Bootstrapping a Crosslingual Semantic Parser

Tom Sherborne, Yumo Xu and **Mirella Lapata**
Institute for Language, Cognition and Computation
School of Informatics, University of Edinburgh
10 Crichton Street, Edinburgh EH8 9AB
{tom.sherborne,yumo.xu}@ed.ac.uk, mlap@inf.ed.ac.uk

Abstract

Recent progress in semantic parsing scarcely considers languages other than English but professional translation can be prohibitively expensive. We adapt a semantic parser trained on a single language, such as English, to new languages and multiple domains with minimal annotation. We query if machine translation is an adequate substitute for training data, and extend this to investigate bootstrapping using joint training with English, paraphrasing, and multilingual pre-trained models. We develop a Transformer-based parser combining paraphrases by ensembling attention over multiple encoders and present new versions of ATIS and Overnight in German and Chinese for evaluation. Experimental results indicate that MT can approximate training data in a new language for accurate parsing when augmented with paraphrasing through multiple MT engines. Considering when MT is inadequate, we also find that using our approach achieves parsing accuracy within 2% of complete translation using only 50% of training data.[1]

1 Introduction

Semantic parsing is the task of mapping natural language utterances to machine-interpretable expressions such as SQL or a logical meaning representation. This has emerged as a key technology for developing natural language interfaces, especially in the context of question answering (Kwiatkowski et al., 2013; Berant et al., 2013; Liang, 2016; Kollar et al., 2018), where a semantically complex question is translated to an executable *query* to retrieve an answer, or *denotation*, from a knowledge base.

Sequence-to-sequence neural networks (Sutskever et al., 2014) are a popular approach to semantic parsing, framing the task as *sequence transduction* from natural to formal languages (Jia and Liang, 2016; Dong and Lapata, 2016). Recent proposals include learning intermediate logic representations (Dong and Lapata, 2018; Guo et al., 2019), constrained decoding (Yin and Neubig, 2017; Krishnamurthy et al., 2017; Lin et al., 2019), and graph-based parsing (Bogin et al., 2019; Shaw et al., 2019).

Given recent interest in semantic parsing and the data requirements of neural methods, it is unsurprising that many challenging datasets have been released in the past decade (Wang et al., 2015; Zhong et al., 2017; Iyer et al., 2017; Yu et al., 2018, 2019). However, these widely use English as synonymous for natural language. English is neither linguistically typical (Dryer and Haspelmath, 2013) nor the most widely spoken language worldwide (Eberhard et al., 2019), but is presently the *lingua franca* of both utterances and knowledge bases in semantic parsing. Natural language interfaces intended for international deployment must be adaptable to multiple locales beyond prototypes for English. However, it is uneconomical to create brand new datasets for every new language and domain.

In this regard, most previous work has focused on *multilingual* semantic parsing i.e., learning from multiple natural languages in parallel assuming the availability of multilingual training data. Examples of multilingual datasets include GeoQuery (Zelle and Mooney, 1996), ATIS (Dahl et al., 1994) and NLMaps (Haas and Riezler, 2016) but each is limited to one domain. For larger datasets, professional translation can be prohibitively expensive and require many man-hours from experts and native speakers. Recently, Min et al. (2019) reproduced the public partitions of the SPIDER dataset (Yu et al., 2018) into Chinese, but this required three expert annotators for verification and agreement. We posit there exists a more efficient strategy for expanding semantic parsing to a new language.

In this work, we consider *crosslingual semantic parsing*, adapting a semantic parser trained on English, to another language. We expand executable

[1] Our code and data can be found at github.com/tomsherborne/bootstrap.

Findings of the Association for Computational Linguistics: EMNLP 2020, pages 499–517
November 16 - 20, 2020. ©2020 Association for Computational Linguistics

semantic parsing to new languages and multiple domains by bootstrapping from in-task English datasets, task-agnostic multilingual resources, and publicly available machine translation (MT) services, in lieu of expert translation of training data. We investigate a core hypothesis that MT can provide a noisy, but reasonable, approximation of training data in a new source language. We further explore the benefit of augmenting noisy MT data using pre-trained models, such as BERT (Devlin et al., 2019), and multilingual training with English. Additionally, we examine approaches to ensembling multiple machine translations as approximate paraphrases. This challenge combines both *domain adaptation* and *localization*, as a parser must generalize to the locale-specific style of queries using only noisy examples to learn from.

For our evaluation, we present the first *multi-domain*, executable semantic parsing dataset in three languages and an additional locale for a *single-domain* dataset. Specifically, we extend ATIS (Dahl et al., 1994), pairing Chinese (ZH) utterances from Susanto and Lu (2017a) to SQL queries and create a parallel German (DE) human-translation of the full dataset. Following this, we also make available a new version of the multi-domain Overnight dataset (Wang et al., 2015) where only development and test sets are translations from native speakers of Chinese and German. This is representative of the real-world scenario where a semantic parser needs to be developed for new languages without gold-standard training data.

Our contributions can be summarized as follows: (1) new versions of ATIS (Dahl et al., 1994) and Overnight (Wang et al., 2015) for generating executable logical forms from Chinese and German utterances; (2) a combined encoder-decoder attention mechanism to ensemble over multiple Transformer encoders; (3) a cost-effective methodology for bootstrapping semantic parsers to new languages using minimal new annotation. Our proposed method overcomes the paucity of gold-standard training data using pre-trained models, joint training with English, and paraphrasing through MT engines; and (4) an investigation into practical minimum gold-standard translation requirements for a fixed performance penalty when MT is unavailable.

2 Related Work

Across logical formalisms, there have been several proposals for multilingual semantic parsing which employ multiple natural languages in parallel (Jones et al., 2012; Andreas et al., 2013; Lu, 2014; Susanto and Lu, 2017b; Jie and Lu, 2018).

Jie and Lu (2014) ensemble monolingual parsers to generate a single parse from < 5 source languages for GeoQuery (Zelle and Mooney, 1996). Similarly, Richardson et al. (2018) propose a polyglot automaton decoder for source-code generation in 45 languages. Susanto and Lu (2017a) explore a multilingual neural architecture in four languages for GeoQuery and three languages for ATIS by extending Dong and Lapata (2016) with multilingual encoders. Other work focuses on multilingual representations for semantic parsing based on universal dependencies (Reddy et al., 2017) or embeddings of logical forms (Zou and Lu, 2018).

We capitalize on existing semantic parsing datasets to bootstrap from English to another language, and therefore, do not assume that multiple languages are available as parallel input. Our work is closest to Duong et al. (2017), however they explore how to parse both English and German simultaneously using a multilingual corpus. In contrast, we consider English data only as an augmentation to improve parsing in Chinese and German and do not use "real" utterances during training. Recently, Artetxe et al. (2020) studied MT for crosslingual entailment, however, our results in Section 5 suggest these prior findings may not extend to semantic parsing, owing to the heightened requirement for factual consistency across translations.

Our work complements recent efforts in crosslingual language understanding such as XNLI for entailment (Conneau et al., 2018), semantic textual similarity (Cer et al., 2017) or the XTREME (Hu et al., 2020) and XGLUE (Liang et al., 2020) benchmarks. There has also been interest in parsing into interlingual graphical meaning representations (Damonte and Cohen, 2018; Zhang et al., 2018), spoken language understanding (Upadhyay et al., 2018) and λ-calculus expressions (Kwiatkowski et al., 2010; Lu and Ng, 2011; Lu, 2014). In contrast, we focus on logical forms grounded in knowledge-bases and therefore do not consider these approaches further.

3 Problem Formulation

Throughout this work, we consider the real-world scenario where a typical developer wishes to develop a semantic parser to facilitate question answering from an existing commercial database to

Noun/Adjective Ambiguity ("first-class fares" is a noun object)	
EN	Show me the first class fares from Baltimore to Dallas
DE$_{\text{MT}}$	Zeigen Sie mir die **erstklassigen** Tarife von Baltimore nach Dallas
DE$_{\text{H}}$	Zeige mir die Preise in der ersten Klasse von Baltimore nach Dallas
Entity Misinterpretation (Airline names aren't preserved)	
EN	Which Northwest and United flights go through Denver before noon?
DE$_{\text{MT}}$	Welche **Nordwesten** und **Vereinigten** Flüge gehen durch Denver vor Mittag
DE$_{\text{H}}$	Welche Northwest und United Flüge gehen durch Denver vor Mittag
Question to Statement Mistranslation (rephrased as "You have a...")	
EN	Do you have an 819 flight from Denver to San Francisco?
ZH$_{\text{MT}}$	**你有**一个从丹佛到旧金山的819 航班
ZH$_{\text{H}}$	有没有从丹佛到旧金山的819 航班
Contextual Misinterpretation ("blocks" translated to "街区" [street blocks)])	
EN	What seasons did Kobe Bryant have only three blocks?
ZH$_{\text{MT}}$	什么季节科比布莱恩特只有三个**街区**
Referential Ambiguity (他[he] refers to either players or Kobe Bryant)	
EN	Which players played more games than Kobe Bryant the seasons he played?
ZH$_{\text{MT}}$	在**他**打球的那些赛季中,哪些球员比科比布莱恩特打得更多

Table 1: Examples from ATIS (Dahl et al., 1994) and Overnight (Wang et al., 2015). Utterances are translated into Chinese and German using both machine translation (L$_{\text{MT}}$) and crowdsourcing with verification (L$_{\text{H}}$). We highlight issues with the noisy MT data (underlined and bolded) compared to improved human translations (underlined) for ATIS.

customers in a new locale. For example, an engineer desiring to extend support to German speakers for a commercial database of USA flights in English. Without the resources of high-valued technology companies, costs for annotation and machine learning resources must be minimized to maintain commercial viability. To economize this task, the developer must minimize new annotation or professional translation and instead bootstrap a system with public resources. At a minimum, a test and development set of utterances from native speakers are required for evaluation. However, the extent of annotation and the utility of domain adaptation for training are unknown. Therefore, our main question is *how successfully can a semantic parser learn with alternative data resources to generalize to novel queries in a new language?*

Crosslingual semantic parsing presents a unique challenge as an NLU task. It demands the generation of precise utterance semantics, aligned across languages while ensuring an accurate mapping between logical form and the idiomatic syntax of questions in every language under test. In comparison to NLU classification tasks such as XNLI (Conneau et al., 2018), our challenge is to **preserve** and **generate** meaning, constrained under a noisy MT channel. The misinterpretation of entities, relationships, and relative or numerical expressions can all result in an incorrect parse.

Lexical translation in MT, however accurate it may be, is insufficient alone to represent queries from native speakers. For example, the English expression "dinner flights" can be directly translated to German as "Abendessenflug" [dinner flight], but "Flug zur Abendszeit" [evening flight] better represents typical German dialogue. This issue further concerns question phrasing. For example, the English query "do you have X?" is often mistranslated to a statement "你有一个X" [you have one X] but typical Chinese employs a positive-negative pattern ("有没有一个X?" [have not have one X?]) to query possession. Our parser must overcome each of these challenges without access to gold data.