and refer readers to McKenna et al. (2021) and Appendix H for more details and examples.

In our QA task, the positive / negative triples are concatenated into query propositions, the entire partition of articles in each query's time-span[14] is used as context pool, and a confidence score for each query is produced by each method. As in §6, Precision-Recall curves are drawn, AUC values with >50% precision are reported.

We compare our EG_{Zh} with the DDPORE baseline in §6 as well as BERT baselines in 3 different setups. Note that our QA dataset is monolingual in Chinese so ensembles are not involved. For all methods, the confidence score of a query is defined as its highest score w.r.t. any context sentences.

[13] Articles seen in Webhose are excluded for fairness of the experiment.

[14] Except the sentence hosting the query or its positive.

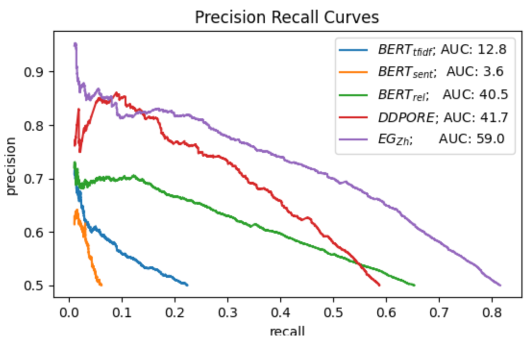


Figure 2: P-R Curves on QA evaluation test set for EG_{Zh} and baselines; AUC values are annotated in the legend.

BERT_{tfidf}: retrieves the top 5 most relevant articles by TF-IDF following (Chen et al., 2017), calculates cosine similarity between each retrieved sentence and the query at [CLS] token;
BERT_{sent}: retrieves the *host-sentences* of the relation triples involving the same arguments as the query, calculates cosine similarity between each host sentence and the query at [CLS] token;
BERT_{rel}: retrieves the *relation triples* involving the same arguments as the query, calculates cosine similarity between each retrieved triple (concatenated into a sentence) and the query at [CLS] token;
EG_{Zh} / DDPORE: retrieve the relation triples involving the same arguments as the query, return entailment scores from each retrieved context triple to the query triple (note that these are directional).

Results are shown in Figure 2. Our EG_{Zh} is again far above all baselines, further stressing the strength of our approach and the necessity of developing entailment graphs for directional inference. BERT_{rel} outperforms the other two BERT baselines, because of its focused context input of concatenated triples, in contrast to the more noisy news sentences for BERT_{tfidf} and BERT_{sent}.

8 Conclusion

We have presented a pipeline for building Chinese entailment graphs. Along the way, we proposed a novel high-recall open relation extraction method, and built a fine-grained entity-typing dataset via label mapping. As our main result, we have shown that: our Chinese entailment graph is comparable with English graphs, where unsupervised BERT baseline did poorly; an ensemble between Chinese and English entailment graphs substantially outperforms both monolinguals, and sets a new SOTA for unsupervised entailment detection. Directions for future work include multilingual EG alignment and alternative predicate disambiguation.

References

Gabor Angeli, Melvin Jose Johnson Premkumar, and Christopher D. Manning. 2015. 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)*, pages 344–354, Beijing, China. Association for Computational Linguistics.

Jonathan Berant, Noga Alon, Ido Dagan, and Jacob Goldberger. 2015. Efficient global learning of entailment graphs. *Computational Linguistics*, 41(2):249–291.

Jonathan Berant, Ido Dagan, and Jacob Goldberger. 2011. Global Learning of Typed Entailment Rules. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 610–619, Portland, Oregon, USA. Association for Computational Linguistics.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to Answer Open-Domain Questions. *arXiv:1704.00051 [cs]*. ArXiv: 1704.00051.

Tongfei Chen, Yunmo Chen, and Benjamin Van Durme. 2020. Hierarchical Entity Typing via Multi-level Learning to Rank. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8465–8475, Online. Association for Computational Linguistics.

Lei Cui, Furu Wei, and Ming Zhou. 2018. Neural Open Information Extraction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 407–413, Melbourne, Australia. Association for Computational Linguistics.

Yiming Cui, Ting Liu, Wanxiang Che, Li Xiao, Zhipeng Chen, Wentao Ma, Shijin Wang, and Guoping Hu. 2019. A Span-Extraction Dataset for Chinese Machine Reading Comprehension. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5883–5889, Hong Kong, China. Association for Computational Linguistics.

Ido Dagan, Lillian Lee, and Fernando C. N. Pereira. 1999. Similarity-Based Models of Word Cooccurrence Probabilities. *Machine Learning*, 34(1):43–69.

Wayne Davis. 2019. Implicature. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, fall 2019 edition. Metaphysics Research Lab, Stanford University.

Oren Etzioni, Anthony Fader, Janara Christensen, Stephen Soderland, and Mausam Mausam. 2011. Open information extraction: the second generation. In *Proceedings of the Twenty-Second international joint conference on Artificial Intelligence - Volume Volume One*, IJCAI'11, pages 3–10, Barcelona, Catalonia, Spain. AAAI Press.

Anthony Fader, Stephen Soderland, and Oren Etzioni. 2011. Identifying Relations for Open Information Extraction. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1535–1545, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Xingyu Fu, Weijia Shi, Xiaodong Yu, Zian Zhao, and Dan Roth. 2020. Design Challenges in Low-resource Cross-lingual Entity Linking. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6418–6432, Online. Association for Computational Linguistics.

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. PPDB: The Paraphrase Database. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 758–764, Atlanta, Georgia. Association for Computational Linguistics.

Maayan Geffet and Ido Dagan. 2005. The Distributional Inclusion Hypotheses and Lexical Entailment. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 107–114, Ann Arbor, Michigan. Association for Computational Linguistics.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation Artifacts in Natural Language Inference Data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Aurélie Herbelot and Mohan Ganesalingam. 2013. Measuring semantic content in distributional vectors. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 440–445, Sofia, Bulgaria. Association for Computational Linguistics.

Xavier Holt. 2019. Probabilistic Models of Relational Implication. *arXiv:1907.12048 [cs, stat]*. ArXiv: 1907.12048.

Mohammad Javad Hosseini, Nathanael Chambers, Siva Reddy, Xavier R. Holt, Shay B. Cohen, Mark Johnson, and Mark Steedman. 2018. Learning Typed Entailment Graphs with Global Soft Constraints. *Transactions of the Association for Computational Linguistics*, 6:703–717.

Mohammad Javad Hosseini, Shay B. Cohen, Mark Johnson, and Mark Steedman. 2019. Duality of Link Prediction and Entailment Graph Induction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4736–4746, Florence, Italy. Association for Computational Linguistics.

Mohammad Javad Hosseini, Shay B. Cohen, Mark Johnson, and Mark Steedman. 2021. Open-Domain Contextual Link Prediction and its Complementarity with Entailment Graphs. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2790–2802, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Shengbin Jia, Shijia E, Maozhen Li, and Yang Xiang. 2018. Chinese Open Relation Extraction and Knowledge Base Establishment. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 17(3):1–22.

Dimitri Kartsaklis and Mehrnoosh Sadrzadeh. 2016. Distributional Inclusion Hypothesis for Tensor-based Composition. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2849–2860, Osaka, Japan. The COLING 2016 Organizing Committee.

Keshav Kolluru, Vaibhav Adlakha, Samarth Aggarwal, Mausam, and Soumen Chakrabarti. 2020. OpenIE6: Iterative Grid Labeling and Coordination Analysis for Open Information Extraction. *arXiv:2010.03147 [cs]*. ArXiv: 2010.03147.

Chin Lee, Hongliang Dai, Yangqiu Song, and Xin Li. 2020. A Chinese Corpus for Fine-grained Entity Typing. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4451–4457, Marseille, France. European Language Resources Association.

Omer Levy and Ido Dagan. 2016. Annotating Relation Inference in Context via Question Answering. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 249–255, Berlin, Germany. Association for Computational Linguistics.

Dekang Lin. 1998. Automatic Retrieval and Clustering of Similar Words. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 2*, pages 768–774, Montreal, Quebec, Canada. Association for Computational Linguistics.

Xiao Ling and Daniel S. Weld. 2012. Fine-grained entity recognition. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*, AAAI'12, pages 94–100, Toronto, Ontario, Canada. AAAI Press.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. 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. Association for Computational Linguistics.

Nick McKenna, Liane Guillou, Mohammad Javad Hosseini, Sander Bijl de Vroe, Mark Johnson, and Mark Steedman. 2021. Multivalent Entailment Graphs for Question Answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10758–10768, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xiaoman Pan, Thamme Gowda, Heng Ji, Jonathan May, and Scott Miller. 2019. Cross-lingual Joint Entity and Word Embedding to Improve Entity Linking and Parallel Sentence Mining. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 56–66, Hong Kong, China. Association for Computational Linguistics.

Hoifung Poon and Pedro Domingos. 2009. Unsupervised Semantic Parsing. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1–10, Singapore. Association for Computational Linguistics.

Likun Qiu and Yue Zhang. 2014. ZORE: A Syntax-based System for Chinese Open Relation Extraction. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1870–1880, Doha, Qatar. Association for Computational Linguistics.

Siva Reddy, Mirella Lapata, and Mark Steedman. 2014. Large-scale Semantic Parsing without Question-Answer Pairs. *Transactions of the Association for Computational Linguistics*, 2:377–392.

Martin Schmitt and Hinrich Schütze. 2021. Language Models for Lexical Inference in Context. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1267–1280, Online. Association for Computational Linguistics.

Sonse Shimaoka, Pontus Stenetorp, Kentaro Inui, and Sebastian Riedel. 2017. Neural Architectures for Fine-grained Entity Type Classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1271–1280, Valencia, Spain. Association for Computational Linguistics.

Gabriel Stanovsky, Julian Michael, Luke Zettlemoyer, and Ido Dagan. 2018. 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)*, pages 885–895, New Orleans, Louisiana. Association for Computational Linguistics.

Kai Sun, Dian Yu, Dong Yu, and Claire Cardie. 2019. Investigating Prior Knowledge for Challenging Chinese Machine Reading Comprehension. *arXiv:1904.09679 [cs]*. ArXiv: 1904.09679.

Idan Szpektor and Ido Dagan. 2008. Learning Entailment Rules for Unary Templates. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 849–856, Manchester, UK. Coling 2008 Organizing Committee.

Daniel Tse and James R. Curran. 2012. The Challenges of Parsing Chinese with Combinatory Categorial Grammar. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 295–304, Montréal, Canada. Association for Computational Linguistics.

Shan Wang and Francis Bond. 2013. Building the Chinese Open Wordnet (COW): Starting from Core Synsets. In *Proceedings of the 11th Workshop on Asian Language Resources*, pages 10–18, Nagoya, Japan. Asian Federation of Natural Language Processing.

Sabine Weber and Mark Steedman. 2019. Construction and Alignment of Multilingual Entailment Graphs for Semantic Inference. pages 77–79.

Julie Weeds and David Weir. 2003. A General Framework for Distributional Similarity. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, pages 81–88.

Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, Yin Tian, Qianqian Dong, Weitang Liu, Bo Shi, Yiming Cui, Junyi Li, Jun Zeng, Rongzhao Wang, Weijian Xie, Yanting Li, Yina Patterson, Zuoyu Tian, Yiwen Zhang, He Zhou, Shaoweihua Liu, Zhe Zhao, Qipeng Zhao, Cong Yue, Xinrui Zhang, Zhengliang Yang, Kyle Richardson, and Zhenzhong Lan. 2020. CLUE: A Chinese Language Understanding Evaluation Benchmark. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4762–4772, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Dani Yogatama, Daniel Gillick, and Nevena Lazic. 2015. Embedding Methods for Fine Grained Entity Type Classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 291–296, Beijing, China. Association for Computational Linguistics.

Shuai Zhang, Lijie Wang, Ke Sun, and Xinyan Xiao. 2020. A Practical Chinese Dependency Parser Based on A Large-scale Dataset. *arXiv:2009.00901 [cs]*. ArXiv: 2009.00901.

Chujie Zheng, Minlie Huang, and Aixin Sun. 2019. ChID: A Large-scale Chinese IDiom Dataset for Cloze Test. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 778–787, Florence, Italy. Association for Computational Linguistics.

A A Brief Summary of Jia et al. (2018)

In Table 4 are the 7 rules from Jia et al. (2018) which they call Dependency Structure Normal Forms. The first rule corresponds to nominal compounds which we elaborated in constructions **D** in §3.1; the second rule corresponds to direct S-V-O relations; the third rule attends to the semantic objects hidden in adjuncts, which are always preverbs in Chinese; the fourth rule subsumes complements of head verbs into the predicate; the fifth rule handles the coordination of subjects, the sixth handles coordination of object, and the seventh handles coordination of predicates. These rules are reflected in our ORE method as well, but for the sake of brevity, only the constructions which have never been covered by previous work are listed in §3.1.

德国 总理 默克尔 。 German Chancellor Merkel . (German, Chancellor, Merkel)
我 看到 你 。 I see you . (I, see, you)
他 在 家 玩 游戏 。 He at home play game . (He, play-game, home)
我 走 到 图书馆 。 I walk to library . (I, walk-to, library)
我 和 你 去 商店 。 I and you go-to shop . (I, go-to, shop) (you, go-to, shop)
我 吃 汉堡 和 薯条 。 I eat burger and chips . (I, eat, burger) (I, eat, chips)
罪犯 击中 、 杀死 了 他 。 Criminal shot, kill -ed him . (criminal, shot, him) (criminal, kill, him)

Table 4: Set of DSNFs from Jia et al. (2018) exemplified. In each box, at top is an example sentence, presented in Chinese and its English metaphrase (inflection ignored); below are the relations they extract.

B Detailed Statistics of the CFIGER dataset

To test our assumption that each ultra-fine-grained type can be unambiguously mapped to a single FIGER type, we inspect the number of FIGER type labels to which each ultra-fine-grained type is mapped through manual labelling. Among the 6273 ultra-fine-grained types in total, 5622 of them are mapped to exactly one FIGER type, another 510 are not mapped to any FIGER types; only 134 ultra-fine-grained types are mapped to 2 FIGER types, and 7 mapped to 3 FIGER types. No ultra-fine-grained types are mapped to more than 3 FIGER types. Therefore, it is safe to say that our no-ambiguity assumption roughly holds.

We further inspected the number of FIGER types each mention is attached with. It turns out that among the 1,913,197 mentions in total, 59,517 of them are mapped to no FIGER types, 1,675,089 of them are mapped to 1 FIGER type, 160,097 are mapped to 2 FIGER types, 16,309 are mapped to 3 FIGER types, 1,952 are mapped to 4 FIGER types, 200 are mapped to 5 FIGER types, and 33 are mapped to 6 FIGER types. No mentions are mapped to more than 6 FIGER types. Note that each mention can be mapped to more than one ultra-fine-grained types from the start, so these numbers are not in contradiction with the above numbers.

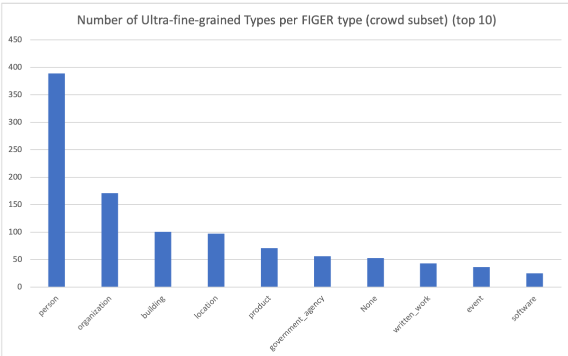


Figure 3: Number of ultra-fine-grained types in crowd-annotated subset mapped to each FIGER type; only the FIGER types with top 10 number of ultra-fine-grained types are displayed.

We also looked at the number of ultra-fine-grained types each FIGER type is mapped to, so as to understand the skewness of our mapping. Results are shown in Figure 3 and 4. Unsurprisingly, the most popular ultra-fine-grained labels are highly correlated with the ones that tend to appear in coarse-grained type sets, with “PERSON” label taking up a large portion. This distribution is largely consistent between crowd-annotated and Wikipedia subsets.

Another set of stats are the number of mentions that corresponds to each FIGER type, shown in Figure 5 and 6. The winners in terms of the number of mentions are consistent with that of the number of

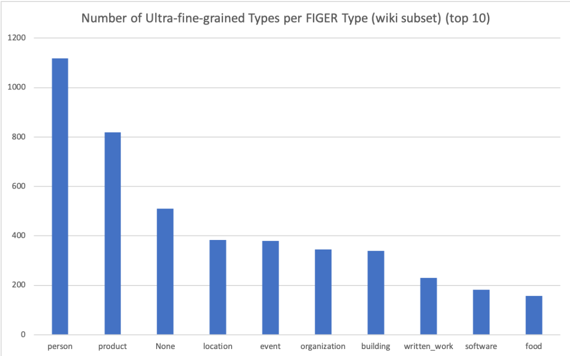


Figure 4: Number of ultra-fine-grained types in wikipedia distantly supervised subset mapped to each FIGER type; only the FIGER types with top 10 number of ultra-fine-grained types are displayed.

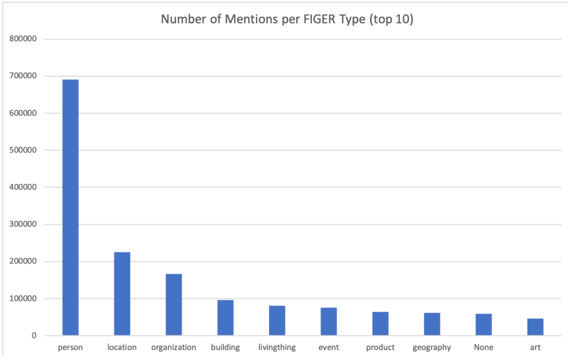


Figure 6: Number of mentions in wikipedia distantly supervised subset labelled as each FIGER type; only the FIGER types with top 10 number of mentions are displayed.

ultra-fine-grained types, and also consistent among themselves (between the two subsets).

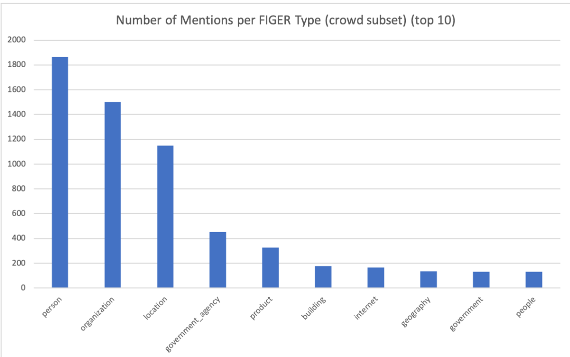


Figure 5: Number of mentions in crowd-annotated subset labelled as each FIGER type; only the FIGER types with top 10 number of mentions are displayed.

C Selecting Relation Triples for Translated Levy-Holt

To retrieve the relation triple most likely reflecting the meaning of the whole sentence, we follow this order when determining which relation triple to select:

- For the amended relations, if the predicate of any of them cover the word with HEAD token in DDParser dependency parse, we randomly choose one of these;
- If none is found, but the predicate of any non-amended relations cover the word with HEAD token in DDParser dependency parse, we randomly choose one of these;
- If none is found, but there are any other relations, we randomly choose one of these;
- Finally, if still none is found, we assign PREMISE_PLACEHOLDER to the premise and HYPOTHESIS_PLACEHOLDER to the hypothesis, so that no entailment relation would ever be detected between them.

D Implementation Details for Entailment Graph Construction

D.1 Corpus and Preprocessing

The original Webhose news corpus that we use consists of 316K news articles. We cut the articles into sentences by punctuations, limiting the maximum sentence length to 500 characters (the maximum sequence length for Chinese Bert). We discard the sentences shorter than 5 characters, and the articles whose sentences are all shorter than 5 characters. After applying the filter, we are left with 313,718 articles, as shown in Table 2.

In the process of entity typing, following previous work, we consider only the first-layer FIGER types; when multiple type labels are outputted, we consider all combinations as valid types for that predicate.

We have also considered using another, larger corpus for building the Chinese entailment graphs, but couldn't finish due to limits on computational resources. We have referred to the larger corpus as the CLUE corpus in §7: the larger corpus is developed by Xu et al. (2020). It is eight times the size of the Webhose corpus and originally intended for training Chinese language models. We provide the typed relation triples extracted from the CLUE corpus as a part of our release, and encourage interested readers to build their own Chinese entailment graph on this larger corpus, as we expect it to exhibit stronger performance, and present an interesting comparison to the language-model driven models pre-trained with the same corpus.

D.2 Entailment Graph Construction

We have used the same entailment graph construction algorithm as Hosseini et al. (2018) to build our Chinese entailment graph from the pool of typed relation triples. When building our entailment graphs, we only feed in the relation triples whose predicate and arguments both appear at least 2 times[15]. Their approach of building entailment graphs comes in two steps, in the paragraphs below we will briefly summarize each step and discuss our implementation details.

The first step is local learning. In this step, instances of relation triples are grouped into clusters based on the arguments they take. Relations (predicates) that are seen with the same arguments of the same types are considered to have co-occurred. For each pair of predicates, based on the co-occurrence information, a few different entailment scores have been proposed, of which the BInc score (Szpektor and Dagan, 2008) was found to have the best empirical performance in (Hosseini et al., 2018). Following them, we also use the BInc score in the local learning step of our Chinese entailment graphs. Note that after the local learning step, the entailment scores between each pair of predicates are independently calculated, and there are no interactions between entailment subgraphs of different type pairs, thus the name **local** learning.

The second step is global learning. In this step, global transitivity constraint is "softly" applied to the local graphs as an optimization problem: paraphrase predicates are encouraged to have the same pattern of entailment; different typed subgraphs are encouraged to have the same entailment score for the same (ignoring type) pair of predicates; finally, the global scores are encouraged to stay similar to the local scores as a measure of regularization. In *Jia* baseline, the local graphs are too weak for global learning to be helpful; in *DDPORE* baseline, the best dev set AUC (as reported in Table 3) is achieved after 2 epochs; in EG_{Zh}, the best dev set AUC is achieved after 3 epochs.

E Case Study for Entailment Detection

In order to further verify the source of our improvements, we analyse our ensemble with a case study: we compare the predictions of our Ensemble_AVG to that of the English monolingual EG_{En}, both thresholded over 65% precision. We categorize the prediction differences into 4 classes: *True Positives*, *False Positives*, *True Negatives*, *False Negatives*. *Positives* are cases where the ensemble switched the prediction label from negative to positive, vice versa for *negatives*; *True* means that the switch is correct, *False*, that the switch is incorrect.

In Table 5, we break down each class of differences according to the direct cause of EG_{Zh} making a different prediction than EG_{En}[16][17]:

- **same sentence after translation**: The premise and hypothesis become identical in relation structure; this can only happen with *positives*;
- **translation error**: The premise or hypothesis becomes unparsable into relations due to translation error; this can only happen with *negatives*;
- **lexicalization**: The difference in predictions is attributed to the cross-lingual difference in the lexicalization of complex relations;
- **ORE error**: After translation, the true relations in premise and hypothesis have the same arguments, but are mistaken due to ORE error;
- **evidence of entailment**: The difference is attributed to the different evidence of entailment in the two graphs; this is most relevant to our EG_{Zh}.

As shown, the majority of our performance gain comes from the additional evidence of entailment in EG_{Zh}; surprisingly, translation played a positive role in the ensemble, though not a major contributor. We attribute this to the fact that MT systems tend to translate semantically similar sentences to the same target sentence, though this similarity is still symmetric, not directional. We have singled out this effect in the "BackTrans Esb" ablation study in §6.3, and have confirmed that this effect is marginal to our success.

In Table 5, for both the differences from evidence of entailment, and differences in TOTAL, the precision of *positives* is lower than that of *negatives*. Namely, $TP/(TP + FP)$ is lower than $TN/(TN + FN)$. This is no surprise, as *positives* and *negatives* have different baselines to start with: *Positives* attempt to correct the false negatives from EG_{En}, where 17% of all negatives are false; *Negatives* attempt to correct the false positives, where 35% of all positives are false (as dictated in the

[15] We experimented with 2-2, 2-3, 3-2 and 3-3, among which this 2-2 setting is empirically favoured.

[16] since the switch in the ensemble is driven by EG_{Zh}.

[17] examples of each class of cause are given in Appendix F.

Direct causes of EG_{Zh}'s different prediction	TP (+)	FP (-)	TN (+)	FN (-)	+/-
translation-related causes, among which:	+52	-28	+42	-47	+19
· *same sentence after translation*	*+52*	*-28*	*0*	*0*	*+24*
· *translation error*	*0*	*0*	*+42*	*-47*	*-5*
lexicalization	+29	-54	+16	-12	-21
ORE error	+8	-20	+8	-5	-9
evidence of entailment	**+109**	**-95**	**+86**	**-40**	**+60**
TOTAL	+198	-197	+152	-104	+49

Table 5: Breakdown of the different predictions between our ensembles and English monolingual graph. "TP", "FP", "TN", "FN" represent *True Positive*, *False Positive*, *True Negative* and *False Negative* respectively; in the column "+/-" is the overall impact of each factor.

setting of our case study). In this context, it is expectable that our evidence of entailment gets $109/(109+95) = 53\%$ correct for *positives*, while a much better $86/(86+40) = 68\%$ correct for *negatives*. These results support the solidarity of our contributions.

F Examples of Different Predictions in Case Study by Category of Direct Cause

In this section, we provide one example for each class of direct cause, as described in the above Appendix E. Chinese sentences and relations in the examples are presented in the same format as §3.1.

Same sentence after translation

- Premise - English: (magnesium sulfate, relieves, headache)
- Hypothesis - English: (magnesium sulfate, alleviates, headaches)
- Premise - Chinese translation: "硫酸镁(magnesium) 缓解(relieves) 头痛(headache)"
- Hypothesis - Chinese translation: "硫酸镁(magnesium) 缓解(alleviates) 头痛(headache)"

The two sentences are translated to the same surface form in Chinese, as the predicates are in many cases synonyms. There are more true positives than false positives, because synonyms are simultaneously more likely true entailments and more likely translated to the same Chinese word.

Translation Error

- Premise - English: (Refuge, was attacked by, terrorists)
- Hypothesis - English: (Terrorists, take, refuge)
- Premise - Chinese translation: "避难所(refuge) 遭到(suffered) 恐怖分子(terrorists) 袭击(attack); Refuge suffered attack from terrorists."
- Hypothesis - Chinese translation: "恐怖分子(terrorists) 避难(take-shelter); Terrorists take shelter."

The hypothesis is supposed to mean "The terrorists took over the refuge". However, with translation, the hypothesis in Chinese is mistaken as an intransitive relation where take-refuge is considered a predicate.

Lexicalization

- Premise - English: (Granada, is located near, mountains)
- Hypothesis - English: (Granada, lies at the foot of, mountains)
- Premise - Chinese translation: "格拉纳达(Granada) 靠近(is-near) 山脉(mountains)"
- Hypothesis - Chinese translation: "格拉纳达(Granada) 位于(is-located-at) 山脚下(hillfoot)"

When the hypothesis is translated into Chinese, the lexicalization of the relation changed, the part of the predicate hosting the meaning of 'the foot of' is absorbed into the object. Therefore, while in English "is located near" does not entail "lies at the foot of", in Chinese "is-near" is considered to entail "is-located-at". In this way, an instance of *false positive* comes into being.

ORE Error

- Premise - English: (A crow, can eat, a fish)
- Hypothesis - English: (A crow, feeds on, fish)
- Premise - Chinese translation: "乌鸦(crow) 可以(can) 吃(eat) 鱼(fish)"
- Hypothesis - Chinese translation: "乌鸦(crow) 以(take) 鱼(fish) 为(as) 食(food)"

- Premise - extracted Chinese relation: (crow, eat, fish)
- Hypothesis - extracted Chinese relation: (crow, take·X·as·food, fish)

While the translations for this pair of relations is correct, in the subsequent Chinese open relation extraction, our ORE method failed to recognize "可以(can)" as an important part of the predicate. To avoid sparsity, most adjuncts of the head verb are discarded, and modals are part of them. While the original premise "can eat" does not entail "feeds on", the Chinese premise "eat" does in a way entail "feeds on", where another instance of *false positive* arises.

Evidence of Entailment

- Premise - English: (quinine, cures, malaria)
- Hypothesis - English: (quinine, is used for the treatment of, malaria)
- Premise - Chinese translation: "奎宁(quinine) 治疗(cure) 疟疾(malaria)"
- Hypothesis - Chinese translation: "奎宁(quinine) 用于(is-used-to) 治疗(cure) 疟疾(malaria)"
- Premise - extracted Chinese relation: (quinine, cure, malaria)
- Hypothesis - extracted Chinese relation: (quinine, is-used-to·cure, malaria)

In the above example, sufficiently strong evidence for "cure" entailing "is used for the treatment of" is not found in the English graph, whereas strong evidence for "治疗(cure)" entailing "用于·治疗(is-used-to·cure)" is found in the Chinese graph. In this way we get an instance of *true positive*.

G More Precision-Recall Curves

In this section, we present more precision-recall curves from the baselines and ablation studies in Table 3. These curves contain more details explaining the AUC values in the table.

Figure 7 contains the curves for the ablation study of DataConcat. Here all three models ultimately come from the same corpus, so the performance difference can be fully attributed to the cross-lingual complementarity of entailment graphs.

Figure 8 contains the curves for two sets of ablation studies: EG_{Zh} with or without entity typing; EG_{En} ensembled with back-translation predictions

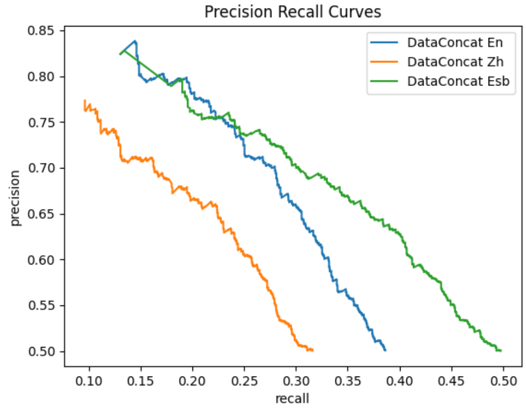


Figure 7: P-R Curves on Levy-Holt test set for DataConcat ablation study.

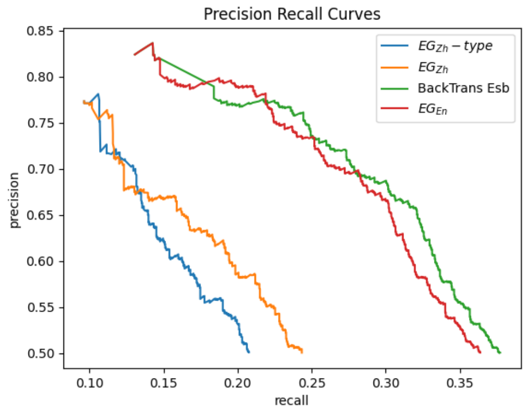


Figure 8: P-R Curves on Levy-Holt test set for EG_{Zh} $-type$, BackTrans Esb, in comparison to EG_{Zh} and EG_{En} respectively.

or not. The former study shows the clear benefit of our entity typing system, while the latter study shows that ensembling with back-translated predictions only results in a marginal gain, therefore the synonym effect from translation is not a major contributor to the success of our ensembling method.

H Implementation Details for Our Question Answering Evaluation

Our Chinese boolean QA dataset is constructed from the CLUE Chinese news corpus (Xu et al., 2020), a huge Chinese news corpus of 2.4M news articles. The CLUE corpus has 8 times the number of articles as the Webhose corpus[18].

We partition the corpus into 122 disjoint 3-day time spans. We look for frequent predicates between frequent typed-argument-pairs in each partition. Since we want the typed-argument-pairs that

[18] the one from which we built our Chinese entailment graphs.

are reported by multiple sources, we look for typed-argument-pairs that appear in at least 15 articles and with at least 15 predicates within a partition; for predicates, we just want to make sure they are felicitous, so we look for those predicates that appear at least 30 times in an arbitrary number of articles anywhere in the corpus. The motivation for these thresholds is the following trade-off: lower thresholds lead to noisier datasets; higher thresholds lead to more biased datasets.

Relation triples satisfying the above criteria are reformatted into textual propositions, and selected as "positives": these predicates are frequently-mentioned, so they are felicitous relations; these argument-pairs are mentioned in many articles within the time-span, so the "positive" predicates should be inferrable from the other mentions of these argument-pairs in the time-span. In order to balance the dataset, at most one positive is chosen from each sentence.

One naive approach to generating negatives, is to substitute random predicates into the positive propositions. However, that is not adversarial enough as a test for directional inference: unrelated words can be easily detected by symmetric similarity measures like Bert similarities, without involving any wisdom in directionality.

Following (McKenna et al., 2021), we replace the positive predicates with their hyponyms / troponyms in Chinese WordNet (Wang and Bond, 2013). These replacements are semantically related to the positives, but do not logically follow the positives. We select those replacements that are absent with the argument pair of its corresponding positive in the corresponding partition. We also require that the replacements appear elsewhere in the corpus at least 5 times. Based on the Gricean cooperative principle of communication (Davis, 2019), we assume that the collection of news articles would report all and only the facts that are known. It is then implied, that these selected replacements are felicitous predicates, but untrue (or, not confirmed to be true) in the contexts of their corresponding positives. Thus, they can be used as adversarial negatives. As discussed in §7, we look for substitution of spans rather than entire predicates, to deal with the multi-token and discontinuous feature of Chinese predicates; to balance the dataset, at most two negatives are chosen from each sentence; in order for quality control, only those positives from which some negatives can be generated are kept.

For instance, for a positive proposition "约翰(John) 在(at) 乐购(Tesco) 购物(shop); John shopped at Tesco", the predicate in this positive is "在·X·购物 (at·X·shop)". By replacing random predicates, for example "起诉(sue)" into the positive proposition, we would get negatives like "约翰(John) 起诉(sue) 乐购(Tesco); John sued Tesco", which is irrelevant to the positive and easy to guess for symmetric measures. On the other hand, using Chinese WordNet, for the subspan "购物(shop)" in the positive predicate, we can find a troponym "买日用品(go marketing)", thus we can get negatives like "约翰(John) 在(at) 乐购(Tesco) 买日用品(go marketting); John went marketing at Tesco". This replacement is much more semantically related than the random one, and, unless otherwise mentioned in the context, it can be assumed that **we don't know John went marketing at Tesco**. Therefore, this latter replacement is still a negative, and a much more challenging one.

Notably, both for this evaluation and for the entailment detection evaluation in §6, we use the actual arguments in the sentences for BERT similarities, not the types of the arguments. This is because we empirically find that by replacing the actual arguments with their types, the language models get confused, and their performances drop.

As our QA task is a *machine reading at scale* task, each method uses the entire partition of news articles as context, and produces a score of whether the queried proposition is true according to the context. Method design choices can be categorized into two dimensions: how to retrieve the relevant context from the context pool, and how to calculate a truthfulness score for the query based on the retrieved relevant context.

Along the first dimension, our Bert$_{tfidf}$ baseline uses TF-IDF matching to retrieve the relevant articles, and use the sentences in these articles as relevant context; the other 4 methods in our experiment use exact match of argument-pairs to identify the "related relation triples", among them, BERT$_{sent}$ baseline retrieves the host-sentences of these relation triples, while Bert$_{rel}$, EG$_{Zh}$ and DDPORE use these relation triples themselves.

Along the second dimension, all methods take each context sentence or relation individually, and calculate the score as "whether the query proposition can be inferred from any context sentence/relation retrieved". The three BERT baselines calculate the cosine similarity between the

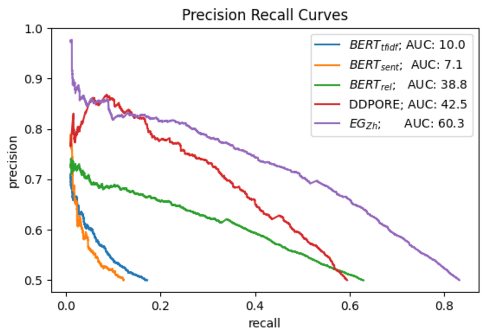


Figure 9: P-R Curves on QA evaluation **dev** set for EG_{Zh} and baselines; AUC values are annotated in the legend.

QA eval AUC (%)	dev	test
BERT_{tfidf}	10.0	12.8
BERT_{sent}	7.1	3.6
BERT_{rel}	38.8	40.5
DDPORE	42.5	41.7
EG_{Zh}	**60.3**	**59.0**

Table 6: Area Under Curve (AUC) on QA evaluation, for Chinese entailment graph (EG_{Zh}) and its baselines.

BERT representations of each context sentence and the query at the [CLS] token; the DDPORE baseline and our EG_{Zh} retrieve the entailment scores from each context triple to the query triple, from the corresponding typed entailment sub-graphs.

In addition to the test set Precision-Recall curves reported in §7, we present the dev set Precision-Recall curves as well, in Figure 9; we also summarize the AUC values in Table 6. The observations here are consistent with the conclusions in §7. It is to be noticed, that we did not do any hyper-parameter tuning on the dev set, the best settings in Levy-Holt dev set are directly applied here. Nevertheless, we present this dev set along with the test set, to form a complete dataset on the task of Chinese boolean machine reading at scale, which we advocate as a solid benchmark for directional inference.

I Manual Examination of Chinese Levy-Holt

In order to provide a quantified evaluation for the quality of our Chinese Levy-Holt dataset from a human perspective, we manually labelled 100 proposition pairs in the Chinese Levy-Holt dev set (1-29, 1124-1136, 2031-2059, 3091-3122, 4061-4089, excluding the entries which are not parsed back into binary relation triples).

In this evaluation, we aim to answer the question of "how accurate is the translate-then-parse procedure when it claims to have successfully converted an evaluation entry". We label each Chinese entry along two dimensions: semantic consistency, whether it has preserved the meaning of the English entry; label consistency, whether the entailment label remains correct.

Along the first dimension of semantic consistency, we summarize our findings as follows:

- Correct: 74/100. These are the Chinese entries whose Chinese **predicates** precisely reflects the meaning of the English entry[19];
- Metaphors: 3/100. These are the cases where the English entry involves metaphorical word-senses of predicates, but such metaphorical senses of these words are infelicitous in Chinese context;
- Adjuncts: 9/100. These are the cases where a part of an English predicate is translated into an adjunct to the Chinese head-verb, and is not included in the Chinese predicate (as in the example for ORE Errors in Appendix F); examples of missed-out adjuncts are 'widely', 'should' and 'may';
- Lexical: 5/100. These are the cases where the word-segmentation of the Chinese sentence is incorrect (as Chinese sentences come with no spaces between words);
- Errors: 7/100. These are the cases where, although the Chinese ORE method outputs some binary relation triples for the translation, that relation triple is not the true relation for the sentence;
- Translation: 2/100. These are the cases where, although the translation can be parsed into some binary relation triples by our Chinese ORE method, the translation is incorrect, thus everything downstream is wrong.

Along the second dimension of label consistency, we find that: in 89 / 100 entries, the actual labels in Chinese are consistent with the English labels; in 10 / 100 entries, the actual labels in Chinese are inconsistent with the English labels; in the remaining 1 / 100 entry, the actual label in Chinese is consistent with the actual label in English, but the provided English label is corrupted.

In summary, for the portion where the conversion is successful, the entries in Chinese Levy-Holt

[19] Arguments are allowed to be translated to different senses of the words, as long as the entailment label between the predicates is not affected.

preserves the meaning of the English entries reasonably well; more importantly, the labels of the Chinese Levy-Holt dataset remains robust.

J Diagram Illustrations of Our Syntactic Analysis

In this section, we present for interested readers a set of diagram illustrations of the set of constructions, as involved in our syntactic analysis in §3.1. For each construction, we draw a diagram to illustrate its dependency structure, an example to instantiate the dependency structure, and in the following lines, all the relations that we extract from this construction (one relation per line). Each relation comes in the form of triple-of-types (consistent with the diagram) and triple-of-words (as in the example), separated by semi-colons. The diagrams are presented in Table 8, Table 9 and Table 10.

K Ethics Considerations

Below we discuss the ethics considerations in our work.

The limitation to our work is two-fold. Firstly, our Chinese entailment graphs focus on the task of predicate entailment detection, and does not attempt to independently solve the more general problem of reasoning and inference: this more general task would also involve other resources including argument hypernymy detection, quantifier identification and co-reference resolution. These are out of the scope of this work. Secondly, while we have shown the effect of cross-lingual complementarity, adding in more languages to the ensemble is not directly straight-forward: this would require linguistic expertise and NLP infrastructure in the respective languages; including more languages, and eventually including arbitrary languages, is one of the directions for our future work.

The risk of our work mostly stems from our use of large-scale news corpora: if the media coverage itself is biased toward certain aspects of the world or certain groups of people, then these biases would be inherited by our entailment graphs. Our response to this is to include as many diverse news sources as possible to reduce such biases to the minimum: our source corpus for building Chinese entailment graphs includes 133 different news sources from a variety of countries and regions.

For the computational cost of building Chinese entailment graphs, the algorithm for open relation extraction takes roughly 140 CPU hours to process the entirety of Webhose corpus; the entity typing model takes roughly 180 GPU hours on NVidia 1080Ti GPUs to do inference on the entirety of Webhose corpus; the local learning process takes less than one hour, and, the global learning process, our major computational bottleneck, takes roughly 800 CPU hours to finish.

The major datasets of use, namely, Webhose corpus, CLUE dataset and the CFET dataset, are open corpora with no specified licenses, thus our academic use is allowed; no license was specified for the Levy-Holt dataset as well; our own CFIGER dataset as well as the constructed entailment graphs can be distributed under the MIT license.

Stats	Webhose	Levy-Holt
AVG sentence length (in # of Chinese characters)	24.9	10.1
AVG # of relations per sentence	15.6	2.72
Percentage of relations from our additional patterns in §3.1	48%	32%

Table 7: Some key statistics of Webhose corpus and Chinese Levy-Holt dataset.

L Comparison Between Webhose Corpus and Levy-Holt Dataset

In this section, we report some key statistics of the Webhose corpus in comparison to the Levy-Holt dataset, which highlight their difference in genre.

As shown in Table 7, the Webhose corpus has much longer sentences than the Chinese Levy-Holt dataset, and on average, a much larger number of open relations can be extracted from the sentences in Webhose corpus. More importantly, the relation patterns which we additionally identified in §3.1 are much better represented (constituting 48% of all relations) than in Chinese Levy-Holt (32%). Thus, it is clear that: 1) our ORE method in §3 was not tuned on the test data, namely Chinese Levy-Holt; 2) tuning on Chinese Levy-Holt would not help with building better ORE methods for news corpora. On the other hand, as a large-scale multi-source news corpus of 5 million sentences, Webhose corpus can be believed to accurately reflect the distribution of linguistic patterns in the entirety of the news genre.

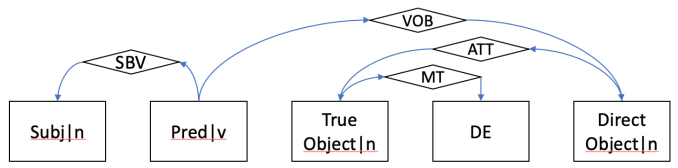
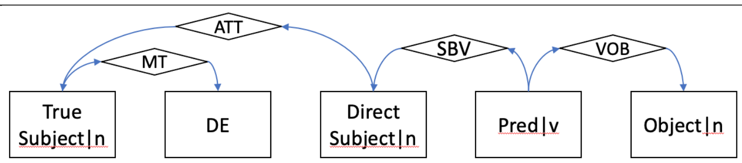
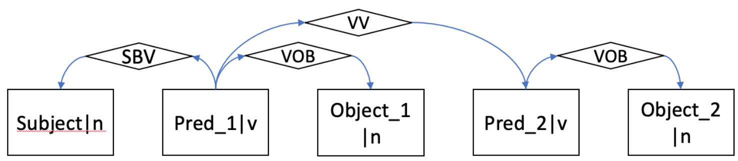
Construction ID	Diagrams and Examples
A.1	
	Example: *"咽炎(pharyngitis) 成为(becomes) 发热(fever) 的(De) 原因(cause); Pharyngitis becomes the cause of fever"*
	Relation 1: (Subj, Pred, Direct_Object); (咽炎(pharyngitis), 成为(becomes), 原因(cause))
	Relation 2: (Subj, Pred·X·DE·Direct_Object, True_Object); (咽炎(pharyngitis), 成为·X·的·原因(becomes·X·DE·cause), 发烧(fever))
A.2	
	Example: *"苹果(Apple) 的(De) 创始人(founder) 是(is) 乔布斯(Jobs); The founder of Apple is Jobs"*
	Relation 1: (Direct_Subject, Pred, Object); (创始人(founder), 是(is), 乔布斯(Jobs))
	Relation 2: (True_Subject, Direct_Subject·Pred, Object); (苹果(Apple), 创始人·是(founder·is), 乔布斯(Jobs))
B.1	
	Example: *"我(I) 去(go-to) 诊所(clinic) 打(take) 疫苗(vaccine); I go to the clinic to take the vaccine"*
	Relation 1: (Subject, Pred_1, Object_1); (我(I), 去(go-to), 诊所(clinic))
	Relation 2: (Subject, Pred_2, Object_2); (我(I), 打(take), 疫苗(vaccine))

Table 8: The syntactic analysis in §3.1 illustrated with diagrams, examples and their extracted relations.

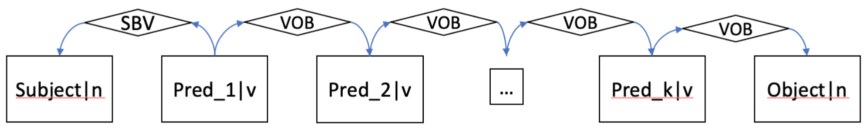
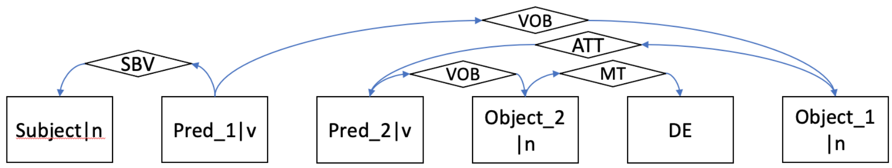
Construction ID	Diagrams and Examples
B.2	
	Example: *"我(I) 想(want) 试图(try) 开始(begin) 写(write) 一个(a) 剧本(play); I want to try to begin to write a play"*
	Relation 1: (Subject, Pred_1, Pred_2); (我(I), 想(want-to), 试图(try))
	Relation 2: (Subject, Pred_1·Pred_2, Pred_3); (我(I), 想·试图(want-to·try), 开始(begin))
	......
	Relation K: (Subject, Pred_1·...·Pred_K, Object); (我(I), 想·试图·开始·写(want-to·try·begin·write), 一个剧本(A play))
C	
	Example: *"他(he) 解决(solve) 了(-ed) 困扰(puzzle) 大家(everyone) 的(De) 问题(problem); He solved the problem that puzzled everyone"*
	Relation 1: (Subject, Pred_1, Object_1); (他(He), 解决(solved), 问题(problem))
	Relation 2: (Object_1, Pred_2, Object_2); (问题(Problem), 困扰(puzzled), 大家(everyone))
D	Analysis in construction D removes the infelicitous instances of the **Nominal Compound** construction; for the illustration of this construction, we refer readers to Jia et al. (2018) and do not repeat here.

Table 9: More syntactic analysis in §3.1 illustrated with diagrams, examples and their extracted relations.

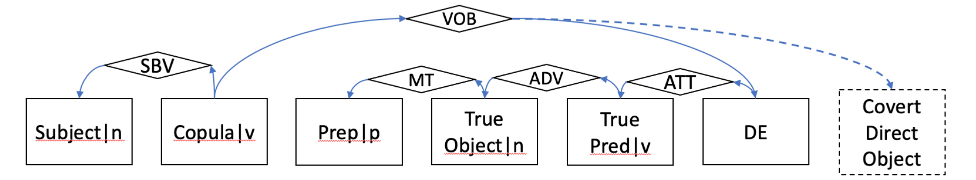
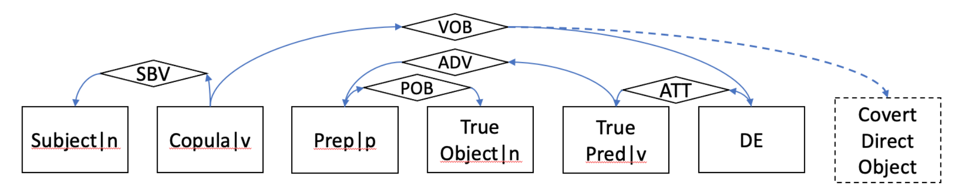
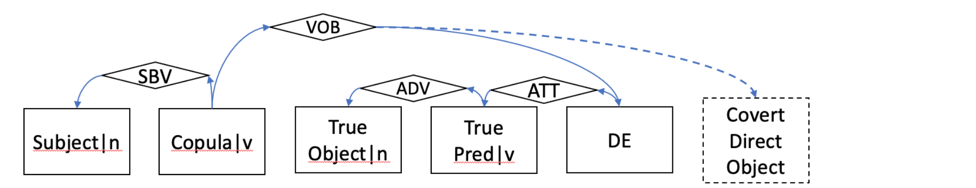
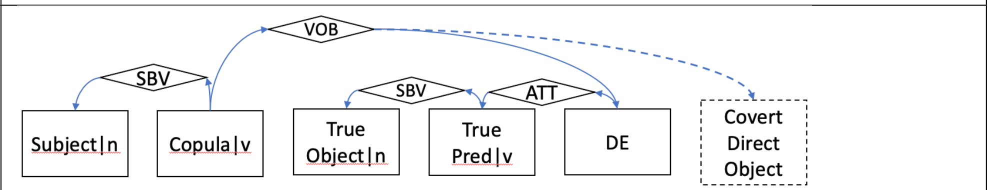
<table>
<tr><th>Construction ID</th><th>Diagrams and Examples</th></tr>
<tr><td rowspan="3">E.1</td><td></td></tr>
<tr><td>Example: “玉米(Corn) 是(is) 从(from) 美国(US) 引进(introduce) 的(De); Corn is introduced from US”</td></tr>
<tr><td>Relation 1: (Subject, Copula·Prep·X·True_Pred·DE, True_Object); (玉米(Corn), 是·从·X·引进·的(is·from·X·introduced·DE), 美国(US))</td></tr>
<tr><td rowspan="3">E.2</td><td></td></tr>
<tr><td>Example: “设备(device) 是(is) 用(from) 木头(wood) 做(make) 的(De); The device is made of wood”</td></tr>
<tr><td>Relation 1: (Subject, Copula·Prep·X·True_Pred·DE, True_Object); (设备(device), 是·用·X·做·的(is·from·X·made), 木头(wood))</td></tr>
<tr><td rowspan="3">E.3</td><td></td></tr>
<tr><td>Example: “设备(device) 是(is) 木头(wood) 做(make) 的(De); The device is made of wood”</td></tr>
<tr><td>Relation 1: (Subject, Copula·X·True_Pred·DE, True_Object); (设备(device), 是·X·做·的(is·X·made), 木头(wood))</td></tr>
<tr><td rowspan="3">E.4</td><td></td></tr>
<tr><td>Example: “设备(device) 是(is) 木匠(carpenter) 做(make) 的(De); The device is made by a carpenter”</td></tr>
<tr><td>Relation 1: (Subject, Copula·X·True_Pred·DE, True_Object); (设备(device), 是·X·做·的(is·X·made·DE), 木匠(carpenter))</td></tr>
</table>

Table 10: Yet more syntactic analysis in §3.1 illustrated with diagrams, examples and their extracted relations.