Semantic Rescoring Framework of Spoken Language Understanding

Wencan Luo
Department of Computer Science
University of Pittsburgh
PA 15260, USA
wencan@cs.pitt.edu

Abstract

In this proposal, we are going to propose a model to resocre the N-Best given by automatic speech recognition (ASR). However, the objective function is to maximizing the semantic meaning of candidate sentences but to minimize the Word Error Rate (WER).

1 Introduction

Let a computer understand what people say is a long term goal. However, ASR is not perfect in current stage. In fact, compared to manual transcriptions, ASR errors have a significant performance decrease in many NLP Tasks, such as Question Answering (Turmo et al., 2007), Speaker Role Recognition (Garg et al., 2008), Natural Language Understanding (NLU) (Raymond and Riccardi, 2007), etc.

In dialogue systems, a NLU component produces a semantic representation that is appropriate for the dialogue task (Jurafsky and Martin, 2000). In this paper, we use frames to represent the semantic knowledge, grounded in the theory of frame semantics (Fillmore, 1982).

For Spoken Language Understanding (SLU), one challenge that has to deal with is the ASR errors, which is the focus in this paper.

2 Related Work

2.1 NLU

NLU is a well-study field and many techniques have been proposed to improve the performance. For example, Eun et al.(2005) showed that combining several different classifiers promoted the performance of natural language understanding.

Besides, both generative and discriminative models work pretty well for spoken language understanding (Raymond and Riccardi, 2007), such as Finite State Transducers, CRF, SVM.

However, the majority of people work on human-transcribed text but rather than directly on speech recognition results. However, without ignoring the speech recognition errors, the performance is expected to decrease a lot.

2.2 Rescoring N-Best

Rescoring N-Best has shown to be effective to decrease the Word Error Rate (WER) (Zhang and Rudnicky, 2004; Zhou et al., 2006).

However, decreasing the WER is not our goal. Lower WER doesn't mean a better understanding, because, some words are more important than others in term of understanding. A better way to rescore the N-Best might be to directly focus on the meaning of the sentences. Quan et al. (2005) proposed an extrinsic measurement to score N-Best with the ad-hoc task of Spoken Language Translation.

A recent work by Morbini et al. (2012) is most similar to us. They rescored the N-Best with the objective to increase the performance of NLU component by combining different ASR engines. However, it proposed a classification model to evaluate different ASR engines, but not the semantic meaning of the sentences.

3 The Corpus

There are several benchmark corpora that are commonly used by researchers to evaluate a SLU component.

3.1 The ATIS

The Air Travel Information System (ATIS) corpus (Dahl et al., 1994) has been used for the last decade to evaluate models of Automatic Speech Recognition and Understanding. It is made of single turns acquired with a Wizard of Oz (WOZ) approach, where users ask for fight information.

However, the ATIS is licensed by LDC, which costs $1500.

3.2 French MEDIAN

The corpus MEDIA was collected within the French project MEDIA-EVALDA (Bonneau-Maynard et al., 2006) for development and evaluation of spoken understanding models and linguistic studies. The corpus is composed of 1,257 dialogs (from 250 different speakers) acquired with a Wizard of Oz (WOZ) approach in the context of hotel room reservations and tourist information.

However, the MEDIAN is licensed by ELRA, which costs 750 EURO.

3.3 Polish LUNA

The data for the Polish corpus has been collected at the Warsaw Transportation call center (Marasek and Gubrynowicz, 2008). This corpus covers the domain of transportation information like e.g. transportation routes, itinerary, stops, or fare reductions.

The LUNA corpus is public available[1]. It has 12,908 sentences for training and 3,005 for testing.

3.4 The Bosch Corpus

This data is collected by Amazon Mechanical Turk by the Bosch Research and Technology Center. The turkers task is to response what they want to say when given a topic[2]. Totally, there are 3,364 distinct ways[3] for 11 topics and their frequencies follow the power law. The number of ways for each topic is shown in Table 1 and the topics are grouped into four domains. These topics are chosen because they believe these are the most popular scenarios when people driving. Among them, 1,870 of sentences are transcribed to speech by a native speaker. 700 of them are used as testing. Two speech recognition engines are used to recognize the speech to text: Google and Vocon. Both the two engines give N-Best outputs with confidence scores.

The topics and slots in this dataset are annotated by human.

Domain	Topic	#
Social	friendsearch	268
	friendactivitysearch	207
	friendlocation	102
	friendactivity	82
	updatesocialstatus	434
POI	localsearch	793
	propertyquery	244
Route	planroute	800
	addmidpoint	181
	removemidpoint	98
Weather	checkweather	155
	Total	3364

Table 1: Number of Topics and Domains in the Bosch Corpus

4 Methodology

All the results below are based on experiments on the Bosch Corpus when I was an intern there, but I believe the results can be reproduced on other data too.

4.1 Basic NLU

For each frame, we have several slots associated with it. The NLU task is to identify slots given an utterance, which can be formed as a sequence labeling problem. For example, “where is the best taco bell in palo alto” will be annotated as “where/O is/O the/O best/B-psrh taco/B-ppn bell/I-ppn in/O palo/B-lcn alto/I-lcn”. The BIO tags are used here, where ‘B’ indicates the beginning of a slot; ‘I’ means the inside of a slot; ‘O’ means the ending of a slot. The slot name “psrh” is short of “property sorting rating high”, “ppn” is short of “poiname” and “lcn” is short for “locationconstraint cityname”.

[1] http://zil.ipipan.waw.pl/LUNA

[2] A topic is described as a frame

[3] A way to say something about a topic is a sentence.

We used CRF (Lafferty et al., 2001) to do the sequence labeling. The accuracies are shown in Table 2. A slot prediction is correct if and only if all the slots are extracted and the values are correct too.

Manual Transcription	VoCon SR	Google SR
0.944	0.486	0.810

Table 2: Slot Prediction Accuracy on the Bosch Data, test on the manual transcription, Vocon ASR and Google ASR

4.2 Combined N-Best

As shown in Table 2, the accuracy on manual transcriptions is much higher than on ASR. It tells us that ASR is the major issue here.

Thus, it is reasonable to combine N-Best ASR to improve the performance.

We have tried three voting methods to combine the N-Best results.

Upper Bound: If the correct one appears in any of the N-Best, it is a correct prediction.

Majority Voting: Use the majority as the prediction. Choose a random one if there are more than one of them.

Weighted Voting: "Majority Voting" does not consider the speech recognition ranking. The top ASR should get more weight. In this approach, each of the ASR gets the weight $(i + 1)/i$, i is the ASR rank. In this way, the top rank gets more weights.

The results are shown in Table 3. As we can see, there is still a big gap between the "Upper Bound" and "Weighted Voting", especially for the Google ASR. That's the motivation for this proposal.

The basic idea is to re-rank the N-Best with the objective function of maximizing the semantic meaning of sentences. Moreover, we are not trying to find the best recognition result which has the least word error rate, but to find a recognition that has a better semantics. Furthermore, we are not trying to find one sentence, but to select possible slot combination among the N-Best which maximizes the semantic meaning.

A possible solution might be Slot N-Gram. Train an n-gram model on the slot only, and select the candidate in the N-Best which has the lowest perplexity.

Another idea might be probabilistic frame-semantic parsing proposed in (Das et al., 2010).

5 Timeline

Sep 09 - Sep 22

- survey the related work regarding frame-semantic model
- pick up a corpus as the test bank

Sep 23 - Oct 20

- extract and format the data
- do Speech Recognition (SR)
- get the N-Best candidates on the chosen corpus

Oct 21 - Nov 9

- implement Slot N-Gram model to rescore the N-Best

Nov 10 - Dec 12

- try other models such as probabilistic frame-semantic

Acknowledgments

Do not number the acknowledgment section.

References

C. J. Fillmore. 1982. *Frame semantics*. In Linguistics in the Morning Calm, pages 111137. Hanshin Publishing Co., Seoul, South Korea.

D. Jurafsky and J. H. Martin. 2000. *Speech and Language Processing: An Introduction to Natural Language Processing*. Computational Linguistics, and Speech Recognition, Prentice-Hall, Upper Saddle River, NJ, 2000.

J. Turmo, P. Comas, C. Ayache, D. Mostefa, S. Rosset, L. Lamel. 2007. *Overview of QAST 2007*. In Working Notes of CLEF 2007 Workshop, Budapest, Hungary.

J. Eun, M. Jeong, G. Geunbae Lee 2005. *A Multiple Classifier-based Concept-Spotting Approach for Robust Spoken Language Understanding.* Eurospeech, Lisbon, Portugal, pp. 3441-3444

C. Raymond and G. Riccardi. 2007. *Generative and discriminative algorithms for spoken language understanding.* In Interspeech, pp. 16051608, Antwerp, Belgium, Aug. 2007.

N. Garg, S. Favre, H. Salamin, D. Hakkani-Tur, and A. Vinciarelli. 2008. *Role recognition for meeting participants:an approach based on lexical information and social network analysis*. Proceedings ACM International Conference on Multimedia, 693-696.

	Upper Bound		Majority Voting		Weighted Voting	
TopK	Vocon	Google	Vocon	Google	Vocon	Google
1	0.486	0.810	0.486	0.810	0.486	0.810
2	0.500	0.859	0.470	0.694	0.486	0.810
3	0.503	0.874	0.459	0.693	0.469	0.724
4		0.893		0.670		0.729
5		0.897		0.671		0.693

Table 3: Slot prediction accuracy with N-Best speech recognition result

R. Zhang, and A.I. Rudnicky. 2004. *Apply N-best list reranking to acoustic model combinations of Boosting Training*. Proc. ICSLP, 2004.

V. H. Quan, et al. 2005. *Integrated n-best re-ranking for spoken language translation*. in Proc. of Interspeech, Lisbon, Portugal.

Z. Zhou, J. Gao, F. K. Soong, and H. Meng. 2006. *A comparative study of discriminative methods for reranking lvcsr n-best hypotheses in domain adaptation and generalization*. in Proceedings of ICASSP, 2006, vol. 1, pp. 141-144.

F. Morbini, K. Audhkhasi, R. Artstein, M. Van Segbroeck, K. Sagae, P. Georgiou, D. R. Traum, S. Narayanan. 2012. *A reranking approach for recognition and classification of speech input in conversational dialogue systems.* Spoken Language Technology Workshop (SLT), 2012 IEEE

Hélène Bonneau-Maynard, Christelle Ayache, F. Bechet, A Denis, A Kuhn, Fabrice Lefèvre, D. Mostefa, M. Qugnard, S. Rosset, and J. Servan, S. Vilaneau. 2006. *Results of the french evalda-media evaluation campaign for literal understanding*. In LREC, pages 2054-2059, Genoa, Italy, May 2006.

D.A. Dahl, M. Bates, M. Brown, W. Fisher, K. Hunicke-Smith, D. Pallett, C. Pao, A. Rudnicky, and E. Shriberg. 1994. *Expanding the scope of the atis task: the atis-3 corpus.* In Proceedings of Human Language Technologies, page 4348, 1994.

K. Marasek and R. Gubrynowicz. 2008. *Design and Data Collection for Spoken Polish Dialogs Database.* In Proc. of the Sixth Int. Conf. on Language Resources and Evaluation (LREC), Marrakech, Morocco, May 2008.

J. Lafferty, A. McCallum, and F. Pereira. 2001. *Conditional random fields: Probabilistic models for segmenting and labeling sequence data.* In Proc. of ICML, pp.282-289.

D. Das, N. Schneider , D. Chen , N. A. Smith. 2010. *Probabilistic frame-semantic parsing.* Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, p.948-956, June 02-04, 2010, Los Angeles, California