Response-based Learning for Machine Translation of Open-domain Database Queries

Carolin Haas
Computational Linguistics
Heidelberg University
69120 Heidelberg, Germany
haas1@cl.uni-heidelberg.de

Stefan Riezler
Computational Linguistics
Heidelberg University
69120 Heidelberg, Germany
riezler@cl.uni-heidelberg.de

Abstract

Response-based learning allows to adapt a statistical machine translation (SMT) system to an extrinsic task by extracting supervision signals from task-specific feedback. In this paper, we elicit response signals for SMT adaptation by executing semantic parses of translated queries against the Freebase database. The challenge of our work lies in scaling semantic parsers to the lexical diversity of open-domain databases. We find that parser performance on incorrect English sentences, which is standardly ignored in parser evaluation, is key in model selection. In our experiments, the biggest improvements in F1-score for returning the correct answer from a semantic parse for a translated query are achieved by selecting a parser that is carefully enhanced by paraphrases and synonyms.

1 Introduction

In response-based learning for SMT, supervision signals are extracted from an extrinsic response to a machine translation, in contrast to using human-generated reference translations for supervision. We apply this framework to a scenario in which a semantic parse of a translated database query is executed against the Freebase database. We view learning from such task-specific feedback as adaptation of SMT parameters to the task of translating open-domain database queries, thereby grounding SMT in the task of multilingual database access. The success criterion for this task is F1-score in returning the correct answer from a semantic parse of the translated query, rather than BLEU. Since the semantic parser provides feedback to the response-based learner and defines the final evaluation criterion, the challenge of the presented work lies in scaling the semantic parser to the lexical diversity of open-domain databases such as Freebase. Riezler et al. (2014) showed how to use response-based learning to adapt an SMT system to a semantic parser for the Geoquery domain. The state-of-the-art in semantic parsing on Geoquery achieves a parsing accuracy of over 82% (see Andreas et al. (2013) for an overview), while the state-of-the-art in semantic parsing on the Free917 data (Cai and Yates, 2013) achieves 68.5% accuracy (Berant and Liang, 2014). This is due to the lexical variability of Free917 (2,036 word types) compared to Geoquery (279 word types).

In this paper, we compare different ways of scaling up state-of-the-art semantic parsers for Freebase by adding synonyms and paraphrases. First, we consider Berant and Liang (2014)'s own extension of the semantic parser of Berant et al. (2013) by using paraphrases. Second, we apply WordNet synonyms (Miller, 1995) for selected parts of speech to the queries in the Free917 dataset. The new pairs of queries and logical forms are added to the dataset on which the semantic parsers are retrained. We find that both techniques of enhancing the lexical coverage of the semantic parsers result in improved parsing performance, and that the improvements add up nicely. However, improved parsing performance does not correspond to improved F1-score in answer retrieval when using the respective parser in a response-based learning framework. We show that in order to produce helpful feedback for response-based learning, parser performance on incorrect En-

glish queries needs to be taken into account, which is standardly ignored in parser evaluation. That is, for the purpose of parsing translated queries, a parser should retrieve correct answers for correct English queries (true positives), and must not retrieve correct answers for incorrect translations (false positives). In order to measure false discovery rate, we prepare a test set of manually verified incorrect English in addition to a standard test set of original English queries. We show that if false discovery rate on incorrect English queries is taken into account in model selection, the semantic parser that yields best results for response-based learning in SMT can be found reliably.

2 Related Work

Our work is most closely related to Riezler et al. (2014). We extend their application of response-based learning for SMT to a larger and lexically more diverse dataset and show how to perform model selection in the environment from which response signals are obtained. In contrast to their work where a monolingual SMT-based approach (Andreas et al., 2013) is used as semantic parser, our work builds on existing parsers for Freebase, with a focus on exploiting paraphrasing and synonym extension for scaling semantic parsers to open-domain database queries.

Response-based learning has been applied in previous work to semantic parsing itself (Kwiatowski et al. (2013), Berant et al. (2013), Goldwasser and Roth (2013), *inter alia*). In these works, extrinsic responses in form of correct answers from a database are used to alleviate the problem of manual data annotation in semantic parsing. Saluja et al. (2012) integrate human binary feedback on the quality of an SMT system output into a discriminative learner.

Further work on learning from weak supervision signals has been presented in the machine learning community, e.g., in form of coactive learning (Shivaswamy and Joachims, 2012), reinforcement learning (Sutton and Barto, 1998), or online learning with limited feedback (Cesa-Bianchi and Lugosi, 2006).

3 Response-based Online SMT Learning

We denote by $\phi(x, y)$ a joint feature representation of input sentences x and output translations $y \in Y(x)$, and by $s(x, y; w) = \langle w, \phi(x, y) \rangle$ a linear scoring function for predicting a translation $\hat{y}$. A response signal is denoted by a binary function $\mathrm{e}(y) \in \{1, 0\}$ that executes a semantic parse against the database and checks whether it receives the same answer as the gold standard parse. Furthermore, a cost function $\mathrm{c}(y^{(i)}, y) = (1 - \mathrm{BLEU}(y^{(i)}, y))$ based on sentence-wise BLEU (Nakov et al., 2012) is used. Algorithm 1, called "Response-based Online Learning" in Riezler et al. (2014), is based on contrasting a "positive" translation y^+ that receives positive feedback, has a high model score, and a low cost of predicting y instead of $y^{(i)}$, with a "negative" translation y^- that leads to negative feedback, has a high model score, and a high cost:

$$y^+ = \underset{y \in Y(x^{(i)}): \mathrm{e}(y)=1}{\arg\max} \Big(\mathrm{s}(x^{(i)}, y; w) - \mathrm{c}(y^{(i)}, y) \Big),$$

$$y^- = \underset{y \in Y(x^{(i)}): \mathrm{e}(y)=0}{\arg\max} \Big(\mathrm{s}(x^{(i)}, y; w) + \mathrm{c}(y^{(i)}, y) \Big).$$

The central algorithm operates as follows: The SMT system predicts translation $\hat{y}$, and in case of positive task feedback, the prediction is accepted and stored as positive example by setting $y^+ \leftarrow \hat{y}$. In that case, y^- needs to be computed in order to perform the stochastic gradient descent update of the weight

Algorithm 1 Response-based Online Learning

repeat
  for $i = 1, \ldots, n$ **do**
    Receive input string $x^{(i)}$
    Predict translation $\hat{y}$
    Receive task feedback $\mathrm{e}(\hat{y}) \in \{1, 0\}$
    if $\mathrm{e}(\hat{y}) = 1$ **then**
      $y^+ \leftarrow \hat{y}$
      Store $\hat{y}$ as reference $y^{(i)}$ for $x^{(i)}$
      Compute y^-
    else
      $y^- \leftarrow \hat{y}$
      Receive reference $y^{(i)}$
      Compute y^+
    end if
    $w \leftarrow w + \eta(\phi(x^{(i)}, y^+) - \phi(x^{(i)}, y^-))$
  end for
until Convergence

vector. If the feedback is negative, the prediction is treated as y^- and y^+ needs to be computed for the update. If either y^+ or y^- cannot be computed, the example is skipped.

4 Scaling Semantic Parsing to Open-domain Database Queries

The main challenge of grounding SMT in semantic parsing for Freebase lies in scaling the semantic parser to the lexical diversity of the open-domain database. Our baseline system is the parser of Berant et al. (2013), called SEMPRE. We first consider the approach presented by Berant and Liang (2014) to scale the baseline to open-domain database queries: In their system, called PARASEMPRE, pairs of logical forms and utterances are generated from a given query and the database, and the pair whose utterance best paraphrases the input query is selected. These new pairs of queries and logical forms are added as ambiguous labels in training a model from query-answer pairs.

Following a similar idea of extending parser coverage by paraphrases, we extend the training set with synonyms from WordNet. This is done by iterating over the queries in the FREE917 dataset. To ensure that the replacement is sensible, each sentence is first POS tagged (Toutanova et al., 2003) and WordNet lookups are restricted to matching POS between synonym and query words, for nouns, verbs, adjectives and adverbs. Lastly, in order to limit the number of retrieved words, a WordNet lookup is performed by carefully choosing from the first three synsets which are ordered from most common to least frequently used sense. Within a synset all words are taken. The new training queries are appended to the training portion of FREE917.

5 Model Selection

The most straightforward strategy to perform model selection for the task of response-based learning for SMT is to rely on parsing evaluation scores that are standardly reported in the literature. However, as we will show experimentally, if precision is taken as the percentage of correct answers out of instances for which a parse could be produced, recall as the percentage of total examples for which a correct answer could be found, and F1 score as their harmonic mean, the metrics are not appropriate for model selection in our case. This is because for our goal of learning the language of correct English database queries from positive and negative parsing feedback, the semantic parser needs to be able to parse and retrieve correct answers for correct database queries, but it must not do so for incorrect queries.

However, information about incorrect queries is ignored in the definition of the metrics given above. In fact, retrieving correct answers for incorrect database queries hurts response-based learning for SMT. The problem lies in the incomplete nature of semantic parsing databases, where terms that are not parsed into logical forms in one context make a crucial difference in another context. For example in Geoquery, the gold standard queries "People in Boulder?" and "Number of people in Boulder?" parse into the same logical form, however, the queries "Give me the cities in Virginia" and "Give me the number of cities in Virginia" have different parses and different answers. While in the first case, for example in German-to-English translation of database queries, the German "Anzahl" may be translated incorrectly without consequences, it is crucial to translate the term into "number" in the second case. On an example from Free917, the SMT system translates the German "Steinformationen" into "kind of stone", which is incorrect in the geological context, where it should be "rock formations". If during response-based learning, the error slips through because of an incomplete parse leading to the correct answer, it might hurt on the test data. Negative parser feedback for incorrect translations is thus crucial for learning how to avoid these cases in response-based SMT.

In order to evaluate parsing performance on incorrect translations, we need to extend standard evaluation data of correct English database queries with evaluation data of incorrect English database queries. For this purpose, we took translations of an out-of-domain SMT system that were judged either grammatically or semantically incorrect by the authors to create a dataset of negative examples. On this dataset, we can define *true positives (TP)* as correct English queries that were given a correct answer by the semantic parser, and *false positives (FP)* as wrong English queries that obtained the correct answer. The crucial evaluation metric is the *false*

Model	#data	F1	FDR
S	620	56.8	28.00
P	620	66.54	25.22
P1	3,982	65.38	**24.89**
P2	6,740	**66.92**	26.38
P3	8,465	66.15	25.97

Table 1: Parsing F1 scores and False Discovery Rate (FDR) for SEMPRE (S), PARASEMPRE (P), and extensions of the latter with synonyms from first one (P1), first two (P2) and first three (P3) synsets, evaluated on the FREE917 test set of correct database queries for F1 and including the test set of incorrect database queries for FDR, and trained on #data training queries. Best results are indicated in **bold face**.

discovery rate (FDR) (Murphy, 2012), defined as *FP/FP+TP*, i.e., as the ratio of false positives out of all positive answer retrieval events.

6 Experiments

We use a data dump of Freebase[1] which was has been indexed by the Virtuoso SPARQL engine[2] as our knowledge base. The corpus used in the experiments is the FREE917 corpus as assembled by Cai and Yates (2013) and consists of 614 training and 276 test queries in English and corresponding logical forms.[3] The dataset of negative examples, i.e., incorrect English database queries that should receive incorrect answers, consists of 166 examples that were judged either grammatically or semantically incorrect by the authors.

The translation of the English queries in FREE917 into German, in order to provide a set of source sentences for SMT, was done by the authors. The SMT framework used is CDEC (Dyer et al., 2010) with standard dense features and additional sparse features as described in Simianer et al. (2012)[4]. Training of the baseline SMT system was performed on the COMMON CRAWL[5] (Smith et al., 2013) dataset consisting of 7.5M parallel English-German segments extracted from the web. Response-based learning for SMT uses the code described in Riezler et al. (2014)[6].

For semantic parsing we use the SEMPRE and PARASEMPRE tools of Berant et al. (2013) and Berant and Liang (2014) which were trained on the training portion of the FREE917 corpus[7]. Further models use the training data enhanced with synonyms from WordNet as described in Section 4. Following Jones et al. (2012), we evaluate semantic parsers according to *precision*, defined as the percentage of correctly answered examples out of those for which a parse could be produced, *recall*, defined as the percentage of total examples answered correctly, and *F1-score*, defined as harmonic mean of precision and recall. Furthermore, we report *false discovery rate (FDR)* on the combined set of 276 correct and 166 incorrect database queries.

Table 1 reports standard parsing evaluation metrics for the different parsers SEMPRE (S), PARASEMPRE (P), and extensions of the latter with synonyms from the first one (P1), first two (P2) and first three (P3) synsets which are ordered according to frequency of use of the sense. As shown in the second column, the size of the training data is increased up to 10 times by using various synonym extensions. As shown in the third column, PARASEMPRE improves F1 by nearly 10 points over SEMPRE. Another 0.5 points are added by extending the training data using two synsets. The third column shows that the system P1 that scored second-worst in terms of F1 score, scores best under the FDR metric[8].

Table 2 shows an evaluation of the use of different parsing models to retrieve correct answers from the FREE917 test set of correct database queries. The systems are applied to translated queries, but evaluated in terms of standard parsing metrics. Statistical significance is measured using an Approximate Randomization test (Noreen, 1989; Riezler and Maxwell, 2005). The baseline system is CDEC as described above. It never sees the FREE917 data during training. As a second baseline method we use a stochastic (sub)gradient descent variant of RAMPION (Gimpel and Smith, 2012). This system is

[1] http://www.freebase.com/

[2] http://virtuoso.openlinksw.com/

[3] Note that we filtered out 33 questions (21 from the training set and 12 from the test set) because their logical forms only returned an empty string as an answer.

[4] https://github.com/pks/cdec-dtrain

[5] http://www.statmt.org/wmt13/training-parallel-commoncrawl.tgz

[6] https://github.com/pks/rebol

[7] www-nlp.stanford.edu/software/sempre

[8] Note that in case of FDR, smaller is better.

	1 CDEC	2 RAMPION	3 REBOL
S	40.0	40.36	42.92[12]
P	42.92	44.59	45.85
P1	42.92	46.36[1]	**48.8**[1]
P2	43.81	45.92	47.06
P3	43.36	45.92	47.49

Table 2: Parsing F1 score on FREE917 test set of translated database queries using different parser models to provide response for translated queries. Best results are indicated in **bold face**. Statistical significance of result differences at $p < 0.05$ are indicated by algorithm number in superscript.

	CDEC	RAMPION	REBOL
F1	0.85	0.29	0.1
FDR	-0.21	-0.58	-0.7

Table 3: Spearman correlation between F1 / FDR from Table 1 and CDEC / RAMPION / REBOL F1 from Table 2.

trained by using the correct English queries in the FREE917 training data as references. Neither CDEC nor RAMPION use parser feedback in training. REBOL (**Re**sponse-**b**ased **O**nline **L**earning) is an implementation of Algorithm 1 described in Section 3. This algorithm makes use of positive parser feedback to convert predicted translation into references, in addition to using the original English queries as references. Training for both RAMPION and REBOL is performed for 10 epochs over the FREE917 training set, using a constant learning rate η that was chosen via cross-validation. All methods then proceed to translate the FREE917 test set. Best results in Table 2 are obtained by using an extension of PARASEMPRE with one synset as parser in response-based learning with REBOL. This parsing system scored best under the FDR metric in Table 1.

Table 3 shows the Spearman rank correlation (Siegel and Castellan, 1988) between the F1 / FDR ranking of semantic parsers from Table 1 and their contribution to F1 scores in Table 2 for parsing query translations of CDEC, RAMPION or REBOL. The system CDEC cannot learn from parser performance based on query translations, thus best results on translated queries correlate positively with good parsing F1 score per se. RAMPION can implicitly take advantage of parsers with good FDR score since learning to move away from translations dissimilar to the reference is helpful if they do not lead to correct answers. REBOL can make the best use of parsers with low FDR score since it can learn to prevent incorrect translations from hurting parsing performance at test time.

7 Conclusion

We presented an adaptation of SMT to translating open-domain database queries by using feedback of a semantic parser to guide learning. Our work highlights an important aspect that is often overlooked in parser evaluation, namely that parser model selection in real-world applications needs to take the possibility of parsing incorrect language into account. We found that for our application of response-based learning for SMT, the key is to learn to prevent cases where the correct answer is retrieved despite the translation being incorrect. This can be avoided by performing model selection on semantic parsers that parse and retrieve correct answers for correct database queries, but do not do retrieve correct answers for incorrect queries.

In our experiments, we found that the parser that contributes most to response-based learning in SMT is one that is carefully extended by paraphrases and synonyms. In future work, we would like to investigate additional techniques for paraphrasing and synonym extension. For example, a good fit for our task of response-based learning for SMT might be Bannard and Callison-Burch (2005)'s approach to paraphrasing via pivoting on SMT phrase tables.

Acknowledgments

This research was supported in part by DFG grant RI-2221/2-1 "Grounding Statistical Machine Translation in Perception and Action".

References

Jacob Andreas, Andreas Vlachos, and Stephen Clark. 2013. Semantic parsing as machine translation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL'13)*, Sofia, Bulgaria.

Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In *Proceed-*

ings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05), Ann Arbor, MI.

Jonathan Berant and Percy Liang. 2014. Semantic parsing via paraphrasing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014)*, Baltimore, MD.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on freebase from question-answer pairs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP'13)*, Seattle, WA.

Qingqing Cai and Alexander Yates. 2013. Large-scale semantic parsing via schema matching and lexicon extenstion. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL'13)*, Sofia, Bulgaria.

Nicolò Cesa-Bianchi and Gàbor Lugosi. 2006. *Prediction, Learning, and Games*. Cambridge University Press.

Chris Dyer, Adam Lopez, Juri Ganitkevitch, Johnathan Weese, Ferhan Ture, Phil Blunsom, Hendra Setiawan, Vladimir Eidelman, and Philip Resnik. 2010. cdec: A decoder, alignment, and learning framework for finite-state and context-free translation models. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL'10)*, Uppsala, Sweden.

Kevin Gimpel and Noah A. Smith. 2012. Structured ramp loss minimization for machine translation. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (HLT-NAACL'12)*, Stroudsburg, PA.

Dan Goldwasser and Dan Roth. 2013. Learning from natural instructions. *Machine Learning*, 94(2):205–232.

Bevan K. Jones, Mark Johnson, and Sharon Goldwater. 2012. Semantic parsing with bayesion tree transducers. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL'12)*, Jeju Island, Korea.

Tom Kwiatowski, Eunsol Choi, Yoav Artzi, and Luke Zettlemoyer. 2013. Scaling semantic parsers with on-the-fly ontology matching. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP'13)*, Seattle, WA.

George A. Miller. 1995. Wordnet: A lexical database for english. *Communications of the ACM*, 38(11):39–41.

Kevin P. Murphy. 2012. *Machine Learning. A Probabilistic Perspective*. The MIT Press.

Preslav Nakov, Francisco Guzmán, and Stephan Vogel. 2012. Optimizing for sentence-level bleu+1 yields short translations. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012)*, Bombay, India.

Eric W. Noreen. 1989. *Computer Intensive Methods for Testing Hypotheses: An Introduction*. Wiley, New York.

Stefan Riezler and John Maxwell. 2005. On some pitfalls in automatic evaluation and significance testing for MT. In *Proceedings of the ACL-05 Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*, Ann Arbor, MI.

Stefan Riezler, Patrick Simianer, and Carolin Haas. 2014. Response-based learning for grounded machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014)*, Baltimore, MD.

Avneesh Saluja, Ian Lane, and Ying Zhang. 2012. Machine translation with binary feedback: A large-margin approach. In *Proceedings of the 10th Biennial Conference of the Association for Machine Translation in the Americas (AMTA'12)*, San Diego, CA.

Pannaga Shivaswamy and Thorsten Joachims. 2012. Online structured prediction via coactive learning. In *Proceedings of the 29th International Conference on Machine Learning (ICML'12)*, Scotland, UK.

Sidney Siegel and John Castellan. 1988. *Nonparametric Statistics for the Behavioral Sciences. Second Edition*. MacGraw-Hill, Boston, MA.

Patrick Simianer, Stefan Riezler, and Chris Dyer. 2012. Joint feature selection in distributed stochastic learning for large-scale discriminative training in SMT. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL'12)*, Jeju Island, South Korea.

Jason Smith, Herve Saint-Amand, Magdalena Plamada, Philipp Koehn, Chris Callison-Burch, and Adam Lopez. 2013. Dirt cheap web-scale parallel text from the Common Crawl. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL'13)*, Sofia, Bulgaria.

Richard S. Sutton and Andrew G. Barto. 1998. *Reinforcement Learning. An Introduction*. The MIT Press.

Kristina Toutanova, Dan Klein, Christopher Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (HLT-NAACL'03)*. Edmonton, Canada.