Discourse Representation Structure Parsing for Chinese

Chunliu Wang, Xiao Zhang, Johan Bos
CLCG, University of Groningen
{chunliu.wang, xiao.zhang, johan.bos}@rug.nl

Abstract

Previous work has predominantly focused on monolingual English semantic parsing. We, instead, explore the feasibility of Chinese semantic parsing in the absence of labeled data for Chinese meaning representations. We describe the pipeline of automatically collecting the linearized Chinese meaning representation data for sequential-to-sequential neural networks. We further propose a test suite designed explicitly for Chinese semantic parsing, which provides fine-grained evaluation for parsing performance, where we aim to study Chinese parsing difficulties. Our experimental results show that the difficulty of Chinese semantic parsing is mainly caused by adverbs. Realizing Chinese parsing through machine translation and an English parser yields slightly lower performance than training a model directly on Chinese data.

1 Introduction

Semantic parsing is the task of transducing natural language text into semantic representations, which are expressed in logical forms underlying various grammar formalisms, such as abstract meaning representations (AMR, Wang et al. 2020; Bevilacqua et al. 2021), minimal recursion semantics (MRS, Horvat et al. 2015), and Discourse Representation Theory (DRT, Kamp and Reyle 1993). In this work, we explore the feasibility of parsing Chinese text to semantic representation based on Discourse Representation Structures (DRSs, Bos 2015a; van Noord et al. 2018), which are meaning representations proposed from DRT, a recursive first-order logic representation comprising of discourse referents (the entities introduced in the discourse) and relations between them.

Several neural parsers for DRS have been recently developed (Fancellu et al., 2019; Evang, 2019; van Noord et al., 2019; Liu et al., 2019; Wang et al., 2021; van Noord et al., 2020a) and reached remarkable performance, but mostly focused on monolingual English or some language using the Latin alphabet. Meaning representations are considered to be language-neutral, and texts with the same semantics but in different languages have the same meaning representation. The literature presents several examples of parsing multilingual text by training on monolingual English semantic representations (Ribeiro et al., 2021).

For the reason of relatively limited amounts of labeled gold-standard multilingual meaning representation data, multilingual text parsing relies on the source of silver English meaning representation data. As long as the meanings are expressed in a language-neutral way, this is a valid approach. However, named entities aren't usually, because they can (a) have different orthography for different languages using the same alphabet (in particular for location names, e.g., Berlin, Berlijn, Berlino, Berlynas) or (b) be written with a completely different character set, as is the case for Chinese.

Figure 1 shows a (nearly) language-neutral meaning representation for a simple English sentence. For non-English Latin alphabet languages, the named entities in the text are usually consistent with English, and the meaning in the form of a graph structure of the corresponding Discourse Representation (Discourse Representation Graph, DRG) would be identical to these languages (Bos, 2021), as shown in Figure 1. However, it would be rather absurd to expect a semantic parser for Chinese to produce meaning representations (with interlingual WordNet synsets) where proper names are anchored using the Latin alphabet using English (or any other language for that matter) orthography. We need to keep this important aspect in mind when evaluating semantic parsers for languages other than English.

However, for non-Latin alphabet languages, such as the widely used language of Chinese, is it feasible to use English meaning representation as the meaning representation of Chinese? Our objective

is to investigate whether Chinese semantic parsing can achieve the same performance as English semantic parsing while using the same amount of data. We try to investigate whether it is necessary to develop a dedicated parser for Chinese, or whether it is possible to achieve a similar performance using an English parser by leveraging machine translation (MT) on Chinese. We provide inexpensively acquired silver-standard Chinese DRS data to implement our exploration: (1) We collect Chinese and English aligned texts from the Parallel Meaning Bank (PMB, Abzianidze et al. 2017), which provides parallel multilingual corpora including corresponding English meaning representation expressed in DRSs. (2) We leverage GIZA++ (Och and Ney, 2003) to align the word-segmented Chinese and English to obtain Chinese-English named entity alignment pairs, the resulting named entities are used to replace the named entities in our English semantic representation. (3) We train two monolingual parsers on the two languages separately, and then provide a set of fine-grained evaluation metrics to make better comparison between parsers. We aim to answer the following questions:

1. Can existing DRS parsing models achieve good results for Chinese? (RQ1)
2. What are the difficulties in semantic parsing for Chinese? (RQ2)
3. Is it feasible to use machine translation and an English parser to parse Chinese? How is it different from designing a special parser for Chinese? (RQ3)
4. How to conduct more fine-grained evaluation of experimental results and reduce the workload of manual evaluation? (RQ4)

2 Background

2.1 Discourse Representation Structure

DRS, as a kind of formal meaning representation, can be used to represent the semantic meaning of sentences and discourse. For the wide coverage of linguistic phenomena at quantification, negation, reference resolution, comparatives, discourse relations, and presupposition, DRT and DRS possess stronger semantic representation power than AMR. A DRS comprises discourse referents and conditions. However, some variants of DRS formats have been introduced in recent years, the format we employ throughout our work being one

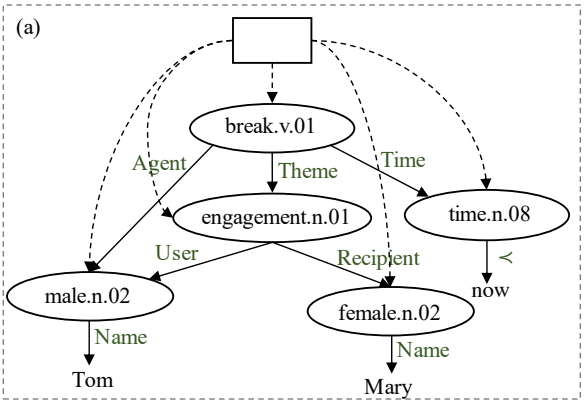

Figure 1: DRS in (a) graph format, (b) sequential box notation and (c) corresponding multilingual texts for English, German and Chinese.

of them. We use a simplified DRS, which can be called Discourse Representation Graph (DRG) or Simplified Box Notation (SBN; Bos 2021). It discards explicit discourse references and variables while maintaining the same expressive power, as shown in Figure 2.

As introduced by Bos (2021), DRS allows two kinds of representations: graph and sequential notation (Figure 1). There are five types of semantic information involved in DRS: concepts (`read.v.01`, `paper.n.02`, `new.a.01`, `...`), roles (`Agent`, `Theme`, `Time`, `...`), constants (`speaker`, `hearer`, `now`, `...`), comparison operators (=, $\prec$, $\sim$, `...`) and discourse relations (`NEGATION`, `CONTINUATION`, `CONTRAST`, `...`), where concepts and roles are represented by WordNet synsets (Fellbaum, 2000) and VerbNet thematic relations (Kipper et al., 2006) respectively.

2.2 DRS parsing

DRS parsing was originally applied to English and has been continuously extended to other Latin languages. Initially, rule-based systems were predominantly utilized by early parsers for analyzing small English texts (Johnson and Klein, 1986; Asher and Wada, 1988; Bos, 2004, 2008, 2015b). The first version of GMB (Basile et al., 2012) which pro-

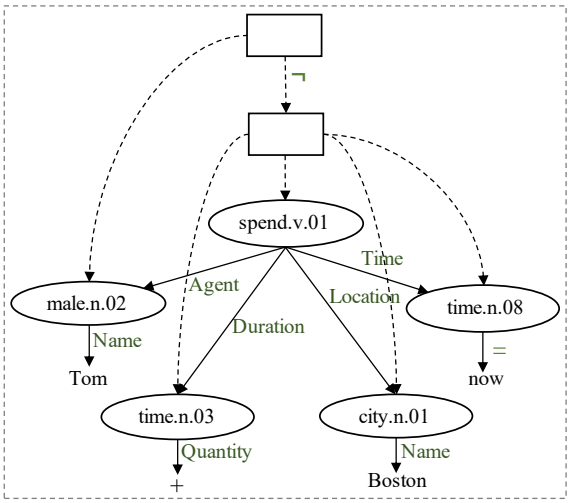

Figure 2: An example of a DRG with negation for sentence: "Tom doesn't spend much time in Boston."

vides English texts with DRS, is built on Boxer (Bos, 2008). With the release of PMB (Abzianidze et al., 2017) and the propose of the first shared tasks (Abzianidze et al., 2019), related research keeps growing, with a focus on deep learning models (Evang, 2019; Fancellu et al., 2019; van Noord et al., 2018, 2020a; Liu et al., 2019). The target languages have also expanded to other languages: German, Italian, Dutch and Chinese (Shen and Evang, 2022; Poelman et al., 2022a; Wang et al., 2021; Liu et al., 2021). Translation has been utilized in two manners when dealing with cross-lingual parsing: the first involves translating other languages into English and then employing an English parser, while the second involves translating English into other languages and training a parser specific to that language (Liu et al., 2021). In this paper, we use the existing Chinese-English parallel corpus to design a specific parser for Chinese, and compare the performance of the parser with the first method.

3 Data Creation

In previous work, for non-English parsing tasks, the semantic representation of English is usually directly used as the semantic representation of the target language, but most of these works focus on Latin languages (Fancellu et al., 2019; Ribeiro et al., 2021). For non-Latin languages such as Chinese, named entities are not language-neutral, as illustrated in the work of Wang et al. (2021), and are quite different from named entities in English texts. To design a more reasonable Chinese parser, we first focus on replacing the named entities in the English semantic representation with Chinese, so that the parser can parse out the Chinese named entities corresponding to the text content according to different texts.

To achieve our goal, we use the data of PMB, the largest parallel corpus of DRS data available, as our experimental object. From the PMB, English-Chinese parallel texts and DRS data for English texts are collected. Based on that, we propose a pipeline to obtain Chinese DRS for Chinese text. Our pipeline has three steps: (1) using tokenizers tools to segment Chinese and English text data; (2) utilizing the English-Chinese alignment tool to obtain the alignment tokens between Chinese and English texts; (3) replacing named entities in English DRS with Chinese named entities. Figure 3 shows our processing pipeline.

3.1 Text Tokenizers

Preprocessing data with a tokenizer is an important step in the pipeline because the alignment of Chinese and English texts needs to act on the data after tokenization. At the same time, since the quality of upstream results directly affects downstream performance, the quality of text segmentation also directly affects the correctness of Chinese and English text alignment. In this work, we use Moses (Koehn et al., 2007) for English, which is advanced and widely used. It is a collection of complex normalization and segmentation logic that works very well for structured languages like English. For Chinese, we choose HanLP (He and Choi, 2021), which is an efficient, user-friendly and extendable tokenizer. Different from a widely used Jieba tokenizer, HanLP is based on the CRF algorithm. It takes into account word frequency and context at the same time, and can better identify ambiguous words and unregistered words.

3.2 English-Chinese Alignment

In order to realize the replacement of named entities in English semantic representation with Chinese named entities, it is very important to obtain the correct alignment of Chinese and English texts, especially the alignment of named entities in the two texts. In order to quickly and effectively obtain the alignment data in Chinese and English, we choose the GIZA++ word aligning tool. GIZA++ is the most popular statistical alignment and MT toolkit (Och and Ney, 2000), which implements the lexical translation models of Brown et al. (1993) (IBM Models), and the Hidden-Markov

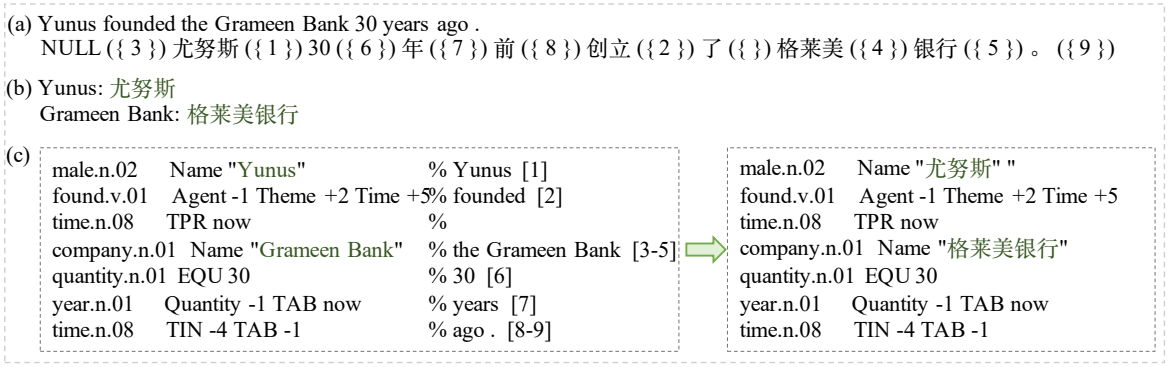

Figure 3: (a) Alignment tokens obtained by GIZA++ tool for English (*"Yunus founded the Grameen Bank 30 years ago."*) and Chinese ("尤努斯30年前创立了格莱美银行。"), (b) aligned named entities dictionary in above texts, (c) same meaning representations with different named entities for English text and Chinese text respectively.

alignment Model (Vogel et al., 1996), trained using expectation-maximization (EM). GIZA++ is highly effective at aligning frequent words in a corpus, but error-prone for infrequent words.

3.3 Replacing Named Entities

The last step to obtain the Chinese semantic representation is to replace the named entities in the English DRS with Chinese named entities. First, the English named entities in DRS data can be easily obtained according to the edge types between two nodes. When the edge type is `Name`, the output nodes are named entities in the DRG. After processing the Chinese and English texts with the GIZA++ tool in the second step, we can obtain alignment tokens between Chinese and English. On this basis, a named entity alignment dictionary can be obtained, and then the English named entities in the DRS data can be replaced with Chinese named entities based on this dictionary.

4 Methodology

4.1 Neural Models

We adopt Recurrent Neural Networks (RNN) equipped with Long Short-Term Memory units (LSTM; Hochreiter and Schmidhuber 1997) as our baseline models. Following the work of van Noord et al. (2020b), we use frozen mBERT (Devlin et al., 2019) embeddings to initialize the encoder. An attention-based LSTM architecture is used for the decoder, where the attention memory is the concatenation of the attention vectors among all the input tokens. In addition, the copy mechanism (Gu et al., 2016; Gulcehre et al., 2016) is added to the decoder, which can integrate the attention distribution into the final vocabulary distribution. The copy mechanism favors copying tokens from the source text into the target text instead of generating all target tokens only from the target vocabulary.

4.2 Evaluation

Given a document to the DRS parser, it will generate variable-free sequential notation DRS as shown in Figure 1(b). The evaluation tool for DRS parsing task was recently proposed by Poelman et al. (2022b) and is based on the AMR standard evaluation tool Smatch (Cai and Knight, 2013). By converting a sequential DRS into DRG, Penman notation format data (Kasper, 1989) can be obtained, as shown in Figure 4 (b), and then Smatch can be used to compute F-scores based on matching triples between system output and gold meanings.

However, we note that the scores given by the above evaluation tool have two flaws: (1) the evaluation scores are too inflated, and it is difficult to detect the differences between different parsers. (2) the evaluation tool only gives an overall score without evaluating the different types of constituent elements in the DRS, it is difficult to quantitatively determine what is the difficulty of the parser in the parsing process. Based on that, we propose to compress evaluation scores to improve the above evaluation methods and further propose fine-grained evaluation metrics for different subtasks according to different types of components in DRS.

4.2.1 Overall Evaluation

Our improvement strategy is mainly aimed at the representation of the Penman format of DRG. We mainly improve on two points, one is WordNet synsets representation, and the other is constants representation.

In the previous evaluation method, the WordNet

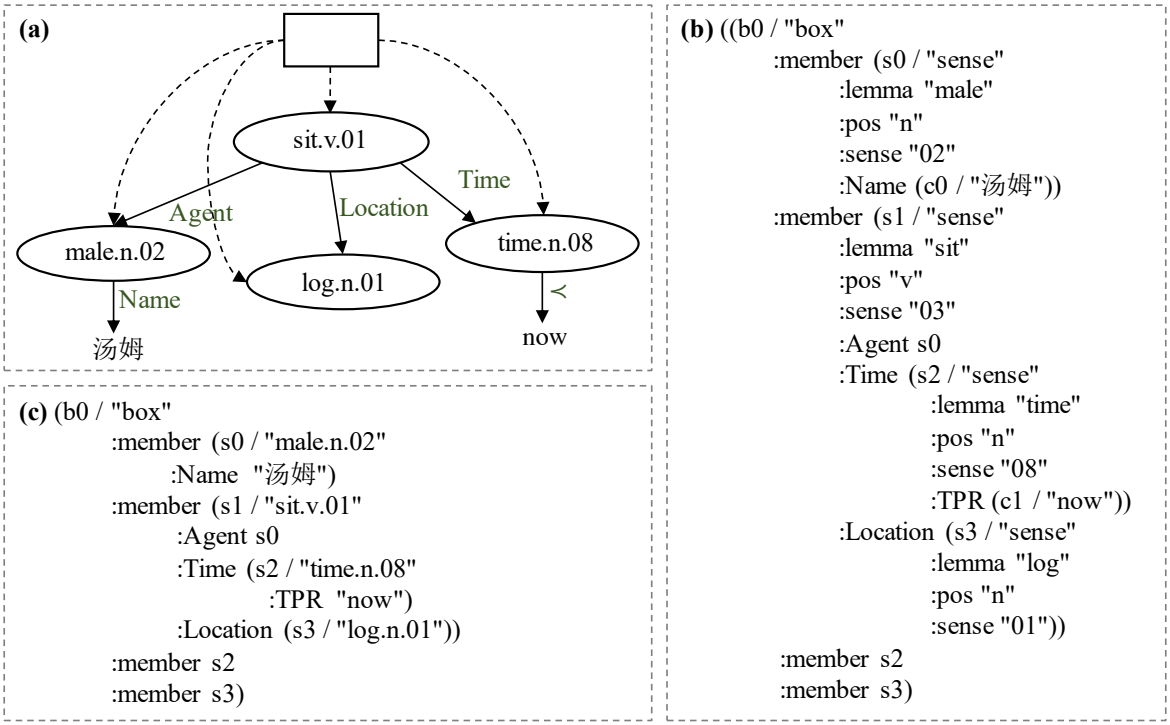

Figure 4: (a) Graph structured DRS for Chinese sentence:" 汤姆坐在一根圆木上。". (b) Penman format of DRG with fine-grained WordNet synsets used for evaluation (Poelman et al., 2022b). (c) Penman format of DRG with coarse-grained WordNet synsets used for evaluation (Ours).

synsets in Penman format are fine-grained during the evaluation process, and the WordNet synsets are divided into three parts (lemma, pos, number) according to their constituents. On this basis, even if the parser generates wrong concepts, such as `time.n.08` and `time.n.01`, the Smatch still obtains a similar inflated F1 score. To this end, we change the WordNet synsets in the Penman format to a coarse-grained representation to strictly evaluate WordNet synsets qualities generated by parsers, as shown in Figure 4 (c). In addition, we have also modified the constant representation in Penmen format, such as the constant `now` shown in the figure, because the variable *c* is added to the constant, making the triples in Penman format redundant, which also makes the F1-score higher to a certain extent. By omitting the *c* variable as shown in Figure 4 (c), we further compress the F1-score.

4.2.2 Fine-grained Evaluation

To evaluate the quality of specific subtasks in DRS parsing, we imitate the fine-grained metrics for AMR parsing task (Damonte et al., 2017; Zhang et al., 2019) to DRS parsing. In order to make them compatible with DRS, we make some changes based on the data characteristics of DRS. Our fine-grained metrics consist of three categories in total: *graph-level*, *node-level* and *edge-level*. Each category includes more fine-grained evaluation metrics. All the metrics are proposed based on the semantic information types involved in DRS (see Section 2.1).

In ***graph-level*** evaluation, `No Roles`, `No Discourse`, `No Operators` and `No Senses` are used to represent the Smatch scores of the DRG in Penman format ignoring Roles, Discourse, Operators and Senses respectively. In theory, they are Smatch's coarse-grained scores, which are higher than the original Smatch scores.

In ***node-level*** evaluation, we compute F-score on the list of parsed information types (such as roles, constants, and discourse relations) instead of using Smatch. Note that different from the metrics in the AMR parsing task, concepts in DRS are represented by WordNet synsets, so `Concepts` can be evaluated more finely by part-of-speech (`noun`, `adjective`, `adverb` and `verb`). `Discourse` detects all discourse relation labels except *NEGATION* since it is more common and specific in DRS than other discourse relations labels, the `Negation` metric is used for evaluation to detect *NEGATION* edge label alone. In addition, `Member` metric is added to evaluate the ratio of the generated concepts. In DRG, *member* represents the edge label

Alignment Error	Data Type	Example
Dislocation	Chinese	梅尔·卡玛津 是 天狼星 的 执行官。
	English	Mel Karmazin is an executive of Sirius .
	Wrong	male.n.02 Name "梅尔·卡玛津" be.v.08 Theme -1 Time +1 Co-Theme +2 time.n.08 EQU now person.n.01 Role +1 executive.n.01 Of +1 company.n.01 Name "执行官"
	Corrected	male.n.02 Name "梅尔·卡玛津" be.v.08 Theme -1 Time +1 Co-Theme +2 time.n.08 EQU now person.n.01 Role +1 executive.n.01 Of +1 company.n.01 Name "天狼星"
Character Exclusion	Chinese	什么 乐队 唱 了 “ 快乐 在 一起 ” 这 首 歌 ?
	English	What group sang the song " Happy Together " ?
	Wrong	group.n.01 Name ? sing.v.02 Agent -1 Time +1 Theme +2 time.n.08 TPR now song.n.01 EQU +1 music.n.01 Name "快乐一起"
	Corrected	group.n.01 Name ? sing.v.02 Agent -1 Time +1 Theme +2 time.n.08 TPR now song.n.01 EQU +1 music.n.01 Name "快乐在一起"
Character Inclusion	Chinese	卢瑟福·海斯 1822 年 出生 于 俄亥俄州 。
	English	Rutherford Hayes was born in Ohio in 1822 .
	Wrong	male.n.02 Name "卢瑟福·海斯1822" time.n.08 TPR now bear.v.02 Patient -2 Location +1 Time +2 state.n.01 Name "俄亥俄州" time.n.08 YearOfCentury 1822 TIN -3
	Corrected	male.n.02 Name "卢瑟福·海斯" time.n.08 TPR now bear.v.02 Patient -2 Location +1 Time +2 state.n.01 Name "俄亥俄州" time.n.08 YearOfCentury 1822 TIN -3
Nationality	Chinese	我 不 是 爱尔兰人 。
	English	I am not Irish .
	Wrong	person.n.01 EQU speaker NEGATION <1 time.n.08 EQU now be.v.03 Theme -2 Time -1 Source +1 country.n.02 Name ""
	Corrected	person.n.01 EQU speaker NEGATION <1 time.n.08 EQU now be.v.03 Theme -2 Time -1 Source +1 country.n.02 Name "ireland"

Table 1: Alignment errors illustrated by four examples. In Chinese and English texts, words of the same color indicate correct alignment between them. Inbformation marked in red is the wrong named entity obtained by the GIZA++ tool. Text in green indicates the correct named entity in the corrected DRS.

connecting the *BOX* node and the concepts node, i.e., the dashed line as shown in Figure 4 (a).

For ***edge-level*** evaluation, we focus on calculating the F-score based on the number of matching triples in the parsed DRG and the gold DRG. For example, `Names` in edge-level is a metric that considers the relations between concepts nodes and named entities, which differs from the metric of `Names` in node-level, which only considers the concepts labeled with *Name* and ignores the accuracy of named entities themselves. [1]

5 Experiments

5.1 Dataset

We collect all Chinese-English text pairs in the PMB. According to the quality label of English DRS, we divide the data into gold data and silver data, and randomly split the test set and development set from the gold data. Since PMB data may contain duplicate data, before splitting, we first filter the duplicate data. Then we merge the remaining gold data and silver data as our training set, and get a total of 137,781 training instances, 1,000 development instances and 1,000 test instances, each instance contains English DRS data, corresponding English text, and Chinese text. [2]

After splitting the data, we use the pipeline introduced in Section 3 to process our Chinese and English texts to get the Chinese and English word alignment data, and then replace the named entities in the English DRS with Chinese. However, we noticed that not all replacements were successful. We classified the wrong replacement types into four types, as shown in Table 1. These errors are mainly caused by GIZA++ alignment errors when aligning Chinese and English text words. Among them, the fourth type of error is quite special. In our experiment, we directly ignore the location named entities used to refer to nationality and do not replace them with Chinese named entites. In order to reduce the work of manual correction and make the work reproducible, We only fix incorrect named entity replacements in the test set, where 26 of the

[1] Our evaluation suite is available at: https://github.com/wangchunliu/SBN-evaluation-tool.

[2] Our data and code are available at: https://github.com/wangchunliu/Chinese-SBN-parsing.

Alignment Error	Reason	Example
Named-entities	Jieba	English: Melanie killed a spider with her hand . Chinese: 媚兰用 (1 5 6) 手 (7) 杀死 (2) 了 ( ) 一只 (3) 蜘蛛 (4) 。 (8)
	HanLP	English: Melanie killed a spider with her hand . Chinese: 媚兰 (1 6) 用 (5) 手 (7) 杀死 (2) 了 ( ) 一 (3) 只 ( ) 蜘蛛 (4) 。 (8)
Information units	gold data	The ground floor was flooded . Chinese: 一楼 (1) 被 ( ) 淹 (2 3 4 5) 了 ( ) 。 (6)
	all data	English: The ground floor was flooded . Chinese: 一楼 (1 2 3) 被 (4) 淹 (5) 了 ( ) 。 (6)

Table 2: Impact of different tokenizers and data sizes on GIZA++ performance.

1000 test set instances require manual correction of named entities.

5.2 Settings

For tokenizers, we use Moses (Koehn et al., 2007) and HanLP (He and Choi, 2021) on English and Chinese respectively. We observe that the HanLP tokenizer outperforms Jieba[3], a tokenizer widely used in Chinese, in segmenting text containing named entities. This is an important indicator for selecting a tokenizer, because getting the correct Chinese and English named entity pairs is our main goal. In addition, we observed that HanLP's segmentation results also outperformed Jieba's tokenizer on text containing traditional Chinese characters, while the Chinese data in PMB contains traditional Chinese characters. This is also one of the reasons for choosing the HanLP tokenizer. At the top of Table 2, we show the difference in name entities between the Jieba tokenizer and the HanLP tokenizer. In addition, we give an example of the impact of different sizes of training data on the alignment performance of GIZA++ at the bottom of Table 2, and the results show that it is almost impossible to achieve correct alignment using only gold data.

	Document-level			Word-level	
Data	**Train**	**dev**	**test**	**src**	**tgt**
English	137,781	1,000	1,000	38,441	39,761
Chinese	137,781	1,000	1,000	42,446	41,734

Table 3: Document statistics and vocabulary sizes.

All experiments are implemented based on OpenNMT (Klein et al., 2017). For the vocabulary, we construct vocabularies from all words, the vocabulary sizes as shown in Table 3. The hyperparameters are set based on performance on the development set. We use SGD optimizer with the initial learning rate set to 1 and decay 0.8. In addition, we set the dropout to 0.5 at the decoder layer to avoid overfitting with batch size 32.

[3]https://github.com/fxsjy/jieba

Metric	**EN**	**ZH**	**ZH→EN$_{zh}$**
Smatch$_1$	91.0	86.0	84.7
Smatch$_2$	88.9	83.8	81.7
Well-formed	99.8	99.7	99.7
Graph-level			
No Roles	90.0	85.5	84.2
No Discourse	89.5	83.9	82.7
No Operators	89.5	84.7	83.4
No Senses	91.9	85.6	84.7

Table 4: F-scores with Smatch on the test set of semantic parsers. Note: Smatch$_1$ and Smatch$_2$ represent the original evaluation (Poelman et al., 2022b) and our improved evaluation.

5.3 Main Results

Table 4 shows the results obtained by the parsers with Smatch, which gives the overall performance for different parsers. The first parser (**EN**) is trained on the English dataset based on the model introduced in Section 4.1. The Smatch$_1$ result of our English parser is slightly lower than the results of Poelman et al. (2022b), which we believe is due to slightly different training, development and test set instances. The result of Smatch$_2$ is significantly lower than the result of Smatch$_1$, indicating that the F1-score has been significantly compressed and will not be too inflated (see Section 4).

The Chinese parser (**ZH**) is trained on the data created by the pipeline introduced in Section 3. The results show that the performance of the Chinese parser is lower than the English parser in all overall evaluation metrics. **ZH→EN$_{zh}$** shows the perfor-

	Metric	EN	ZH	ZH→EN$_{zh}$
Node	Names	70.8	66.0	67.7
	Negation	92.3	88.7	88.8
	Discourse	86.0	80.4	75.2
	Roles	89.2	84.0	84.9
	Members	97.5	95.4	95.9
	Concepts	81.2	73.3	74.4
	noun	87.1	82.1	83.3
	adj	73.3	54.2	52.5
	adv	76.8	35.3	45.5
	verb	59.7	45.5	47.2
Edge	Roles	81.0	73.3	73.7
	Names	79.4	74.0	45.5
	Members	90.9	86.4	87.0
	Operators	92.9	87.7	87.7
	Discourse	86.2	79.6	75.3

Table 5: F-scores of fine-grained evaluation on the test set of semantic parsers. The evaluation metrics in the table are all based on the Penman format DRG with coarse-grained WordNet synsets.

mance by using the English parser on English text translated from Chinese text instead of training a dedicated model for Chinese text. The only unreasonable point is that the model will generate English named entities, which may not be recognized as the correct Chinese semantic representation.

The smatch$_1$ scores and the smatch$_2$ scores show that the Chinese parser outperforms using the ZH→EN$_{zh}$ approach. For the metrics `No Senses` and `No Roles`, the evaluation results have been significantly improved compared with Smatch$_2$. This shows that Concepts and Roles have a greater impact on evaluation results than Discourse and Operators. It is worth noting that the performance difference between the Chinese and English parsers is about five percentage points across all metrics, while the difference between the ZH and the ZH→EN$_{zh}$ narrows at the graph-level metrics compared to Smatch$_2$ score.

5.4 Fine-grained Results and Analysis

To further explore the performance of parsers, we apply our proposed fine-grained evaluation metrics to the results of two parsers. Tabel 5 shows the fine-grained evaluation performance of different component types based on DRG at node-level and edge-level.

`Names`: From the results, we observe that the metric `Names` gives completely opposite results at different evaluation levels. On the node-level, the `Names` metric in ZH parser scores the lowest, but on the edge-level, `Names` metric in ZH→EN$_{zh}$ gives the lowest scores. This is reasonable and expected because the node-level `Names` metric only evaluates whether the parser can parse concepts to contain named entities, so the results of ZH→EN$_{zh}$ parser should be similar to those of the English parser. However, the edge-level `Names` metric evaluates whether the generated named entities completely match the original text, and the ZH→EN$_{zh}$ parser completely loses the Chinese named entity information.

`Discourse`: An important observation is that the metric `Discourse` has very low F1 scores on both the node-level and the edge-level for the Chinese parser. Using machine translation and an English parser to parse Chinese (ZH→EN$_{zh}$) will further degrade the performance of the metric `Discourse`. Based on the text data and parsed output, we find that discourse relations in Chinese are inconspicuous, and even disappears after being translated into English (see Table 6 for examples).

`Concepts`: Table 5 shows the `Concepts` scores of ZH parser are lower than those for ZH→EN$_{zh}$ except for the *adj* category. This is an interesting finding, because the performance of other parts of speech in the ZH parser is worse than that of ZH→EN$_{zh}$, while *adj* is special. We observe that the expressions of adjectives in Chinese translated into English are diverse and may not match the original English text (see Table 6 and Appendix B for relevant examples).

For the English parser, *verbs* are the most difficult words to parse, scoring significantly lower than other parts of speech. However, the difficulty of Chinese semantic parsing is mainly reflected in *adv*. In addition, the accuracy of ZH→EN$_{zh}$ in parsing concepts of *adv* is significantly better than that of the ZH parser, but it is still the lowest results in four types of parts of speech for ZH→EN$_{zh}$. On the one hand, the corpus containing adverb data is smaller, which makes the training insufficient. On the other hand, the adverbs in Chinese are usually not obvious and diverse.

For *noun* and *verb*, ZH has the worst performance, with the ZH→EN$_{zh}$ method, the performance of *noun* and *verb* is slightly improved, but it is much worse than the **EN** parser. A typical reason is that the English text translated from Chinese may not be consistent with the original English text. We

Information Type	Example	Lost/Changed in Translation
Discourse	EN: A parrot can mimic a person's voice. ZH: 鹦鹉会模仿人的声音。 ZH→EN: Parrots mimic human voices.	POSSIBILITY Lost
	EN: Tom asks his mother if she can buy him a new toy. ZH: 汤姆请求他母亲给他买新玩具。 ZH→EN: Tom begged his mother to buy him new toys.	ATTRIBUTION Lost
Concepts	EN: That guy is completely nuts! ZH: 那家伙真是疯了！ ZH→EN: That guy is crazy!	Adverb Lost
	EN: She's very handy with a saw. ZH: 她很会用锯子。 ZH→EN: She is good with a saw.	Adjective Changed
	EN: I'm awake. ZH: 我醒了。 ZH→EN: I woke up.	Adjective Lost
	EN: Tom is suffering from a bad headache. ZH: 汤姆头痛得厉害。 ZH→EN: Tom has a bad headache.	Verb Changed
Operators	EN: I slept on the boat. ZH: 我睡在船上。 ZH→EN: I sleep on the boat.	Tense Lost
Negation	EN: The music lured everyone. ZH: 音乐吸引了所有人。 ZH→EN: Music appeals to all.	NEGATION Lost
	EN: The printer doesn't work. ZH: 打印机坏了。 ZH→EN:The printer is broken.	

Table 6: Examples of translated English texts with loss of information.

observe that the DRS sequences parsed using the translated text are overall shorter than those parsed using the original English text, some noun concepts are missing, and the verb concepts may be inconsistent with the reference DRS (see Appendix B for examples).

Operators & Negation: Our fine-grained results obtained by using machine translation and the English parser are not always worse than training a Chinese parser alone. For the metrics Negation and Operators, both methods have similar scores at both the node-level and the edge-level. However, when we compare the results of ZH→EN$_{zh}$ with EN parser, we find that all the results of ZH→EN$_{zh}$ are significantly lower than those of the EN parser. We found that tense information is usually lost in the process of English-Chinese translation, but almost no tense information is lost in the process of Chinese-English translation. This explains why the result of the Chinese parser operator is significantly lower than that of the English parser, while the result of ZH→EN$_{zh}$ is the same as that of the ZH parser. For Negation, we can observe something interesting. As the connector NEGATION in English DRss can also express universal quantification (using nesting of two negation operators) for words such as "every" and "always", this information is missing in the translation process, and as a result not picked up by the parser.

Members & Roles: For this metric, ZH→EN$_{zh}$ even slightly outperforms the ZH parser, but they are both lower than the EN parser. On the one hand, a free translation may lead to a different ordering of semantic information. Although texts with the same meaning but realised with different word order have the same semantic graph, a parser based on sequence-to-sequence neural networks may get the wrong graph structure leading to a lower evaluation score of the Roles evaluation metric. On the other hand, both evaluation metrics are affected by the correctness of Concepts, and in our results, the Chinese parser scored lower than the other two parsers for Concepts.

6 Conclusion

Given an annotated meaning bank primarily designed for English, it is feasible to develop a semantic parser for Chinese by pairing the "English" meaning representation with Chinese translations, reaching good results. Most difficulties in Chinese parsing are caused by adverbs, while the diversity of Chinese verbs and adjectives also has a big impact on parsing performance. Using Machine Translation as an alternative to approach semantic parsing for Chinese yields slightly lower results. Our fine-grained graph evaluation gives better insight when comparing different parsing approaches.

Acknowledgments

This work was funded by the NWO-VICI grant "Lost in Translation—Found in Meaning" (288-89-003) and the China Scholarship Council (CSC). We thank the anonymous reviewers for detailed comments that improved this paper. We would also like to thank the Center for Information Technology of the University of Groningen for their support and for providing access to the Peregrine high performance computing cluster.

References

Lasha Abzianidze, Johannes Bjerva, Kilian Evang, Hessel Haagsma, Rik van Noord, Pierre Ludmann, Duc-Duy Nguyen, and Johan Bos. 2017. The Parallel Meaning Bank: Towards a multilingual corpus of translations annotated with compositional meaning representations. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 242–247, Valencia, Spain. Association for Computational Linguistics.

Lasha Abzianidze, Rik van Noord, Hessel Haagsma, and Johan Bos. 2019. The first shared task on discourse representation structure parsing. In *Proceedings of the IWCS Shared Task on Semantic Parsing*, Gothenburg, Sweden. Association for Computational Linguistics.

Nicholas Asher and Haijime Wada. 1988. A Computational Account of Syntactic, Semantic and Discourse Principles for Anaphora Resolution. *Journal of Semantics*, 6(1):309–344.

Valerio Basile, Johan Bos, Kilian Evang, and Noortje Venhuizen. 2012. A platform for collaborative semantic annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 92–96, Avignon, France.

Michele Bevilacqua, Rexhina Blloshmi, and Roberto Navigli. 2021. One SPRING to rule them both: Symmetric AMR semantic parsing and generation without a complex pipeline. In *Proceedings of AAAI*.

Johan Bos. 2004. Computational semantics in discourse: Underspecification, resolution, and inference. *Journal of Logic, Language and Information*, 13:139–157.

Johan Bos. 2008. Wide-coverage semantic analysis with Boxer. In *Semantics in Text Processing. STEP 2008 Conference Proceedings*, pages 277–286. College Publications.

Johan Bos. 2015a. Open-domain semantic parsing with boxer. In *Proceedings of the 20th Nordic Conference of Computational Linguistics, NODALIDA 2015, May 11-13, 2015, Institute of the Lithuanian Language, Vilnius, Lithuania*, pages 301–304. Linköping University Electronic Press / ACL.

Johan Bos. 2015b. Open-domain semantic parsing with boxer. In *Nordic Conference of Computational Linguistics*.

Johan Bos. 2021. Variable-free discourse representation structures. *Semantics Archive*.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Shu Cai and Kevin Knight. 2013. Smatch: an evaluation metric for semantic feature structures. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 748–752, Sofia, Bulgaria. Association for Computational Linguistics.

Marco Damonte, Shay B. Cohen, and Giorgio Satta. 2017. An incremental parser for Abstract Meaning Representation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 536–546, Valencia, Spain. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Kilian Evang. 2019. Transition-based DRS parsing using stack-LSTMs. In *Proceedings of the IWCS Shared Task on Semantic Parsing*, Gothenburg, Sweden. Association for Computational Linguistics.

Federico Fancellu, Sorcha Gilroy, Adam Lopez, and Mirella Lapata. 2019. Semantic graph parsing with recurrent neural network DAG grammars. 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 2769–2778, Hong Kong, China. Association for Computational Linguistics.

Christiane D. Fellbaum. 2000. Wordnet : an electronic lexical database. *Language*, 76:706.

Jiatao Gu, Zhengdong Lu, Hang Li, and Victor O.K. Li. 2016. Incorporating copying mechanism in sequence-to-sequence learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1631–1640, Berlin, Germany. Association for Computational Linguistics.

Caglar Gulcehre, Sungjin Ahn, Ramesh Nallapati, Bowen Zhou, and Yoshua Bengio. 2016. Pointing the unknown words. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 140–149, Berlin, Germany. Association for Computational Linguistics.

Han He and Jinho D. Choi. 2021. The stem cell hypothesis: Dilemma behind multi-task learning with transformer encoders. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5555–5577, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

Matic Horvat, Ann Copestake, and Bill Byrne. 2015. Hierarchical statistical semantic realization for Minimal Recursion Semantics. In *Proceedings of the 11th International Conference on Computational Semantics*, pages 107–117, London, UK. Association for Computational Linguistics.

Mark Johnson and Ewan Klein. 1986. Discourse, anaphora and parsing. In *Proceedings of the 11th Coference on Computational Linguistics*, COLING '86, page 669–675, USA. Association for Computational Linguistics.

Hans Kamp and U. Reyle. 1993. From discourse to logic: Introduction to model theoretic semantics of natural language, formal logic and discourse representation theory. *Language*, 71(4).

Robert T. Kasper. 1989. A flexible interface for linking applications to Penman's sentence generator. In *Speech and Natural Language: Proceedings of a Workshop Held at Philadelphia, Pennsylvania, February 21-23, 1989*.

Karin Kipper, Anna Korhonen, Neville Ryant, and Martha Palmer. 2006. Extending VerbNet with novel verb classes. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Jiangming Liu, Shay B. Cohen, and Mirella Lapata. 2019. Discourse representation structure parsing with recurrent neural networks and the transformer model. In *Proceedings of the IWCS Shared Task on Semantic Parsing*, Gothenburg, Sweden. Association for Computational Linguistics.

Jiangming Liu, Shay B. Cohen, Mirella Lapata, and Johan Bos. 2021. Universal Discourse Representation Structure Parsing. *Computational Linguistics*, 47(2):445–476.

Franz Josef Och and Hermann Ney. 2000. Improved statistical alignment models. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, ACL '00, page 440–447, USA. Association for Computational Linguistics.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Wessel Poelman, Rik van Noord, and Johan Bos. 2022a. Transparent semantic parsing with Universal Dependencies using graph transformations. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4186–4192, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Wessel Poelman, Rik van Noord, and Johan Bos. 2022b. Transparent semantic parsing with Universal Dependencies using graph transformations. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4186–4192, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Leonardo F. R. Ribeiro, Jonas Pfeiffer, Yue Zhang, and Iryna Gurevych. 2021. Smelting gold and silver for improved multilingual AMR-to-Text generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 742–750, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Minxing Shen and Kilian Evang. 2022. DRS parsing as sequence labeling. In *Proceedings of the 11th Joint Conference on Lexical and Computational Semantics*, pages 213–225, Seattle, Washington. Association for Computational Linguistics.

Rik van Noord, Lasha Abzianidze, Antonio Toral, and Johan Bos. 2018. Exploring neural methods for parsing discourse representation structures. *Transactions of the Association for Computational Linguistics*, 6:619–633.

Rik van Noord, Antonio Toral, and Johan Bos. 2019. Linguistic information in neural semantic parsing with multiple encoders. In *Proceedings of the 13th International Conference on Computational Semantics*

- *Short Papers*, pages 24–31, Gothenburg, Sweden. Association for Computational Linguistics.

Rik van Noord, Antonio Toral, and Johan Bos. 2020a. Character-level representations improve DRS-based semantic parsing even in the age of BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4587–4603, Online. Association for Computational Linguistics.

Rik van Noord, Antonio Toral, and Johan Bos. 2020b. Character-level representations improve DRS-based semantic parsing even in the age of BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4587–4603, Online. Association for Computational Linguistics.

Stephan Vogel, Hermann Ney, and Christoph Tillmann. 1996. HMM-based word alignment in statistical translation. In *COLING 1996 Volume 2: The 16th International Conference on Computational Linguistics*.

Chunliu Wang, Rik van Noord, Arianna Bisazza, and Johan Bos. 2021. Input representations for parsing discourse representation structures: Comparing English with Chinese. In *ACL/IJCNLP (2)*, pages 767–775. Association for Computational Linguistics.

Tianming Wang, Xiaojun Wan, and Hanqi Jin. 2020. AMR-to-text generation with graph transformer. *Transactions of the Association for Computational Linguistics*, 8:19–33.

Sheng Zhang, Xutai Ma, Kevin Duh, and Benjamin Van Durme. 2019. AMR parsing as sequence-to-graph transduction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 80–94, Florence, Italy. Association for Computational Linguistics.

A Result Plots

According to the fine-grained evaluation results, for both English and Chinese DRS parsing, relatively low f1 scores tend to appear in `Names` and `Concepts`. The performance of parser declined by approximately five percent after the named entity was converted to Chinese, especially the `adj` and `adv`, comparing **EN** with **ZH**.

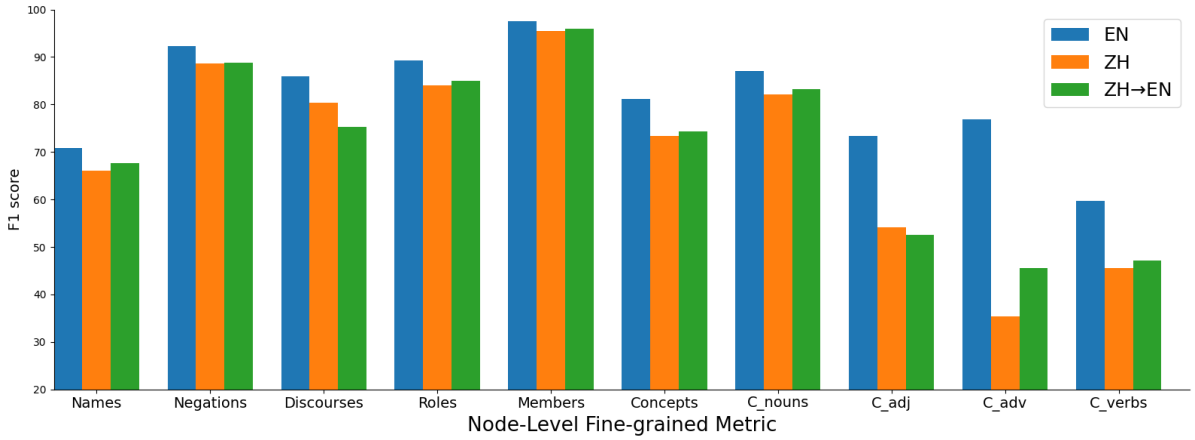

Figure 5: Fine-grained results among **EN**, **ZH**, and **ZH**→**EN**$_{zh}$ in Node-level.

B Output DRS

Number	Type	Example
No.1	EN Text	The music **lured everyone**.
	ZH Text	音乐吸引了所有人。
	ZH→EN	Music appeals to all.
	EN	music.n.01 NEGATION <1 person.n.01 NEGATION <1 surprise.v.02 Stimulus -2 Experiencer -1 Time +1 time.n.08 EQU now
	ZH	music.n.01 NEGATION <1 person.n.01 NEGATION <1 appeal.v.01 Agent -2 Theme -1 Time +1 time.n.08 TPR now
	ZH→EN$_{zh}$	event.v.01 Participant +1 music.n.01 appeal.v.01 Theme -1
	Gold DRS	music.n.01 NEGATION <1 person.n.01 NEGATION <1 lure.v.01 Agent -2 Patient -1 Time +1 time.n.08 TPR now
No.2	EN Text	She's **very handy** with a saw.
	ZH Text	她很会用锯子。
	ZH→EN	She is good with a saw.
	EN	female.n.02 time.n.08 EQU now very.r.01 handy.a.01 AttributeOf -3 Time -2 Degree -1 Instrument +1 saw.n.02
	ZH	female.n.02 time.n.08 TSU now use.v.01 Agent -2 Time -1 Theme +1 Instrument +2 entity.n.01 saw.n.02
	ZH→EN$_{zh}$	female.n.02 time.n.08 EQU now good.a.01 AttributeOf -2 Time -1 Instrument +1 saw.n.02
	Gold DRS	female.n.02 time.n.08 EQU now very.r.01 handy.a.03 AttributeOf -3 Time -2 Degree -1 Instrument +1 saw.n.02

Table 7: Examples of output DRSs by different parsers.