Improving Black-box Speech Recognition using Semantic Parsing

Rodolfo Corona and **Jesse Thomason** and **Raymond J. Mooney**
Department of Computer Science, University of Texas at Austin
{rcorona, jesse, mooney}@cs.utexas.edu

Abstract

Speech is a natural channel for human-computer interaction in robotics and consumer applications. Natural language understanding pipelines that start with speech can have trouble recovering from speech recognition errors. Black-box automatic speech recognition (ASR) systems, built for general purpose use, are unable to take advantage of in-domain language models that could otherwise ameliorate these errors. In this work, we present a method for re-ranking black-box ASR hypotheses using an in-domain language model and semantic parser trained for a particular task. Our re-ranking method significantly improves both transcription accuracy and semantic understanding over a state-of-the-art ASR's vanilla output.

1 Introduction

Voice control makes robotic and computer systems more accessible in consumer domains. Collecting sufficient data to train ASR systems using current state of the art methods, such as deep neural networks (Graves and Jaitly, 2014; Xiong et al., 2016), is difficult. Thus, it is common to use well-trained, cloud-based ASR systems. These systems use general language models not restricted to individual application domains. However, for an ASR in a larger pipeline, the expected words and phrases from users will be biased by the application domain. The general language model of a black-box ASR leads to more errors in transcription. These errors can cause cascading problems in a language understanding pipeline.

In this paper, we demonstrate that an in-domain language model and semantic parser can be used to improve black-box ASR transcription and downstream semantic accuracy. We consider a robotics domain, where language understanding is key for ensuring effective performance and positive user experiences (Thomason et al., 2015). We collect a dataset of spoken robot commands paired with transcriptions and semantic forms to evaluate our method.[1] Given a list of ASR hypotheses, we re-rank the list to choose the hypothesis scoring best between an in-domain trained semantic parser and language model (Figure 1). This work is inspired by other re-ranking methods which have used prosodic models (Ananthakrishnan and Narayanan, 2007), phonetic post-processing (Twiefel et al., 2014), syntactic parsing (Zechner and Waibel, 1998; Basili et al., 2013), as well as features from search engine results (Peng et al., 2013).

Other work has similarly employed semantics to improve ASR performance, for example by assigning semantic category labels to entire utterances and re-ranking the ASR n-best list (Morbini et al., 2012), jointly modeling the word and semantic tag sequence (Deoras et al., 2013), and learning a semantic grammar for use by both the ASR system and semantic parser (Gaspers et al., 2015). Closest to our work is that of Erdogan et al. (2005), which uses maximum entropy modeling to combine information from the semantic parser and ASR's language model for re-ranking. Although their method could be adapted for use with a black-box ASR, their parsing framework employs a treebanked dataset of parses (Davies et al., 1999; Jelinek et al., 1994). In contrast, the Combinatory Categorial Grammar (CCG) framework which we use in this work only requires that the root-level semantic form be given along with groundings for a small number of words (see section 2.2), significantly reducing the cost of data collection. Further, although they also experiment with an out-of-the-box language model, they only

[1] Our dataset will be made available upon request. Source code can be found in: https://github.com/thomason-jesse/nlu_pipeline/tree/speech

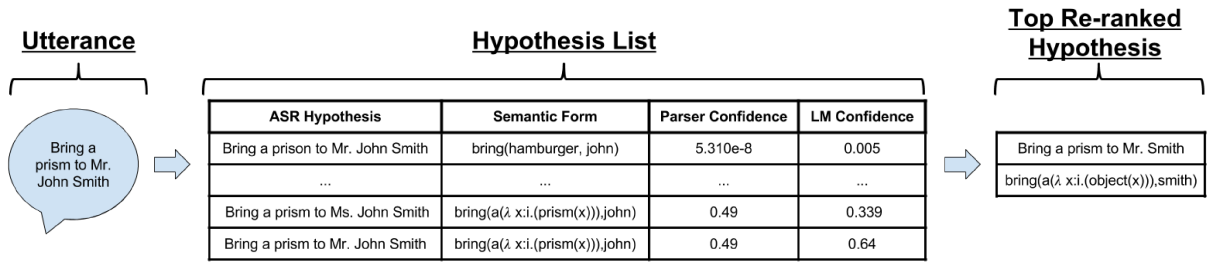

Figure 1: Our proposed methodology. The black-box ASR outputs an ordered list of its top hypotheses. Each hypothesis is given confidence scores by an in-domain semantic parser and language model, which are then used to re-rank the list. In this example, the parser has learned that "Mr." and "Ms." are functionally equivalent, while the language model has learned that "Mr." co-occurs with "John" more than "Ms." does. Together, they guide us to select the correct transcription.

measure for improvements in transcription accuracy, which may not entail improvements in language understanding (Wang et al., 2003).

To the best of our knowledge, our method is the first to improve language understanding by employing a low-cost semantic parser and language model post-hoc on a high-cost, black-box ASR system. This significantly lowers word error rate (WER) while increasing semantic accuracy.

2 Methodology

Given a user utterance U, the black-box ASR system generates a list of n-best hypotheses H. For each hypothesis $h \in H$, we produce an interpolated score[2] $S(h)$ from its language model score $S_{lm}(h)$ and semantic parser score $S_{sem}(h)$.[3] Parser confidence scores vary by orders of magnitude between hypotheses, making it difficult to find a meaningful interpolation weight α between the language model and semantic parser. We therefore normalize over the sum of scores in each hypothesis list for each model. We then choose the highest scoring hypothesis h^*:

$$h^* = \arg\max_{h \in H} \left(S(h) \right); \quad (1)$$

$$S(h) = (1 - \alpha) \cdot S_{lm}(h) + \alpha \cdot S_{sem}(h). \quad (2)$$

[2]In order to avoid underflow errors, all computations are done in log space.

[3]We do not assume a black-box ASR system will provide confidence scores for its n-best list. Google Speech API, for example, often only shows confidence for the top hypothesis. Preliminary experiments using proxy scores based on rank did not improve performance.

2.1 Language Model

We implement an in-domain language model using the SRI Language Modeling Toolkit (Stolcke et al., 2002). We use a trigram back-off model with Witten-Bell discounting (Witten and Bell, 1991) and an open vocabulary. We use perplexity-based, length-normalized scores to compare hypotheses with different numbers of word tokens.

2.2 Semantic Parsing Model

For semantic parsing, we used a CCG (Steedman and Baldridge, 2011) based probabilistic CKY parser.

The parser consists of a lexicon whose entries are words paired with syntactic categories and semantic forms (see Table 1 for example lexical entries). CCG categories may be atomic or functional, with functional categories specifying combinatory rules for adjacent categories. These may be expressed logically by representing semantic forms using a formalism such as lambda calculus. For example, consider the combination between the functional category (NP/N) and the atomic category N, along with its pertaining lambda cal-

Surface Form	CCG Category	Semantic Form
walk	S/PP	$\lambda x.(\text{walk}(x))$
to	PP/NP	$\lambda x.(x)$
john	N	john

Table 1: Example lexical entries in our domain. Given an initial lexicon, additional entries are induced by the parser during training for use at test time.

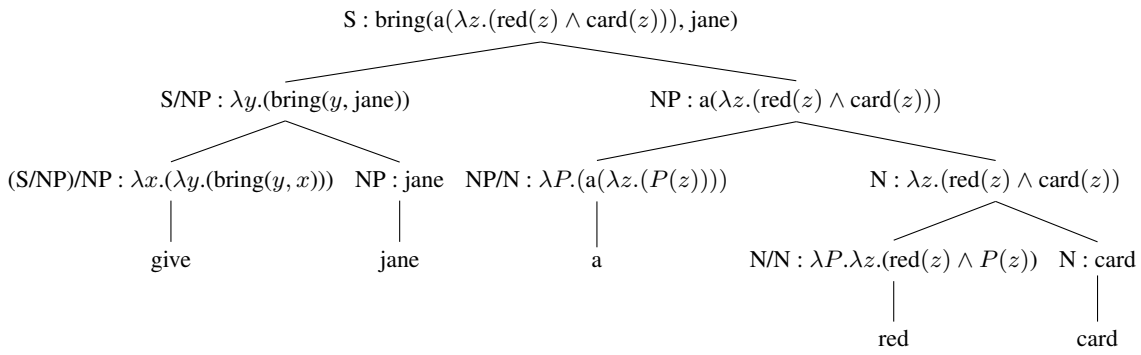

Figure 2: A parse tree of the phrase "give jane a red card." The token *give* is an imperative, taking two noun phrases on its right which represent the recipient and the patient of the action (the robot is the implicit agent in the command). *jane* immediately resolves to a noun phrase. *red* is an adjectival predicate, consuming the noun predicate *card* on its right, the result of which is consumed by the determiner *a* in order to form a complete noun phrase.

culus expression:

$$(NP/N)\ N \implies NP$$
$$(\lambda x.(x))\ y \implies y$$

The combinatory rules of a CCG implicitly define a grammar. An example CCG parse tree may be seen in Figure 2.

Following Zettlemoyer and Collins (2005), gold labels for parsing contain only root-level semantic forms, and only a small set of bootstrapping lexical entries are provided. This necessitates that latent parse trees be inferred and that additional lexical entries be induced during training.

Given a corpus of training examples T of sentences paired with their semantic forms, we follow the framework proposed by (Liang and Potts, 2015) and train a perceptron model to greedily search for the maximally scoring parse of each hypothesis. We bootstrap the parser's lexicon entries with mappings for words from 20 randomly selected examples from our validation set, which were parsed by hand to obtain the latent trees. Sample templates used to create our dataset are shown in Table 2.

To normalize likelihoods between hypotheses of different lengths, we calculate average likelihoods for CCG productions and semantic *null* nodes, then expand the semantic parse trees to accommodate the maximum token length for utterances when scoring.

Because our application is human-robot interaction, we give the parser a budget of 10 seconds per hypothesis during the re-ranking process.[4] If a valid parse is not found in time, the hypothesis is given a confidence score of zero. If no hypotheses from a list are parsed, the re-ranking decision falls solely to the language model.

3 Experimental Evaluation

We evaluate chosen hypotheses by word error rate (WER), semantic form accuracy (whether the chosen hypothesis' semantic parse exactly matched the gold parse), and semantic form F1 score, the average harmonic mean of the precision and recall of hypotheses' semantic predicates with their corresponding gold predicates (see Table 3 for example F1 computations). In the robotic command domain, higher F1 can mean shorter clarification dialogs with users when there are misunderstandings, since the intended (gold) semantic parse's predicates are better represented for parses with higher F1. We compare the ASR's top hypothesis to re-ranking (Eq. 2) using only the language model ($\alpha = 0$), only the semantic parser ($\alpha = 1$), and a weighted combination of the two ($\alpha = 0.7$).

3.1 Dataset

We collected a corpus of speech utterances from 32 participants, consisting of both male and female, native and non-native English speakers. Participants were asked to read sentences from a computer screen for 25 minutes each, resulting in a total of 5,161 utterances. The sentences read were

[4] We found that hypotheses successfully parsed within the budget were parsed in 1.94 seconds on average, suggesting that a stricter budget can be used.

Template	Example Sentences	Corresponding Semantic Form
(f) (w) to (p)'s office	roll over to dr bell's office can you please walk to john's office run over to professor smith's office	walk(the($\lambda x.$(office(x) $\wedge$ possesses(x, tom)))) walk(the($\lambda x.$(office(x) $\wedge$ possesses(x, john)))) walk(the($\lambda x.$(office(x) $\wedge$ possesses(x, john))))
(f) (d) (i) to (p)	go and bring coffee to jane please deliver a red cup to tom would you take the box to jack	bring(coffee, jane) bring(a($\lambda x.$(red(x) $\wedge$ cup(x))), tom) bring(box, jack)
(f) (s) (p) in (l)	please look for ms. jones in the lab can you find jack in room 3.512 search for the ta in the kitchen	searchroom(3414b, jane) searchroom(3512, jack) searchroom(kitchen, jack)

Table 2: Example templates used to generate our dataset. Our template parameterization includes items (i), people (p), locations (l), filler words (f), and actions such as walk (w), delivery(d), and search (s). Parameter instances had multiple referring expressions (e.g. "john" and "professor smith" both refer to the person *john*). Eight distinct templates were used across the 3 actions, with 70 items, 69 adjectives, over 20 referents for people, and a variety of wordings for actions and filler, resulting in over 400 million possible utterances.

generated using templates for commanding a robot in an office domain (Table 2). The use of templates allowed for the automatic generation of ground truth transcriptions and semantic forms for each spoken utterance.

3.2 Experimental Setup and Results

To test our methodology, we used the Google Speech API,[5] a state-of-the-art, black-box ASR system which has been used in recent robotics tasks (Arumugam et al., 2017; Kollar et al., 2013). For each utterance, 10 hypotheses were requested from Google Speech.[6] An average of 9.2 hypotheses were returned per utterance (the API sometimes returns fewer than requested). We held out 2 speakers from our dataset as validation for hyperparameter tuning, leaving 30 speakers for a 27/3 (90%/10%) training and test set split using 10-fold cross validation.

We set the language model and semantic parser hyperparameters using the held-out validation set. Performance of the ASR's top hypothesis (**ASR**) was tested against re-ranking solely based on semantic-parser scores (**SemP**), solely on language model scores (**LM**), and on an interpolation of these with $\alpha = 0.7$ which maximized semantic form accuracy on the validation set (**Both**).

Table 4 summarizes the results of these models on the test set. All of our model's scores are statistically significantly better than the ASR baseline ($p < 0.05$ with a Student's independent paired t-test). Additionally, **SemP** and **Both** perform significantly better than **LM** in F1 while the **Both** condition performs significantly better than **LM** in semantic accuracy without a significant loss in WER or F1 performance against **LM** and **SemP**, respectively.

3.3 Discussion

All re-ranking conditions significantly improve word error rate, semantic parsing accuracy, and semantic form F1 scores against using the ASR's top hypothesis.

When examining the overall parsing accuracy of our models, we found that 37.5% of the ASR hypothesis lists generated for test utterances had at least 1 out of vocabulary (OOV) word per hypothesis. Our semantic parser is closed-vocabulary at test time, ignoring OOV words, which can contain valuable semantic information.

Consistent with intuition, using a language model alone decreases WER most. Semantic accuracy increases when interpolating confidences from the semantic parser and language model, meaning there are cases where the hypothesis the semantic parser most favors has an incorrect semantic form even while another hypothesis in the list gives the correct one. In this case, a lower confidence parse from a better-worded transcript is more likely to be correct, and we need both the semantic parser and the language model to select it.

There is no significant difference in semantic accuracy performance between solely using the language model or semantic parser, but interpolating the two gives a significant improvement over just using a language model. The semantic parser and interpolation conditions give significantly bet-

[5] https://cloud.google.com/speech/

[6] Preliminary experiments showed diminishing returns for hypothesis lists of size $n > 10$. Therefore, n was set to 10 for the accuracy vs. runtime tradeoff.

Semantic Form	P	R	F1
bring(cup, jane)	$\frac{3}{3}$	$\frac{3}{3}$	1.0
bring(a(λx.(red(x)$\wedge$ cup(x))), jane)	$\frac{3}{4}$	$\frac{3}{3}$	0.857
bring(jane, jane)	$\frac{3}{3}$	$\frac{2}{3}$	0.8

Table 3: Example F1 computations for the phrase "Bring Jane a cup". Here, the relevant (gold) predicates are *bring*, *cup*, and *jane*. F1 is the harmonic mean of the precision (P) and recall (R): F1$= 2 \cdot \frac{P \cdot R}{P+R}$

Model	WER	Acc	F1
Oracle	13.4 ± 4.2	27.9 ± 3.8	39.3 ± 3.9
ASR	30.8 ± 4.6	7.38 ± 1.9	15.9 ± 3.0
SemP	20.8 ± 5.3	24.8 ± 3.9	$\mathbf{38.3} \pm 4.1$
LM	$\mathbf{15.7} \pm 4.7$	22.7 ± 3.3	31.7 ± 4.1
Both	16.8 ± 4.6	$\mathbf{26.3} \pm 3.7$	$\mathbf{38.1} \pm 4.1$

Table 4: Average performance of re-ranking with standard deviation using semantic parsing (**SemP**), language model (**LM**), and **Both** against the black-box **ASR**'s top hypothesis. **Oracle** denotes the best possible performance achievable through re-ranking per metric (i.e. choosing the hypothesis from the ASR that optimizes for each metric in turn).

ter F1 performance over a language model alone. These results show that the integration of semantic information into the speech recognition pipeline can significantly improve language understanding.

4 Conclusion and Future Work

We have shown that post-hoc re-ranking of a black-box ASR's hypotheses using an in-domain language model and a semantic parser can significantly improve the accuracy of transcription *and* semantic understanding. Using both re-ranking components together improves parsing accuracy over either alone without sacrificing WER reduction.

A natural extension to this work would be to test re-ranking using a neural language model, which has been shown to encode some semantic information in addition to capturing statistical regularities in word sequences (Bengio et al., 2003).

Our approach should improve language understanding in robotics applications. The increase in F1 should help expedite dialogues because it would entail fewer predicates needing clarification from the user. Additionally, due to the large proportion of OOV words that we encountered from ASR, in the future we will use an open-vocabulary semantic parser, perhaps through leveraging distributional semantic representations in order to induce the meaning of novel words. By adapting existing work on learning semantic parsers for robots through dialog (Thomason et al., 2015) to incorporate ASR, a robot equipped with our pipeline could iteratively learn the meaning of new words and expressions it encounters in the wild.

Acknowledgments

We thank our reviewers for their helpful feedback and requests for clarification. This work is supported by an NSF EAGER grant (IIS-1548567), an NSF NRI grant (IIS-1637736), and a National Science Foundation Graduate Research Fellowship to the second author.

References

Sankaranarayanan Ananthakrishnan and Shrikanth Narayanan. 2007. Improved speech recognition using acoustic and lexical correlates of pitch accent in a n-best rescoring framework. In *Proc. Int. Conf. Acoust., Speech, Signal Process., Honolulu, HI*, volume 4, pages 873–876. IEEE.

Dilip Arumugam, Siddharth Karamcheti, Nakul Gopalan, Lawson L. S. Wong, and Stefanie Tellex. 2017. Accurately and Efficiently Interpreting Human-Robot Instructions of Varying Granularities. In *Robotics: Science and Systems*.

Roberto Basili, Emanuele Bastianelli, Giuseppe Castellucci, Daniele Nardi, and Vittorio Perera. 2013. Kernel-based discriminative re-ranking for spoken command understanding in HRI. In *Congress of the Italian Association for Artificial Intelligence*, pages 169–180. Springer.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155.

K. Davies, Robert E. Donovan, Mark Epstein, Martin Franz, Abraham Ittycheriah, Ea-Ee Jan, Jean-Michel LeRoux, David Lubensky, Chalapathy Neti, Mukund Padmanabhan, Kishore Papineni, Salim

Roukos, Andrej Sakrajda, Jeffrey S. Sorensen, Borivoj Tydlitát, and Todd Ward. 1999. The IBM conversational telephony system for financial applications. In *Sixth European Conference on Speech Communication and Technology, EUROSPEECH 1999, Budapest, Hungary, September 5-9, 1999*.

Anoop Deoras, Gokhan Tur, Ruhi Sarikaya, and Dilek Hakkani-Tür. 2013. Joint discriminative decoding of word and semantic tags for spoken language understanding. In *IEEE Transactions on Audio, Speech, and Language Processing*, pages 1612–1621. IEEE.

Hakan Erdogan, Ruhi Sarikaya, Stanley F Chen, Yuqing Gao, and Michael Picheny. 2005. Using semantic analysis to improve speech recognition performance. *Computer Speech & Language*, 19(3):321–343.

Judith Gaspers, Philipp Cimiano, and Britta Wrede. 2015. Semantic parsing of speech using grammars learned with weak supervision. In *HLT-NAACL*, pages 872–881.

Alex Graves and Navdeep Jaitly. 2014. Towards end-to-end speech recognition with recurrent neural networks. In *ICML*, volume 14, pages 1764–1772.

Frederick Jelinek, John Lafferty, David Magerman, Robert Mercer, Adwait Ratnaparkhi, and Salim Roukos. 1994. Decision tree parsing using a hidden derivation model. In *Proceedings of the workshop on Human Language Technology*, pages 272–277. Association for Computational Linguistics.

Thomas Kollar, Vittorio Perera, Daniele Nardi, and Manuela Veloso. 2013. Learning environmental knowledge from task-based human-robot dialog. In *Robotics and Automation (ICRA), 2013 IEEE International Conference on*, pages 4304–4309. IEEE.

Percy Liang and Christopher Potts. 2015. Bringing machine learning and compositional semantics together. *Annu. Rev. Linguist.*, 1(1):355–376.

Fabrizio Morbini, Kartik Audhkhasi, Ron Artstein, Maarten Van Segbroeck, Kenji Sagae, Panayiotis Georgiou, David R Traum, and Shri Narayanan. 2012. A reranking approach for recognition and classification of speech input in conversational dialogue systems. In *Spoken Language Technology Workshop (SLT), 2012 IEEE*, pages 49–54. IEEE.

Fuchun Peng, Scott Roy, Ben Shahshahani, and Françoise Beaufays. 2013. Search results based n-best hypothesis rescoring with maximum entropy classification. In *ASRU*, pages 422–427.

Mark Steedman and Jason Baldridge. 2011. Combinatory categorial grammar. *Non-Transformational Syntax: Formal and Explicit Models of Grammar. Wiley-Blackwell*, pages 181–224.

Andreas Stolcke et al. 2002. Srilm-an extensible language modeling toolkit. In *Intl. Conf. on Spoken Language Processing*, pages 901– 904.

Jesse Thomason, Shiqi Zhang, Raymond Mooney, and Peter Stone. 2015. Learning to interpret natural language commands through human-robot dialog. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1923–1929.

Johannes Twiefel, Timo Baumann, Stefan Heinrich, and Stefan Wermter. 2014. Improving domain-independent cloud-based speech recognition with domain-dependent phonetic post-processing. In *Twenty-Eighth AAAI Conference on Artificial Intelligence. Quebec City, Canada*, pages 1529–1536.

Ye-Yi Wang, Alex Acero, and Ciprian Chelba. 2003. Is word error rate a good indicator for spoken language understanding accuracy. In *IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 577–582, St. Thomas, US Virgin Islands. Institute of Electrical and Electronics Engineers, Inc.

Ian H Witten and Timothy C Bell. 1991. The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression. *IEEE Transactions on Information Theory*, 37(4):1085–1094.

Wayne Xiong, Jasha Droppo, Xuedong Huang, Frank Seide, Mike Seltzer, Andreas Stolcke, Dong Yu, and Geoffrey Zweig. 2016. Achieving human parity in conversational speech recognition. *arXiv preprint arXiv:1610.05256*.

Klaus Zechner and Alex Waibel. 1998. Using chunk based partial parsing of spontaneous speech in unrestricted domains for reducing word error rate in speech recognition. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 2*, pages 1453–1459. Association for Computational Linguistics.

Luke S Zettlemoyer and Michael Collins. 2005. Learning to map sentences to logical form: Structured classification with probabilistic categorial grammars. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 658–666.