Semantic Parsing Using Word Confusion Networks With Conditional Random Fields

Gokhan Tur, Anoop Deoras, Dilek Hakkani-Tür

Microsoft Silicon Valley, USA

gokhan.tur@ieee.org, anoop.deoras@microsoft.com, dilek@ieee.org

Abstract

A challenge in large vocabulary spoken language understanding (SLU) is robustness to automatic speech recognition (ASR) errors. The state of the art approaches for semantic parsing rely on using discriminative sequence classification methods, such as conditional random fields (CRFs). Most dialog systems employ a cascaded approach where the best hypotheses from the ASR system are fed into the following SLU system. In our previous work, we have proposed the use of lattices towards joint recognition and parsing. In this paper, extending this idea, we propose to exploit word confusion networks (WCNs), compiled from ASR lattices for both CRF modeling and decoding. WCNs provide a compact representation of multiple aligned ASR hypotheses, without compromising recognition accuracy. For slot filling, we show significant semantic parsing performance improvements using WCNs compared to ASR 1-best output, approximating the oracle path performance.

Index Terms: conditional random field, semantic parsing, word confusion network, natural language understanding

1. Introduction

Spoken language understanding (SLU) in goal-oriented dialog systems aims to automatically identify the domain and intent of the user, as expressed in natural language (NL), and to extract associated arguments or slots [1]. The pioneering DARPA-sponsored ATIS (Air Travel Information System) Project [2], has coined the term SLU. In this task, users request flight information, such as "*I want to fly to Boston from New York next week*". In this case, understanding was reduced to the problem of extracting task specific arguments in a given frame-based semantic representation involving, for example, *Destination* and *Departure Date*. Another example of semantic parsing from the movies domain is presented in Table 1. While the concept of using semantic frames is motivated by the case frames used in artificial intelligence research, in this instance the slots are very specific to the target domain, and most SLU systems focus on targeted understanding.

The state-of-the-art approach for training frame (slot) filling models relies on statistical machine learning methods. These approaches use generative models such as hidden Markov models [3] and probabilistic context free grammars [4, 5] or discriminative classification methods, such as conditional random fields (CRFs) [6, 7, 8]. An exhaustive survey of SLU methods can be found in [1].

Most systems simply train SLU models using textual data or manual transcriptions of collected utterances, and then ASR 1-Best hypotheses are fed into these models. A big challenge for SLU is finding target values within automatically recognized spoken utterances due to automatic speech recognition (ASR) errors. This paper tackles this challenge, investigating the use of word confusion networks (WCNs) for more robust semantic parsing in a CRF framework.

Utterance	*show me recent action movies by cameron*
Domain:	Movie
Genre:	*action*
Date:	*recent*
Director:	*cameron*

Table 1: An example input sentence with semantic annotations.

WCNs have first been proposed to improve ASR quality [9] and used for many spoken language processing tasks, including SLU [10, 11, among others], but to the best of our knowledge, this is the first study using WCNs in a CRF framework.

In our earlier work on using WCNs for call-type classification and named entity extraction [10], we have only improved the classification decoding algorithm so as to exploit WCN. In this study, going on step further, we also propose a novel technique for training CRF models using WCN.

In the next section, we present the CRF-based slot filling framework we employ. Detailed motivation is presented in Section 3. Then in Section 4, we present the related work on WCNs and their use in spoken language processing. Then Section 5 presents the method for employing WCNs with CRF. Finally, in Section 6, we present the experiments and results.

2. Semantic Parsing

Following the state-of-the-art approaches for slot filling [7, 8, 12, among others], we use discriminative statistical models, namely conditional random fields, (CRFs) [13], for modeling. More formally, slot filling is framed as a sequence classification problem to obtain the most probable slot sequence:

$$\hat{Y} = \underset{Y}{\operatorname{argmax}}\, P(Y|X)$$

where $X = x_1, ..., x_N$ is the input word sequence and $Y = y_1, ..., y_N, y_t \in C$ is the sequence of associated class labels, C.

CRFs are shown to outperform other classification methods for semantic parsing [1], since the training can be done discriminatively over a sequence. The baseline model relies on word n-gram based linear chain CRF, imposing the first order Markov constraint on the model topology. Similar to maximum entropy models, in this model, the conditional probability, $P(Y|X)$ is defined as:

$$P(Y|X) = \frac{1}{Z(X)} exp\left(\sum_k \lambda_k f_k(y_{t-1}, y_t, x_t)\right)$$

with the difference that both X and Y are sequences instead of individual local decision points given a set of linear prediction

Words					
find	recent	comedies	by	james	cameron
↓	↓	↓	↓	↓	↓
O	B-date	B-genre	O	B-director	I-director

Figure 1: An example utterance semantically annotated in IOB format.

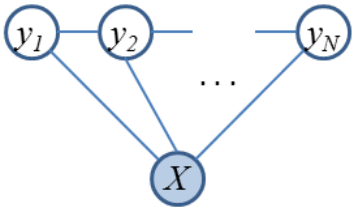

Figure 2: Linear chain CRF model.

functions f_k (such as n-gram lexical features, state transition features, or others) with associated weights λ_k. $Z(X)$ is the normalization term [13]. After the transition and emission probabilities are optimized, the most probable state sequence, $\hat{Y}$, can be determined using the well-known Viterbi algorithm. Figure 2 depicts the standard linear chain CRF model. Note that, the tag sequence is depending on the observation sequence, X, instead of corresponding observations, x_t. This is the main difference to local models like Maximum Entropy Markov Model or Hidden Markov Model [13].

In this study, we follow the popular IOB (in-out-begin) format in representing the data as shown in Figure 1.

3. Understanding ASR Output

Typically spoken language processing systems are composed of sequential and independent components, starting from an automatic speech recognizer (ASR). Further components, such as spoken language understanding (SLU) use the best hypothesis output of ASR (ASR 1-best). More formally, let:

$$\hat{X} = \underset{X}{\operatorname{argmax}}\, P(X|A)$$

where A is the input utterance, and $\hat{X}$ is the most probable ASR hypothesis, which is then fed into semantic parsing:

$$\hat{Y} = \underset{Y}{\operatorname{argmax}}\, P(Y|\hat{X})$$

In such systems, it is very important to be robust to ASR errors. Especially with spontaneous conversational speech with potential background noise, the typical word error rate (WER) for ASR 1-best output is around 20%-30%; in other words, one in every three-four words is misrecognized [14, 15, 16]. Misrecognizing a word may result in misunderstanding the whole utterance, even though all other words are correct.

Furthermore, robustness to ASR output is more critical for slot filling, compared to intent determination, which can tolerate some amount of noise due to redundancy in natural language. One distinct characteristic of slot filling, unlike intent determination is that, it heavily relies on prepositions. The whole semantics change when one says "*united flights to boston*" versus "*united flights from boston*". When such key prepositions are dropped from the ASR best hypotheses, semantic parsing is adversely affected. The same argument holds for proper names like actor or director names.

One immediate solution to end up with more robust systems is using ASR n-best hypotheses instead of just using the best one [17, 18, 19]. One notable study by Yaman *et al.,* rescored ASR n-best hypotheses for joint intent determination and recognition re-ranking for the ATIS corpus. Another solution is to use ASR word confidence scores during understanding tasks to prevent errors caused by misrecognized words [20, 21, 22].

Figure 3: Conceptual process of typical spoken dialog systems with cascaded speech recognition and understanding.

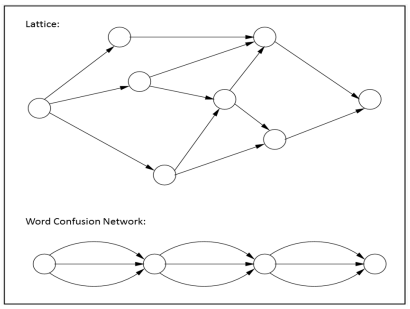

Figure 4: Typical structures of lattices and WCNs.

Going one step further, another option is using the whole word lattice output of the ASR instead of only using the ASR 1-best or n-best output [23, 24, 16]. The oracle accuracy of lattices is much higher than the word accuracy of ASR 1-best hypotheses [25, 20]. The oracle accuracy is the accuracy of the path in a lattice closest to the reference transcriptions. Word lattices are used to approximate the word search space in large vocabulary continuous speech recognition (LVCSR) systems. Usually they are acyclic and have no a-priori structures. Their transitions are weighted by the acoustic and language model probabilities.

In our previous work [16, 26], we proposed a Maximum Entropy Markov Model based framework to decode ASR lattices for joint speech recognition and slot filling. This approach resulted in significant improvements in slot filling, beating a CRF model using ASR 1-Best.

4. Word Confusion Networks

A compact and normalized class of word lattices, called word confusion networks (or position specific posterior lattices or sausages) have been proposed initially for improving ASR performance [9] and later for SLU [27] and speech retrieval [28]. These confusion networks are more efficient than canonical word lattices, in terms of size and structure, without compromising recognition accuracy. They also provide an alignment for all the strings in the word lattices. The general structure of these lattices and WCNs are shown in Figure 4. Since WCNs force the competing words to be in the same group, they enforce the alignment of the words that occur at the same approximate time interval. This time alignment may be very useful in language processing. The words in the WCNs have posterior probabilities, which can be used as their confidence scores. These are basically the sum of the probabilities of all paths which contain that word at around that approximate time frame. WCNs are much smaller than ASR lattices and they still have better or comparable word accuracy and oracle accuracy.

Word Confusion Bins				
a	t	with	ashton	kutcher
tv	series	wet	aston	
the	tv		astion	
↓	↓	↓	↓	↓
B-type	I-type	O	B-stars	I-stars

Figure 5: An example word confusion network with semantically annotated bins.

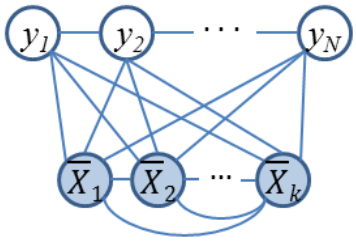

Figure 6: Linear chain CRF model using WCNs with k bins.

It has been shown that WCNs are extremely effective for many spoken language processing tasks such as call-type classification [29, 20], named entity extraction [10], speech translation [30], or speech summarization [31].

In our previous work, we have employed WCNs for other understanding tasks, such as call classification and named entity extraction [10]. In this work, we extend that work and earlier work on using lattices for slot filling in an ME-MM framework [16], to decode WCNs using CRFs.

A recent work by Henderson et al. proposed using WCN for slot filling using a local SVM based classification framework [11]. While the main idea is very similar to this study, that work uses unaligned training data of utterances and semantic frames. In other words, the semantic parsing task is not framed as a sequence classification task. In that respect, it is similar to the Chanel system using semantic classification trees [6]. During decoding, instead of using n-grams in ASR best path, they simply use all n-grams in the WCN.

In another related study, Kurata *et al.* employed WCNs for named entity extraction using a Maximum Entropy framework [32]. Instead of training and decoding with confusing words in each bin, they clustered confusable words (like "*two*" and "*to*") and used cluster IDs for modeling.

5. Using Word Confusion Networks for Semantic Parsing

The main idea in this paper is that, in order to make the model more robust to ASR noise, the model must be trained also with ASR noise. This has been a known phenomena for spoken language, but requires the availability of spoken training data.

The process starts with manual transcription of the data and semantic annotation of these transcriptions. Then the ASR output is aligned to the manual transcriptions using the NIST Sclite toolkit[1]. The semantic annotations can then be transferred to the corresponding ASR output. If a word is deleted, one can do one of the two solutions: either put an epsilon token for deletion, or totally drop the semantic tag (which may result in partial slots).

An example word confusion network is shown in Figure 5. The manual transcription reads "*tv series with ashton kutcher*". Note the similarity with Figure 1. The only difference is that,

[1]http://www.nist.gov/speech/tools

	No. Utt.	No. Words	No. Slots
Training	7,519	25,614	6,321
Test	1,764	6,485	1,742

Table 2: Data sets used in the experiments.

the semantic tags are assigned to WCN bins instead of words. In other words, the goal of CRF model is now using features from these bins and assign a slot for them:

$$\hat{Y} = \underset{Y}{\operatorname{argmax}}\, P(Y|\bar{X})$$

where $\bar{X} = \bar{x}_1, ..., \bar{x}_N$ is the input WCN bin sequence and $Y = y_1, ..., y_N, y_t \in C$ is the sequence of associated class labels, C of size N. $\bar{x}_t = w_t^1, ..., w_t^k$ is a set of confusing word hypotheses for a bin size of k words.

Training and decoding algorithms are changed due to the change in non-transitional prediction functions, f_k. Instead of binary indicators of word n-grams as lexical features, f_k is extended so as to use all possible n-grams in the neigboring bins. More formally:

$$P(Y|\bar{X}) = \frac{1}{Z(\bar{X})} exp\left(\sum_k \lambda_k f_k(y_{t-1}, y_t, \bar{x}_t)\right)$$

Figure 6 depicts the extension of standard linear chain CRF model using WCN of size k bins. Note that, similar to Figure 2, the tag sequence depends on the global observation sequence, but instead of a single word sequence, we have k sequences for a WCN of bin size k. Furthermore, the words in these bins also depend on each other, hence the links inbetween them.

In this work, we used word hypotheses using bin trigram features, a total of up to k^3 word ngrams for each bin $\bar{x}_t$. During training the confidence scores are ignored in this study, and only the thresholding is applied at the bin size level, k.

During decoding, one can directly use the CRF model as is, ignoring the word confidences. However, as an extension to this idea, similar to our previous work [29], it is possible to weigh the word n-gram features, $\bar{w}$, with respect to their confidences.

$$\hat{f}_k(\bar{w}) = f_k(\bar{w}) \times P(\bar{w})$$

As a suboptimal approximation of the formulation above, in this study we have done decision tag level weighted linear interpolation, using posterior probabilities of tags of $\bar{x}_t$ for each word hypotheses, $P_{CRF}(w_t^j|\bar{X})$:

$$P(y_t|\bar{X}) = \sum_{j=1}^{N} P_{CRF}(w_t^j|\bar{X}) \times P(w_t^j)$$

Another extension of this model would be also including the manual transcription of the utterances into the WCNs during training. This will enable the model to see both correct and noisy input and help greatly for robustness.

6. Experiments and Results

Experiments are performed using a conversational understanding system, with real users for the entertainment domain. The users present queries about various movies, such as "*who is the director of avatar*", "*show me some action movies with academy awards*", or "*when is the next harry potter gonna be released*". The semantic space consists of 22 slot types, such as named

Train/Test	Manual Transcriptions	Lattice 1-Best	WCN 1-Best	Oracle Path	WCN
Manual transcriptions (Man)	88.75%	77.72%	79.15%	84.24%	-
Lattice 1-Best	-	79.74%	-	83.84%	-
WCN 1-Best	-	-	81.93%	83.90%	-
WCN	-	-	-	-	82.28%
WCN + Man.	-	-	-	-	83.28%
WCN + Man. with confidences					**83.73%**

Table 3: Slot filling performances in F-Measure for various training and test conditions

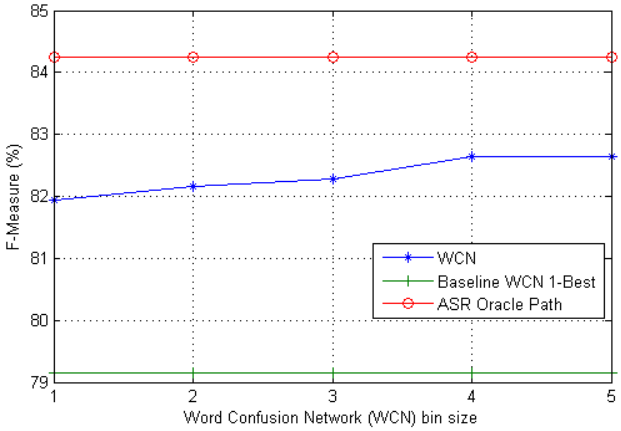

Figure 7: Effect of using different WCN bin sizes.

ones (movie or actor names) or unnamed ones (genre or language).

Table 2 shows the properties of the data sets. We only used spoken utterances (instead of written sentences) for training and test. No separate held-out set is used as no parameter is optimized, including WCN bin size. On average, there are about 3.5 words and 1 slot per utterance. Off-the-shelf Microsoft ASR with generic acoustic model and domain-adapted language model (using an earlier textual data) is employed for recognition. The word error rate (WER) in this data set is found to be 18.5%, with 10.6% substitution, 5.9% deletion, and 2.1% insertion errors. The WER of the oracle path is found to be 10.4%. The word confusion networks are built using the SRILM toolkit [33], which uses a method similar to AT&T pivot algorithm [27]. The word error rate of the best WCN hypotheses is significantly lower, at 16.9%.

For semantic parsing evaluation, the slot F-measure is used, following the literature [8] using the CoNLL evaluation script.[2] The baseline performance is obtained using only word n-grams with a linear chain CRF using the CRF++ toolkit[3] using default parameters with word level IOB format.

Table 3 presents the results showing the effectiveness of using whole WCN for training and testing. The oracle path is the word sequence which minimizes the ASR error rate in each utterance (instead of SLU). Between ASR 1-Best and Oracle path, the F-Measure difference is about 6.5%. The first immediate thing to notice is that, using ASR output for training, even ASR 1-Best, helps significantly, improving F-Measure by 2% absolute, from 77.72% to 79.74%. This figure increases 2% more when WCN 1-Best is used for training and test. This is in line with having better ASR accuracy.

When WCN with a bin size of 3 is used, we see an additional improvement, but adding manual transcriptions to the training data by aligning the manual transcriptions to the WCN gives 1% absolute F-Measure on top. The final improvement comes by using the second extension of this technique by exploiting the word confidences, reaching 83.73%. This is equivalent to closing 92% of the performance difference between ASR 1-Best and Oracle path decoding.

[2]http://www.cnts.ua.ac.be/conll2000/chunking/output.html

[3]http://crfpp.sourceforge.net

Slot	1-Best (77.72%)		WCN (83.73%)	
	Precision	Recall	Precision	Recall
Overall	79.31%	76.19%	86.40%	81.22%
Movie Name	60.62%	74.84%	76.23%	71.49%
Actor	88.44%	82.23%	89.18%	88.79%
Genre	92.55%	78.73%	92.57%	88.63%
Director	81.40%	79.55%	85.00%	82.93%

Table 4: An example input sentence with semantic annotations.

Figure 7 shows the effect of using different WCN bin sizes, k, without adding the manual transcriptions or the word confidences. The difference between using $k = 1$ and $k = 5$ is about 0.7% absolute (from 81.93% to 82.65%). So, most of the improvement comes from training using WCN, instead of decoding WCN.

The last analysis we have performed is checking the slot level recall/precision figures using ASR 1-Best and WCN. Table 4 presents these figures. For most slots, we see good recall improvements, as expected. Interestingly for movie names precision has improved significantly while recall has dropped a bit. This may be due to the fact that movie names are typically longer phrases.

7. Conclusions and Future Work

We have proposed the use of word confusion networks (WCNs) with conditional random fields (CRFs) for semantic parsing in a conversational understanding system. While the approach is very straightforward, the performance improvements are impressive. Compared to the established technique of training with manual transcriptions and testing on ASR 1-Best, we observed 6% absolute F-Measure improvement, which is almost equivalent to the performance with ASR Oracle path.

It is clear to conclude that in order to build more robust understanding systems, one must train also using ASR output of the training data (better yet with WCNs). Furthermore, using word confidences on the WCNs resulted in extra gains.

Future work involves adding additional features into this framework, inspired from our earlier work on call classification [34]. These include syntactic features, such as part of speech tags and semantic features such as named entities.

Acknowledgments We would like to thank our colleagues at Microsoft, especially Ruhi Sarikaya, Asli Celikyilmaz, and Ashley Fidler, for many helpful discussions and their help for the experimental setup.

8. References

[1] G. Tur and R. D. Mori, Eds., *Spoken Language Understanding: Systems for Extracting Semantic Information from Speech*. New York, NY: John Wiley and Sons, 2011.

[2] P. J. Price, "Evaluation of spoken language systems: The ATIS domain," in *Proceedings of the DARPA Workshop on Speech and Natural Language*, Hidden Valley, PA, June 1990.

[3] R. Pieraccini, E. Tzoukermann, Z. Gorelov, J.-L. Gauvain, E. Levin, C.-H. Lee, and J. G. Wilpon, "A speech understanding system based on statistical representation of semantics," in *Proceedings of the ICASSP*, San Francisco, CA, March 1992.

[4] S. Seneff, "TINA: A natural language system for spoken language applications," *Computational Linguistics*, vol. 18, no. 1, pp. 61–86, 1992.

[5] W. Ward and S.Issar, "Recent improvements in the CMU spoken language understanding system," in *Proceedings of the ARPA HLT Workshop*, March 1994, pp. 213–216.

[6] R. Kuhn and R. D. Mori, "The application of semantic classification trees to natural language understanding," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, pp. 449–460, 1995.

[7] Y.-Y. Wang and A. Acero, "Discriminative models for spoken language understanding," in *Proceedings of the ICSLP*, Pittsburgh, PA, September 2006.

[8] C. Raymond and G. Riccardi, "Generative and discriminative algorithms for spoken language understanding," in *Proceedings of the Interspeech*, Antwerp, Belgium, 2007.

[9] L. Mangu, E. Brill, and A. Stolcke, "Finding consensus in speech recognition: word error minimization and other applications of confusion networks," *Computer Speech and Language*, vol. 14, no. 4, pp. 373–400, 2000.

[10] D. Hakkani-Tür, F. Bechet, G. Riccardi, and G. Tur, "Beyond asr 1-best: Using word confusion networks in spoken language understanding," *Computer Speech and Language*, vol. 20, no. 4, pp. 495–514, 2006.

[11] M. Henderson, M. Gasic, B. Thomson, P. Tsiakoulis, K. Yu, and S. Young, "Discriminative spoken language understanding using word confusion networks," in *In Prooceedings of the IEEE SLT Workshop*, Miami, FL, December 2012.

[12] G. Tur, D. Hakkani-Tür, and L. Heck, "What is left to be understood in ATIS?" in *Proceedings of the IEEE SLT Workshop*, Berkeley, CA, 2010.

[13] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proceedings of the ICML*, Williamstown, MA, 2001.

[14] B. Kingsbury, L. Mangu, G. Saon, G. Zweig, S. Axelrod, V. Goel, K. Visweswariah, and M. Picheny, "Toward domain-independent conversational speech recognition," in *Proceedings of the Eurospeech*, Geneva, Switzerland, September 2003, pp. 1881–1884.

[15] G. Riccardi and D. Hakkani-Tür, "Active and unsupervised learning for automatic speech recognition," in *Proceedings of the EUROSPEECH*, Geneva, Switzerland, September 2003.

[16] A. Deoras, R. Sarikaya, G. Tur, and D. Hakkani-Tür, "Joint decoding for speech recognition and semantic tagging," in *In Prooceedings of the Interspeech*, Portland, OR, September 2012.

[17] A. Stolcke, Y. Konig, and M. Weintraub, "Explicit word error rate minimization in N-best list rescoring," in *Proceedings of the EUROSPEECH*, Rhodes, Greece, September 1997.

[18] V. Goel, W. Byrne, and S. Khudanpur, "LVCSR rescoring with modified loss functions: A decision theoretic perspective," in *Proceedings of the ICASSP*, Seattle, WA, May 1998, pp. 425–428.

[19] Y. He and S. Young, "A data-driven spoken language understanding system," in *Proceedings of the IEEE ASRU Workshop*, U.S. Virgin Islands, December 2003, pp. 583–588.

[20] G. Tur, J. Wright, A. Gorin, G. Riccardi, and D. Hakkani-Tür, "Improving spoken language understanding using word confusion networks," in *Proceedings of the ICSLP*, Denver, CO, September 2002.

[21] T. J. Hazen, S. Seneff, and J. Polifroni, "Recognition confidence scoring and its use in speech understanding systems," *Computer Speech and Language*, no. 16, pp. 49–67, 2002.

[22] S. Cox and G. Cawley, "The use of confidence scores in vector based call-routing," in *Proceedings of the EUROSPEECH*, Geneva, Switzerland, September 2003, pp. 633–636.

[23] M. Saraclar and R. Sproat, "Lattice-based search for spoken utterance retrieval," in *Proceedings of the HLT-NAACL*, Boston, MA, 2004.

[24] S. Saleem, S.-C. Jou, S. Vogel, and T. Schulz, "Using word lattice information for a tighter coupling in speech translation systems," in *Proceedings of the ICSLP*, Jeju-Island, Korea, October 2004.

[25] M. Oerder and H. Ney, "Word graphs: an efficient interface between continuous-speech recognition and language understanding," in *Proceedings of the ICASSP*, 1993.

[26] A. Deoras, G. Tur, R. Sarikaya, and D. Hakkani-Tur, "Joint Discriminative Decoding of Words and Semantic Tags for Spoken Language Understanding," *IEEE Transactions on Audio, Speech and Language Processing*, 2013.

[27] D. Hakkani-Tür and G. Riccardi, "A general algorithm for word graph matrix decomposition," in *Proceedings of the ICASSP*, Hong Kong, May 2003.

[28] C. Chelba and A. Acero, "Position specific posterior lattices for indexing speech," in *Proceedings of the ACL*, Ann Arbor, MI, 2005.

[29] G. Tur, D. Hakkani-Tür, and G. Riccardi, "Extending boosting for call classification using word confusion networks," in *Proceedings of the ICASSP*, Montreal, Canada, May 2004.

[30] N. Bertoldi and M. Federico, "A new decoder for spoken language translation based on confusion networks," in *Proceedings of the IEEE ASRU Workshop*, Puerto Rico, November 2005.

[31] S. Xie and Y. Liu, "Using confusion networks for speech summarization," in *Proceedings of the HLT-NAACL*, Los Angeles, CA, June 2010.

[32] G. Kurata, N. Itoh, M. Nishimura, A. Sethy, and B. Ramabhadran, "Named entity recognition from conversational telephone speech leveragingword confusion networks for training and recognition," in *Proceedings of the ICASSP*, Prague, Czech Republic, 2011.

[33] A. Stolcke, "SRILM – An Extensible Language Modeling Toolkit," in *Proceedings of the ICSLP*, Denver, CO, September 2002.

[34] D. Hakkani-Tür, G. Tur, and A. Chotimongkol, "Using syntactic and semantic graphs for call classification," in *Proceedings of the ACL Workshop on Feature Engineering for Machine Learning in Natural Language Processing*, Ann Arbor, MI, June 2005.