Semantic Parsing for English as a Second Language

Yuanyuan Zhao[1,2], **Weiwei Sun**[1,3], **Junjie Cao**[1*] **and Xiaojun Wan**[1]
[1]Wangxuan Institute of Computer Technology, Peking University
[1]The MOE Key Laboratory of Computational Linguistics, Peking University
[2]Academy for Advanced Interdisciplinary Studies, Peking University
[3]Center for Chinese Linguistics, Peking University
{zhao_yy,ws,wanxiaojun}@pku.edu.cn
junjie.junjiecao@alibaba-inc.com

Abstract

This paper is concerned with semantic parsing for English as a second language (ESL). Motivated by the theoretical emphasis on the learning challenges that occur at the syntax-semantics interface during second language acquisition, we formulate the task based on the divergence between literal and intended meanings. We combine the complementary strengths of English Resource Grammar, a linguistically-precise hand-crafted deep grammar, and TLE, an existing manually annotated ESL UD-TreeBank with a novel reranking model. Experiments demonstrate that in comparison to human annotations, our method can obtain a very promising SemBanking quality. By means of the newly created corpus, we evaluate state-of-the-art semantic parsing as well as grammatical error correction models. The evaluation profiles the performance of neural NLP techniques for handling ESL data and suggests some research directions.

1 Introduction

There are more people around the world learning English as a second language (ESL) than there are native speakers of English with this gap continually and steadily expanding (Crystal, 2012). Accordingly, an extremely large volume of non-native English texts are generated every day. We need an automatic machinery to annotate such large-scale *atypical* data with in-depth linguistic analysis. High-performance automatic annotation of learner texts, from an engineering point of view, enables it possible to derive high-quality information by structuring the specific type of data, and from a scientific point of view, facilitates quantitative studies for Second Language Acquisition (SLA), which is complementary to hands-on experiences in interpreting interlanguage phenomena (Gass, 2013). This direction has been recently explored by the NLP community (Nagata and Sakaguchi, 2016; Berzak et al., 2016a; Lin et al., 2018).

Different from *standard* English, ESL may preserve many features of learners' first languages[1]. The difference between learner texts and benchmark training data, e.g. Penn TreeBank (PTB; Marcus et al., 1993), is more related to linguistic competence, rather than performance (Chomsky, 2014). This makes processing ESL different from almost all the existing discussions on domain adaptation in NLP.

Despite the ubiquity and importance of interlanguages at both the scientific and engineering levels, it is only partially understood how NLP models perform on them. In this paper, we present, to the best of our knowledge, the first study on Semantic Parsing for English as a Second Language. Motivated by the *Interface Hypothesis* (Sorace, 2011) in SLA, we emphasize on the divergence between literal and intended meanings. To obtain reliable semantic analyses in order to represent the two types of meanings, we propose to combine English Resource Grammar (Flickinger, 2000), which is a wide-coverage, linguistically-precise, hand-crafted grammar and TLE, which is a manually annotated syntactic treebank for ESL in the Universal Dependency (UD; Berzak et al., 2016b) framework. In particular, we introduce a reranking model which utilizes the partial constraints provided by gold syntactic annotations to disambiguate among the grammar-licensed candidate analyses. Experiments on DeepBank (Flickinger et al., 2012) demonstrates the effectiveness of our proposed model.

By means of the newly created corpus, we study semantic parsing for ESL, taking Elementary De-

*Now works at Alibaba Group.

[1]Henceforth, the first and second language are referred to as L1 and L2, respectively.

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 6783–6794
July 5 - 10, 2020.

pendency Structure (EDS; Oepen and Lønning, 2006) as the target representation. We probe the semantic parsing of multiple state-of-the-art neural parsers for literal meaning and intended meaning, and investigate how grammatical error correction (GEC) can contribute to the parsing. In addition, we give a detailed analysis of the effect from grammatical errors. Results reveal three facts: 1) semantic parsing is sensitive to non-canonical expressions, and the distribution as well as types of grammatical errors have an effect on parsing performance; 2) Factorization-based parser is the most effective and robust parser to process learner English; and 3) automatic GEC has a positive, but limited influence on the parsing of intended meaning.

2 Related Work

Early work regarding the collection of learner corpora mainly concentrates on tagging alleged errors (Rozovskaya and Roth, 2010; Nagata et al., 2011). The past decade has seen a tendency to directly annotate the linguistic properties in learner sentences (Dickinson and Ragheb, 2009; Dıaz-Negrillo et al., 2010; Rastelli, 2013). The lack of precisely annotated data has limited the systematic analysis of interlanguages.

There are several attempts to set up annotation schemes for different linguistic layers of learner languages, such as POS tags and syntactic information (Hirschmann et al., 2007; Dıaz-Negrillo et al., 2010; Rosen et al., 2014; Nagata and Sakaguchi, 2016; Berzak et al., 2016b). But it is challenging to elucidate the exact definition of "syntax" for learner languages. Ragheb and Dickinson (2012) defines multiple layers (morphological dependencies, distributional dependencies, and subcategorization) based on different evidence to capture non-canonical properties. Similarly, motivated by the *Interface Hypothesis* (Sorace, 2011), we employ a principled method to create parallel semantic representations for learner English by discriminating between the literal and intended meanings.

With regard to the semantic analysis for learner languages, Lin et al. (2018) takes the first step in this direction. Based on a parallel semantic role labeling (SRL) corpus, they prove the importance of syntactic information to SRL for learner Chinese. In this paper, we provide a much deeper semantic analysis for learner English.

3 Literal versus Intended Meaning

There is a classic distinction between two aspects of meaning: the literal meaning (conventional meaning or sentence meaning) versus the intended meaning (speaker meaning or interpretation). The former puts an emphasis on the linguistic code features appearing in the sentence, while the latter is derived from the author's intention. When we consider an interlanguage, the divergence between literal and intended meanings is much larger due to various cross-lingual influences. It is reasonable to consider both aspects to develop a principled method to process outputs from L2 learners.

3.1 SLA at the Syntax-Semantics Interface

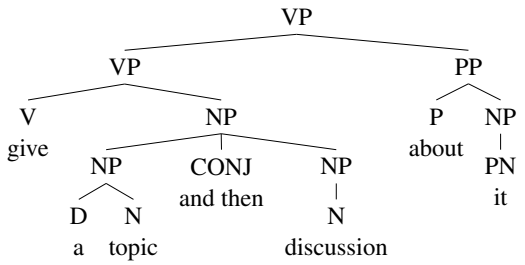
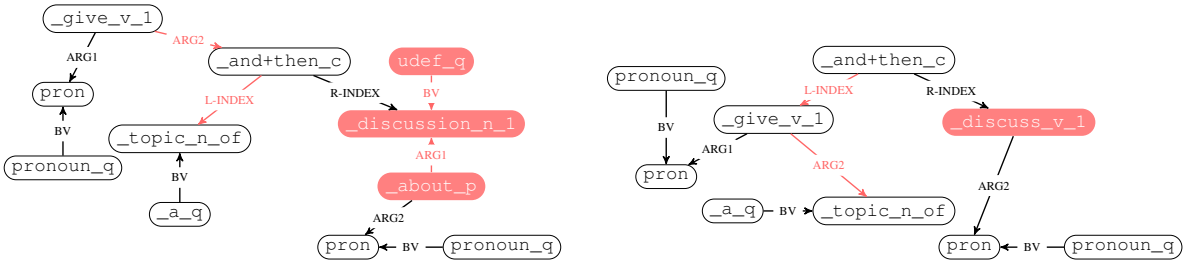


Figure 1: A plausible syntactic analysis of *give a topic and then discussion about it*. The example is from the *TLE* corpus. The corrected counterpart of this fragment in *TLE* is *Give a topic and then discuss it.*

Contemporary research on SLA has extensively argued and empirically supported the claim that linguistic properties pertaining to the interface between syntax and other linguistic modules are vulnerable in L2 and integrating linguistic phenomena relevant to such interfaces imposes much difficulty to L2 learners (Sorace, 2006; White, 2011). According to this view, the interaction or mapping between syntactic and semantic representations is less likely to be acquired completely than structures within one single module, either syntactic or semantic. With respect to outputs of L2 learners, mismatches between syntactic structures and intended meanings are frequently observable.

Figure 1 presents an example from the TLE corpus. Although *discussion* is misused, the whole fragment is grammatical and thus interpretable according to syntactic analysis. However, the literal meaning along with a sound syntactic analysis is far from the intended meaning that a native speaker can infer from intra- and inter-sentence contexts. It is quite obvious that *discussion* should be regarded as a verb coordinating with *give*.



(a) Literal meaning

(b) Intended meaning

Figure 2: Semantic analysis of the fragment *give a topic and then discussion about it*, where the contrastive parts are colored. The analysis is based on English Resource Semantics. Nodes represent concepts, while edges represent semantic dependencies. Following morphosyntax, "discussion" acts as the conjunct of the previous noun "topic". However, according to discourse, it should be juxtaposed with the verb "give" because these are two successive actions.

3.2 Importance of Parallel Representations

The application scenarios of both literal and intended meanings are practiced in accordance with their different emphases. For example, extracting literal meanings according to the morphosyntactic forms are more useful for text quality assessment tasks in computer-assisted language learning, such as content-based automatic essay scoring. On the contrary, the intended meaning-centric representations help figure out logical relationships and may benefit text mining applications like relation extraction.

3.3 Building a Parallel L2-L1 SemBank

In order to comprehensively study the issue, we consider both literal and intended meanings. To conduct quantitative research, we create two versions of high-quality *silver* data and provide a two-sided evaluation for the semantic parsing on learner English.

3.3.1 Target Meaning Representation

English Resource Semantics (ERS; Flickinger et al., 2016) is an important resource of semantic representations produced by the English Resource Grammar (ERG; Flickinger, 1999), a broad-coverage, linguistically motivated precision Head-Driven Phrase Structure Grammar (HPSG; Pollard and Sag, 1994) of English (Flickinger, 2000, 2011). It provides rich semantic representations including the semantic roles and other detailed information such as the scope of quantifiers and scopal operators including negation, as well as semantic representations of linguistically complex phenomena such as time and date expressions, conditionals, and comparatives (Flickinger et al., 2014). ERS helps to reveal much deeper semantic analysis than other shallow target structures such as the predicate-argument relations in the semantic role labeling (SRL) task. Moreover, it can be derived into several different forms, like the logical-form-based representation Minimal Recursion Semantics (MRS) and the graph-shaped structure Elementary Dependency Structures (EDS). We resort to this resource to build an informative analysis for learner English and choose EDS as the target structure.

Figure 2 shows the two kinds of semantic analysis of our running example.

3.3.2 SemBanking with ERG

As there is no gold semantics-annotated corpus for learner English and building such a corpus from scratch is tedious and time-consuming, we exploit ERG to establish a large-scale sembanking with informative semantic representations. To be specific, for each input sentence S, we generate K-best semantic graphs $G_1, G_2, ..., G_K$ with an ERG-based processor, i.e. ACE[2]. The created grammar-licensed analyses contain both a derivation tree recording the used grammar rules and lexical entries, and the associated semantic representation constructed compositionally via this derivation (Bender et al., 2015). The elaborate grammar rules enable sembanking reusable, automatically derivable and task-independent, and it can benefit many NLP systems by incorporating domain-specific knowledge and reasoning.

[2] http://sweaglesw.org/linguistics/ace/

3.3.3 Reranking ERG Analyses with Gold UD

Previous work has proved that high-quality syntax makes a large impact on semantic parsing tasks such as SRL (Hermann and Blunsom, 2013; He et al., 2017; Qian et al., 2017). The exploratory work in Lin et al. (2018) draws the same conclusion in an L2 situation. We assume that the incorporation of syntactic trees helps improve the quality of our evaluation data.

We conduct a reranking procedure on the K-best candidates derived under the ERG framework with the aid of gold Universal Dependencies (UD; Berzak et al., 2016b) trees and select the graph which best fits into the gold syntactic tree (represented as T). Our reranking model can be formulated into:

$$\hat{G} = \arg\max_{1 \leqslant i \leqslant K} \ \textsc{score}(G_i, T)$$

where $\textsc{score}(G_i, T)$ is a numerical measurement of the matching between G_i and T. Here, we define it as follows:

$$\textsc{score}(G_i, T) = W^T \mathcal{F}(f_{G_i}, f_T)$$

where W refers to the parameter matrix and $\mathcal{F}$ is the function to calculate the coherency between feature vectors f_{G_i} and f_T, which can resort to neural encoders or feature engineering. Here, we use feature engineering which outperformed Graph Neural Network (GNN; Scarselli et al., 2008) in the pilot study to encode the discrete properties in the graph and the UD tree. During the training process, there is a gold semantic graph G_g for S. By going through all the K graphs, we can pick out graph G_p with the highest score $\textsc{score}(G_p, T)$. Our goal is to ensure $\textsc{score}(G_g, T) \geqslant \textsc{score}(G_p, T)$, which can be achieved with the help of the averaged structured perceptron learning algorithm.

3.3.4 Effectiveness of the Reranking Model

To evaluate the capability of our proposed reranking model, we randomly extract 10,000 and 2,476 sentences from DeepBank (Flickinger et al., 2012) as the training and validation data respectively. The gold UD analyses are derived from the original PTB (Marcus et al., 1993) annotations. With regard to evaluation metrics, we use SMATCH (Cai and Knight, 2013) and Elementary Dependency Matching (EDM; Dridan and Oepen, 2011). Results are shown in Table 1. The first three rows demonstrates that the parsing performance has been greatly improved after reranking, proving the power of the proposed model. The larger K is set to, the greater the improvement will be, since the search space has been expanded. Results of "Oracle" provide the upper bound. The high numerical value demonstrates the potential of reranking method. The results also prove that syntactic information does facilitate the semantic analysis, which is in line with previous studies.

	SMATCH			EDM
	Node	Edge	All	All
Top-1	92.8	90.0	91.4	87.8
Rerank (50)	94.7	93.4	94.1	92.0
Rerank (500)	95.1	93.9	94.5	92.7
Oracle (50)	97.6	96.9	97.2	95.6
Oracle (500)	98.7	98.5	98.6	97.6
Inter-Annotator Agreement	–	–	–	94-95

Table 1: Results of reranking. "Top-1" means the most preferable graph generated by the ACE parser. "Rerank (50)" and "Rerank (500)" means that K is set to 50 and 500 during reranking respectively. "Oracle" means directly selecting the best-performing graph for each sentence from the K-best list. The inter-annotator agreement of EDM is reported in Bender et al. (2015).

3.3.5 The Data

The Treebank of Learner English (TLE; Berzak et al., 2016a) is a collection of 5,124 ESL sentences, manually annotated with POS tags and dependency trees according to Universal Dependencies (UD; Nivre et al., 2016) framework. Both *original* sentences which contain grammatical errors and *corrected* sentences which are revised by native speakers are provided to constitute a parallel corpus. The *corrected* sentences are reconstructed based on a target hypothesis. Following the idea of parallel semantic representations, we produce two versions of *silver* semantic annotation for learner English. The first version of annotation is obtained by processing the *original* sentences in TLE with the sembanking-reranking pipeline. Henceforth, this will be called *L-silver*. It concentrates on the morphosyntactic features encoded in the sentences. Then we process the *corrected* sentences in the same way and call the produced semantic graphs *I-silver*, henceforth. In this case, we give priority to the intended meaning.

Node	Edge	All
86.27	86.68	86.48

Table 2: SMATCH scores between the parallel meaning representations.

During the process of building the corpus, a part of the sentences from TLE are excluded. With the elaborate semantic representations, ERG fails to analyse sentences which are too long or contain particular unknown words/constructions within a certain time limit. The coverage of ACE on *original* sentences and *corrected* sentences from TLE is 55.39% and 79.63%, respectively. In addition, a further reduction of coverage is caused by the inconsistent tokenization between the ERG-licensed analysis and the TLE annotation, such as the different treatment of apostrophes. Ultimately, 52.50% *original* sentence and 73.54% *corrected* sentences are processed, forming the final data. This may introduce bias, and how to include the rest part of sentences is left for future research.

3.4 A Quantitative Analysis of the Divergence

The parallel meaning representations focus on different linguistic layers. Previous studies on the relevance of the two kinds of meanings are mostly based on psycholinguistic methods. We propose to measure the similarity in a quantitative manner with a corpus-based approach. The literal and intended meanings are represented as the semantic graphs in *L-silver* and *I-silver*, respectively. Since the sentences are parallel, we can compare the graph structures directly. We use SMATCH (Cai and Knight, 2013) as the evaluation metric which provides the token-wise evaluation along with effective explorations of variable alignments. The numerical results are displayed in Table 2. The modest SMATCH scores indicate the existence of great divergence between the literal and intended meaning representations.

4 Two State-of-the-art Parsers

Existing work in data-driven semantic graph parsing can be roughly divided into four types, namely composition-, factorization-, transition- and translation-based ones (Koller et al., 2019). According to experimental results obtained on benchmark datasets with various target structures including Abstract Meaning Representation(AMR; Langkilde and Knight, 1998; Banarescu et al., 2013), Elementary Dependency Structures (EDS; Oepen and Lønning, 2006), Semantic Dependency Parsing (SDP) as well as Universal Conceptual Cognitive Annotatio (UCCA; Abend and Rappoport, 2013), the composition- and factorization-based approaches are the leading approaches obtained by now (Lindemann et al., 2019; Zhang et al., 2019). In this paper, we use these two kinds of parsers (composition- and factorization-based parsers) described in Chen et al. (2019) as state-of-the-art representatives.

Following the principle of compositionality, a semantic graph can be viewed as the result of a derivation process, in which a set of lexical and syntactico-semantic rules are iteratively applied and evaluated. The core engine of the composition-based parser is a graph rewriting system that explicitly explores the syntactico-semantic recursive derivations that are governed by a Synchronous Hyperedge Replacement Grammar (SHRG; Chen et al., 2018b). The parser constructs DMRS graphs by explicitly modeling such derivations. It utilizes a constituent parser to build a syntactic derivation, and then selects semantic HRG rules associated to syntactic CFG rules to generate a graph. When multiple rules are applicable for a single phrase, a neural network is used to rank them. We use the parser in Chen et al. (2019) based on both the lexicalized grammar and the constructional grammar (refer to Chen et al. (2018b) for the distinction). Henceforth, they are called lexicalized and constructional composition-based parsers respectively.

Figure 3 shows an example of the SHRG-based syntactico-semantic derivation from the constructional composition-based parser. The derivation can be viewed as a syntactic tree enriched with semantic interpretation rules that are defined by an HRG. Each phrase in the syntactic tree is assigned with a sub-graph of the final semantic structure. Moreover, some particular nodes in a sub-graph are marked as *communication channels* to other meaning parts in the same sentence. In HRG, these nodes are summarized as a hyperedge. Two sub-graphs are glued according to a construction rule following the graph substitution principle of HRG.

The factorization-based parser explicitly models the target semantic structures by defining a score function that is able to evaluate the *goodness* of any candidate graph. It needs to know how to find the highest-scoring graph from a large set of

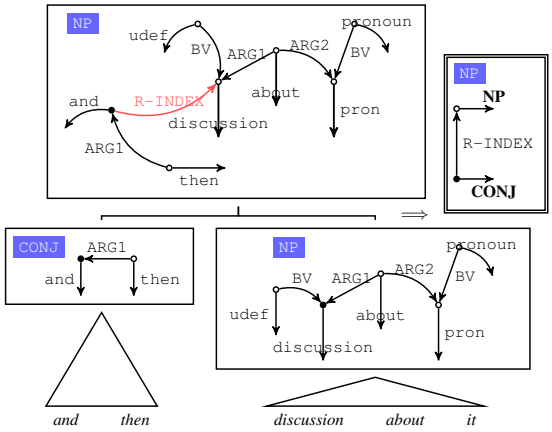


Figure 3: An SHRG-based syntactico-semantic derivation from the composition-based parser. Each phrase in the syntactic tree ("and then" and "discussion about it") is assigned with a sub-graph of the final semantic structure, as illustrated in the boxes. Some particular nodes (filled nodes) in a sub-graph are marked as *communication channels* to other meaning parts. According to the construction rule (shown in double-framed box), we glue the two sub-parts via the filled nodes, forming a larger graph with the syntactic label "NP". More details are illustrated in Chen et al. (2019)

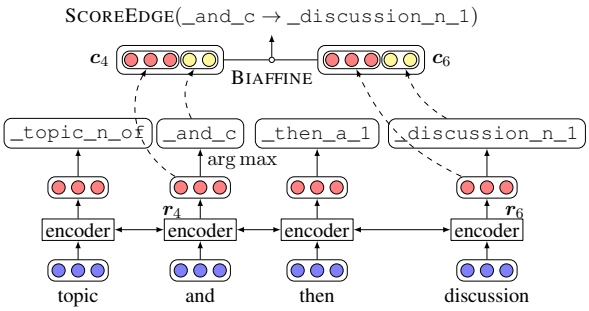


Figure 4: The network architecture for the factorization-based parser. Textual embeddings (in red) are used to identify concepts and also to determine dependency relations together with the resulted conceptual embeddings (in yellow).

possible candidates. The parser works with a two-stage pipeline structure, for concept identification and relation detection, as illustrated in Figure 4. In the first phase, sequence labeling models are used to predict nodes, and in the second phase, we utilize the dependency model introduced by Dozat and Manning (2018) to link nodes. The two models in both stages use a multi-layer BiLSTM to encode tokens. In the first stage, another softmax layer is utilized to predict concept-related labels, while in the second stage, the dependency model is utilized to calculate a score for selecting token pairs.

5 Parsing to Literal Meanings

5.1 Robustness of Parsing Models

We experiment with three different parsers introduced in last section, i.e., lexicalized and constructional composition-based parsers and the factorization-based parser. We train these parsers on DeepBank version 1.1, corresponding to ERG 1214, and use the standard data split. In order to examine the robustness of parsing models, we test on both L1 and L2 sentences.

Detailed results are shown in Table 3. The parsing performances are depicted by SMATCH scores with regard to nodes, edges and the overall view. Comparing different models, we can see that the factorization-based approach performs better on all setups, which is consistent with previous studies (Koller et al., 2019). The gap between results on DeepBank and the other two datasets demonstrates the existence of cross-domain effect, which has been observed in plenty of NLP tasks, including but not limited to semantic parsing (Chen et al., 2018a; Lindemann et al., 2019; Blitzer and Pereira, 2007; Ben-David et al., 2010; Elsahar and Gallé, 2019). Furthermore, it is clear that there is a drop from L1 to L2 data. The gap is marked in the last row, the average of which is about 4 points, indicating the insufficiency of using standard models to parse learner texts.

Still, the factorization-based model yields a little bit more robust results on non-native data. We hold that the poor performance of composition-based model is caused by the explicit syntactico-semantic derivation process. Since the interface between syntax and semantics of learner languages is somewhat unclear, directly applying rewriting rules extracted from L1 data may be partly misleading.

5.2 Relatedness to Grammatical Errors

It is crucial to understand whether and to what extent parsers are indeed robust to learner errors. We re-analyse the results from two aspects. First, we modify the original SMATCH evaluation metric and enable it to be sensitive to distances from errors. Then we make a distinction among typical error types proposed in CoNLL-2014 Shared Task (Ng et al., 2014). Results show that standard parsers can not handle learner errors well enough and their behaviors vary among different

Data	LEX			CXG			FAC		
	Node	Edge	All	Node	Edge	All	Node	Edge	All
DeepBank	94.05	92.96	93.50	95.83	92.87	94.34	96.85	95.19	96.01
L1	88.41	86.44	87.41	90.32	86.04	88.14	92.28	89.12	90.91
L2	84.38	82.23	83.29	86.47	81.70	84.04	88.68	84.45	86.91
Δ	4.03	4.21	4.12	3.85	4.34	4.10	3.60	4.67	4.00

Table 3: SMATCH scores of semantic parsing on different test data. Henceforth, *LEX*, *CXG* and *FAC* refer to lexicalized and constructional composition-based parsers and the factorization-based parser, respectively. Δ refers to the gap between *L1* and *L2*.

Model	Data	Node	Edge	All
LEX	✓	86.31	81.95	84.23
	×	68.94	79.81	75.74
CXG	✓	89.04	82.14	85.75
	×	71.46	79.77	76.66
FAC	✓	90.96	84.48	87.86
	×	73.55	80.27	77.75

Table 4: ✓ refers to error-ignored ($\sigma_k = 0$ when the kth triple in G_g is involved with errors, $\sigma_k = 1$ otherwise) SMATCH scores while × refers to error-oriented ($\sigma_k = 1$ when the kth triple in G_g is involved with errors, $\sigma_k = 0$ otherwise) SMATCH scores.

error types.

It should be noticed that only several points in a sentence are occupied by errors while most of the structure is still well-formed. The scores of *L2* in Table 3 may be not able to exactly reflect the robustness of models. Therefore, we modify the original SMATCH evaluation metric by paying additional attention to erroneous points. The original metric can be formulated into an Integer Linear Programming (ILP) problem. Suppose there are gold and predicted graphs G_g (m variables) and G_p (n variables). Semantic relations in graphs are represented as triples which can illustrate both the concepts (represented as (`variable`, `concept`, `relation`)) and edges (represented as (`variable1`, `variable2`, `relation`)). We define $v_{ij} = 1$ iff the ith variable in G_g is mapped to the jth variable in G_p in the current alignment, $v_{ij} = 0$ otherwise. We have $t_{kl} = 1$ iff the kth triple (`x`, `y`, `relation1`) in G_g and the lth triple (`w`, `z`, `relation2`) in G_p are matched, which means $v_{xw} = 1$, $v_{yz} = 1$ and `relation1=relation2`. In the original metric, t_{kl} takes the value of 1 or 0 and all triple pairs are treated equally. In order to focus on the erroneous points, we put various weights on different triple pairs depending on their distance from errors. Then the optimization problem can be stated as:

$$
\begin{aligned}
\max \quad & \sum_{kl} \sigma_k t_{kl} \\
s.t. \quad & \textstyle\sum_j v_{ij} \leq 1, && i = 1, 2, 3 \ldots, m \\
& \textstyle\sum_i v_{ij} \leq 1, && j = 1, 2, 3 \ldots, n \\
& t_{r_{xy} r_{wz}} \leq v_{xw}, \\
& t_{r_{xy} r_{wz}} \leq v_{yz}, && r_{xy} r_{wz} \in \mathcal{R}
\end{aligned}
$$

Here, r_{xy} means the triple describing the relationship between x and y, and $\mathcal{R}$ means the set of all triple pairs. σ_k refers to the weight of the kth triple in G_g. If we want to explore the performance on erroneous points, triples related to these points will be assigned a larger weight. If we want to find out the performance on *good* part, we can just set the weight of triples involved with errors to zero.

Table 4 compares the error-oriented and error-ignored results. We can see that although the average gap in Table 3 is about 4 points, the actual performance pertaining to the ill-formed part is much lower. Especially, the F-score of nodes drops heavily. The gray line in Figure 6 illustrates the tendency of scores changing with the distance from abnormal points. It clearly shows that farther nodes suffer less.

Moreover, we explore the relationship between learner errors (LEs) and parsing errors (PEs). We find that **21.40%** PEs are caused by LEs and **66.80%** LEs cause at least one PE. It indicates that parsing models are really struggling with learner errors.

Furthermore, we look into the produced graphs with regard to different error types. We refer to the list of error types introduced in the CoNLL-2014 Shared Task (Ng et al., 2014). Detailed results are illustrated in Figure 5. This dia-

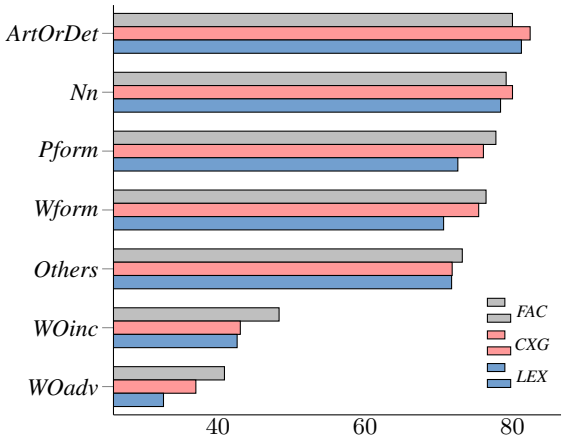


Figure 5: Overall SMATCH scores with regard to different grammatical error types. Detailed descriptions of errors are provided in Ng et al. (2014).

Model	$F_{0.5}$
Chollampatt and Ng (2018)	45.36
Zhao et al. (2019)	61.15

Table 5: Performances of the two GEC models on CoNLL-2014 test set.

gram reflects a clear comparison among different error types. The four best-performing types are *ArtOrDet*, *Nn*, *Pform* and *Wform*, referring to errors of article or determinier, noun number, pronoun form and general word form, respectively. We can see that most of them are related to morphological variations and can be disambiguated at the word level. In contrast, *WOadv* and *WOinc*, meaning incorrect adjective/adverb order and other word order errors, are much more complex. They are involved with reorganizations of the sentence structure and hence more difficult to handle. Factorization-based model is more robust to these hard cases than composition-based models since it is grounded on graph structures and can reduce the influence from broken sequential syntax.

6 Parsing to Intended Meanings

Previous evaluation indicates the difficulty to adopt a *standard* semantic parsing model to handle competence errors. Motivated by this fact, we are concerned with whether it is feasible to automatically normalize the texts first. Specifically, our strategy is correcting the grammatical errors contained in the input sentences, and then parsing the revised texts into semantic structures with standard models. The first step can resort to Grammatical Error Correction (GEC), the task of correcting different kinds of errors in text. It has attracted a lot of attention and considerable effort has been made to promote the performance on specific benchmark data. We utilize two off-the-shelf GEC models. One is a multilayer convolutional encoder-decoder neural network proposed in Chollampatt and Ng (2018). We choose the basic model introduced in the paper. The other model copies the unchanged words from the source sentence to the target sentence using a pretrained copy-augmented architecture with a denoising auto-encoder (Zhao et al., 2019). It achieves the state-of-the-art performance without extra pseudo data. Performances of the two GEC models on CoNLL-2014 test set are shown in Table 5.

We train the factorization-based model on DeepBank and examine the performance on L2 and L1 sentences as well as the revised sentences by two GEC models. The produced graphs are compared with *I-silver* which represents the intended meaning. We notice that during the computation of SMATCH, some disagreements of nodes result from the discrepancy of morphological variation or different collocations between the input and the standard sentence. Hence the node score may be underestimated. Therefore, we relax the standards of matching nodes. We establish a paraphrase table based on the statistical machine translation between a parallel learner corpus[3]. As long as the labels of two aligned nodes have the same stem or they form a paraphrase pair in our table, then the two nodes can be considered "matching". We call the new evaluation metric as "node-relaxed SMATCH".

Table 6 summarizes the results. The gap between the first and the last rows demonstrates that it may be difficult to automatically infer the intended meaning based on the literal representation. GEC does help us to understand the learner English, but it seems to be a small step on the progress bar. Although the second GEC model (Zhao et al., 2019) outperforms the first model (Chollampatt and Ng, 2018) a lot on benchmark data (Table 5), its superiority on semantic parsing is not so obvious. There is still a long way to go before automatically capturing the intended mean-

[3] https://sites.google.com/site/naistlang8corpora/

Test Data	Standard			Error-oriented			Node-relaxed		
	Node	Edge	All	Node	Edge	All	Node	Edge	All
L2 sentence	83.91	84.86	84.39	45.91	78.93	66.73	57.18	78.45	70.59
Chollampatt and Ng (2018)	84.98	85.13	85.06	53.06	80.06	70.08	62.05	79.26	72.90
Zhao et al. (2019)	86.10	85.85	85.98	58.73	80.56	72.49	65.39	80.09	74.66
L1 sentence	92.28	89.60	90.92	86.08	85.72	85.85	89.64	85.48	87.02

Table 6: Results of SMATCH scores compared to *I-silver*. Chollampatt and Ng (2018) and Zhao et al. (2019) mean the revised sentences with GEC models introduced in the two studies. "Error-oriented" means only focusing on the parts aligned to grammatical errors in *I-silver*. "Node-relaxed" is an error-oriented metric that relax the standards of matching nodes.

ing like humans.

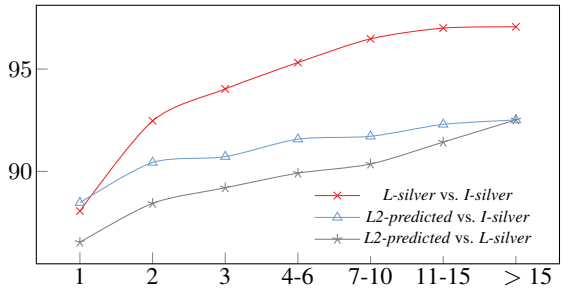


Figure 6: Overall SMATCH scores with regard to the distance from errors. *L-silver* and *I-silver* mean the silver standards of literal and intended meanings, respectively. *L2-predicted* refers to the predicted semantic graphs produced by neural parsers on L2 sentences.

In order to figure out to what extent grammatical errors influence the *good* part and hence the whole sentence structure, we draw curves concerning distance from errors, which is displayed in Figure 6. The red line compares the two kinds of silver representations, which indicates the deviation from the intended meaning due to ungrammatical expressions. It appears as a smooth curve which goes steadily up. The overall trend indicates that the damage to farther parts from errors is less extensive. We assume that the propagation process is limited by the syntactic architecture. However, the situation of automatically predicted graphs by neural models is slightly different. It is depicted by the blue line in Figure 6 and the gradient is much smaller. We suggest it results from the great power of neural models to encode contextual information. In the L2 circumstance, while such characteristic enables the encoder to capture long-distance dependencies, it also expands the scope of errors' influence.

7 Conclusion and Future Work

In this paper, we formulate the ESL semantic parsing task based on the divergence on literal and intended meanings. We establish parallel meaning representations by combining the complementary strengths of knowledge-intensive ERG-licensed analysis and dependency tree annotations through a new reranking model. For literal meaning, we probe the semantic parsing of multiple state-of-the-art neural parsers and give detailed analysis of effects from grammatical errors. For intended meaning, we investigate how grammatical errors affect the understanding of sentences as well as how grammatical error correction (GEC) can contribute to the parsing. Results reveal three facts: 1) semantic parsing is sensitive to non-canonical expressions, and the parsing performance varies with regard to the distribution as well as types of grammatical errors; 2) Factorization-based parser is the most promising parser to process learner English; and 3) GEC has a positive, but limited influence on the parsing of intended meaning.

This paper shows a pilot study on the semantic parsing for learner language. Future research may involve tailoring existing parsers to learner data, combining literal and intended meanings in a unified framework, evaluating GEC models in terms of speakers' intention and parsing for other languages.

Acknowledgement

This work is supported in part by the National Hi-Tech R&D Program of China (No. 2018YFB1005100). We thank the anonymous reviewers and the area chair for their useful feedback and suggestions, Yufei Chen and Yajie Ye for providing the parsers and Ben Roberts for proofreading. Weiwei Sun is the corresponding author.

References

Omri Abend and Ari Rappoport. 2013. Universal conceptual cognitive annotation (ucca). In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 228–238.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for sembanking. In *Proceedings of the 7th linguistic annotation workshop and interoperability with discourse*, pages 178–186.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. 2010. A theory of learning from different domains. *Machine learning*, 79(1-2):151–175.

Emily M Bender, Dan Flickinger, Stephan Oepen, Woodley Packard, and Ann Copestake. 2015. Layers of interpretation: On grammar and compositionality. In *Proceedings of the 11th international conference on Computational Semantics*, pages 239–249.

Yevgeni Berzak, Jessica Kenney, Carolyn Spadine, Jing Xian Wang, Lucia Lam, Keiko Sophie Mori, Sebastian Garza, and Boris Katz. 2016a. Universal dependencies for learner english. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 737–746, Berlin, Germany. Association for Computational Linguistics.

Yevgeni Berzak, Jessica Kenney, Carolyn Spadine, Jing Xian Wang, Lucia Lam, Keiko Sophie Mori, Sebastian Garza, and Boris Katz. 2016b. Universal dependencies for learner english. *arXiv preprint arXiv:1605.04278*.

John Blitzer and Fernando Pereira. 2007. Domain adaptation of natural language processing systems. *University of Pennsylvania*, pages 1–106.

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)*, volume 2, pages 748–752.

Yufei Chen, Sheng Huang, Fang Wang, Junjie Cao, Weiwei Sun, and Xiaojun Wan. 2018a. Neural maximum subgraph parsing for cross-domain semantic dependency analysis. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 562–572.

Yufei Chen, Weiwei Sun, and Xiaojun Wan. 2018b. Accurate shrg-based semantic parsing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 408–418. Association for Computational Linguistics.

Yufei Chen, Yajie Ye, and Weiwei Sun. 2019. Peking at mrp 2019: Factorization-and composition-based parsing for elementary dependency structures. In *Proceedings of the Shared Task on Cross-Framework Meaning Representation Parsing at the 2019 Conference on Natural Language Learning*, pages 166–176.

Shamil Chollampatt and Hwee Tou Ng. 2018. A multilayer convolutional encoder-decoder neural network for grammatical error correction. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*.

Noam Chomsky. 2014. *Aspects of the Theory of Syntax*, volume 11. MIT press.

David Crystal. 2012. *English as a global language*. Cambridge university press.

Ana Dıaz-Negrillo, Detmar Meurers, Salvador Valera, and Holger Wunsch. 2010. Towards interlanguage pos annotation for effective learner corpora in sla and flt. In *Language Forum*, volume 36, pages 139–154.

Markus Dickinson and Marwa Ragheb. 2009. Dependency annotation for learner corpora. In *Proceedings of the Eighth Workshop on Treebanks and Linguistic Theories (TLT-8)*, pages 59–70.

Timothy Dozat and Christopher D. Manning. 2018. Simpler but more accurate semantic dependency parsing. page fromto484490.

Rebecca Dridan and Stephan Oepen. 2011. Parser evaluation using elementary dependency matching. In *Proceedings of the 12th International Conference on Parsing Technologies*, pages 225–230.

Hady Elsahar and Matthias Gallé. 2019. To annotate or not? predicting performance drop under domain shift. 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 2163–2173.

Dan Flickinger. 1999. The english resource grammar.

Dan Flickinger. 2000. On building a more effcient grammar by exploiting types. *Natural Language Engineering*, 6(1):15–28.

Dan Flickinger. 2011. Accuracy vs. robustness in grammar engineering. *Language from a cognitive perspective: Grammar, usage, and processing*, 201:31–50.

Dan Flickinger, Emily M Bender, and Stephan Oepen. 2014. Towards an encyclopedia of compositional semantics: Documenting the interface of the english resource grammar. In *LREC*, pages 875–881.

Dan Flickinger, Emily M. Bender, and Woodley Packard. 2016. English resource semantics. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, pages 1–5, San Diego, California. Association for Computational Linguistics.

Dan Flickinger, Yi Zhang, and Valia Kordoni. 2012. Deepbank. a dynamically annotated treebank of the wall street journal. In *Proceedings of the 11th International Workshop on Treebanks and Linguistic Theories*, pages 85–96.

Susan M Gass. 2013. *Second language acquisition: An introductory course*. Routledge.

Luheng He, Kenton Lee, Mike Lewis, and Luke Zettlemoyer. 2017. Deep semantic role labeling: What works and whatâĂŹs next. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 473–483.

Karl Moritz Hermann and Phil Blunsom. 2013. The role of syntax in vector space models of compositional semantics. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 894–904.

Hagen Hirschmann, Seanna Doolittle, and Anke Lüdeling. 2007. Syntactic annotation of non-canonical linguistic structures.

Alexander Koller, Stephan Oepen, and Weiwei Sun. 2019. Graph-based meaning representations: Design and processing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 6–11, Florence, Italy. Association for Computational Linguistics.

Irene Langkilde and Kevin Knight. 1998. Generation that exploits corpus-based statistical knowledge. In *Proceedings of the 17th international conference on Computational linguistics-Volume 1*, pages 704–710. Association for Computational Linguistics.

Zi Lin, Yuguang Duan, Yuanyuan Zhao, Weiwei Sun, and Xiaojun Wan. 2018. Semantic role labeling for learner chinese: the importance of syntactic parsing and l2-l1 parallel data. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3793–3802.

Matthias Lindemann, Jonas Groschwitz, and Alexander Koller. 2019. Compositional semantic parsing across graphbanks.

Mitchell Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of english: The penn treebank.

Ryo Nagata and Keisuke Sakaguchi. 2016. Phrase structure annotation and parsing for learner english. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1837–1847.

Ryo Nagata, Edward Whittaker, and Vera Sheinman. 2011. Creating a manually error-tagged and shallow-parsed learner corpus. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 1210–1219. Association for Computational Linguistics.

Hwee Tou Ng, Siew Mei Wu, Ted Briscoe, Christian Hadiwinoto, Raymond Hendy Susanto, and Christopher Bryant. 2014. The conll-2014 shared task on grammatical error correction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–14.

Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666.

Stephan Oepen and Jan Tore Lønning. 2006. Discriminant-based mrs banking. In *LREC*, pages 1250–1255.

Carl Pollard and Ivan A. Sag. 1994. *Head-Driven Phrase Structure Grammar*. The University of Chicago Press, Chicago.

Feng Qian, Lei Sha, Baobao Chang, LuChen Liu, and Ming Zhang. 2017. Syntax aware lstm model for semantic role labeling. In *Proceedings of the 2nd Workshop on Structured Prediction for Natural Language Processing*, pages 27–32.

Marwa Ragheb and Markus Dickinson. 2012. Defining syntax for learner language annotation. In *Proceedings of COLING 2012: Posters*, pages 965–974, Mumbai, India.

Stefano Rastelli. 2013. Learner corpora without error tagging. *Linguistik online*, 38(2).

Alexandr Rosen, Jirka Hana, Barbora Štindlová, and Anna Feldman. 2014. Evaluating and automating the annotation of a learner corpus. *Language Resources and Evaluation*, 48(1):65–92.

Alla Rozovskaya and Dan Roth. 2010. Annotating esl errors: Challenges and rewards. In *Proceedings of the NAACL HLT 2010 fifth workshop on innovative use of NLP for building educational applications*, pages 28–36. Association for Computational Linguistics.

Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. 2008. The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1):61–80.

Antonella Sorace. 2006. Gradedness and optionality in mature and developing grammars. *Gradience in grammar: Generative perspectives*, pages 106–123.

Antonella Sorace. 2011. Pinning down the concept of âĂIJinterfaceâĂİ in bilingualism. *Linguistic approaches to bilingualism*, 1(1):1–33.

Lydia White. 2011. Second language acquisition at the interfaces. *Lingua*, 121(4):577–590.

Sheng Zhang, Xutai Ma, Kevin Duh, and Benjamin Van Durme. 2019. Amr parsing as sequence-to-graph transduction.

Wei Zhao, Liang Wang, Kewei Shen, Ruoyu Jia, and Jingming Liu. 2019. Improving grammatical error correction via pre-training a copy-augmented architecture with unlabeled data. *arXiv preprint arXiv:1903.00138*.