Multilingual Semantic Parsing : Parsing Multiple Languages into Semantic Representations

Zhanming Jie
University of Electronic Science & Technology of China
allanmcgrady@gmail.com

Wei Lu
Information Systems Technology and Design
Singapore University of Technology and Design
luwei@sutd.edu.sg

Abstract

We consider *multilingual semantic parsing* – the task of simultaneously parsing semantically equivalent sentences from multiple different languages into their corresponding formal semantic representations. Our model is built on top of the *hybrid tree* semantic parsing framework, where natural language sentences and their corresponding semantics are assumed to be generated jointly from an underlying generative process. We first introduce a variant of the joint generative process, which essentially gives us a new semantic parsing model within the framework. Based on the different models that can be developed within the framework, we then investigate several approaches for performing the multilingual semantic parsing task. We present our evaluations on a standard dataset annotated with sentences in multiple languages coming from different language families.

1 Introduction

Semantic parsing, the task of parsing natural language sentences into their formal semantic representations (Mooney, 2007) is one of the most important tasks in the field of natural language processing and artificial intelligence. This area of research recently has received a significant amount of attention (Zettlemoyer and Collins, 2005; Kate and Mooney, 2006; Wong and Mooney, 2006; Lu et al., 2008; Jones et al., 2012b). Consider these example sentence-semantics pairs:

English:	Which states have points that are higher than the highest point in Texas ?
Semantics:	$answer(state(loc_1(place(higher_2(highest(place(loc_2(stateid('TX'))))))))$
English:	What rivers do not run through Tennessee ?
Semantics:	$answer(exclude(river(all), traverse_2(stateid('TN'))))$

In the typical setting, the semantic parser learns from a collection of such sentence-semantics pairs a model that can parse novel input sentences into their respective semantic representations. Such semantic representations can then be used to interact with certain downstream components to perform interesting tasks. For example, retrieving of answers from an underlying database, or performing certain actions based on the generated executable semantic instructions.

Note that in the training data, although complete sentence-semantics pairs are given, specific word-level semantic information is not explicitly provided. The model therefore needs to automatically learn such latent mappings between natural language words/phrases and semantic units.

One natural assumption is that the semantics exhibit certain restricted structures, such as the recursive tree structures. Under such an assumption, one can convert the second semantics appeared above as the tree structure illustrated in Figure 1. More details about such tree structured representations will be given in Section 2.1.

Currently, researchers only focused on the semantic parsing task under a single language setting where the input is a sentence from one particular language. However, natural language is highly ambiguous, and identifying the correct semantics associated with words with limited background information is a

Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 1291–1301, Dublin, Ireland, August 23-29 2014.

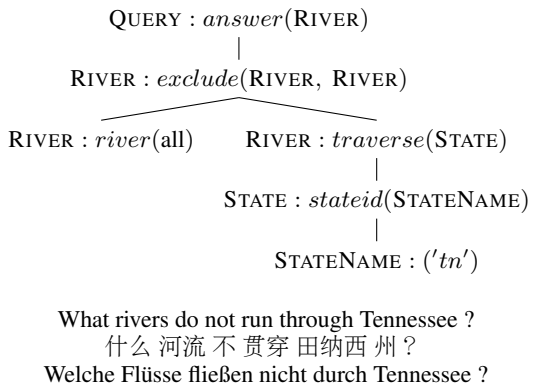

Figure 1: An example tree-structured semantic representation (above) and its corresponding natural language sentences (in English, Chinese and German).

challenging task. Researchers resorted to performing context-dependent semantic parsing to alleviate such an issue (Zettlemoyer and Collins, 2009).

On the other hand, researchers have successfully exploited parallel texts for improved word-level semantic processing (Chan and Ng, 2005). This is because words from different languages that convey the same semantics can be used to disambiguate each other's semantics. In fact, texts from different languages that convey the same semantic information becomes increasingly available nowadays. Web crawlers such as Google and Yahoo! are able to rapidly aggregate a large volume of news stories every day. One crucial fact is that many such news articles written in different languages are actually all discussing the same underlying story and therefore convey similar or identical semantic information. To build better automatic systems for improved natural language understanding, it is therefore helpful to develop algorithms that can simultaneously process the underlying semantic information associated with all these documents coming from different language sources together. For example, consider the following example taken from the multilingual version of the dataset, which shows semantically equivalent sentences from three different languages and their corresponding semantics:

English:	What rivers do not run through Tennessee ?
Chinese:	什么 河流 不 贯穿 田纳西？
German:	Welche Flüsse fließen nicht durch Tennessee ?
Semantics:	$answer(exclude(river(all), traverse_2(stateid('TN'))))$

As a step towards the above-mentioned goal, this work focuses on the development of an automated system that is capable of simultaneously parsing semantically equivalent natural language texts in different languages into their underlying semantics.

Specifically, in this work, we first introduce a new variant of a semantic parsing model under an existing framework. This new variant can be used together with other models for jointly making semantic parsing predictions, leading to an improved multilingual semantic parsing system. We demonstrate the effectiveness of this new variant through experiments. Although bilingual parsing has been extensively studied in fields such as statistical machine translation (Wu, 1997; Chiang, 2007), to the best of our knowledge, bilingual or multilingual semantic parsing that focuses on parsing sentences from multiple different languages into their formal semantic representations has not yet been studied. We present the very first work on performing multilingual semantic parsing that simultaneously parses semantically equivalent sentences from multiple different languages into their semantics. We believe this line of work can potentially lead to further developments and advancements in areas such as multilingual semantic processing and semantics-based machine translations (Jones et al., 2012a).

2 Background

2.1 Semantics

Researchers have focused on various semantic formalisms for semantic parsing. Popular examples include the tree-structured semantic representations (Wong and Mooney, 2006; Kate and Mooney,

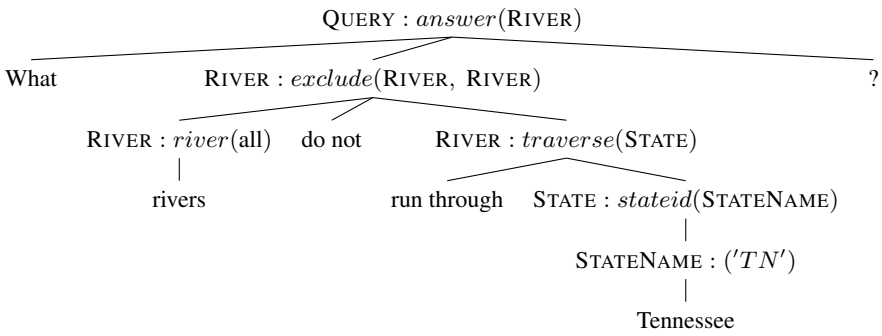

Figure 2: An example hybrid tree. Such a hybrid tree is generated from the generative process, and captures the correspondences between natural language words and semantic units.

2006), the lambda calculus expressions (Zettlemoyer and Collins, 2005; Wong and Mooney, 2007), and dependency-based semantic representations (DCS) (Liang et al., 2013). In this work, we specifically focus on the tree-structured representations for semantics.

Each semantic representation consists of semantic units as its tree nodes, where each semantic unit is of the following form:

$$m_a \equiv \tau_a : p_\alpha(\tau_b*) \tag{1}$$

Here m_a is used to denote a complete semantic unit, which consists of its semantic type τ_a, its function symbol p_α, as well as a list of types for argument semantic units τ_b* (here $*$ means 0, 1, or 2; we assume there are at most two arguments for each semantic unit). In other words, each semantic unit can be regarded as a function which takes in other semantic representations of specific types as arguments, and returns a new semantic representation of a particular type. For example, in Figure 1, the semantic unit at the root has a type QUERY, a function name $answer$, and a single argument type RIVER.

2.2 Related Work

Substantial research efforts have focused on building monolingual semantic parsing systems. We survey in this section several of them.

WASP (Wong and Mooney, 2006) is a model motivated by statistical synchronous parsing-based machine translation (Chiang, 2007), which essentially casts the semantic parsing problem as a phrase-based translation problem (Koehn et al., 2003). KRISP (Kate and Mooney, 2006) makes use of Support Vector Machines with string kernels (Lodhi et al., 2002) to recursively map contiguous word sequences into semantic units to construct a tree structure. The SCISSOR model (Ge and Mooney, 2005) performs integrated semantic and syntactic parsing. The model parses natural language sentences into semantically augmented parse trees whose nodes consist of both semantic and syntactic labels and then builds semantic representations based on such augmented trees. The *hybrid tree* model (Lu et al., 2008; Lu et al., 2009), whose code is publicly available, makes the assumption that there exists an underlying generative process for jointly producing both the language and semantics. The model employs efficient dynamic programming algorithms for learning a distribution over the latent *hybrid trees* which jointly encode both language and semantics. An example hybrid tree representation is shown in Figure 2. Jones et al. (2012b) recently proposed a framework that performs semantic parsing with tree transducers. The model learns representations that are similar to the hybrid tree structures using a generative process under a Bayesian framework.

Besides these approaches, recently there are also several works that take alternative learning approaches for semantic parsing which do not require annotated semantic representations (Poon and Domingos, 2009; Clarke et al., 2010; Goldwasser et al., 2011; Liang et al., 2013; Artzi and Zettlemoyer, 2013). Most of such approaches rely on either weak supervision or certain forms of indirect supervision. Some of these works also focus on optimizing specific downstream tasks rather than the semantic parsing task itself.

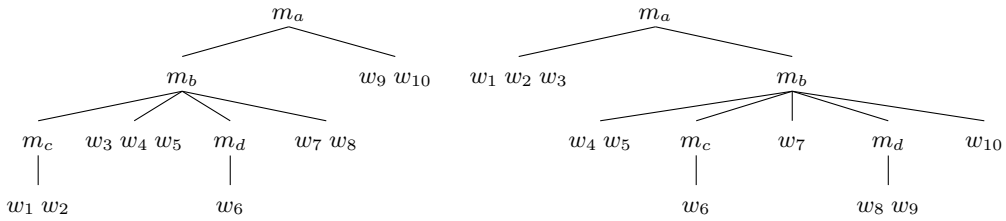

Figure 3: Two example hybrid trees. Their leaves are natural language words, and the internal nodes are semantic units. Both hybrid trees correspond to the same $\mathbf{n}$-$\mathbf{m}$ pair $\langle \overline{w_1\ w_2\ w_3\ w_4\ w_5\ w_6\ w_7\ w_8\ w_9\ w_{10}}, m_a(m_b(m_c, m_d))\rangle$. Thus they can be viewed as two different ways of generating such a pair from the joint generative process.

We note there also exist various multilingual or cross-lingual semantic processing works. Most of such works focus on semantic role labeling(SRL), the task of recovery of shallow meaning. Examples include multilingual semantic role labeling (Björkelund et al., 2009), multilingual joint syntactic and semantic dependency parsing (Henderson et al., 2013), and cross-lingual transfer of semantic role labeling models (Kozhevnikov and Titov, 2013). Researchers also looked into exploiting semantic information for bilingual processing such as machine translations (Chan et al., 2007; Carpuat and Wu, 2007; Jones et al., 2012a).

In this work, we focus on the task of multilingual semantic parsing under the setting where the input consists of semantically equivalent sentences from multiple different languages, and the outputs are formal semantic representations. We specifically focus on the hybrid tree model, a state-of-the-art framework for semantic parsing. We first make an extension to the model, and investigate methods for performing such a multilingual semantic parsing task by aggregating a few variants of the models under such a framework.

3 Approach

In this section, we first discuss the hybrid tree model of Lu et al. (2008), and introduce a novel extension. Next we discuss the approach used for multilingual semantic parsing.

3.1 The Hybrid Tree Model

For a given $\mathbf{n}$-$\mathbf{m}$ pair (where $\mathbf{n}$ is a complete natural language sentence, and $\mathbf{m}$ is a complete semantic representation), the hybrid tree model assumes that both $\mathbf{n}$ and $\mathbf{m}$ are generated from an underlying generative process in a top-down, left-to-right, level-by-level, recursive manner. The joint generative process for the pair results in a new tree-structured representation called a *hybrid tree*, which consists of natural language words as leaves, and semantic units as internal nodes.

There are three types of model parameters involved in the generative process. The meaning representation model parameters (ρ) are used for generating one semantic unit from its parent semantic unit. The hybrid pattern parameters (ϕ) are used for deciding how natural language words and semantic units are organized together to form the next level of the nodes of the hybrid tree structure. The emission parameters (θ) are used for generating natural language words from its corresponding semantic unit.

For a given $\mathbf{n}$-$\mathbf{m}$ pair, there are multiple possible hybrid trees that can jointly represent such a pair. See Figure 3 for two possible hybrid trees that contain the same $\mathbf{n}$-$\mathbf{m}$ pair. Consider the first example hybrid tree illustrated there. The probability of generating such a hybrid tree $\mathbf{h}$ (i.e., jointly generating both the natural language sentence $\mathbf{n}$ and the semantics $\mathbf{m}$) is:

$$\begin{aligned} P(\mathbf{n}, \mathbf{m}, \mathbf{h}) = {} & \rho(m_a) \times \phi(\overline{\mathbf{X}\mathbf{w}}|m_a) \times \theta(\mathbf{X}|m_a, \Lambda) \times \theta(w_9|m_a, \Lambda) \times \theta(w_{10}|m_a, \Lambda) \\ & \times \rho(m_b|m_a, \arg = 1) \times \phi(\overline{\mathbf{X}\mathbf{w}\mathbf{Y}\mathbf{w}}|m_b) \times \theta(\mathbf{X}|m_b, \Lambda) \times \theta(w_3|m_b, \Lambda) \\ & \times \theta(w_4|m_b, \Lambda) \times \theta(w_5|m_b, \Lambda) \times \theta(\mathbf{Y}|m_b, \Lambda) \times \theta(w_7|m_b, \Lambda) \times \theta(w_8|m_b, \Lambda) \\ & \times \rho(m_c|m_b, \arg = 1) \times \phi(\overline{\mathbf{w}}|m_c) \times \theta(w_1|m_c, \Lambda) \times \theta(w_2|m_c, \Lambda) \\ & \times \rho(m_d|m_b, \arg = 2) \times \phi(\overline{\mathbf{w}}|m_d) \times \theta(w_6|m_d, \Lambda) \end{aligned} \quad (2)$$

Note that $\overline{\mathbf{Xw}}$ refers to a pattern which says the next level of the hybrid tree is expected to consist of the first child semantic unit, followed by a contiguous sequence of natural language words. Similar definitions can be given to the patterns $\overline{\mathbf{XwYw}}$ and $\overline{\mathbf{w}}$, where $\mathbf{X}$ and $\mathbf{Y}$ refer to the first and second child semantic unit, respectively. The symbols $\mathbf{X}$ and $\mathbf{Y}$ appear in emission parameters are used to denote placeholders for the first and second child semantic unit, respectively.

The hybrid tree model then focuses on the learning of these model parameters from the training data using maximum likelihood estimation. In other words, the model tries to maximize:

$$\sum_i \log P(\mathbf{n}_i, \mathbf{m}_i; \rho, \phi, \theta) = \sum_i \log \sum_{\mathbf{h}} P(\mathbf{n}_i, \mathbf{m}_i, \mathbf{h}; \rho, \phi, \theta) \tag{3}$$

Since the correct hybrid tree associated with $\mathbf{n}$-$\mathbf{m}$ pair is unknown, we marginalize over the hidden variable $\mathbf{h}$. The model parameters will then be estimated using the Expectation-Maximization (EM) algorithm (Dempster et al., 1977). Specifically, an inside-outside style algorithm (Baker, 2005) is used where an additional layer of dynamic programming algorithms are used for efficient inference (Lu et al., 2008). The complexity of the inference algorithm is $O(mn^3)$, where m is the size of the semantic representation (number of semantic units), and n is the number of words in the input sentence.

Note that the generation of natural language words involves the context Λ. Specifically, if the context is empty, the model is regarded as the *unigram model*. If the context is the previously generated word, the model is called a *bigram model*. For example, consider the generation of the natural language word w_4 in the left hybrid tree in Figure 2. The probability for generating this word is $\theta(w_4|m_b)$ and $\theta(w_4|m_b, w_3)$, under the unigram and the bigram model, respectively. In Lu et al. (2008), the mixgram model (an interpolation between the unigram model and the bigram model) was also considered when parsing novel sentences, which yielded a better performance.

Once the model parameters are learned, we will be able to use them to parse novel sentences. Specifically, for each novel input sentence, we first find the most probable hybrid tree that contains the sentence $\mathbf{n}$, and then extract its internal nodes to form the semantic representation. Efficient dynamic programming algorithms similar to the ones used for training can also be employed here. In addition, the algorithm can also be extended to support exact top-k decoding, which will be useful later for combining multiple lists of outputs with rank aggregation (to be discussed in Sec. 3.3).

3.2 The Backward Bigram Model

One assumption associated with the original hybrid tree model is that nodes at each level of the hybrid tree are generated from the left to the right. An alternative assumption would be that the nodes at each level are generated in the reverse order – from the right to the left. While this alternative assumption will not introduce any difference in the unigram model (since each node is generated from its respective parent semantic unit only, regardless of its context), such a new assumption will lead to a completely new generative process under the bigram assumption.

To see this, again consider the emission probability for generating the word w_4 in the hybrid tree on the left of Figure 3. Under the assumption of our new model, the probability of generating this word is $\theta(w_4|m_b, w_5)$, since now the context Λ becomes the word to the right of the current word. The parameter estimation and parsing (decoding) procedures are largely similar to those of the original bigram model, where similar efficient dynamic programming algorithms can be employed.

3.3 Multilingual Semantic Parsing

In multilingual semantic parsing, the input consists of multiple semantically equivalent sentences, each of which is from a different language. One approach for building such a multilingual semantic parsing system is to develop a joint generative process from which both the semantic representations and the sentences in different languages are generated simultaneously. However, building such a joint model is non-trivial. Typically, sentences from different languages exhibit very different syntactic structures and word orderings. It is also non-trivial to design efficient dynamic programming algorithms for this case where multiple languages are involved in the joint generative process. Furthermore, the difficulty of building such a joint generative model becomes higher as the number of input languages increases.

Previous research efforts show that it can be beneficial to learn individual models independently, and then combine the learned models only during the inference stage (Punyakanok et al., 2005; Chang et al., 2012). Motivated by this, we take the approach that learns a separate semantic parser for each different language first. Next, we combine these semantic parsers for different languages into a single multilingual semantic parser only during the inference stage.

One common approach for combining different outputs from different systems is to perform *majority voting* based on optimal predictions from each parser. We first obtain the best output semantic representation from each individual semantic parser, and then count the number of occurrences for each possible output. The most frequent output semantic representation is returned as the final output of our system. Naturally, this approach is only applicable when there are at least three systems/models.

An alternative approach is to allow each system to produce a ranked list of k most probable outputs, each is associated with a score. Our system then aggregates these ranked lists to select the best output. This problem is known as *rank aggregation* and has been extensively studied in fields such as data mining and information retrieval (Dwork et al., 2001; Gleich and Lim, 2011; Li, 2011). For our task, we first let each semantic parser (for each language) generate a ranked list of the top-k most probable outputs (hybrid trees) for the given input. Next, based these hybrid trees we find a ranked list of most probable semantic representations. Each such semantic representation is also associated with a score, which is the log-likelihood of the hybrid tree, i.e., $\log P(\mathbf{n}, \mathbf{m}, \mathbf{h})$. Note that for each semantic representation, we only consider the score associated with the most probable hybrid tree that contains such a semantic representation. We use the standard approach for combining two ranked lists with scores. Consider a ranked list from the j-th model/system that consists of n distinct items. Let's use $s_i^{(j)}$ to denote the original score associated with the i-th semantic representation in the j-th ranked list. We normalize the score $s_i^{(j)}$ in the following way to obtain the new score $\tilde{s}_i^{(j)}$ (normalized score, divided by the standard deviation associated with the sample):

$$\tilde{s}_i^{(j)} = \frac{s_i^{(j)}}{n\mu^{(j)}\delta^{(j)}} \quad \text{where} \quad \mu^{(j)} = \frac{1}{n}\sum_{k=1}^{n} s_k^{(j)}, \quad \delta^{(j)} = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(s_k^{(j)} - \mu^{(j)})^2}$$

Such new scores will then be used for aggregating the results to form a new ranked list. How do we find the best output from multiple lists? Two useful sources of information that we may use include: 1) the number of times each output appears in these lists; 2) the combined score $\sum_j \tilde{s}^{(j)}$ for each output s. We believe the more frequent an output appears in these lists (i.e., more systems/models predict such an output in their top-k lists), the more likely it can be a good candidate. Therefore we first find the set of most frequent outputs, next from such a set we select the output with the highest overall score $\sum_j \tilde{s}^{(j)}$ as the final output of our system.

4 Experiments

4.1 Data and Setup

We conducted our experiments on the multilingual GEOQUERY dataset released by Jones et al. (2012b). This dataset consists of 880 instances of natural language queries related to US geography facts. Each query is coupled with its corresponding semantic representation originally written in Prolog. The original GEOQUERY dataset (Wong and Mooney, 2006; Kate and Mooney, 2006) contains natural language queries in English only. Additional Chinese annotations were provided by Lu and Ng (2011) when performing a natural language generation task. Jones et al. (2012b) further provided the following three additional language annotations to this dataset: German, Greek and Thai. Thus, this dataset is now fully annotated with five different languages, two of which (Chinese, Thai) are Sino-Tibetan languages, and the rest are all Indo-European languages.

Following previous works on semantic parsing (Kwiatkowski et al., 2010; Jones et al., 2012b), we split the dataset into two portions. The training set consists of 600 instances, and we report evaluation results on the portion consisting of the remaining 280 instances. We used the identical split provided by Jones et al. (2012b) for all the experiments. Following previous works, we used the standard approach for

	EN	DE	EL	TH	CN
Unigram	70.0	59.6	70.0	68.9	68.9
Bigram	75.4	56.1	65.4	70.7	68.9
Bigram (inv)	74.3	57.1	65.4	71.1	66.8
Mixgram	76.1	**62.5**	69.3	73.2	70.7
Voting (u,b,m)	76.1	61.1	70.4	73.6	70.0
Voting (u,b,bi)	76.4	61.4	71.8	**74.3**	72.1
Aggregation	**78.6**	60.0	**72.1**	71.4	**73.2**

Table 1: Monolingual semantic parsing results on all five languages (EN:English, DE:German, EL:Greek, TH:Thai, CN:Chinese.). We report accuracy percentages in this table.

	ENDE	ENEL	ENTH	ENCN	DEEL	DETH	DECN	ELTH	ELCN	THCN
Unigram	74.6	76.1	76.4	75.0	76.8	72.1	74.3	80.4	79.6	74.0
Bigram	80.0	77.9	**87.1**	78.2	72.1	75.0	76.4	81.4	76.8	79.6
Bigram (inv)	78.2	76.8	86.4	75.7	72.5	75.7	76.1	82.1	75.7	79.3
Mixgram	77.9	76.4	82.5	81.1	76.1	75.7	74.3	81.1	80.7	77.9
Voting (u,b)	80.0	79.6	83.6	82.1	**77.1**	74.6	74.6	82.1	78.6	79.6
Voting (u,b,bi)	**82.1**	79.3	86.4	82.1	76.8	**77.1**	76.4	**85.4**	78.9	**80.7**
Aggregation	78.9	**82.1**	85.7	**83.6**	76.4	73.6	**76.8**	83.9	**81.4**	79.3

Table 2: Semantic parsing results when two different input languages are considered (for example, the column ENDE gives the results when each input to our system consists of a pair of semantically equivalent sentences written in English and German.). Scores are accuracy percentages.

evaluation on the multilingual GEOQUERY dataset. Specifically, we first let our semantic parsers produce semantic representations from multilingual input sentences. The resulting semantic representations are then converted into Prolog queries in a deterministic manner, which can be used to interact with the underlying knowledge base to retrieve answers. A predicted semantic representation is considered correct if and only if it retrieves identical results as the correct reference semantic representation when both are used for retrieving answers from the underlying database.

4.2 Results and Discussions

We performed experiments on the conventional monolingual semantic parsing task first. We report accuracy scores, which are defined as the number of correctly parsed inputs (i.e., the total number of correct semantic representations) divided by the total number of input sentences. Baseline results for unigram, bigram, and mixgram models, which are originally introduced in Lu et al. (2008) are reported under "Unigram", "Bigram", and "Mixgram" respectively in Table 1. The results for backward bigram models are reported under "Bigram(inv)".

To assess the effectiveness of our methods for combining different outputs, we first conducted experiments on voting over the outputs from the three models originally introduced in the work of Lu et al. (2008) (Voting(u,b,m)). Next we performed voting over outputs from unigram model, bigram model, as well as the backward bigram model introduced in this paper (Voting(u,b,bi)). These voting-based approaches yielded better results over the first voting-based approach. Specifically, we compared this new voting-based approach against the previous best model reported in Lu et al. (2008) – mixgram model, which was also based on a combination of unigram and bigram models. We used the paired t-test to assess the significance of the overall improvements across different languages when using our new method. When comparing the approach "Voting(u,b,m)" over "Mixgram", we obtained a one-tailed p-value of 0.40. When comparing the approach "Voting(u,b,bi)" over "Mixgram", we obtained a one-tailed p-value of 0.11. We also investigated the effectiveness of the aggregation-based approach. This approach is based on aggregating the two top-100 lists generated by unigram, bigram and backward bigram models. When comparing this approach over "Mixgram", we obtained a one-tailed p-value of

	EnDe El	EnDe Th	EnDe Cn	EnEl Th	EnEl Cn	EnTh Cn	DeEl Th	DeEl Cn	DeTh Cn	ElTh Cn
Unigram	79.6	78.2	79.3	83.2	83.2	79.3	81.8	79.6	77.1	81.4
Bigram	82.1	85.7	81.8	87.5	81.8	86.4	82.5	80.7	79.6	83.6
Bigram (inv)	82.9	85.4	79.6	86.8	81.1	85.4	82.1	80.4	78.9	83.2
Mixgram	81.4	83.2	81.8	85.0	83.2	84.3	82.9	80.7	79.3	82.9
Voting (u,b)	83.2	85.0	84.3	87.9	84.0	85.0	84.0	**83.6**	**81.1**	84.6
Voting (u,b,bi)	**84.0**	**86.1**	**85.4**	**89.6**	84.3	**86.8**	**85.0**	82.5	**81.1**	84.6
Aggregation	83.6	85.0	**85.4**	88.9	**87.1**	85.7	82.5	82.5	80.0	**85.4**

Table 3: Semantic parsing results when three different input languages are considered (for example, the column EnDeEl gives the results when each input to our system consists of three semantically equivalent sentences, which are written in English, German and Greek, respectively.). Scores are accuracy percentages.

	EnDe ElTh	EnDe ElCn	EnDe ThCn	EnEl ThCn	DeEl ThCn	EnDeEl ThCn
Unigram	82.9	82.1	81.1	85.0	82.1	84.0
Bigram	86.1	83.6	84.3	87.1	85.0	86.1
Bigram (inv)	86.4	82.5	84.0	86.8	85.4	85.0
Mixgram	84.0	82.1	83.2	86.4	84.0	85.7
Voting (u,b)	87.5	86.1	86.4	89.6	**86.4**	89.3
Voting (u,b,bi)	**88.6**	86.8	**87.1**	**90.0**	85.7	**89.6**
Aggregation	87.1	**87.1**	86.1	88.9	86.1	88.6

Table 4: Semantic parsing results when four or five different input languages are considered (for example, the column EnDeElTh gives the results when each input to our system consists of four semantically equivalent sentences, which are written in English, German, Greek, and Thai respectively.). Scores are accuracy percentages.

0.29 under the paired t-test. These results indicate that the approach based on voting over the unigram, bigram and backward bigram models gives the most promising results for monolingual semantic parsing, demonstrating the usefulness of our proposed backward bigram model.

Next we move to the multilingual setting where we would like to simultaneously process more than two languages. Specifically, we considered multilingual semantic parsing where there are two, three, four and five input languages. Table 2, Table 3, and Table 4 summarize these results. Table 2 shows the results for bilingual semantic parsing where we have two different input languages. The results reported under "Unigram" are based on the aggregation approach over unigram models. Similarly for "Bigram", "Bigram(inv)", and "Mixgram" (we also tried the voting-based approach for combining such baseline systems, which yielded slightly worse results). From this table we can see that generally speaking by considering two different languages as the input, our system is able to do better semantic parsing. We compared the voting-based approaches against the baseline approaches. For the approach "Voting(u,b)" (we excluded mixgram models in voting since now we have four models, two from each language, which are sufficient for voting, and preliminary results show that the inclusion of the mixgram models is not helpful), it does not outperform the bigram baseline approach (which is the most competitive amongst all baseline approaches) significantly ($p = 0.19$). When comparing the aggregation approach against the bigram baseline approach, we obtain $p = 0.04$. In contrast, the approach "Voting(u,b,bi)" outperforms all the baseline systems significantly ($p < 0.005$). These results again demonstrate the effectiveness of our newly proposed backward bigram model.

We can see from the results presented in Table 3 and Table 4 that, in general, the performance of the multilingual semantic parser tends to improve as the number of input languages increases. However this is not always the case. For example, consider the final system where we use all five languages as the input (refer to the results in the column of EnDeElThCn in Table 4); interestingly, when we remove German (De) from the inputs, we are able to build a better system in terms of accuracy (refer to the results in the column of EnElThCn). We believe this is partly due to the fact that the monolingual semantic parsing

task with German as the input language (see DE in Table 1) is relatively more challenging. Nevertheless, when all the languages are considered, the overall system is able to obtain an accuracy of 89.6% with the voting-based approach where our proposed backward bigram model is incorporated. This is significantly higher than any other monolingual system's performance reported in the literature. According to Jones et al. (2012b), the results of state-of-the-art monolingual semantic parsing systems on four of these five languages considered here are: 82.1%(EN), 75.0%(DE), 75.4%(EL), and 78.2%(TH). Note that to date, no single system reported in the literature can dominate all other systems across all these languages on this dataset in terms of accuracy performance. We hypothesize that this is because semantic information conveyed by the sentences from a single language tends to be highly ambiguous, and various linguistic phenomenons can be difficult to capture under a monolingual setting for any existing monolingual semantic parsing system. The multilingual semantic parsing system introduced in this work, in contrast, can exploit richer information from multiple languages to successfully disambiguate the semantics associated with the inputs for improved semantic parsing.

5 Conclusions

In this work, we focused on *multilingual semantic parsing*, the task of simultaneously parsing sentences from various different languages into their corresponding formal semantic representations. Our work is built on top of the *hybrid tree* framework where different generative process can be developed for jointly modelling the generation of both language and semantics. We first introduced a variant of the generative process, leading to a new semantic parsing model. Next we presented methods for combining and aggregating outputs from different models within the framework to build our multilingual semantic parsing system. Our results demonstrate the effectiveness of our approaches for such a task. To the best of our knowledge, this is the first work that tackles such a multilingual semantic parsing task which simultaneously parses sentences from multiple languages into formal semantic representations. Future work include explorations on applications of our system in areas such as multilingual semantic processing, cross-lingual semantic processing, and semantics-based machine translations (Jones et al., 2012a).

Acknowledgements

We would like to thank the anonymous reviewers for comments. This work was conducted during the first author's internship at SUTD. This work was supported by SUTD grant SRG ISTD 2013 064.

References

Yoav Artzi and Luke Zettlemoyer. 2013. Weakly supervised learning of semantic parsers for mapping instructions to actions. *TACL*.

James K Baker. 2005. Trainable grammars for speech recognition. *The Journal of the Acoustical Society of America*, 65(S1):S132–S132.

Anders Björkelund, Love Hafdell, and Pierre Nugues. 2009. Multilingual semantic role labeling. In *Proceedings of CONLL'09*, pages 43–48.

Marine Carpuat and Dekai Wu. 2007. Improving statistical machine translation using word sense disambiguation. In *Proceedings of EMNLP-CoNLL '07*, pages 61–72.

Yee Seng Chan and Hwee Tou Ng. 2005. Scaling up word sense disambiguation via parallel texts. In *Proceedings of AAAI '05*, pages 1037–1042.

Yee Seng Chan, Hwee Tou Ng, and David Chiang. 2007. Word sense disambiguation improves statistical machine translation. In *Proceedings of ACL '07*.

Ming-Wei Chang, Lev Ratinov, and Dan Roth. 2012. Structured learning with constrained conditional models. *Machine learning*, 88(3):399–431.

David Chiang. 2007. Hierarchical phrase-based translation. *Computational Linguistics*, 33(2):201–228.

James Clarke, Dan Goldwasser, Ming-Wei Chang, and Dan Roth. 2010. Driving semantic parsing from the world's response. In *Proceedings of CONLL '10*, pages 18–27.

Arthur P Dempster, Nan M Laird, Donald B Rubin, et al. 1977. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal statistical Society*, 39(1):1–38.

Cynthia Dwork, Ravi Kumar, Moni Naor, and Dandapani Sivakumar. 2001. Rank aggregation methods for the web. In *Proceedings of WWW '01*, pages 613–622.

Ruifang Ge and Raymond J. Mooney. 2005. A statistical semantic parser that integrates syntax and semantics. In *Proceedings of CONLL '05*, pages 9–16.

David F Gleich and Lek-Heng Lim. 2011. Rank aggregation via nuclear norm minimization. In *Proceedings of KDD '11*, pages 60–68.

Dan Goldwasser, Roi Reichart, James Clarke, and Dan Roth. 2011. Confidence driven unsupervised semantic parsing. In *Proceedings of ACL '11*, pages 1486–1495.

James Henderson, Paola Merlo, Ivan Titov, and Gabriele Musillo. 2013. Multi-lingual joint parsing of syntactic and semantic dependencies with a latent variable model. *Computational Linguistics*, 39(4).

Bevan Jones, Jacob Andreas, Daniel Bauer, Karl Moritz Hermann, and Kevin Knight. 2012a. Semantics-based machine translation with hyperedge replacement grammars. In *Proceedings of COLING '12*, pages 1359–1376.

Bevan Keeley Jones, Mark Johnson, and Sharon Goldwater. 2012b. Semantic parsing with bayesian tree transducers. In *Proceedings of ACL '12*, pages 488–496.

Rohit J. Kate and Raymond J. Mooney. 2006. Using string-kernels for learning semantic parsers. In *Proceedings of COLING/ACL '06*, pages 913–920.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of NAACL '03*, pages 48–54.

Mikhail Kozhevnikov and Ivan Titov. 2013. Cross-lingual transfer of semantic role labeling models. In *Proceedings of ACL '13*.

Tom Kwiatkowski, Luke Zettlemoyer, Sharon Goldwater, and Mark Steedman. 2010. Inducing probabilistic ccg grammars from logical form with higher-order unification. In *Proceedings of EMNLP'10*, pages 1223–1233.

Hang Li. 2011. Learning to rank for information retrieval and natural language processing. *Synthesis Lectures on Human Language Technologies*, 4(1):1–113.

Percy Liang, Michael I Jordan, and Dan Klein. 2013. Learning dependency-based compositional semantics. *Computational Linguistics*, 39(2):389–446.

Huma Lodhi, Craig Saunders, John Shawe-Taylor, Nello Cristianini, and Chris Watkins. 2002. Text classification using string kernels. *The Journal of Machine Learning Research*, 2:419–444.

Wei Lu and Hwee Tou Ng. 2011. A probabilistic forest-to-string model for language generation from typed lambda calculus expressions. In *Proceedings of EMNLP '11*, pages 1611–1622.

Wei Lu, Hwee Tou Ng, Wee Sun Lee, and Luke S. Zettlemoyer. 2008. A generative model for parsing natural language to meaning representations. In *Proceedings of EMNLP '08*, pages 783–792.

Wei Lu, Hwee Tou Ng, and Wee Sun Lee. 2009. Natural language generation with tree conditional random fields. In *Proceedings of EMNLP '09*, pages 400–409.

Raymond J. Mooney. 2007. Learning for semantic parsing. In *Computational Linguistics and Intelligent Text Processing*, pages 311–324. Springer.

Hoifung Poon and Pedro Domingos. 2009. Unsupervised semantic parsing. In *Proceedings of EMNLP '09*, pages 1–10.

Vasin Punyakanok, Dan Roth, Wen-tau Yih, and Dav Zimak. 2005. Learning and inference over constrained output. In *Proceedings of IJCAI '05*, pages 1124–1129.

Yuk Wah Wong and Raymond J. Mooney. 2006. Learning for semantic parsing with statistical machine translation. In *Proceedings of HLT/NAACL '06*, pages 439–446.

Yuk Wah Wong and Raymond J. Mooney. 2007. Learning synchronous grammars for semantic parsing with lambda calculus. In *Proceedings of ACL '07*.

Dekai Wu. 1997. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics*, 23(3):377–403.

Luke S Zettlemoyer and Michael Collins. 2005. Learning to map sentences to logical form: Structured classification with probabilistic categorial grammars. In *Proceedings of UAI '05*.

Luke S Zettlemoyer and Michael Collins. 2009. Learning context-dependent mappings from sentences to logical form. In *Proceedings of ACL/IJCNLP '09*, pages 976–984.