An Analysis of Using Semantic Parsing for Speech Recognition

Rodolfo Corona

Outline

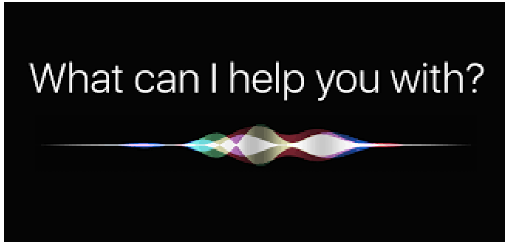
- Introduction
  - Background
  - Related Work
- Methodology
- Experiment
  - Dataset
  - Experimental Set-up
  - Experiments & Results
- Conclusion
  - Future Work
  - Concluding Remarks

Outline

- Introduction
  - Background
  - Related Work
- Methodology
- Experiment
  - Dataset
  - Experimental Set-up
  - Experiments & Results
- Conclusion
  - Future Work
  - Concluding Remarks

Introduction

- Automatic Speech Recognition (ASR) becoming more prominent.
- Performance beginning to allow wider adoption.
- There is still room to grow.



Introduction

- **<u>Motivation</u>**: Would like a language-understanding pipeline in BWI lab.
- Speech would allow for greater user-friendliness.

Introduction

- **Utterance**: The speech signal given by the user.
- **Transcription**: The correct text representation of the utterance.
- **Hypothesis**: The ASR approximation of the transcription.

Introduction

- **Our approach**: Use semantic parsing to re-rank the n-best list from ASR.
- Additionally, use re-ranking scheme to generate new training examples for re-training system.
- Most "meaningful" parse likely to be correct hypothesis.

Introduction

- **Results**: We show that language understanding is improved despite decrease in transcription performance.

ASR

- Process user utterance U and compute a hypothesis H of it from candidates W in our language (i.e. English).
- Uses *language* and *acoustic* models in tandem.
- Formally: $H = argmax_w P(U|W)P(W)$

ASR

Utterance

Please take the tea to Dan

Hypotheses

Hypothesis
Please take the pizza Dan
Please take the tea to Dan
Please take the Peter Dan
Please take the tea to Dan a

Output

Please take the tea to Dan

Semantic Parsing

- Derive computer-interpretable representation of user transcript.
- Use formalisms such as first order logic and typed lambda calculus.
- Output referred to as *semantic form*.

Semantic Parsing

$$\dfrac{\dfrac{\text{John}}{\begin{array}{c}\mathrm{NP}\\ John\end{array}} \qquad \dfrac{\dfrac{\text{is}}{\begin{array}{c}\mathrm{S\backslash NP/ADJ}\\ \lambda f.\lambda x.f(x)\end{array}} \qquad \dfrac{\text{happy}}{\begin{array}{c}\mathrm{ADJ}\\ \lambda x.happy(x)\end{array}}}{\begin{array}{c}\mathrm{S\backslash NP}\\ \lambda x.happy(x)\end{array}}}{\begin{array}{c}\mathrm{S}\\ happy(John)\end{array}}$$

Related Work

- Zechner et al. uses part-of-speech (POS) tagging with a chunk-based parser for re-ranking (Zechner et al. 1998)
- Erdogan et al. uses semantic parsing to re-rank. Does not produce forms that may be immediately executed by system. (Erdogan et al. 2005).
- Peng et al. use Google search on n-best list and extract features from results for re-ranking (Peng et al. 2013)

Outline

- Introduction
  - Background
  - Related Work
- Methodology
- Experiment
  - Dataset
  - Experimental Set-up
  - Experiments & Results
- Conclusion
  - Future Work
  - Concluding Remarks

Re-ranking

- Use ASR to generate list of *n* hypotheses for a given utterance.
- Use parser to compute a parse for each hypothesis on list.
- Use confidence scores from ASR and parser to assign a new score to each hypothesis.
- Re-rank (i.e. sort) based on new scores.

Re-ranking

- Given hypothesis h with ASR score s_{a_i} and parse score s_{p_i}.
- Normalize scores over other hypotheses: $\overline{s_{p_i}} = \log(s_{p_i}) - \log\left(\sum_{j=1}^{N} s_{p_j}\right)$

$$\overline{s_{a_i}} = \log(s_{a_i}) - \log\left(\sum_{j=1}^{N} s_{a_j}\right)$$

- Re-score hypotheses by linearly interpolating ASR and parser confidence scores with a weight β: $score_h = \beta \cdot \overline{s_{p_i}} + (1-\beta) \cdot \overline{s_{a_i}}$

Re-ranking

Please take the tea to Dan

ASR

Hypothesis	ASR Score
Please take the pizza Dan	-1242160
Please take the tea to Dan	-1242202
Please take the Peter Dan	-1242431
Please take the tea to Dan a	-1242544

Parser

Hypothesis	Parse	Parse Score
Please take the pizza Dan	walk(the(λ1:l.(and(possesses(1,jane),office(1)))))	-62.30
Please take the tea to Dan	Bring(tea,dan)	-32.18
Please take the Peter Dan	walk(the(λ1:l.(and(possesses(1,jane),office(1)))))	-62.29
Please take the tea to Dan a	None	$-\infty$

Re-ranking

Hypothesis	Parse	Parse Score
Please take the pizza Dan	walk(the(λ1:l.(and(possesses(1,jane),office(1)))))	-62.30
Please take the tea to Dan	Bring(tea,dan)	-32.18
Please take the Peter Dan	walk(the(λ1:l.(and(possesses(1,jane),office(1)))))	-62.29
Please take the tea to Dan a	None	$-\infty$

Sort

Hypothesis	Parse	Parse Score
Please take the tea to Dan	Bring(tea,dan)	-32.18
Please take the Peter Dan	walk(the(λ1:l.(and(possesses(1,jane),office(1)))))	-62.29
Please take the pizza Dan	walk(the(λ1:l.(and(possesses(1,jane),office(1)))))	-62.30
Please take the tea to Dan a	None	$-\infty$

Re-training

- Compute a hypothesis list for an utterance and re-rank.
- Generate new training pair consisting of utterance and top hypothesis transcription.
- Use set of new examples to adapt ASR acoustic model.

Re-training

ASR

Hypothesis
Please take the pizza Dan
Please take the tea to Dan
Please take the Peter Dan
Please take the tea to Dan a

Re-rank

Hypothesis
Please take the tea to Dan
Please take the Peter Dan
Please take the pizza Dan
Please take the tea to Dan a

Re-training Set	Utterances
___________	
___________	
___________	

Re-training

ASR

Hypothesis
Please take the pizza Dan
Please take the tea to Dan
Please take the Peter Dan
Please take the tea to Dan a

Re-rank

Hypothesis
Please take the tea to Dan
Please take the Peter Dan
Please take the pizza Dan
Please take the tea to Dan a

Re-training Set	Utterances
__________	
__________	
__________	
Please take the tea to Dan	

Outline

- Introduction
  - Background
  - Related Work
- Methodology
- Experiment
  - Dataset
  - Experimental Set-up
  - Experiments & Results
- Conclusion
  - Future Work
  - Concluding Remarks

Dataset

- Collected corpus from 32 participants.
- Tuples of *utterance, transcription,* and *semantic form*.
- Read randomly generated transcriptions for 25 minutes.
- 150 tuples contributed on average.
- 10 word average transcript length.

Action	Arguments
$bring(x, y)$	Bring person y item x
$searchroom(x, y)$	Search room y for person x
$walk(x)$	Walk to location x
$walk_p(x)$	Walk to the office of person x

Action	Template Examples	Number of Templates
$bring(x, y)$	I would like you to please bring *x* to *y* … Please take *y* the *x*	74
$searchroom(x, y)$	Find out if *x* is in *y* … Look for *x* in *y*	43
$walk(x)$	Would you please go to *x* … Run over to *x*	39
$walk_{p(x)}$	Hurry and walk to *x*'s office … Please go to *x*'s office	39

Dataset

- 11 people, 12 location, and 30 item atoms.
- 42 noun and 72 adjective predicates allowed for 110K more items (Noun + up to 2 adjectives).

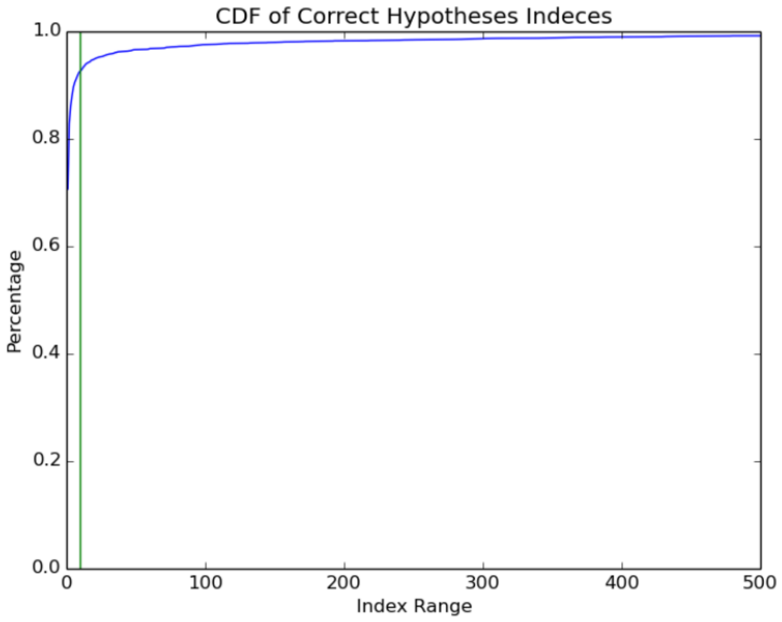
Transcript	Semantic Form
See if Bob is in room thirty-four one eight.	$searchroom(bob, l3\_418)$
Deliver a green cup to Jane	$bring(a(\lambda x : i.(and(green(x), cup(x)))), jane)$
Run over to room three five one six.	$walk(l3\_516)$
Go to John's office.	$walk(the(\lambda x : l.(and(office(x), possesses(x, john)))))$

Experiment Set-up

- Used CMU Sphinx-4 for ASR (Lamere et al.).
- Created in-domain language model and adapted Sphinx acoustic model with our data. Additionally added corpus-specific entries to dictionary.
- Used a CCG-based CKY parser (Liang & Potts, 2015), (Artzi & Zettlemoyer, 2013).
- Split data set into 8 folds by participant in corpus (32 participants).
- (28, 2, 2) dataset split for training, validation, and test sets.

Experiment Set-up

- Originally generated 1K hypotheses per utterance.
- Correct hypotheses lay in top 10 results in 92% of lists.
- Set consequent list lengths to 10.
- Used only transcriptions with fewer than 8 words due to computational cost of parsing.



Transcription Evaluation Metrics

- **Word error rate (WER)**: Measure of alignment between transcripts. Combines *substitutions* s, *deletions* d, *and insertions* i to measure accuracy $(N = |transcription|)$: $WER(p) = \frac{s+d+i}{N}$
- **Recall@1:** Top hypothesis correct.
- **Recall@5:** One of top 5 hypotheses correct.

Semantic Evaluation Metrics

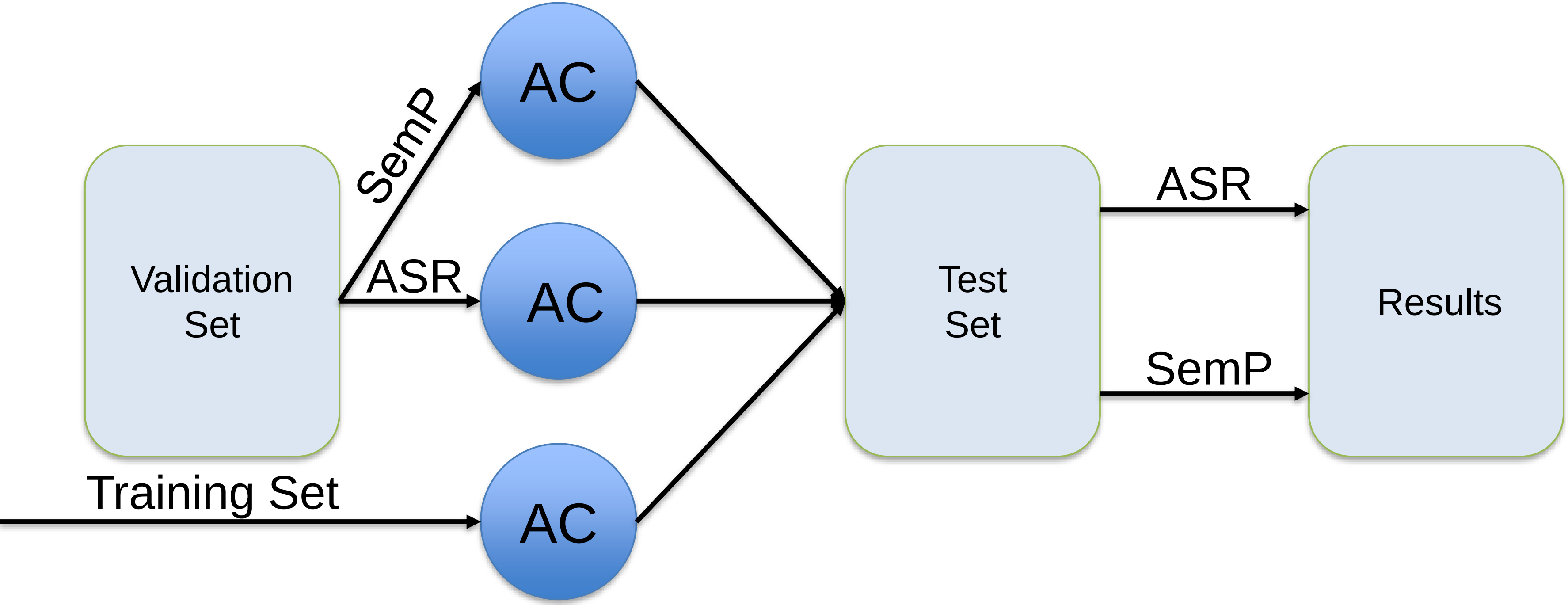
- **Full Semantic Form**: Exact match of predicates in ground truth form.
- **Recall:** $\frac{\#Correct\ predicates\ in\ hypothesis}{\#Correct\ Predicates}$
- **Precision:** $\frac{\#Correct\ predicates\ in\ hypothesis}{\#Predicates\ in\ hypothesis}$
- **F1:** Harmonic mean of precision and recall $\frac{2}{\frac{1}{R}+\frac{1}{P}}$

Main Experiment

- Baseline with no re-ranking (i.e. $\beta = 0$), denoted ASR.
- Main system with no interpolation (i.e. $\beta = 1$), denoted SemP.
- Re-trained using validation set over different combinations of conditions.
- Ran experiments with 8-fold cross validation.

Re-training Condition

Re-ranking Condition



Results

Re-training	Re-ranking	WER	R@1	R@5	SF	F1	R	P
None	ASR	**14.55***	**55.31***	**72.47***	**0.334**	0.482	0.484	0.504

Results

Re-training	Re-ranking	WER	R@1	R@5	SF	F1	R	P
None	ASR	**14.55***	**55.31***	**72.47***	**0.334**	0.482	0.484	0.504
None	SemP	18.46	38.42	65.33	0.299	0.557*	0.564*	0.598*

Results

Re-training	Re-ranking	WER	R@1	R@5	SF	F1	R	P
None	ASR	**14.55***	**55.31***	**72.47***	**0.334**	0.482	0.484	0.504
None	SemP	18.46	38.42	65.33	0.299	0.557*	0.564*	0.598*
ASR	ASR	22.00	45.86	59.12	0.276	0.457	0.456	0.478

Results

Re-training	Re-ranking	WER	R@1	R@5	SF	F1	R	P
None	ASR	**14.55***	**55.31***	**72.47***	**0.334**	0.482	0.484	0.504
None	SemP	18.46	38.42	65.33	0.299	0.557*	0.564*	0.598*
ASR	ASR	22.00	45.86	59.12	0.276	0.457	0.456	0.478
SemP	ASR	22.22	45.92	59.58	0.283	0.440	0.443	0.455

Results

Re-training	Re-ranking	WER	R@1	R@5	SF	F1	R	P
None	ASR	**14.55***	**55.31***	**72.47***	**0.334**	0.482	0.484	0.504
None	SemP	18.46	38.42	65.33	0.299	0.557*	0.564*	0.598*
ASR	ASR	22.00	45.86	59.12	0.276	0.457	0.456	0.478
SemP	ASR	22.22	45.92	59.58	0.283	0.440	0.443	0.455
ASR	SemP	25.57	30.46	52.42	0.302	**0.569**	**0.581**	**0.604**

Results

Re-training	Re-ranking	WER	R@1	R@5	SF	F1	R	P
None	ASR	**14.55***	**55.31***	**72.47***	**0.334**	0.482	0.484	0.504
None	SemP	18.46	38.42	65.33	0.299	0.557*	0.564*	0.598*
ASR	ASR	22.00	45.86	59.12	0.276	0.457	0.456	0.478
SemP	ASR	22.22	45.92	59.58	0.283	0.440	0.443	0.455
ASR	SemP	25.57	30.46	52.42	0.302	**0.569**	**0.581**	**0.604**
SemP	SemP	25.79	29.54	52.55	0.311	0.566	0.573	0.600

Results

- Ran paired Student’s t-tests on results.
- Statistically significant increase in partial semantic performance (P, R, F1) over baseline ($p < 0.05$).
- No significant difference in full semantic performance ($p = 0.12$)
- Significant decrease in transcription performance (WER, T1, T5).
- Re-training has significant adverse effect on transcription.
- No significant difference in partial semantic form performance for re-ranking under different re-training conditions.

Results

- Ultimately interested in semantic parsing performance of system.

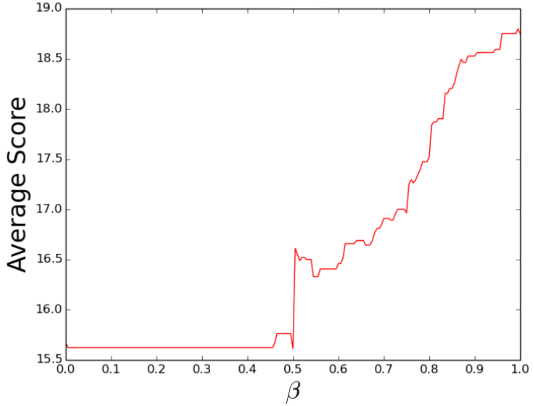
Hypothesis	Semantic Form	Parse Score	ASR Score
Please walk to professor smith a coffee	Walk(l3_516)	-45.40	-476184
Please walk to professor smith's office	walk(the(λx:l.(and(possesses(x,tom),office(x)))))	-38.55	-476359
Please walk to professor smith the coffee	Walk(l3_516)	-46.54	-476378

Hypothesis	Semantic Form	Parse Score	ASR Score
Please walk to professor smith's office	walk(the(λx:l.(and(possesses(x,tom),office(x)))))	-38.55	-476359
Please walk to professor smith a coffee	Walk(l3_516)	-45.40	-476254
Please walk to professor smith the coffee	Walk(l3_516)	-46.54	-476378

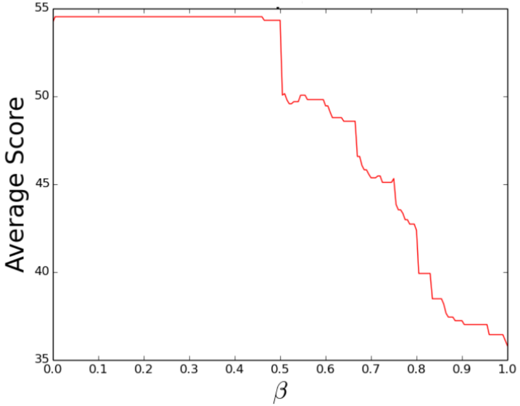
Interpolation Experiments

- Additional experiments run with interpolation of ASR and parse confidence scores.
- Tested $\beta \in [0,1]$ at 0.005 intervals on validation set.

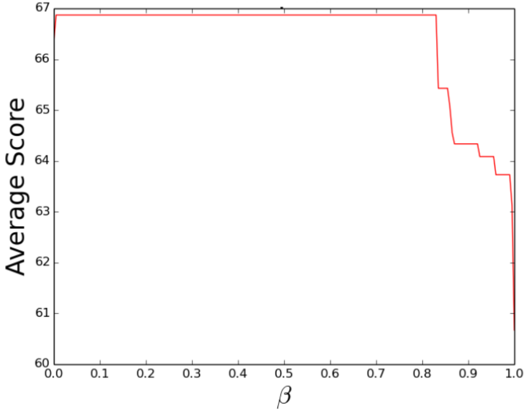
WER


Average Score
β
19.0
18.5
18.0
17.5
17.0
16.5
16.0
15.5
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0

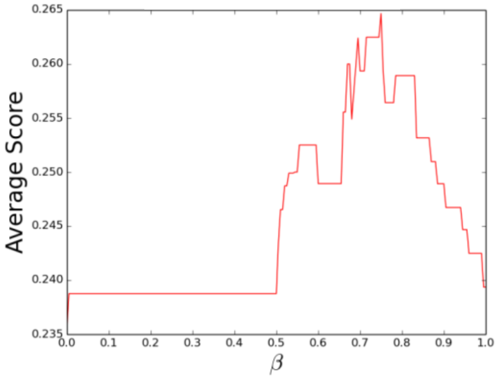
R@1


Average Score
β
55
50
45
40
35
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0

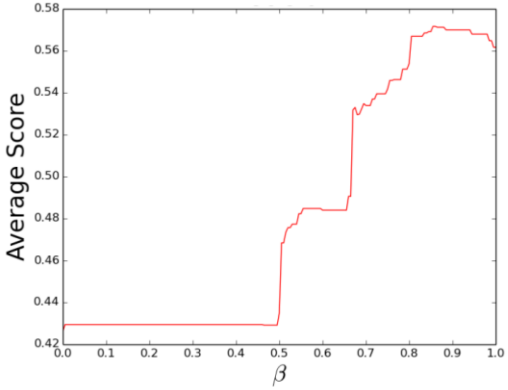
R@5


Average Score
β
67
66
65
64
63
62
61
60
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0

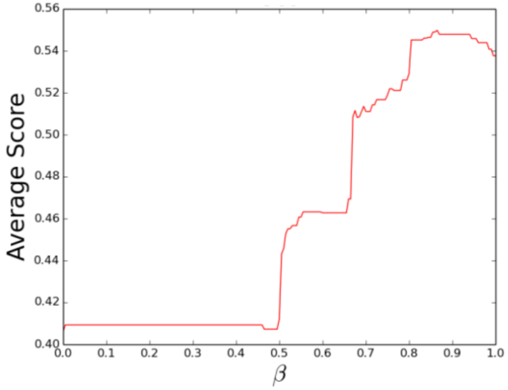
Full Semantic Form



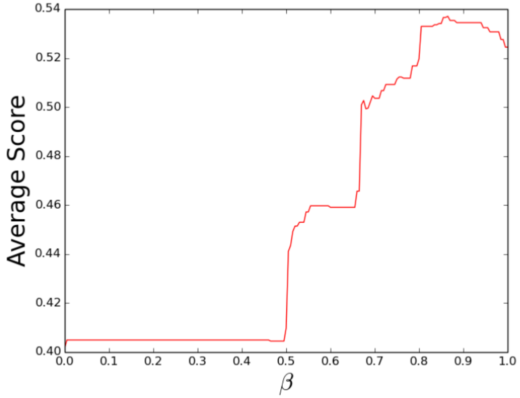
Precision



Recall



F1



Interpolation Experiments

- $\beta = 0.865$ Maximized F1 performance.
- Implies signal from both ASR and parser is useful.
- No statistical significance between $\beta = 0.865$ and $\beta = 1.0$
- Statistical significance results identical to no interpolation case.
- Re-training not pursued due to statistical analysis results.

Outline

- Introduction
  - Background
  - Related Work
- Methodology
- Experiment
  - Dataset
  - Experimental Set-up
  - Experiments & Results
- Conclusion
  - Future Work
  - Concluding Remarks

Future Work

- Deep learning approaches allow for end-to-end ASR (Graves et al. 2014, Xiong et al. 2016)
- Neural parsing technique claims to require less computation time than CKY algorithm (Misra et al. 2016)
- Could replace components in pipeline, train jointly.
- Use pre-trained models with our dataset for fine-tuning.

Future Work

- Current results motivate pursuit of dialogue-based pipeline (Thomason et al. 2015)

Future Work

- Improved F1 scores could result in shorter disambiguation dialogs.

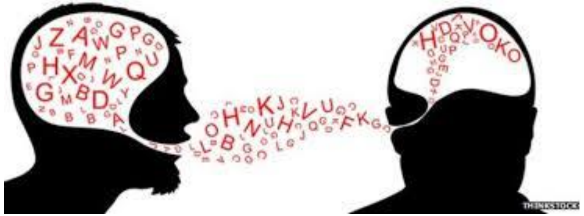
Correct: walk(the(λx:l.(and(possesses(x,smith),office(x)))))

ASR: walk(l3_516)

SemP: walk(the(λx:l.(and(possesses(x,tom),office(x)))))

Conclusion

- Re-ranking significantly improves partial semantic performance.
- Decrease in transcription performance significant.
- Current results encouraging for dialogue pipeline potential.



Acknowledgements

An Analysis of Using Semantic Parsing for Speech Recognition

Rodolfo Corona

References

Artzi, Y., & Zettlemoyer, L. (2013a). UW SPF: The University of Washington Semantic Parsing Framework. arXiv preprint arXiv:1311.3011

Erdogan, H., Sarikaya, R., Chen, S. F., Gao, Y., & Picheny, M. (2005). Using Semantic Analysis to Improve Speech Recognition Performance. Computer Speech & Language, 19(3), 321–343.

Graves, A., & Jaitly, N. (2014). Towards End-To-End Speech Recognition with Recurrent Neural Networks. In ICML (Vol. 14, pp. 1764–1772).

Liang, P., & Potts, C. (2015). Bringing Machine Learning and Compositional Semantics Together. Annu. Rev. Linguist., 1(1), 355–376.

Misra, D. K., & Artzi, Y. (2016). Neural Shift-Reduce CCG Semantic Parsing. In Proceedings of the Conference on Empirical Methods in Natural Language Processing.

Peng, F., Roy, S., Shahshahani, B., & Beaufays, F. (2013). Search Results Based N-best Hypothesis Rescoring with Maximum Entropy Classification. In ASRU (pp. 422–427).

Thomason, J., Zhang, S., Mooney, R., & Stone, P. (2015). Learning to Interpret Natural Language Commands Through Human-Robot Dialog. In Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI)

Xiong, W., Droppo, J., Huang, X., Seide, F., Seltzer, M., Stolcke, A., . . . Zweig, G. (2016). Achieving Human Parity in Conversational Speech Recognition. arXiv preprint arXiv:1610.05256

Zechner, K., & Waibel, A. (1998). Using Chunk Based Partial Parsing of Spontaneous Speech in Unrestricted Domains for Reducing Word Error Rate in Speech Recognition. In Proceedings of the 17th International Conference on Computational Linguistics-Volume 2 (pp. 1453–1459).