Improving Black-box Speech Recognition Using Semantic Parsing

Rodolfo Corona, Jesse Thomason, Raymond J. Mooney
Department of Computer Science, The University of Texas at Austin

TEXAS
The University of Texas at Austin

Utterance

Bring a prism to Mr. John Smith

Hypothesis List

ASR Hypothesis	Semantic Form	Parser Confidence	LM Confidence
Bring a prison to Mr. John Smith	bring(hamburger, john)	5.310e-8	0.005
...	...	...	...
Bring a prism to Ms. John Smith	bring(a(λ x:i.(prism(x))),john)	0.49	0.339
Bring a prism to Mr. John Smith	bring(a(λ x:i.(prism(x))),john)	0.49	0.64

Top Re-ranked Hypothesis

Bring a prism to Mr. Smith
bring(a(λ x:i.(object(x))),smith)

Introduction

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.

Approach

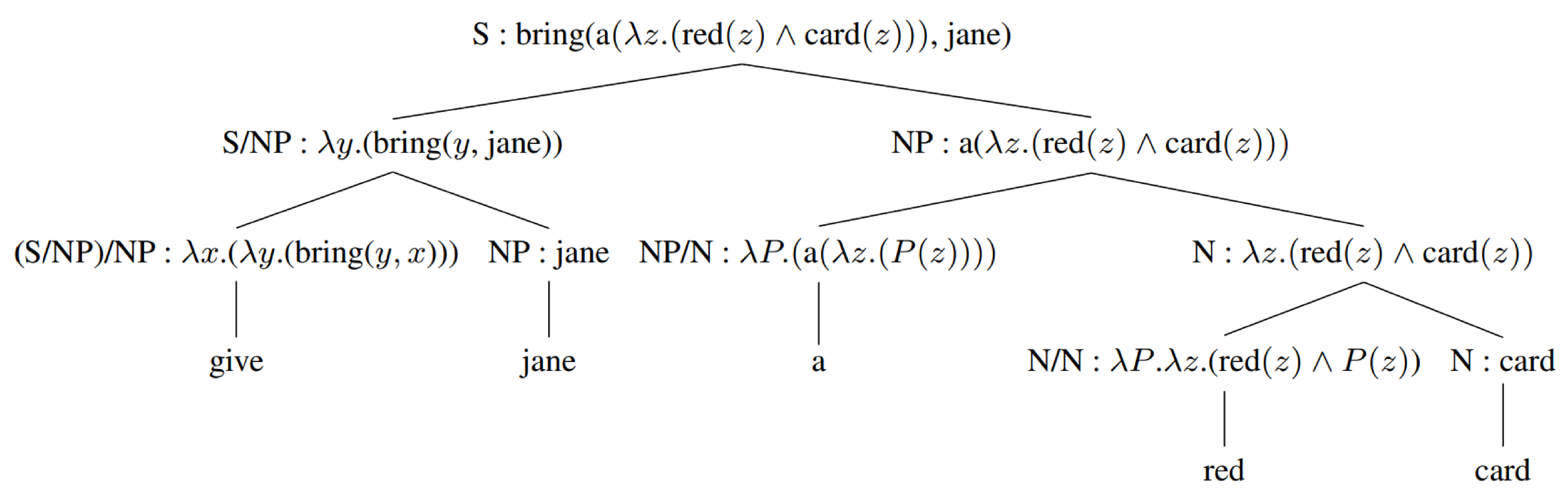
We re-rank the *n*-best hypothesis list from an ASR system by interpolating scores from an in-domain semantic parser and language model.

$$h^* = \arg\max_{h \in H} (S(h))$$

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

Semantic Parsing

Used a Combinatory Categorical Grammar (CCG) based probabilistic CKY parser



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

Language Modeling

Used a trigram back-off language model with Witten-Bell discounting

$$P(w_n | w_1, \ldots, w_{n-1}) = P(w_n | w_{n-2}, w_{n-1})$$

$$P(w_1, \ldots, w_n) = \prod_{i=1}^{N} P(w_i | w_{n-2}, w_{n-1})$$

Acknowledgements

This work is supported by an NSF EAGER grant (IIS-1548567), and NSF NRI grant (IIS-1637736), and a National Science Foundation Graduate Research Fellowship to the second author.

Dataset

We collected a dataset of 5,161 speech utterances paired with their transcriptions and logical semantic forms from 32 participants.

Utterances randomly generated using templates. Eight distinct template were used across 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.

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(λx.(office(x) $\wedge$ possesses(x, tom)))) walk(the(λx.(office(x) $\wedge$ possesses(x, john)))) walk(the(λx.(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(λx.(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)

Results

Tested our methodology using the Google Speech API

- Requested 10 hypotheses per utterance.
- Gave parser budget of 10 seconds per hypothesis.

Measured system performance over 5 different conditions:

- **Oracle**: Best achievable performance from re-ranking.
- **ASR**: System performance without re-ranking.
- **SemP**: Re-ranking using solely semantic parser scores.
- **LM:** Re-ranking using solely language model scores.
- **Both:** Re-ranking using interpolated semantic parser and language model scores.

Evaluated system performance on 3 metrics:

- **Word error rate (WER):** Computes number of *insertions*, *deletions*, and *substitutions* in hypothesis in order to measure transcription accuracy.
- **Semantic form accuracy (ACC)**: Checks for a one-to-one match between hypothesis logical form and correct logical form.
- **Semantic form F1**: Measures harmonic mean of *recall* and *precision* of the predicates in the hypothesis semantic form.

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	$\mathbf{26.3} \pm 3.7$	$\mathbf{38.1} \pm 4.1$

All conditions significantly improve performance over baseline.