Freebase QA: Information Extraction or Semantic Parsing?

Xuchen Yao [1] **Jonathan Berant** [3] **Benjamin Van Durme** [1,2]
[1]Center for Language and Speech Processing
[2]Human Language Technology Center of Excellence
Johns Hopkins University
[3]Computer Science Department
Stanford University

Abstract

We contrast two seemingly distinct approaches to the task of question answering (QA) using Freebase: one based on information extraction techniques, the other on semantic parsing. Results over the same test-set were collected from two state-of-the-art, open-source systems, then analyzed in consultation with those systems' creators. We conclude that the differences between these technologies, both in task performance, and in how they get there, is not significant. This suggests that the semantic parsing community should target answering more compositional open-domain questions that are beyond the reach of more direct information extraction methods.

1 Introduction

Question Answering (QA) from structured data, such as DBPedia (Auer et al., 2007), Freebase (Bollacker et al., 2008) and Yago2 (Hoffart et al., 2011), has drawn significant interest from both knowledge base (KB) and semantic parsing (SP) researchers. The majority of such work treats the KB as a database, to which standard database queries (SPARQL, MySQL, etc.) are issued to retrieve answers. Language understanding is modeled as the task of converting natural language questions into queries through intermediate logical forms, with the popular two approaches including: CCG parsing (Zettlemoyer and Collins, 2005; Zettlemoyer and Collins, 2007; Zettlemoyer and Collins, 2009; Kwiatkowski et al., 2010; Kwiatkowski et al., 2011; Krishnamurthy and Mitchell, 2012; Kwiatkowski et al., 2013; Cai and Yates, 2013a), and dependency-based compositional semantics (Liang et al., 2011; Berant et al., 2013; Berant and Liang, 2014).

We characterize semantic parsing as the task of deriving a representation of meaning from language, sufficient for a given task. Traditional information extraction (IE) from text may be coarsely characterized as representing a certain level of semantic parsing, where the goal is to derive enough meaning in order to populate a database with factoids of a form matching a given schema.[1] Given the ease with which reasonably accurate, deep syntactic structure can be automatically derived over (English) text, it is not surprising that IE researchers would start including such "features" in their models.

Our question is then: what is the difference between an IE system with access to syntax, as compared to a semantic parser, when both are targeting a factoid-extraction style task? While our conclusions should hold generally for similar KBs, we will focus on Freebase, such as explored by Krishnamurthy and Mitchell (2012), and then others such as Cai and Yates (2013a) and Berant et al. (2013). We compare two open-source, state-of-the-art systems on the task of Freebase QA: the semantic parsing system SEMPRE (Berant et al., 2013), and the IE system jacana-freebase (Yao and Van Durme, 2014).

We find that these two systems are on par with each other, with no significant differences in terms of accuracy between them. A major distinction between the work of Berant et al. (2013) and Yao and Van Durme (2014) is the ability of the former to represent, and compose, aggregation operators (such as `argmax`, or `count`), as well as integrate disparate pieces of information. This representational capability was important in previous, closed-domain tasks such as GeoQuery. The move to Freebase by the SP community was meant to

[1]So-called Open Information Extraction (OIE) is simply a further blurring of the distinction between IE and SP, where the schema is allowed to grow with the number of verbs, and other predicative elements of the language.

Proceedings of the ACL 2014 Workshop on Semantic Parsing, pages 82–86,
Baltimore, Maryland USA, June 26 2014. ©2014 Association for Computational Linguistics

provide richer, open-domain challenges. While the vocabulary increased, our analysis suggests that compositionality and complexity decreased. We therefore conclude that the semantic parsing community should target more challenging open-domain datasets, ones that "standard IE" methods are less capable of attacking.

2 IE and SP Systems

jacana-freebase[2] (Yao and Van Durme, 2014) treats QA from a KB as a binary classification problem. Freebase is a gigantic graph with millions of nodes (topics) and billions of edges (relations). For each question, jacana-freebase first selects a "view" of Freebase concerning only involved topics and their close neighbors (this "view" is called a topic graph). For instance, for the question "who is the brother of justin bieber?", the topic graph of Justin Bieber, containing all related nodes to the topic (think of the "Justin Bieber" page displayed by the browser), is selected and retrieved by the Freebase Topic API. Usually such a topic graph contains hundreds to thousands of nodes in close relation to the central topic. Then each of the node is judged as answer or not by a logistic regression learner.

Features for the logistic regression learner are first extracted from both the question and the topic graph. An analysis of the dependency parse of the question characterizes the question word, topic, verb, and named entities of the main subject as the question features, such as qword=who. Features on each node include the types of relations and properties the node possesses, such as type=person. Finally features from both the question and each node are combined as the final features used by the learner, such as qword=who|type=person. In this way the association between the question and answer type is enforced. Thus during decoding, for instance, if there is a who question, the nodes with a person property would be ranked higher as the answer candidate.

SEMPRE[3] is an open-source system for training semantic parsers, that has been utilized to train a semantic parser against Freebase by Berant et al. (2013). SEMPRE maps NL utterances to logical forms by performing bottom-up parsing. First, a lexicon is used to map NL phrases to KB predicates, and then predicates are combined to form a full logical form by a context-free grammar. Since logical forms can be derived in multiple ways from the grammar, a log-linear model is used to rank possible derivations. The parameters of the model are trained from question-answer pairs.

3 Analysis

3.1 Evaluation Metrics

Both Berant et al. (2013) and Yao and Van Durme (2014) tested their systems on the WEBQUESTIONS dataset, which contains 3778 training questions and 2032 test questions collected from the Google Suggest API. Each question came with a standard answer from Freebase annotated by Amazon Mechanical Turk.

Berant et al. (2013) reported a score of 31.4% in terms of accuracy (with partial credit if inexact match) on the test set and later in Berant and Liang (2014) revised it to 35.7%. Berant et al. focused on accuracy – how many questions were correctly answered by the system. Since their system answered almost all questions, accuracy is roughly identical to F_1. Yao and Van Durme (2014)'s system on the other hand only answered 80% of all test questions. Thus they report a score of 42% in terms of F_1 on this dataset. For the purpose of comparing among *all* test questions, we lowered the logistic regression prediction threshold (usually 0.5) on jacana-freebase for the other 20% of questions where jacana-freebase had not proposed an answer to, and selected the best-possible prediction with the highest prediction score as the answer. In this way jacana-freebase was able to answer all questions with a lower accuracy of 35.4%. In the following we present analysis results based on the test questions where the two systems had very similar performance (35.7% vs. 35.4%).[4] The difference is not significant according to the paired permutation test (Smucker et al., 2007).

3.2 Accuracy vs. Coverage

First, we were interested to see the proportions of questions SEMPRE and jacana-freebase jointly and separately answered correctly. The answer to

[2] https://code.google.com/p/jacana/

[3] http://www-nlp.stanford.edu/software/sempre/

[4] In this setting accuracy equals averaged macro F_1: first the F_1 value on each question were computed, then averaged among all questions, or put it in other words: "accuracy with partial credit". In this section our usage of the terms "accuracy" and "F_1" can be exchanged.

SEMPRE		jacana ($F_1 = 1$)		jacana ($F_1 \geq 0.5$)	
		$\surd$	$\times$	$\surd$	$\times$
	$\surd$	153 (0.08)	383 (0.19)	429 (0.21)	321 (0.16)
	$\times$	136 (0.07)	1360 (0.67)	366 (0.18)	916 (0.45)

Table 1: The absolute and proportion of questions SEMPRE and jacana-freebase answered correctly ($\surd$) and incorrectly ($\times$) jointly and separately, running a threshold F_1 of 1 and 0.5.

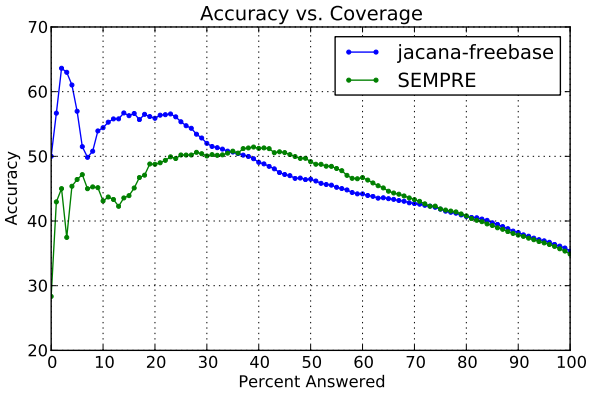

Figure 1: Precision with respect to proportion of questions answered

many questions in the dataset is a set of answers, for example what to see near sedona arizona?. Since turkers did not exhaustively pick out all possible answers, evaluation is performed by computing the F_1 between the set of answers given by the system and the answers provided by turkers. With a strict threshold of $F_1 = 1$ and a permissive threshold of $F_1 \geq 0.5$ to judge the correctness, we list the pair-wise correctness matrix in Table 1. Not surprisingly, both systems had most questions wrong given that the averaged F_1's were only around 35%. With the threshold $F_1 = 1$, SEMPRE answered more questions exactly correctly compared to jacana-freebase, while when $F_1 \geq 0.5$, it was the other way around. This shows that SEMPRE is more accurate in certain questions. The reason behind this is that SEMPRE always fires queries that return exactly one set of answers from Freebase, while jacana-freebase could potentially tag multiple nodes as the answer, which may lower the accuracy.

We have shown that both systems can be more accurate in certain questions, but when? Is there a correlation between the system confidence and accuracy? Thus we took the logistic decoding score (between 0 and 1) from jacana-freebase and the probability from the log-linear model used by SEMPRE as confidence, and plotted an "accuracy vs. coverage" curve, which shows the accuracy of a QA engine with respect to its coverage of all questions. The curve basically answers one question: at a fixed accuracy, what is the proportion of questions that can be answered? A better system should be able to answer more questions correctly with the same accuracy.

The curve was drawn in the following way. For each question, we select the best answer candidate with the highest confidence score. Then for the whole test set, we have a list of (question, highest ranked answer, confidence score) tuples. Running a threshold from 1 to 0, we select those questions with an answer confidence score above the threshold and compute accuracy at this point. The X-axis indicates the percentage of questions above the threshold and the Y-axis the accuracy, shown in Figure 1.

The two curves generally follow a similar trend, but while jacana-freebase has higher accuracy when coverage is low, SEMPRE obtains slightly better accuracy when more questions are answered.

3.3 Accuracy by Question Length and Type

Do accuracies of the two systems differ with respect to the complexity of questions? Since there is no clear way to measure question complexity, we use question length as a surrogate and report accuracies by question length in Figure 2. Most of the questions were 5 to 8 words long and there was no substantial difference in terms of accuracies. The major difference lies in questions of length 3, 12 and 13. However, the number of such questions was not high enough to show any statistical significance.

Figure 3 further shows the accuracies with respect to the question types (as reflected by the WH-word). Again, there is no significant difference between the two systems.

3.4 Learned Features

What did the systems learn during training? We compare them by presenting the top features by weight, as listed in Table 2. Clearly, the type of knowledge learned by the systems in these features is similar: both systems learn to associate certain phrases with predicates from the KB.

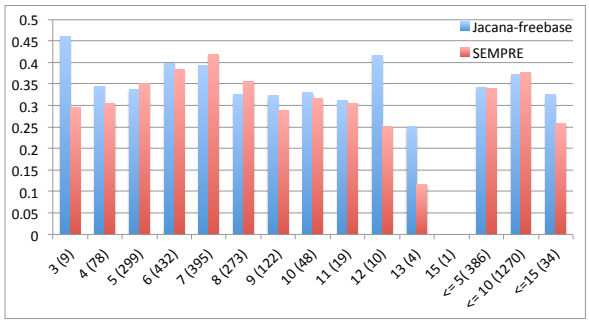

Figure 2: Accuracy (Y-axis) by question length. The X-axis specifies the question length in words and the total number of questions in parenthesis.

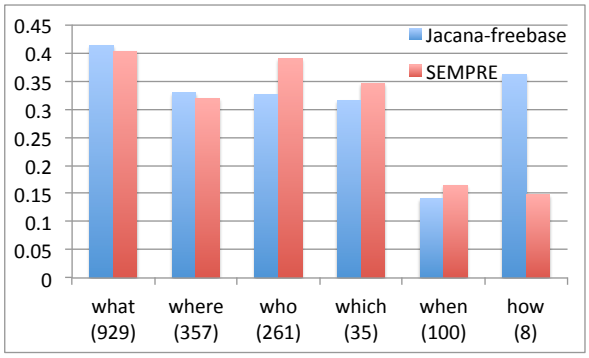

Figure 3: Accuracy by question type (and the number of questions).

We note, however, that SEMPRE also obtains information from the fully constructed logical form. For instance, SEMPRE learns that logical forms that return an empty set when executed against the KB are usually incorrect (the weight for this feature is -8.88). In this respect the SP approach "understands" more than the IE approach.

We did not further compare on other datasets such as GeoQuery (Tang and Mooney, 2001) and FREE917 (Cai and Yates, 2013b). The first one involves geographic inference and multiple constraints in queries, directly fitting the compositional nature of semantic parsing. The second one was manually generated by looking at Freebase topics. Both datasets were less realistic than the WEBQUESTIONS dataset. Both datasets were also less challenging (accuracy/F_1 were between 80% and 90%) compared to WEBQUESTIONS (around 40%).

4 Discussion and Conclusion

Our analysis of two QA approaches, semantic parsing and information extraction, has shown no significant difference between them. Note the similarity between features used in both systems shown in Table 2: the systems learned the same "knowledge" from data, with the distinction that the IE approach acquired this through a direct association between dependency parses and answer properties, while the SP approach acquired this through optimizing on intermediate logic forms.

feature	weight
qfocus=religion|type=Religion	8.60
qfocus=money|type=Currency	5.56
qverb=die|type=CauseOfDeath	5.35
qword=when|type=datetime	5.11
qverb=border|rel=location.adjoins	4.56

(a) jacana-freebase

feature	weight
die from=CauseOfDeath	10.23
die of=CauseOfDeath	7.55
accept=Currency	7.30
bear=PlaceOfBirth	7.11
in switzerland=Switzerland	6.86

(b) SEMPRE

Table 2: Learned top features and their weights for jacana-freebase and SEMPRE.

With a direct information extraction technology easily getting on par with the more sophisticated semantic parsing method, it suggests that SP-based approaches for QA with Freebase has not yet shown its power from a "deeper" understanding of the questions, among questions of various lengths. We suggest that more compositional open-domain datasets should be created, and that SP researchers should focus on utterances in existing datasets that are beyond the reach of direct IE methods.

5 Acknowledgement

We thank the Allen Institute for Artificial Intelligence for assistance in funding this work. This material is partially based on research sponsored by the NSF under grant IIS-1249516 and DARPA under agreements number FA8750-13-2-0017 and FA8750-13-2-0040 (the DEFT program).

References

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. DBPedia: A nucleus for a web of open data. In *The semantic web*, pages 722–735. Springer.

Jonathan Berant and Percy Liang. 2014. Semantic parsing via paraphrasing. In *Proceedings of ACL*.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic Parsing on Freebase from Question-Answer Pairs. In *Proceedings of EMNLP*.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250. ACM.

Qingqing Cai and Alexander Yates. 2013a. Large-scale semantic parsing via schema matching and lexicon extension. In *Proceedings of ACL*.

Qingqing Cai and Alexander Yates. 2013b. Large-scale semantic parsing via schema matching and lexicon extension. In *Proceedings of ACL*.

Johannes Hoffart, Fabian M Suchanek, Klaus Berberich, Edwin Lewis-Kelham, Gerard De Melo, and Gerhard Weikum. 2011. Yago2: exploring and querying world knowledge in time, space, context, and many languages. In *Proceedings of the 20th international conference companion on World Wide Web*, pages 229–232. ACM.

Jayant Krishnamurthy and Tom Mitchell. 2012. Weakly supervised training of semantic parsers. In *Proceedings of EMNLP*.

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*, pages 1223–1233.

Tom Kwiatkowski, Luke Zettlemoyer, Sharon Goldwater, and Mark Steedman. 2011. Lexical generalization in CCG grammar induction for semantic parsing. In *Proceedings of EMNLP*.

Tom Kwiatkowski, Eunsol Choi, Yoav Artzi, and Luke Zettlemoyer. 2013. Scaling Semantic Parsers with On-the-fly Ontology Matching. In *Proceedings of EMNLP*.

Percy Liang, Michael I. Jordan, and Dan Klein. 2011. Learning Dependency-Based Compositional Semantics. In *Proceedings of ACL*.

M.D. Smucker, J. Allan, and B. Carterette. 2007. A comparison of statistical significance tests for information retrieval evaluation. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 623–632. ACM.

Lappoon R Tang and Raymond J Mooney. 2001. Using multiple clause constructors in inductive logic programming for semantic parsing. In *Machine Learning: ECML 2001*. Springer.

Xuchen Yao and Benjamin Van Durme. 2014. Information extraction over structured data: Question answering with freebase. In *Proceedings of ACL*.

Luke S Zettlemoyer and Michael Collins. 2005. Learning to map sentences to logical form: Structured classification with probabilistic categorial grammars. *Uncertainty in Artificial Intelligence (UAI)*.

Luke S. Zettlemoyer and Michael Collins. 2007. Online learning of relaxed CCG grammars for parsing to logical form. In *Proceedings of EMNLP-CoNLL*.

Luke S Zettlemoyer and Michael Collins. 2009. Learning context-dependent mappings from sentences to logical form. In *Proceedings of ACL-CoNLL*.