Response-based Learning for Grounded Machine Translation

Stefan Riezler and **Patrick Simianer** and **Carolin Haas**
Department of Computational Linguistics
Heidelberg University, 69120 Heidelberg, Germany
{riezler,simianer,haas1}@cl.uni-heidelberg.de

Abstract

We propose a novel learning approach for statistical machine translation (SMT) that allows to extract supervision signals for structured learning from an extrinsic response to a translation input. We show how to generate responses by grounding SMT in the task of executing a semantic parse of a translated query against a database. Experiments on the GEOQUERY database show an improvement of about 6 points in F1-score for response-based learning over learning from references only on returning the correct answer from a semantic parse of a translated query. In general, our approach alleviates the dependency on human reference translations and solves the reachability problem in structured learning for SMT.

1 Introduction

In this paper, we propose a novel approach for learning and evaluation in statistical machine translation (SMT) that borrows ideas from response-based learning for grounded semantic parsing. In this framework, the meaning of a sentence is defined in the context of an extrinsic task. Successful communication of meaning is measured by a successful interaction in this task, and feedback from this interaction is used for learning.

We suggest that in a similar way the preservation of meaning in machine translation should be defined in the context of an interaction in an extrinsic task. For example, in the context of a game, a description of a game rule is translated successfully if correct game moves can be performed based only on the translation. In the context of a question-answering scenario, a question is translated successfully if the correct answer is returned based only on the translation of the query.

We propose a framework of response-based learning that allows to extract supervision signals for structured learning from the response of an extrinsic task to a translation input. Here, learning proceeds by "trying out" translation hypotheses, receiving a response from interacting in the task, and converting this response into a supervision signal for updating model parameters. In case of positive feedback, the predicted translation can be treated as reference translation for a structured learning update. In case of negative feedback, a structural update can be performed against translations that have been approved previously by positive task feedback. This framework has several advantages:

- The supervision signal in response-based learning has a different quality than supervision by human-generated reference translations. While a human reference translation is generated independently of the SMT task, conversion of predicted translations into references is always done with respect to a specific task. In this sense we speak of grounding meaning transfer in an extrinsic task.

- Response-based learning can repeatedly try out system predictions by interacting in the extrinsic task. Instead of and in addition to learning from human reference translations, response-based learning allows to convert multiple system translations into references. This alleviates the supervision problem in cases where parallel data are scarce.

- Task-specific response acts upon system translations. This avoids the problem of unreachability of independently generated reference translations by the SMT system.

The proposed approach of response-based learning opens the doors for various extrinsic tasks

in which SMT systems can be trained and evaluated. In this paper, we present a proof-of-concept experiment that uses feedback from a simulated world environment. Building on prior work in grounded semantic parsing, we generate translations of queries, and receive feedback by executing semantic parses of translated queries against the database. Successful response is defined as receiving the same answer from the semantic parses for the translation and the original query. Our experimental results show an improvement of about 6 points in F1-score for response-based learning over standard structured learning from reference translations. We show in an error analysis that this improvement can be attributed to using structural and lexical variants of reference translations as positive examples in response-based learning. Furthermore, translations produced by response-based learning are found to be grammatical. This is due to the possibility to boost similarity to human reference translations by the additional use of a cost function in our approach.

2 Related Work

The key idea of *grounded language learning* is to study natural language in the context of a non-linguistic environment, in which meaning is grounded in perception and/or action. This presents an analogy to human learning, where a learner tests her understanding in an actionable setting. Such a setting can be a simulated world environment in which the linguistic representation can be directly executed by a computer system. For example, in semantic parsing, the learning goal is to produce and successfully execute a meaning representation. Executable system actions include access to databases such as the GEOQUERY database on U.S. geography (Wong and Mooney (2006), *inter alia*), the ATIS travel planning database (Zettlemoyer and Collins (2009), *inter alia*), robotic control in simulated navigation tasks (Chen and Mooney (2011), *inter alia*), databases of simulated card games (Goldwasser and Roth (2013), *inter alia*), or the user-generated contents of FREEBASE (Cai and Yates (2013), *inter alia*). Since there are many possible correct parses, matching against a single gold standard falls short of grounding in a non-linguistic environment. Rather, the semantic context for interpretation, as well as the success criterion in evaluation is defined by successful execution of an action in the extrinsic environment, e.g., by receiving the correct answer from the database or by successful navigation to the destination. Recent attempts to learn semantic parsing from question-answer pairs without recurring to annotated logical forms have been presented by Kwiatowski et al. (2013), Berant et al. (2013), or Goldwasser and Roth (2013). The algorithms presented in these works are variants of structured prediction that take executability of semantic parses into account. Our work builds upon these ideas, however, to our knowledge the presented work is the first to embed translations into grounded scenarios in order to use feedback from interactions in these scenarios for structured learning in SMT.

A recent important research direction in SMT has focused on employing automated translation as an aid to human translators. *Computer assisted translation* (CAT) subsumes several modes of interaction, ranging from binary feedback on the quality of the system prediction (Saluja et al., 2012), to human post-editing operations on a system prediction resulting in a reference translation (Cesa-Bianchi et al., 2008), to human acceptance or overriding of sentence completion predictions (Langlais et al., 2000; Barrachina et al., 2008; Koehn and Haddow, 2009). In all interaction scenarios, it is important that the system learns dynamically from its errors in order to offer the user the experience of a system that adapts to the provided feedback. Since retraining the SMT model after each interaction is too costly, *online adaptation* after each interaction has become the learning protocol of choice for CAT. Online learning has been applied in generative SMT, e.g., using incremental versions of the EM algorithm (Ortiz-Martínez et al., 2010; Hardt and Elming, 2010), or in discriminative SMT, e.g., using perceptron-type algorithms (Cesa-Bianchi et al., 2008; Martínez-Gómez et al., 2012; Wäschle et al., 2013; Denkowski et al., 2014). In a similar way to deploying human feedback, extrinsic loss functions have been used to provide learning signals for SMT. For example, Nikoulina et al. (2012) propose a setup where an SMT system feeds into cross-language information retrieval, and receives feedback from the performance of translated queries with respect to cross-language retrieval performance. This feedback is used to train a reranker on an n-best list of translations order with respect to retrieval performance. In con-

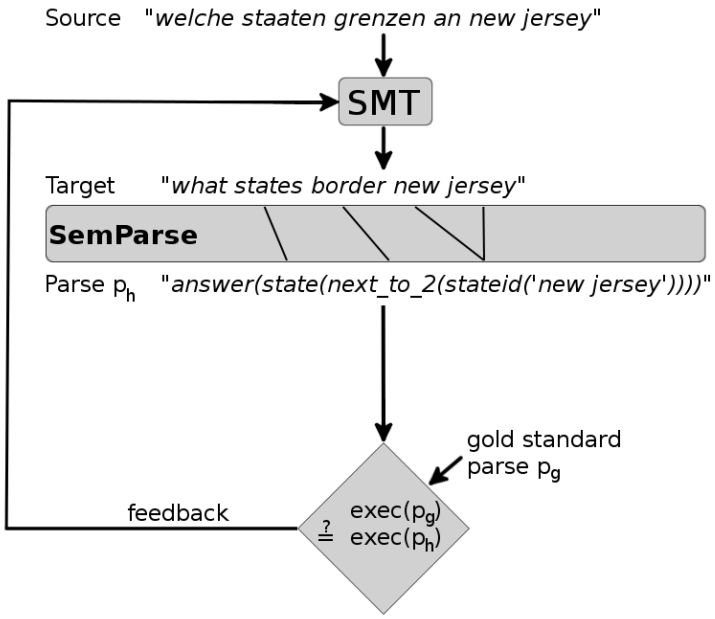

Figure 1: Response-based learning cycle for grounding SMT in virtual trivia gameplay.

trast to our work, all mentioned approaches to interactive or adaptive learning in SMT rely on human post-edits or human reference translations. Our work differs from these approaches in that exactly this dependency is alleviated by learning from responses in an extrinsic task.

Interactive scenarios have been used for evaluation purposes of translation systems for nearly 50 years, especially using *human reading comprehension* testing (Pfafflin, 1965; Fuji, 1999; Jones et al., 2005), and more recently, using face-to-face conversation mediated via machine translation (Sakamoto et al., 2013). However, despite offering direct and reliable prediction of translation quality, the cost and lack of reusability has confined task-based evaluations involving humans to *testing* scenarios, but prevented a use for interactive *training* of SMT systems as in our work.

Lastly, our work is related to *cross-lingual natural language processing* such as cross-lingual question answering or cross-lingual information retrieval as conducted at recent evaluation campaigns of the CLEF initiative.[1] While these approaches focus on improvements of the respective natural language processing task, our goal is to improve SMT by gathering feedback from the task.

[1] http://www.clef-initiative.eu

3 Grounding SMT in Semantic Parsing

In this paper, we present a proof-of-concept of our ideas of embedding SMT into simulated world environments as used in semantic parsing. We use the well-known GEOQUERY database on U.S. geography for this purpose. Embedding SMT in a semantic parsing scenario means to define translation quality by the ability of a semantic parser to construct a meaning representation from the translated query, which returns the correct answer when executed against the database. If viewed as simulated gameplay, a valid game move in this scenario returns the correct answer to a translated query.

The diagram in Figure 1 gives a sketch of response-based learning from semantic parsing in the geographical domain. Given a manual German translation of the English query as source sentence, the SMT system produces an English target translation. This sentence is fed into a semantic parser that produces an executable parse representation p_h. Feedback is generated by executing the parse against the database of geographical facts. Positive feedback means that the correct answer is received, i.e., $\text{exec}(p_g) \stackrel{?}{=} \text{exec}(p_h)$ indicates that the same answer is received from the gold standard parse p_g and the parse for the hypothesis translation p_h; negative feedback results in case a different or no answer is received.

The key advantage of response-based learning

is the possibility to receive positive feedback even from predictions that differ from gold standard reference translations, but yet receive the correct answer when parsed and matched against the database. Such structural and lexical variation broadens the learning capabilities in contrast to learning from fixed labeled data. For example, assume the following English query in the geographical domain, and assume positive feedback from executing the corresponding semantic parse against the geographical database:

```
Name prominent elevations in the
USA
```

The manual translation of the English original reads

```
Nenne prominente Erhebungen in
den USA
```

An automatic translation[2] of the German string produces the result

```
Give prominent surveys in the US
```

This translation will trigger negative task-based feedback: A comparison with the original allows the error to be traced back to the ambiguity of the German word `Erhebung`. Choosing a general domain translation instead of a translation appropriate for the geographical domain hinders the construction of a semantic parse that returns the correct answer from the database. An alternative translation might look as follows:

```
Give prominent heights in the US
```

Despite a large difference to the original English string, key terms such as `elevations` and `heights`, or `USA` and `US`, can be mapped into the same predicate in the semantic parse, thus allowing to receive positive feedback from parse execution against the geographical database.

4 Response-based Online Learning

Recent approaches to machine learning for SMT formalize the task of discriminating good from bad translations as a structured prediction problem. Assume a joint feature representation $\phi(x, y)$ of input sentences x and output translations $y \in Y(x)$, and a linear scoring function $s(x, y; w)$ for predicting a translation $\hat{y}$ (where $\langle \cdot, \cdot \rangle$ denotes the standard vector dot product) s.t.

$$\hat{y} = \arg\max_{y \in Y(x)} \mathrm{s}(x, y; w) = \arg\max_{y \in Y(x)} \langle w, \phi(x, y) \rangle .$$

[2] `http://translate.google.com`

The structured perceptron algorithm (Collins, 2002) learns an optimal weight vector w by updating w on input $x^{(i)}$ by the following rule, in case the predicted translation $\hat{y}$ is different from and scored higher than the reference translation $y^{(i)}$:

$$w = w + \phi(x^{(i)}, y^{(i)}) - \phi(x^{(i)}, \hat{y}).$$

This stochastic structural update aims to demote weights of features corresponding to incorrect decisions, and to promote weights of features for correct decisions.

An application of structured prediction to SMT involves more than a straightforward replacement of labeled output structures by reference translations. Firstly, update rules that require to compute a feature representation for the reference translation are suboptimal in SMT, because often human-generated reference translations cannot be generated by the SMT system. Such "unreachable" gold-standard translations need to be replaced by "surrogate" gold-standard translations that are close to the human-generated translations and still lie within the reach of the SMT system. Computation of distance to the reference translation usually involves cost functions based on sentence-level BLEU (Nakov et al. (2012), *inter alia*) and incorporates the current model score, leading to various ramp loss objectives described in Gimpel and Smith (2012).

An alternative approach to alleviate the dependency on labeled training data is response-based learning. Clarke et al. (2010) or Goldwasser and Roth (2013) describe a response-driven learning framework for the area of semantic parsing: Here a meaning representation is "tried out" by iteratively generating system outputs, receiving feedback from world interaction, and updating the model parameters. Applied to SMT, this means that we predict translations and use positive response from acting in the world to create "surrogate" gold-standard translations. This decreases the dependency on a few (mostly only one) reference translations and guides the learner to promote translations that perform well with respect to the extrinsic task.

In the following, we will present a framework that combines standard structured learning from given reference translations with response-based learning from task-approved references. We need to ensure that gold-standard translations lead to positive task-based feedback, that means they can

be parsed and executed successfully against the database. In addition, we can use translation-specific cost functions based on sentence-level BLEU in order to boost similarity of translations to human reference translations.

We denote feedback by a binary execution function $\mathrm{e}(y) \in \{1, 0\}$ that tests whether executing the semantic parse for the prediction against the database receives the same answer as the parse for the gold standard reference. Our cost function $\mathrm{c}(y^{(i)}, y) = (1 - \mathrm{BLEU}(y^{(i)}, y))$ is based on a version of sentence-level BLEU Nakov et al. (2012). Define y^+ as a surrogate gold-standard translation that receives positive feedback, has a high model score, and a low cost of predicting y instead of $y^{(i)}$:

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

The opposite of y^+ is the translation y^- that leads to negative feedback, has a high model score, and a high cost. It is defined as follows:

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

Update rules can be derived by minimization of the following ramp loss objective:

$$\min_w \Bigg(- \max_{y \in Y(x^{(i)}):\mathrm{e}(y)=1} \Big(\mathrm{s}(x^{(i)}, y; w) - \mathrm{c}(y^{(i)}, y) \Big) + \max_{y \in Y(x^{(i)}):\mathrm{e}(y)=0} \Big(\mathrm{s}(x^{(i)}, y; w) + \mathrm{c}(y^{(i)}, y) \Big) \Bigg) .$$

Minimization of this objective using stochastic (sub)gradient descent (McAllester and Keshet, 2011) yields the following update rule:

$$w = w + \phi(x^{(i)}, y^+) - \phi(x^{(i)}, y^-).$$

The intuition behind this update rule is to discriminate the translation y^+ that leads to positive feedback and best approximates (or is identical to) the reference within the means of the model from a translation y^- which is favored by the model but does not execute and has high cost. This is done by putting all the weight on the former.

Algorithm 1 presents pseudo-code for our response-driven learning scenario. Upon predicting translation $\hat{y}$, in case of positive feedback from the task, we treat the prediction as surrogate reference by setting $y^+ \leftarrow \hat{y}$, and by adding it to the set of reference translations for future use. Then we need to compute y^-, and update by the difference in feature representations of y^+ and y^-, at a learning rate η. If the feedback is negative, we want to move the weights away from the prediction, thus we treat it as y^-. To perform an update, we need to compute y^+. If either y^+ or y^- cannot be computed, the example is skipped.

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

The sketched algorithm allows several variations. In the form depicted above, it allows to use human reference translations in addition to task-approved surrogate references. The cost function can be implemented by different versions of sentence-wise BLEU, or it can be omitted completely so that learning relies on task-based feedback alone, similar to algorithms recently suggested for semantic parsing (Goldwasser and Roth, 2013; Kwiatowski et al., 2013; Berant et al., 2013). Lastly, regularization can be introduced by using update rules corresponding to primal form optimization variants of support vector machines (Collobert and Bengio, 2004; Chapelle, 2007; Shalev-Shwartz et al., 2007).

5 Experiments

5.1 Experimental Setup

In our experiments, we use the GEOQUERY database on U.S. geography as provided by Jones

method	precision	recall	F1	BLEU
1 CDEC	63.67	58.21	60.82	46.53
2 EXEC	70.36	63.57	66.79[1]	48.00[1]
3 RAMPION	75.58	69.64	72.49[12]	**56.64**[12]
4 REBOL	**81.15**	**75.36**	**78.15**[123]	55.66[12]

Table 1: Experimental results using extended parser for returning answers from GEOQUERY (precision, recall, F1) and n-gram match to original English query (BLEU) on 280 re-translated test examples. Best results for each column are highlighted in **bold face**. Superscripts [1234] denote a significant improvement over the respective method.

method	precision	recall	F1	BLEU
1 CDEC	65.59	57.86	61.48	46.53
2 EXEC	66.54	61.79	64.07	46.00
3 RAMPION	67.68	63.57	65.56	**55.67**[12]
4 REBOL	**70.68**	**67.14**	**68.86**[12]	**55.67**[12]

Table 2: Experimental results using the original parser for returning answers from GEOQUERY (precision, recall, F1) and n-gram match to original English query (BLEU) on 280 re-translated test examples.

et al. (2012).[3] The dataset includes 880 English questions and their logical forms. The English strings were manually translated into German by the authors of Jones et al. (2012)), and corrected for typos by the authors of this paper. We follow the provided split into 600 training examples and 280 test examples.

For response-based learning, we retrained the semantic parser of Andreas et al. (2013)[4] on the full 880 GEOQUERY examples in order to reach full parse coverage. This parser is itself based on SMT, trained on parallel data consisting of English queries and linearized logical forms, and on a language model trained on linearized logical forms. We used the hierarchical phrase-based variant of the parser. Note that we do not use GEOQUERY test data in SMT training. Parser training includes GEOQUERY test data in order to be less dependent on parse and execution failures in the evaluation: If a translation system, response-based or reference-based, translates the German input into the gold standard English query it should be rewarded by positive task feedback. To double-check whether including the 280 test examples in parser training gives an unfair advantage to response-based learning, we also present experimental results using the original parser of Andreas et al. (2013) that is trained only on the 600 GEOQUERY training examples.

The bilingual SMT system used in our experiments is the state-of-the-art SCFG decoder CDEC (Dyer et al., 2010)[5]. We built grammars using its implementation of the suffix array extraction method described in Lopez (2007). For language modeling, we built a modified Kneser-Ney smoothed 5-gram language model using the English side of the training data. We trained the SMT system on the English-German parallel web data provided in the COMMON CRAWL[6] (Smith et al., 2013) dataset.

5.2 Compared Systems

Method 1 is the baseline system, consisting of the CDEC SMT system trained on the COMMON CRAWL data as described above. This system does not use any GEOQUERY data for training. Methods 2-4 use the 600 training examples from GEOQUERY for discriminative training only.

Variants of the response-based learning algorithm described above are implemented as a standalone tool that operates on CDEC n-best lists of 10,000 translations of the GEOQUERY training data. All variants use sparse features of CDEC as described in Simianer et al. (2012) that extract rule

[3] http://homepages.inf.ed.ac.uk/s1051107/geoquery-2012-08-27.zip

[4] https://github.com/jacobandreas/smt-semparse

[5] https://github.com/redpony/cdec

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

prediction:	how many inhabitants has new york
reference:	how many people live in new york
prediction:	how big is the population of texas
reference:	how many people live in texas
prediction:	which are the cities of the state with the highest elevation
reference:	what are the cities of the state with the highest point
prediction:	how big is the population of states , through which the mississippi runs
reference:	what are the populations of the states through which the mississippi river runs
prediction:	what state borders california
reference:	what is the adjacent state of california
prediction:	what are the capitals of the states which have cities with the name durham
reference:	what is the capital of states that have cities named durham
prediction:	what rivers go through states with the least cities
reference:	which rivers run through states with fewest cities

Table 3: Predicted translations by response-based learning (REBOL) leading to positive feedback versus gold standard references.

shapes, rule identifiers, and bigrams in rule source and target directly from grammar rules. Method 4, named REBOL, implements REsponse-Based Online Learning by instantiating y^+ and y^- to the form described in Section 4: In addition to the model score s, it uses a cost function c based on sentence-level BLEU (Nakov et al., 2012) and tests translation hypotheses for task-based feedback using a binary execution function e. This algorithm can convert predicted translations into references by task-feedback, and additionally use the given original English queries as references. Method 2, named EXEC, relies on task-execution by function e and searches for executable or non-executable translations with highest score s to distinguish positive from negative training examples. It does not use a cost function and thus cannot make use of the original English queries.

We compare response-based learning with a standard structured prediction setup that omits the use of the execution function e in the definition of y^+ and y^-. This algorithm can be seen as a stochastic (sub)gradient descent variant of RAMPION (Gimpel and Smith, 2012). It does not make use of the semantic parser, but defines positive and negative examples based on score s and cost c with respect to human reference translations.

We report BLEU (Papineni et al., 2001) of translation system output measured against the original English queries. Furthermore, we report precision, recall, and F1-score for executing semantic parses built from translation system outputs against the GEOQUERY database. Precision is defined as the percentage of correctly answered examples out of those for which a parse could be produced; recall is defined as the percentage of total examples answered correctly; F1-score is the harmonic mean of both. Statistical significance is measured using Approximate Randomization (Noreen, 1989) where result differences with a p-value smaller than 0.05 are considered statistically significant.

Methods 2-4 perform structured learning for SMT on the 600 GEOQUERY training examples and re-translate the 280 unseen GEOQUERY test data, following the data split of Jones et al. (2012). Training for RAMPION, REBOL and EXEC was repeated for 10 epochs. The learning rate η is set to a constant that is adjusted by cross-validation on the 600 training examples.

5.3 Empirical Results

We present an experimental comparison of the four different systems according to BLEU and

reference	RAMPION	REBOL
how many colorado rivers are there	how many rivers with the name colorado gives it	how many rivers named colorado are there
what are the populations of states which border texas	how big are the populations of the states , which in texas borders	how big are the populations of the states which on texas border
what is the biggest capital city in the us	what is the largest city in the usa	what is the largest capital in the usa
what state borders new york	what states limits of new york	what states border new york
which states border the state with the smallest area	what states boundaries of the state with the smallest surface area	what states border the state with the smallest surface area

Table 4: Predicted translations by response-based learning (REBOL) leading to positive feedback versus translations by supervised structured learning (RAMPION) leading to negative feedback.

F1, using an extended semantic parser (trained on 880 GEOQUERY examples) and the original parser (trained on 600 GEOQUERY training examples). The extended parser reaches and F1-score of 99.64% on the 280 GEOQUERY test examples; the original parser yields an F1-score of 82.76%.

Table 1 reports results for the extended semantic parser. A system ranking according to F1-score shows about 6 points difference between the respective methods, ranking REBOL over RAMPION, EXEC and CDEC. The exploitation of task-feedback allows both EXEC and REBOL to improve task-performance over the baseline. REBOL's combination of task feedback with a cost function achieves the best results since positively executable hypotheses and reference translations can both be exploited to guide the learning process. Since all English reference queries lead to positively executable parses in the setup that uses the extended semantic parser, RAMPION implicitly also has access to task feedback. This allows RAMPION to improve F1 over the baseline. All result differences are statistically significant.

In terms of BLEU score measured against the original English GEOQUERY queries, the best nominal result is obtained by RAMPION which uses them as reference translations. REBOL performs worse since BLEU performance is optimized only implicitly in cases where original English queries function as positive examples. However, the result differences between these two systems do not score as statistically significant. Despite not optimizing for BLEU performance against references, the fact that positively executable translations include the references allows even EXEC to improve BLEU over CDEC which does not use GEOQUERY data at all in training. This result difference is statistically significant.

Table 2 compares the same systems using the original parser trained on 600 training examples. The system ranking according to F1-score shows the same ordering that is obtained when using an extended semantic parser. However, the respective methods are separated only by 3 or less points in F1 score such that only the result difference of REBOL over the baseline CDEC and over EXEC is statistically significant. We conjecture that this is due to a higher number of empty parses on the test set which makes this comparison unstable.

In terms of BLEU measured against the original queries, the result differences between REBOL and RAMPION are not statistically significant, and neither are the result differences between EXEC and CDEC. The result differences between systems of the former group and the systems of latter group are statistically significant.

5.4 Error Analysis

For a better understanding of the differences between the results produced by supervised and response-based learning, we conducted an er-

reference	RAMPION	REBOL
how many states have a higher point than the highest point of the state with the largest capital city in the us	how many states have a higher nearby point as the highest point of the state with the largest capital in the usa	how many states have a high point than the highest point of the state with the largest capital in the usa
how tall is mount mckinley	how high is mount mckinley	what is mount mckinley
what is the longest river that flows through a state that borders indiana	how is the longest river , which runs through a state , borders the of indiana	what is the longest river which runs through a state of indiana borders
what states does the mississippi river run through	through which states runs the mississippi	through which states is the mississippi
which is the highest peak not in alaska	how is the highest peaks of not in alaska is	what is the highest peak in alaska is

Table 5: Predicted translations where supervised structured learning (RAMPION) leads to positive feedback versus translations by response-based learning (REBOL) leading to negative feedback.

ror analysis on the test examples. Table 3 shows examples where the translation predicted by response-based learning (REBOL) differs from the gold standard reference translation, but yet leads to positive feedback via a parse that returns the correct answer from the database. The examples show structural and lexical variation that leads to differences on the string level at equivalent positive feedback from the extrinsic task. This can explain the success of response-based learning: Lexical and structural variants of reference translations can be used to boost model parameters towards translations with positive feedback, while the same translations might be considered as negative examples in standard structured learning.

Table 4 shows examples where translations from REBOL and RAMPION differ from the gold standard reference, and predictions by REBOL lead to positive feedback, while predictions by RAMPION lead to negative feedback. Table 5 shows examples where translations from RAMPION outperform translations from REBOL in terms of task feedback. We see that predictions from both systems are in general grammatical. This can be attributed to the use of sentence-level BLEU as cost function in RAMPION and REBOL. Translation errors of RAMPION can be traced back to mistranslations of key terms (`city` versus `capital`, `limits` or `boundaries` versus `border`). Translation errors of REBOL more frequently show missing translations of terms.

6 Conclusion

We presented a proposal for a new learning and evaluation framework for SMT. The central idea is to ground meaning transfer in successful interaction in an extrinsic task, and use task-based feedback for structured learning. We presented a proof-of-concept experiment that defines the extrinsic task as executing semantic parses of translated queries against the GEOQUERY database. Our experiments show an improvement of about 6 points in F1-score for response-based learning over structured learning from reference translations. Our error analysis shows that response-based learning generates grammatical translations which is due to the additional use of a cost function that boosts similarity of translations to human reference translations.

In future work, we would like to extend our work on embedding SMT in virtual gameplay to larger and more diverse datasets, and involve human feedback in the response-based learning loop.

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.

Sergio Barrachina, Oliver Bender, Francisco Casacuberta, Jorge Civera, Elsa Cubel, Shahram Khadivi, Antonio Lagarda, Hermann Ney, Jesús Tomás, Enrique Vidal, and Juan-Miguel Vilar. 2008. Statistical approaches to computer-assisted translation. *Computational Linguistics*, 35(1):3–28.

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, Gabriele Reverberi, and Sandor Szedmak. 2008. Online learning algorithms for computer-assisted translation. Technical report, SMART (`www.smart-project.eu`).

Olivier Chapelle. 2007. Training a support vector machine in the primal. *Neural Computation*, 19(5):1155–1178.

David L. Chen and Raymond J. Mooney. 2011. Learning to interpret natural language navigation instructions from observations. In *Proceedings of the 25th AAAI Conference on Artificial Intelligence (AAAI'11)*, pages 859–866, San Francisco, CA.

James Clarke, Dan Goldwasser, Wing-Wei Chang, and Dan Roth. 2010. Driving semantic parsing from the world's response. In *Proceedings of the 14th Conference on Natural Language Learning (CoNLL'10)*, pages 18–27, Uppsala, Sweden.

Michael Collins. 2002. Discriminative training methods for hidden markov models: theory and experiments with perceptron algorithms. In *Proceedings of the conference on Empirical Methods in Natural Language Processing (EMNLP'02)*, Philadelphia, PA.

Ronan Collobert and Samy Bengio. 2004. Links between perceptrons, MLPs, and SVMs. In *Proceedings of the 21st International Conference on Machine Learning (ICML'04)*, Banff, Canada.

Michael Denkowski, Chris Dyer, and Alon Lavie. 2014. Learning from post-editing: Online model adaptation for statistical machine translation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL'14)*, Gothenburg, Sweden.

Chris Dyer, Adam Lopez, Juri Ganitkevitch, Jonathan 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 ACL 2010 System Demonstrations*, Uppsala, Sweden.

Masaru Fuji. 1999. Evaluation experiment for reading comprehension of machine translation outputs. In *Proceedings of the Machine Translation Summit VII*, Singapore.

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

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

Daniel Hardt and Jakob Elming. 2010. Incremental re-training for post-editing SMT. In *Proceedings of the 9th Conference of the Association for Machine Tranlation in the Americas (AMTA'10)*, Denver, CO.

Douglas Jones, Wade Shen, Neil Granoien, Martha Herzog, and Clifford Weinstein. 2005. Measuring translation quality by testing english speakers with a new defense language proficiency test for arabic. In *Proceedings of 2005 International Conference on Intelligence Analysis*, McLean, VA.

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.

Philipp Koehn and Barry Haddow. 2009. Interactive assistance to human translators using statistical machine translation methods. In *Proceedings of MT Summit XII*, Ottawa, Ontario, Canada.

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.

Philippe Langlais, George Foster, and Guy Lapalme. 2000. Transtype: a computer-aided translation typing system. In *Proceedings of the ANLP-NAACL 2000 Workshop on Embedded Machine Translation Systems*, Seattle, WA.

Adam Lopez. 2007. Hierarchical phrase-based translation with suffix arrays. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2007)*, Prague, Czech Republic.

Pascual Martínez-Gómez, Germán Sanchis-Trilles, and Francisco Casacuberta. 2012. Online adaptation

strategies for statistical machine translation in post-editing scenarios. *Pattern Recognition*, 45(9):3193–3202.

David McAllester and Joseph Keshet. 2011. Generalization bounds and consistency for latent structural probit and ramp loss. In *Proceedings of the 25th Annual Conference on Neural Information Processing Sytems (NIPS 2011)*, Granada, Spain.

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.

Vassilina Nikoulina, Bogomil Kovachev, Nikolaos Lagos, and Christof Monz. 2012. Adaptation of statistical machine translation model for cross-lingual information retrieval in a service context. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL'12)*, Avignon, France.

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

Daniel Ortiz-Martínez, Ismal García-Varea, and Francisco Casacuberta. 2010. Online learning for interactive statistical machine translation. In *Proceedings of the Human Language Technologies conference and the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL'10)*, Los Angeles, CA.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. Bleu: a method for automatic evaluation of machine translation. Technical Report IBM Research Division Technical Report, RC22176 (W0190-022), Yorktown Heights, N.Y.

Sheila M. Pfafflin. 1965. Evaluation of machine translations by reading comprehension tests and subjective judgements. *Mechanical Translation and Computational Linguistics*, 8(2):2–8.

Akiko Sakamoto, Nayuko Watanabe, Satoshi Kamatani, and Kazuo Sumita. 2013. Development of a simultaneous interpretation system for face-to-face services and its evaluation experiment in real situation. In *Proceedings of the Machine Translation Summit XIV*, Nice, France.

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.

Shai Shalev-Shwartz, Yoram Singer, and Nathan Srebro. 2007. Pegasos: Primal Estimated sub-GrAdient SOlver for SVM. In *Proceedings of the 24th International Conference on Machine Learning (ICML'07)*, Corvallis, OR.

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 2012)*, Jeju, Korea.

Jason R. 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.

Katharina Wäschle, Patrick Simianer, Nicola Bertoldi, Stefan Riezler, and Marcello Federico. 2013. Generative and discriminative methods for online adaptation in SMT. In *Proceedings of the Machine Translation Summit XIV*, Nice, France.

Yuk Wah Wong and Raymond J. Mooney. 2006. Learning for semantic parsing with statistical machine translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL'06)*, New York City, NY.

Luke S. Zettlemoyer and Michael Collins. 2009. Learning context-dependent mappings from sentences to logical form. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics (ACL-IJCNLP'09)*, Singapore.