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. In this pipeline, we present a novel high-recall open relation extraction (ORE) method and the first Chinese fine-grained entity typing dataset following the FIGER type ontology. Through experiments on the popular Levy-Holt dataset, translated into Chinese, we show that our Chinese entailment graph outperforms a range of strong baselines by large margins. Moreover, an ensemble of Chinese and English entailment graphs sets a new unsupervised SOTA on the original Levy-Holt dataset, surpassing previous SOTA by more than 4 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** holding 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 confirm 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; Hosseini, 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 building entailment graphs has been limited to English, building entailment graphs for other languages is an interesting and challenging goal. The importance is two-fold: for that language, a native entailment graph would facilitate NLU in it; from a multi-lingual perspective, entailment graphs in different languages host complementary information, and the different polyseme mappings are helpful for disambiguation. Thus, entailment graphs in multiple languages open up many possibilities for cross-lingual alignment, as we will showcase with a simple ensemble.

In this paper, we propose a pipeline for building entailment graphs in Chinese, as it is distant enough from English to be rich in complementary information, meanwhile relatively high-resource so that reliable tools can be found. Though being relatively high-resource, building entailment graphs in Chinese is still filled with challenges, where the two toughest ones are open relation extraction (ORE) and fine-grained entity typing (FET).

ORE is crucial for entailment detection, identifying the predicates-argument pairs in sentences. It has been solved with either rule-based methods over syntactic parsers (Fader et al., 2011; Etzioni et al., 2011; Angeli et al., 2015), or neural sequence labellers distantly-supervised by rule-based methods (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 have a higher degree of ambiguity and many linguistic indicators of relations are poorly represented. Previous work on Chinese ORE has resorted to a conservative approach (Qiu

[1] Our codes and data will be released on Github.

and Zhang, 2014; Jia et al., 2018), failing to identify many constructions relevant to relation extraction. In this paper, we propose a novel dependency-based ORE method which, to our best empirical observation, provides a comprehensive account for constructions where relations are involved.

The other challenge, regarding FET, lies mainly in the lack of a suitable dataset over a suitable type ontology for predicate word-sense disambiguation: too coarse a type set would be insufficient for disambiguation, while too granular a type set would result in disastrous sparsity in the entailment graph. We follow Hosseini et al. (2018) on using the popular FIGER type set (Ling and Weld, 2012), and elicit a Chinese FET dataset with FIGER labels via label mapping. Entity typing models built on this dataset are proven to be satisfactory in performance and helpful for predicate disambiguation.

With these challenges solved, we build strong Chinese entailment graphs. Evaluation on the Levy-Holt dataset (Levy and Dagan, 2016; Holt, 2019) (through translation) shows, that our Chinese entailment graph outperforms baselines by large margins, and is comparable with the English graph. By ensembling the prediction scores from English and Chinese graphs, we show a clear advantage over both monolingual graphs, and sets a new SOTA.

Our contribution is 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 cooccurrence-based scores for entailment detection; Berant et al. (2011) proposed to "globalize" the typed entailment graphs by closing them with transitivity constraint; Hosseini et al. (2018) proposed a more scalable global learning approach using soft transitivity constraints; Hosseini et al. (2019); Hosseini (2021) further exploited the duality between entailment graph construction and link prediction to refine the entailment scores.

Our work is closely related to Hosseini et al. (2018), with a few key adaptations for Chinese. First, while they used a CCG parser (Reddy et al., 2014) for ORE, our ORE method is based on dependency parser (Zhang et al., 2020); second, while they typed the entities by linking them to Wikipedia entries, 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 work (Qiu and Zhang, 2014; Jia et al., 2018; Zhang et al., 2020) has built Chinese ORE algorithms from dependency parsers, but their parsers omit many common constructions essential to ORE. In Section 3, we present the most comprehensive Chinese ORE method so far.

Linking-based entity-typing methods can be more accurate than neural entity typing, since the type labels are exact as long as linking is correct. Unfortunately, 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 (FET) has developed rapidly, where Yogatama et al. (2015); Shimaoka et al. (2017); Chen et al. (2020) proposed various methods, sharing a common interest in the FIGER dataset. Lee et al. (2020) built a Chinese ultra-fine-grained entity typing dataset through distant supervision. Based on their dataset, we are able to build our CFIGER dataset by label mapping.

As a relevant task, Ganitkevitch and Callison-Burch (2014) created a multi-lingual database for symmetric paraphrases, in contrast, entailment graphs host directional entailment relations. More recently, Schmitt and Schütze (2021) proposed to fine-tune language models on predicate entailment datasets via handcrafted prompts. In contrast to entailment graph construction, this is a supervised approach, which carries the danger of overfitting to dataset artifacts (Gururangan et al., 2018).

Another related strand of research, exemplified by 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 where external knowledge of predicate entailment is required. Entailment graphs, on the other hand, are focused on providing 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 its 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**: the less felicitous relations occur less frequently, and are less likely to take part in entailments when they do occur, thus they are negligible. As will be shown through Table 3, this approach significantly outperforms previous SOTA on supporting entailment graph construction[2].

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 based on dependency parsing. They defined a set of rules to extract relations from dependency labels, which they call dependency semantic normal forms (DSNFs). We refer readers to Appendix A for a brief summary of their DSNFs.

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. In particular, we highlight 5 important constructions which we additionally identify, explained with examples.

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 roughly seen as 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 predicate "becomes the cause of" is expressed as "becomes·X·De·cause". The direct relation here is "Pharyngitis, becomes, cause", but we *additionally* extract the more informative 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[3].

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[4].

B. Bounded Dependencies In Chinese, bounded dependencies, particularly control structures, are expressed with a covert element of Chomskyan category **T** (typically "to"). We capture the following phenomena in addition to direct relations:

- Sequences of VPs: for sentences such as *"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 indicator of connection. Here we 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, and this time, all verbs but the first one bear a "VOB" dependency label, as the direct object to its antecedent. 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 phenomena are different from conjunction constructions in Table 5: the sequences of events here involve subordination (control) rather than coordination, thus needs a separate rule.

C. Relative Clauses Relative Clauses also take the form of modification structures in Chinese, for

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

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

[4]The legitimacy of such relations depend on the frequency of the verb co-occurring with these direct arguments. Relations with less frequent combinations are less felicitous. However, as in line with the Relation Frequency Assumption, less felicitous relations are also less statistically significant.

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 a noun phrase (NP) has two consecutive "ATT" modifiers. For example, 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 such as **(Germany, Chancellor, Merkel)** for these NPs.

However, they overlooked the fact that prepositional phrases (PPs) in Chinese with omitted "De" take exactly the same form (see constructions **A**). For instance, in NPs with PP modifiers such as *"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: when all three 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 reliably a relation **(Merkel, is·X·De·Chancellor, Germany)**. Otherwise, the NP rarely contains legitimate relations.

E. Copula with Covert Objects Copula are 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 should be 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. At times we depend on Part-of-Speech labels to assist our judgment. We use Stanford CoreNLP (Manning et al., 2014) POS tagger for this purpose. We detect negations by looking for matches of negation keywords in the adjunct modifiers of predicates. We handle negations at the lexical level: for predicates with an odd number of negation matches, we insert a negation indicator, treating them as separate predicates from the non-negated ones.

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.

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 the FIGER type ontology (Ling and Weld, 2012), a commonly used type set: we re-annotate the CFET dataset with the FIGER types through label mapping. Given that there are around 6,000 ultra-fine-grained types and only 112 FIGER types (49 for the first layer), we can reasonably assume that each ultra-fine-grained type can be unambiguously mapped to a single FIGER type. Based on this assumption, we manually create an injective mapping between the two, and obtain a re-annotated CFET dataset, the first in Chinese under the FIGER type ontology. We call the re-annotated dataset **CFIGER**. 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[5].

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

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 for CFET dataset; *HierType* (Chen et al., 2020), a SOTA English entity typing model.

Results are shown in Table 1: we observe that the F1 score for *HierType* model is slightly lower on CFIGER dataset than on FIGER dataset in English; on the other hand, thanks to fewer type labels, *CFET* baseline model sees an increase in F1 score on CFIGER dataset, bringing it on par with the more sophisticated *HierType* model. This indicates that our CFIGER dataset is valid for the Chinese fine-grained entity typing task, 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 dataset[6], a multi-source news corpus of 316K articles, crawled from 133 news websites in October 2016. Similarly to the NewsSpike corpus used in Hosseini et al. (2018, 2019); Hosseini (2021), the Webhose corpus contains non-fiction text from multiple sources in a short period of time. This means it is also rich in reliable and diverse relation triples over a focused set of events, which is ideal for mining entailment relations.

We cut the articles into sentences by punctuations, limiting the maximum sentence length to 500. We discard the sentences shorter than 5 characters, and the articles whose sentences are all discarded. In the end, we are left with 313,718 articles, summing up to a total of 5,065,686 sentences.

We get the POS tags with CoreNLP, then feed the articles and POS tags into our ORE method in Section 3, to extract the corresponding open relations. Then, with the *HierType* model (Chen et al., 2020) on CFIGER, we type all arguments of the extracted relations. Following previous work, we consider only the first-layer FIGER types; we type each predicate with the type-pairs of its subject and object, such as "person-event" or "food-law". When multiple type labels are outputted, we consider all combinations as valid types for that predicate.

	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
|subgraph| > 100	442	115
|subgraph| > 1,000	149	27
|subgraph| > 10,000	26	7

Table 2: Statistics of our Chinese entailment graphs (EG_{Zh}) in comparison to English entailment graphs from Hosseini et al. (2018) (EG_{En}). $|\cdot|$ denotes the number of predicates in a subgraph.

We finally employ the entailment graph construction method in Hosseini et al. (2018), taking in only binary relation triples. We only feed in the relation triples whose predicate and arguments both appear at least 2 times[7]. Resultingly, we obtain a Chinese entailment graph of comparable size to the English graph, with detailed statistics shown in Table 2.

6 Evaluation

Due to the lack of Chinese predicate entailment datasets, we are forced to use the English entailment detection task for evaluation via machine translation: we translate English premise-hypothesis pairs into Chinese, then retrieve entailment scores from Chinese entailment graph as predictions for those pairs. We are painfully aware that translation adds noise, and will return to this point below.

Our experiments are based on Levy-Holt dataset (Levy and Dagan, 2016; Holt, 2019), with the same dev/test configuration as Hosseini et al. (2018). 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 translate Levy-Holt dataset, we concatenate each relation triple into a pseudo-sentence, then use Google Translate to translate the pseudo-sentences into Chinese. For each translated pseudo-sentence, we parse it back into Chinese relation triples, again with our ORE method in Section 3. If multiple relations are returned, to retrieve the most representative relations, we consider only those relations

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

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

whose predicate covers the HEAD word.[8]

To type the translated relation triples, we again use *HierType* model to type their *arguments*. The premise and hypothesis need to take the same types, so we take the intersection of their possible types unless it is empty, in which case we take the union.

These types are used as *predicate types* to specify which typed entailment subgraphs to search when scoring the entailment from premise to hypothesis. When both predicates are found in the right order in the relevant subgraph, we retrieve the entailment score between them. When scores are found in multiple subgraphs, we take their maximum.

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

- *BERT*: We take the raw translations of the pseudo-sentence pairs, and compute the cosine similarity between the pretrained BERT sentence representations of premise and hypothesis, at the [CLS] token. This is a strong associative meaning baseline but symmetric;
- *Jia*: We build entailment graph in the same way as Section 5, but with the more restricted ORE method by Jia et al. (2018); accordingly, Jia et al. (2018) method is also used in evaluation;
- *DDPORE*: Similar to *Jia* baseline, but with the baseline ORE method in DDParser (2020).

Ensembling with English Entailment Graphs In order to examine the complementarity between our Chinese entailment graph and the English graph, we experiment on ensembling the scores from two graphs, $pred_{en}$ and $pred_{zh}$. We take the English graph from Hosseini et al. (2018), and evaluate four ensemble strategies: lexicographic orders from English to Chinese and Chinese to English, max pooling and average pooling:

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

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} (2018), we also ensembled our entailment graph with the SOTA English graph EG_{En}++ (2021). We call the later ones **Ensemble++** here and below.

[8] See Appendix C for more details.

AUC (%)	dev	test
BERT ⋆	5.5	3.2
Jia (2018) ⋆	0.9	2.4
DDPORE (2020) ⋆	9.8	5.9
EG_{Zh} ⋆	**16.1**	**9.1**
EG_{En} (2018) ◇	20.7	16.5
EG_{En}++ (2021) ◇	23.3	19.5
Ensemble En_Zh ◇	**27.9** (γ : 0.5)	**20.8**
Ensemble Zh_En ◇	**27.5** (γ : 0.9)	**21.0**
Ensemble MAX ◇	**29.8** (γ : 0.5)	**21.6** †
Ensemble AVG ◇	**29.8** (γ : 0.3)	**21.7**
Ensemble++ AVG ◇	**31.2** (γ : 0.1)	**24.0** †
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 (AUC) values for Chinese entailment graph (EG_{Zh}), its baselines, ensembles with English graphs, and ablation studies. EG_{En} is the English graph in (Hosseini et al., 2018); EG_{En}++ is the English graph in (Hosseini, 2021). For entries with ⋆, the minimum recall is set by Chinese lemma baseline; for entries with ◇, the minimum recall is set by English lemma baseline; entries with † are the best ensemble strategies according to dev set results.

7 Results and Discussions

To measure the performance of our 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.

For our Chinese entailment graph (EG_{Zh}) and its baselines, we report their AUC calculated with minimum recall set by Chinese lemma baseline. For ensembled models, in order to get commensurable AUC values with previous work instead of being

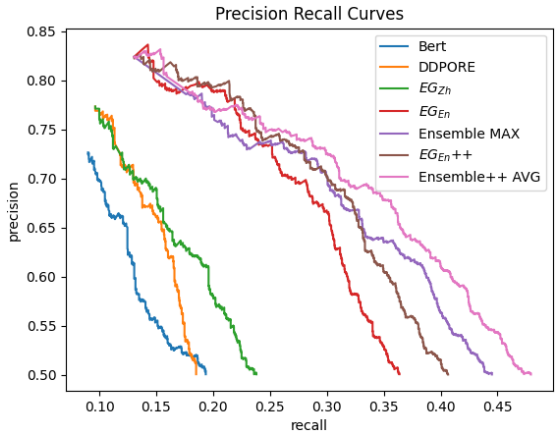

Figure 1: P-R Curves on Levy-Holt test set for EG_{Zh}, ensembles and baselines; *Jia(2018)* baseline is much lower than others, and not displayed for the clarity of the figure.

over-optimistic, we set the minimum recall with English lemma baseline.

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

EG_{Zh} and EG_{En} are built with the same entailment graph induction algorithm (Hosseini et al., 2018), and evaluated on parallel datasets. Learnt from 57% the data, EG_{Zh} achieves an AUC value 55% of its English counterpart. Considering the extra noise from MT in evaluation, it shows that our pipeline is utilizing the source corpus very well.

The ensemble between EG_{Zh} and EG_{En} sets a new SOTA for unsupervised predicate entailment detection. With all 4 ensemble strategies, we observe an improvement upon both monolingual graphs; with **Ensemble MAX**, the best setting on dev set, the margin of test set improvement is more than 5 points. Moreover, with **Ensemble++ AVG**, the best dev set setting when ensembling with EG_{En}++, we get an AUC of 24.0 points on the test set, raising SOTA by more than 4 points.

In an ablation study, the EG_{Zh} *-type* setting, without entity typing, loses 2.1 points in AUC. This means the *HierType* model on **CFIGER** indeed helps entailment graph construction by correctly typing the arguments, thus typing the predicates.

Another ablation study, shown in the fourth section of Table 3, disentangles the effect of ensembling 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, our **Ensemble MAX** is still far ahead of them. This suggests, the success of cross-lingual ensemble cannot be reproduced by sticking in all the data to a monolingual graph via translation. Further, ensembling the two DataConcat graphs delivers a 25.0% AUC, 7 points higher than DataConcat En, an even wider margin than our main setting.

These results show that complementary information is learnt in entailment graphs in the two languages, and the strength of our Chinese entailment graph is sufficient to contribute to the ensemble. The ensemble delivers a huge lift in performance, especially in terms of recall in the moderate precision range (see Figure 1). Thus, we expect that ensembling strong entailment graphs in more languages may result in further improvements.

We further analyse our improvements with a case study: we compare the predictions of our Ensemble_MAX 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 4, we break down each class of differences according to the direct cause of EG_{Zh} making a different prediction than EG_{En}[9][10]:

- **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}.

[9] since the switch in Ensemble_MAX is driven by EG_{Zh}.
[10] examples of each class of cause are given in Appendix D.

Direct causes of EG_{Zh}'s different prediction	TP (+)	FP (-)	TN (+)	FN (-)	+/-
translation-related causes, among which:	+52	-30	+42	-48	+16
· same sentence after translation	*+52*	*-30*	*0*	*0*	*+22*
· translation error	*0*	*0*	*+42*	*-48*	*-6*
lexicalization	+28	-52	+20	-12	-16
ORE error	+8	-17	+8	-7	-8
evidence of entailment	**+108**	**-108**	**+101**	**-51**	**+50**
TOTAL	+196	-207	+171	-118	+42

Table 4: 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.

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. In the "BackTrans Esb" ablation study in Table 3, we single out translation in ensembling: we ensemble predictions on the original and back-translated Levy-Holt dataset, both in English. The performance gain in this case is only marginal, stressing that evidence of entailment is the key to our success, while translation is not. Further, for EG_{Zh} itself, translated datasets is a negative factor overall, as explained later below.

In Table 4, 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 the negatives are false; *Negatives* attempt to correct the false positives, where 35% of the positives are false (as dictated in the setting of our case study). In this context, it is expectable that our evidence of entailment gets $108/(108+108) = 50\%$ correct for *positives*, while a much better $101/(101+51) = 66\%$ correct for *negatives*. These results support the solidarity of our contributions.

The use of translated test data underestimates the power of Chinese entailment graphs in three ways: 1) The quality of machine-translation is imperfect. Without wider context, the translations could drift apart from the meaning of the original relations, and the entailment labels could go wrong. 2) EG_{Zh} is induced purely from native Chinese, while the translated relations bear a translationese language style. This poses a gap in the choice of words, and reduces the chance of finding a match in EG_{Zh}. 3) The original Levy-Holt dataset uses human-annotated relation triples, while for the Chinese version we have to mine them from translated pseudo-sentences with our ORE method, adding an extra source of noise.

While the first two sources of noise are harder to measure, we can crudely quantify the third one by counting the ORE failures. Among the 12,921 relation pairs in Levy-Holt test set, 3,584 of them failed to be translated-then-parsed into binary relations. This means, for Chinese entailment graphs, the hard boundary for recall is not 100%, but rather 72.3%, as is the hard boundary for AUC.

Though hindered by this evaluation setting, our Chinese entailment graph still achieves strong performance. Particularly, in the Data_Concat setting in Table 3, we get a 79.8% ratio of AUC between Chinese and English, which is fully explainable by the 72.3% ratio of hard recall bound. This reaffirms that the strength of our Chinese entailment graph pipeline is on par with its English counterpart.

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 by 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 English monolingual graphs, and sets a new SOTA for unsupervised entailment detection. Directions for future work include multilingual 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.

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.

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

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 and Chris Callison-Burch. 2014. The Multilingual Paraphrase Database. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4276–4283, Reykjavik, Iceland. European Language Resources Association (ELRA).

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. Master's thesis, Macquarie University.

Mohammad Javad Hosseini. 2021. Unsupervised Learning of Relational Entailment Graphs from Text.

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.

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]*.

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.

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.

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.

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.

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.

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]*.

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

In Table 5 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 Section 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 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 that has never been covered by previous work are listed in Section 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 5: 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 injective mapping assumption, we inspect the number of FIGER type labels to which each ultra-fine-grained type is mapped through manual labelling without considering injectivity. 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 label mapping is roughly injective.

We further inspected the number of FIGER types each mention is attached with. It turns out the 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.

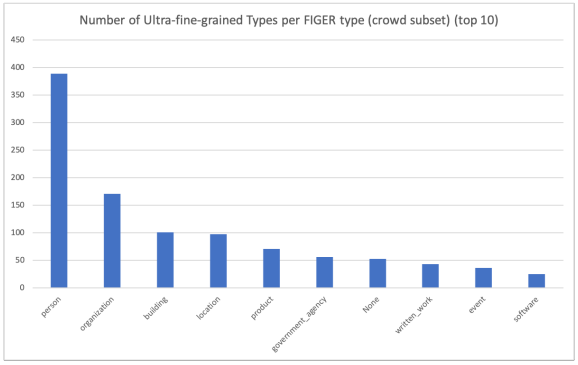

Figure 2: Number of ultra-fine-grained types in crowd-annotated subset mapped to each FIGER type; only the FIGER types with top 10 number of ultra-fine-grained types are displayed.

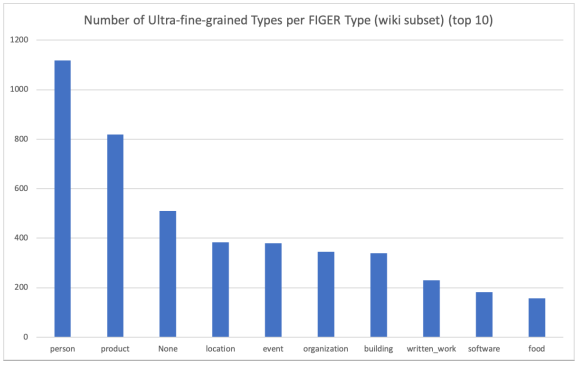

Figure 3: Number of ultra-fine-grained types in wikipedia distantly supervised subset mapped to each FIGER type; only the FIGER types with top 10 number of ultra-fine-grained types are displayed.

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 2 and 3. 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 4 and 5. The winners in terms of the number of mentions are consistent with that of the number of ultra-fine-grained types, and also consistent among themselves (between the two subsets).

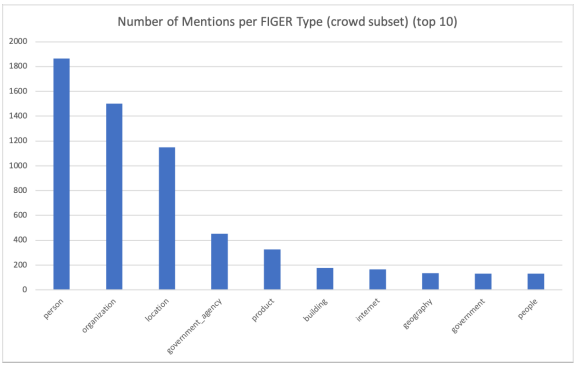

Figure 4: Number of mentions in crowd-annotated subset labelled as each FIGER type; only the FIGER types with top 10 number of mentions are displayed.

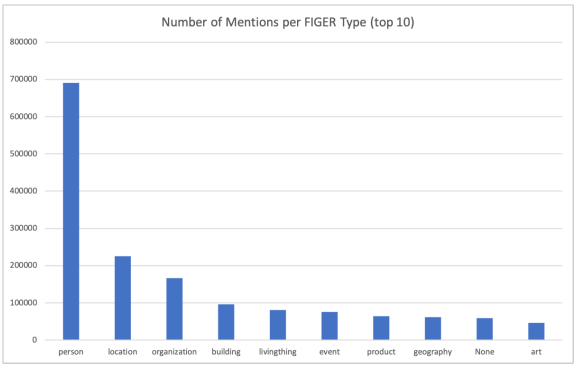

Figure 5: Number of mentions in wikipedia distantly supervised subset labelled as each FIGER type; only the FIGER types with top 10 number of mentions are displayed.

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 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 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 Section 7. Chinese sentences and relations in the examples are presented in the same format as Section 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 simultaneous 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 a 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*.

E 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 6 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 complementarity of entailment graphs in different languages.

Figure 7 contains the curves for two ablation studies: EG_{Zh} with or without entity typing; EG_{En}

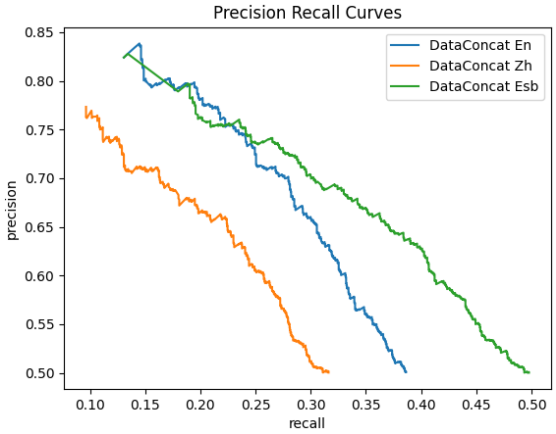

Figure 6: P-R Curves on Levy-Holt test set for DataConcat ablation study.

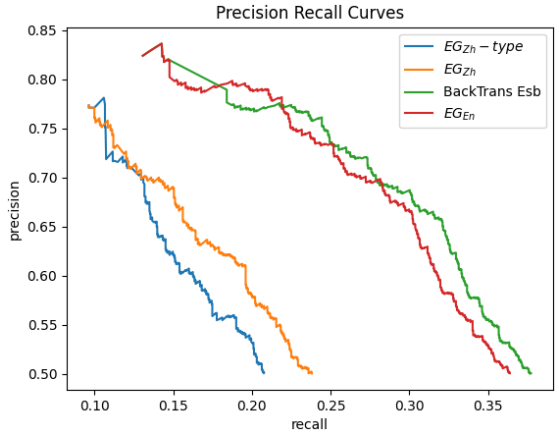

Figure 7: P-R Curves on Levy-Holt test set for $EG_{Zh} - type$, BackTrans Esb, in comparison to EG_{Zh} and EG_{En} respectively.

ensembled with back-translation predictions 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 paraphrases through translation is not a major contributor to the success of our ensembling method.