Error-Corrective Discriminative Joint Decoding of Automatic Spoken Language Transcription and Understanding

Bassam Jabaian and Fabrice Lefèvre

LIA, University of Avignon

{bassam.jabaian,fabrice.lefevre}@univ-avignon.fr

Abstract

Following recent trends in the development of spoken dialogue systems, this paper proposes to improve the performance of the user's intent extraction by means of joint decoding of automatic spoken language transcription and understanding. Gains are expected not only from a better connectivity and mutual awareness of both tasks but also through the use of discriminant models and integration of an error-corrective intermediate mechanism. This latter is based on a statistical post-editing of the speech recognizer word lattice and conditional random fields instantiate the former in our system. An overall absolute reduction of 1.1% is observed by direct application of the proposed techniques on the MEDIA task.

Index Terms: Dialogue System, Speech Understanding, Joint Decoding, Hypothesis Graphs, Error Correction

1. Introduction

Generally a spoken dialogue system (SDS) consists of several components working sequentially to perform oral human-machine interactions. The automatic speech recognition (ASR) module transcribes the user request and transmits the transcribed speech as a written input to the spoken language understanding system (SLU). The SLU process extracts the semantic information from the input text and expresses it in a meaning representation format. This information is used by the dialogue manager to find a response corresponding to the user's request in its related database. This response is thereafter transferred to the user with the help of a natural language generator and a text-to-speech synthesizer.

This pipeline (or cascade) view of an SDS is undoubtedly the most widespread in the state-of-the-art systems. In such an architecture, the SLU system takes the best hypothesis output (1-best) from the ASR system as input to perform semantic tagging. The performance of such system relies greatly on the approach used to derive the semantic information and its robustness to recognition errors. Indeed misrecognized words may imply some misunderstandings.

Recently several statistical approaches have been proposed for the SLU system. Those approaches present an interesting trade-off between their performance and the efforts invested in their development. Several studies investigated and compared probabilistic methods to train a SLU model e.g. [9] and showed that discriminative approaches, and especially conditional random fields (CRF), are the most efficient for sequential semantic tagging.

Despite the good performance of the discriminative approaches, their quality is still a function of the quality of their inputs. To decrease the impact of the recognition errors on the overall system performance, some works proposed to use an ASR n-best hypotheses list instead of the 1-best hypothesis [8, 13]. Others proposed to use ASR word confidence scores [24, 12, 3, 11] during the understanding process in order to prevent errors caused by misrecognized words.

In this paper, we intent to develop a “fat pipeline“ between the SDS components. This transmission must not lie only on a selected abridged information but on the most interesting and promising part of the full hypothesis search space, encoded in a lattice. More particularly our overall objective is to develop a performing approach for SLU with the possibility to operate a joint decoding with the speech transcription step allowing a better information transmission between the two complementary tasks. The joint decoding will allow to select a transcription from the generated word graph (or lattice) taking into account the semantic tagging hypotheses of this transcription. In this line of thought, we no longer seek to derive the best transcription, but the transcription that can be semantically labelled in the best possible way with current semantic models.

An homogeneous representation of models is needed to permit the joint decoding. The output of the ASR system is presented as a word lattice or graph. A similar representation is therefore needed for the understanding module. In that purpose, we propose to have recourse to a CRF-based model for speech understanding. In our approach the semantic tagging is seen as the translation of a word sentence into a tag sentence, and thus can rely on an approach developed for statistical machine translation (SMT). The module is based on finite state transducers into which the different stages of the translation process are composed, as presented in [16].

This formalism for the SLU system offers the advantage to handle graphs intrinsically and so all the well-known operations on graphs (such as composition, minimization, determination etc.) can be applied straightforwardly. In that way, the understanding graph can be composed with the output graph of the ASR system allowing a rich information transmission between the two modules. This composition represents a joint decoding in which all recognition hypotheses are transmitted to the SLU system and their probability can be taken into account during the overall decoding.

Several works dealt with recognition errors by applying a correction mechanism on the output of an ASR system before its transmission to the SLU system. [14] used a maximum entropy language model to simultaneously perform correction and domain adaptation of ASR outputs. Authors of [20] observed that in many cases the second best (2-best) hypothesis of the ASR output is correct and proposed to learn automatically a set of rules to determine when to prefer the 2-best hypothesis.

In a second step, we aim to further increase the robustness of the system to recognition errors. We propose to apply a statistical post edition (SPE) [28] to the ASR output in order to correct some repeated recognition errors that the ASR language model is not able to tackle alone. To do so we seize the opportunity offered by our SLU setup based on a SMT system. Indeed the SPE system itself can be seen as a translation system from noisy "recognized" sentences into clean "manually transcribed" sentences.

The paper is organized as follows: in Section 2 we describe our proposition of a CRF-based transducer understanding model. In Section 3 we present the proposition of a joint decoding between recognition and understanding while in Section 4 we complement it with robust understanding by integrating a post-edition step in the cycle. Finally our experimental study and results are reported with the French dialogue corpus MEDIA in Section 5.

2. A translation-based discriminative approach for speech understanding

Several efficient methods have been proposed for concept tagging [13, 17, 9]. For their good performance we chose CRF [15] as a reference approach. Linear CRF represent a log-linear model, normalized at the sentence level. For $\hat{C} = c_1, \ldots, c_M$ a sequence of M concepts that can be hypothesized by the SLU module from a word sequence of length N $W = w_1 ..., w_N$, CRF model the probability between concepts and words as follows:

$$P(c_1^N | w_1^N) = \frac{1}{Z(w_1^N)} \prod_{m=1}^{N} H(c_{m-1}, c_m, \phi(w_1^N, n)) \quad (1)$$

with

$$H(c_{n-1}, c_n, \phi(w_1^N, n)) = \sum_i \lambda_i h_i(c_{n-1}, c_n, \phi(w_1^N, n))$$

Log-linear models are based on feature functions h_i representing the information extracted from the training corpus, λ are estimated during the training process and Z is a normalization term. $\phi(w_1^N, n)$ represents pattern function that will be used to determined feature functions during training, in our case $\phi(w_1^N, n) = w_{n-2}^{n+2}$ thus delimiting a neighborhood window of length 2 around the current word.

In the optic of having a rich information transmission between the SDS components (possibly by means of the multiple scored hypotheses generated by statistical approaches) and in order to be able to have a joint decoding between these modules we are looking for a similar representation in these modules. A graph can be a rich homogeneous representation for the different module.

A CRF-based graph has been previously proposed for a translation task. In order to obtain an efficient CRF-based graph system for translation, [16] have proposed a model based on a cascaded composition of finite state transducers representing the various stages of the translation process.

The proposed decoder is a composition of weighted finite state transducers (WFST) representing the following steps: reordering and segmentation of the source sentence according to the words tuples, application of the translation model with hypotheses evaluation based on CRF, and composition with a target language model. The understanding of a user utterance can be seen as a translation from a sequence of words (source language) to a sequence of concepts (target language). In that case a similar CRF-based graph model, as proposed for translation, can be obtained for understanding.

This model will be called CRFPB-SMT because it embeds a mechanism for modelling a translation table by sub-sentential segments (called tuples, but analogous to phrase translation table entries) and uses CRF as the probabilistic models providing the hypothesis scores.

This architecture allows to consider the understanding of a sentence as a composition of transducers in the following order:

$$\lambda_{SLU} = \lambda_S \circ \lambda_R \circ \lambda_T \circ \lambda_F \circ \lambda_L$$

with

- λ_S is the acceptor of the source sentence or graph s;
- λ_R implements source segmentation and reordering roles;
- λ_T is based on a dictionary of tuples, combining sequences of the source language and their possible tags based on the tuples inventory;
- λ_F is a feature matcher, which assigns probability scores to tuples using a CRF model;
- λ_L is a language model of the target language (here a concept language model)

3. Joint decoding of speech language transcription and understanding

The joint decoding issue has already been addressed in the past, mostly when system component pipelines are involved for human-machine interaction systems. Several studies have proposed a joint decoding between speech recognition and understanding to take into account the n-best recognition hypotheses during the semantic tagging [30, 27, 10]. These early works have proposed to produce a confusion network out of several recognition graph outputs. The understanding system was represented as a WFST, which weights were obtained by maximum likelihood estimates on the training data. Then joint decoding is obtained as the composition of the recognition graph with the understanding graph.

The positive results obtained by these proposals have encouraged further work in the same line. Given that the most successful models in the SLU state-of-the-art are CRF, [7] proposed to use them instead of WFST for understanding. In the same line, our proposal seeks a joint decoding for recognition and understanding. The CRFPB-SMT model (presented in Section 2) allows to obtain a graph of understanding hypotheses. Since the recognition output vocabulary is the same as for the understanding input, these two graphs can be composed to derive a joint graph:

$$\lambda_{joint} = \lambda_{ASR} \circ \lambda_{SLU}$$

Such a decoding may lead to very complex graphs depending on the targeted domain. Anyhow to avoid technical limitations (machine memory or running time) thresholds can still be introduced to appropriately reduce their size with beam-search or general graph reduction techniques.

Also a normalized class of word lattice, called word confusion network (WCN), has been proposed [19, 11]. WCN force the competing word to be in the same group and impose the alignment of the words that occur at the same approximate interval in the lattice. The posterior probabilities of a word from the WCN are based on the sum of the probabilities of all paths containing this word at the approximate time. WCN have the advantage of being smaller than direct ASR word lattices while

Model	Sub	Del	Ins	CER
CRF-SLU	3.1	8.1	1.8	12.9
CRFPB-SMT	4.2	8.8	2.3	15.3
CRFPB-SMT (+MERT)	3.4	7.7	1.9	13.0

Table 1: Evaluation of CRF-based approaches.

their word occurrence using their best path is comparable to the initial lattice. In the context of joint decoding of recognition and understanding, the use of a WCN in the composition may be a good solution in terms of size and efficiency.

4. Statistical post-edition for ASR error correction

Post-edition is a popular technique in the MT field. It consists in editing the outputs generated by an automatic translation process to improve its quality. This post-edition can be done manually or automatically in order to correct the automatic outputs. A post-editor system can be seen as a monolingual machine translation system. This module translates a sentence from its automatic recognized version to its human written one.

The automatic post-edition is rather largely applied in machine translation. Several recent works in MT, such as [28, 6, 25], proposed a statistical post edition process by using a phrase-based machine translation approach to post edit the output of another machine translation system.

Such a process, while not necessarily intuitive, has been proposed to improve the quality of the translated data sent to human post-editors. To train such a statistical post-editor, the output of a SMT system is used along with its manual post-edition or reference as a parallel training set.

In the same spirit, a SPE process can be applied to correct the ASR output frequent errors. In order to increase the performance of a SDS system by increasing the SLU robustness on recognition errors, we propose to apply an error-corrective mechanism to the ASR outputs based on a statistical post-edition before their transmission to the SLU system.

The advantage of our proposed CRF-based proposed approach for understanding is the fact that it can be generalized to other usages. In order to have a SPE system which can be easily integrated between the ASR system and the SLU system in our proposed joint decoding (presented in Section 3), we propose to obtain a SPE system by the same approach used to obtain the SLU system (as proposed in Section 2).

This system allows to obtain a graph for post-edition. This graph can be composed with the ASR output lattice and the SLU graph as follows:

$$\lambda_{robust\_joint} = \lambda_{ASR} \circ \lambda_{SPE} \circ \lambda_{SLU}$$

This composition takes a user utterance as input, transcribes it, tries to automatically correct this transcription then attributes the corresponding semantic tags. The selection of a transcription and a post-edition is optimized wrt the understanding that can be obtained with this transcription.

5. Experiments and Results

All experiments presented in the paper are based on the MEDIA French dialogue corpus [2]. MEDIA covers the domain of hotel reservation and tourist information. This corpus is annotated with 99 semantic labels that represent the domain semantics. The annotation is segment-based. Though the semantic information extraction can be done only at the sentence-level (e.g. [18]), anchoring the detection of concepts in the uttered words is important with regard to value extraction (i.e. determining the current normalized value for the concept).

The corpus is composed of 1257 dialogues grouped into 3 parts: a training set (13k sentences), a development set (1.3k sentences) and a test set (3.5k sentences). An in-house ASR system is used to obtain the word lattices of all the data. The 1-best path of the test set lattices obtains a 27,3% word error rate comparing to their manual references.

The concept error rate (CER) is the evaluation criterion used to evaluate the understanding task. CER can be defined as the ratio of the sum of concepts deleted, substituted and inserted on the number of concepts in the reference.

As a first step we trained a CRFPB-SMT understanding model. The tools used for the training and the evaluation of this model are presented in Section 5.1. Then we composed the ASR output with the trained SLU model. This experiment is presented in Section 5.2. Finally, we trained an SPE system for ASR error correction and we included this module in the previous composition. The SPE system and ist composition between the ASR and the SLU is detailed in Section 5.3.

5.1. Evaluation of the CRFPB-SMT approach for understanding

A CRFPB-SMT model has been constructed for the understanding task, following description of Section 2. This model was built using the N-CODE tool [5], implemented to train translation models based on n-grams [21].

This tool uses the OPENFST library [1] to built a translation graph by composition of several transducers. The difference between the model initially implemented by this tool and the model we aim to develop lies in the parameters of the so-called translation model. So we adapted the tool to use a CRF model to estimate the translation probabilities and a normalization of the probability scores obtained with the model is done over the different paths in the graph (following [16]).

In N-CODE, the reordering model (proposed by [4]) is based on a set of rules extracted automatically from training data. This approach requires a grammatical labelling of the source training sentences and a word alignment between the source and target sentences to train the λ_R model.

The TreeTagger tool [26] was used for grammatical labelling. Since the MEDIA corpus is tagged at the conceptual segment level, a word alignment between the words and the tags can be obtained using the BIO formalism (Begin Inside Outside) [23]. This formalism ensures that each word in the source sentence is aligned to its corresponding concept and therefore no additional automatic alignment is required. The language model used in our experiments is a trigram model trained on the target side of our training corpus using the SRILM tool [29].

In order to optimize the performance of this model, we propose to use MERT (Minimum Error Rate Training) algorithm [22] in order to optimize the weights of the different scored components of the model (reordering, translation and language models).

A baseline CRF model was also developed in order to compare our proposed approach with a state-of-the-art system (this system will be called CRF-SLU in tables and discussions). Table 1 presents a comparison between the CRF-SLU model and the proposed CRFPB-SMT one.

The results show that the CRFPB-SMT approach based on

Recognition	Understanding		
Output	**Input**	**CER / Oracle**	**WER**
word lattice	1-best(graph)	22.4 / 21.9	27.3
	oracle(graph)	22.2 / 21.7	25.4
	graph	22.0 / 21.6	26.2
confusion network	1-best(graph)	22.5 / 22.1	27.2
	oracle(graph)	22.0 / 21.7	25.2
	graph	**21.9** / 21.4	26.5

Table 2: Evaluation of different configurations of joint decoding (1-best, oracle or graph) using either word lattices or confusion networks.

transducers, after MERT optimization, gives comparable performance to the one obtained by the CRF-SLU approach (13.0% vs. 12.9%). This difference is not statistically significant because it is in the bounds of the confidence interval radius (0.1 at 95% confidence). But weight optimization increases significantly the performance of the CRFPB-SMT system (CER decreases from 15.3% to 13.0%).

On the other hand the results show that the use of graphs in the CRFPB-SMT model is advantageous compared to the use of a basic CRF-SLU approach. Besides the fact that it allows to obtain a similar performance, this approach allows to process input graphs. Subsequently, CRFPB-SMT (+MERT) is used for all reported experiments.

5.2. Evaluation of joint decoding

A joint decoding of recognition and understanding has been applied as proposed in Section 3. For that, we adapted the acceptor of the French understanding model to take graphs as input (instead of a single hypothesis). This transducer generates a weighted understanding graph that takes into account recognition scores.

The two scores (recognition and understanding) are considered during the final decoding (best path extraction in the final graph). We propose that the final score for each path of the graph is the addition of the recognition score and the understanding on this path. This path represents a joint decoding between recognition and understanding (marginalization of random variability caused by intermediate recognition).

We propose to perform the joint decoding in two modes: first, an ASR word lattice output is used while we used a confusion network derived from the lattice in the second. In both cases, the performance of the joint decoding is compared to the one obtained without taking into account the hypothesis graphs in the cascade. In the first case, the 1-best transcription of the graph is transmitted to the understanding system. In the second one, the oracle of the transcription graph is given as input to the understanding module. Oracle scores represent an evaluation based on the closest path to the translation reference.

We also evaluate the oracle scores for understanding on the outputs of the different combination, and we calculated the WER score on the selected transcription by the joint decoding (last column of Table 2).

The results, given in Table 2, show that the recognition graph improves the performance of the system compared to a system using the 1-best (CER 22.0% vs. 22.4% with lattice and 21.9% vs. 22.5% with confusion network). Using a recognition graph also allows better performance compared to the combination with the recognition oracle (CER 22.0% vs. 22.2% with lattice and 21.9% vs. 22.0% with confusion network).

ASR to SPE	SPE to SLU	CER	WER
1-best(graph)	1-best(graph)	21.7	26.5
graph	graph	**21.3**	26.3

Table 3: Evaluation of the error-corrective joint decoding.

It is important to mention that only combinations taking an input graph for understanding allow to select the transcription according to the labelling that will be applied. In other cases the selection of the transcription is done independently.

Also it is worth noting that the WER score of the transcription selected by the joint decoding is higher than the best transcription (26.2 vs. 25.4 with word lattice and 26.5 vs. 25.2 with confusion network) despite that the former is better in terms of CER and therefore CER and WER won't always progress jointly even if it may sound a desirable property.

5.3. Impact of an error-corrective step on the system performance

A SPE system has been obtained as proposed in Section 4. In order to obtain a parallel corpus for this experiment, we transcribed the MEDIA training data with the same ASR engine used for the previous experiments. The manual corpus with its automatic translation are used as a parallel text corpus to train the SPE system.

The composition of the three graphs (ASR, SPE and SLU) represents an error-corrective joint decoding, which tries to correct the output of the ASR system before tagging it with concepts. This decoding is performed in two modes depending on the communication channel: first, 1-best selections are transmitted between models while graphs are transmitted in the second mode. The results of these joint decodings are presented in Table 3.

Table 3 shows that the joint decoding obtained using graphs outperforms 1-best information transmission (21.3% vs. 21.7%). In both cases, the CER decreases compared to a joint decoding without the SPE module (21.7% vs. 22.5% with 1-best and 21.3% vs. 21.9% with graphs). These results confirm the conclusion of the previous experiment on the advantage of using graphs. They also confirm the interest of including a SPE system in the recognition/understanding cycle.

6. Conclusions

In this paper we proposed and evaluated an approach for joint decoding of the recognition and the understanding tasks in the context of a human-machine dialogue system. We have shown that with such a decoding 1-best performance can be beaten. We also proposed to integrate a post-edition step between recognition and understanding in order to improve ASR output correctness before conceptual tagging. It has been shown that such a system can further improve the performance of the overall system. The application of the proposed approaches led to a reduction of CER 1.1% absolute on the MEDIA corpus.

Finally, its is important to mention that the oracle of this robust joint decoding gives a CER of 20.9% and therefore the system performance can be further improved by a proper adjustment of the model weights since better hypotheses are available in the graph. An optimization procedure such as MERT could be employed to determine the weights.

7. References

[1] C. Allauzen, M. Riley, J. Schalkwyk, W. Skut, and M. Mohri. OpenFst: A general and efficient weighted finite-state transducer library. In *Proceedings of the 12th International Conference on Implementation and Application of Automata, (CIAA 2007), volume 4783 of Lecture Notes in Computer Science*, pages 11–23. Springer, 2007.

[2] H. Bonneau-Maynard, S. Rosset, C. Ayache, A. Kuhn, and D. Mostefa. Semantic annotation of the french media dialog corpus. In *EUROSPEECH*, 2005.

[3] S. Cox and S. Dasmahapatra. High-level approaches to confidence estimation in speech recognition. *IEEE Transactions on Speech and Audio Processing*, 10(7):460–471, 2002.

[4] J. M. Crego and J. B. Mariño. Improving statistical mt by coupling reordering and decoding. *Machine Translation*, 20(3):199–215, sep 2006.

[5] J. M. Crego, F. Yvon, and J. B. Mariño. Ncode: an open source bilingual n-gram smt toolkit. *The Prague Bulletin of Mathematical Linguistics*, 96:49–58, 2011.

[6] A. D. de Ilarraza, G. Labaka, and K. Sarasola. Statistical postediting: A valuable method in domain adaptation of rbmt systems for less-resourced languages. *Workshop on Mixing Approaches to Machine Translation*, 2008.

[7] A. Deoras, R. Sarikaya, G. Tur, and D. Hakkani-Tur. Joint decoding for speech recognition and semantic tagging. In *ISCA INTERSPEECH*, 2012.

[8] V. Goel, W. Byrne, and S. Khudanpur. Lvcsr rescoring with modified loss functions: A decision theoretic perspective. In *IEEE ICASSP*, 1998.

[9] S. Hahn, M. Dinarelli, C. Raymond, F. Lefèvre, P. Lehnen, R. De Mori, A. Moschitti, H. Ney, and G. Riccardi. Comparing stochastic approaches to spoken language understanding in multiple languages. *IEEE Transactions in Audio, Speech and Language Processing*, 19(6):1569–1583, 2011.

[10] D. Hakkani-Tür, F. Béchet, G. Riccardi, and G. Tür. Beyond asr 1-best: Using word confusion networks in spoken language understanding. *Computer, Speech and Language*, pages 495–514, 2006.

[11] D. Hakkani-Tür and G. Riccardi. A general algorithm for word graph matrix decomposition. In *IEEE ICASSP*, 2003.

[12] T. J. Hazen, T. Burianek, J. Polifroni, and S. Seneff. Recognition confidence scoring and its use in speech understanding systems. *Computer Speech and Language*, 16(1):49–67, 2002.

[13] Y. He and S. Young. A data-driven spoken language understanding system. In *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2003.

[14] S. Jung, M. Jeong, and G. Geunbae Lee. Speech recognition error correction using maximum entropy language model. In *ISCA INTERSPEECH*, 2004.

[15] J. Lafferty, A. McCallum, and F.C.N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*, 2001.

[16] T. Lavergne, J. M. Crego, A. Allauzen, and F. Yvon. From n-gram-based to crf-based translation models. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 542–553, 2011.

[17] F. Lefèvre. Dynamic bayesian networks and discriminative classifiers for multi-stage semantic interpretation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 4, pages IV–13. IEEE, 2007.

[18] F. Lefèvre, F. Mairesse, and S. Young. Cross-lingual spoken language understanding from unaligned data using discriminative classification models and machine translation. In *ISCA INTERSPEECH*, 2010.

[19] 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*, 14(4):373–400, 2000.

[20] L. Mangu and M. Padmanabhan. Error corrective mechanisms for speech recognition. In *IEEE ICASSP*, 2001.

[21] J. B. Mariño, R. E. Banchs, J. M. Crego, A. de Gispert, P. Lambert, J. A. R. Fonollosa, and M. R. Costa-jussà. N-gram-based machine translation. *Computational Linguistic*, 32(4):527–549, dec 2006.

[22] F. J. Och. Minimum error rate training in statistical machine translation. In *ACL*, 2003.

[23] L. Ramshaw and M. Marcus. Text chunking using transformation-based learning. In *The Workshop on Very Large Corpora*, 1995.

[24] R. C. Rose, H. Yao, G. Riccardi, and J. H. Wright. Integration of utterance verification with statistical language modeling and spoken language understanding. *Speech Communication*, 34(4):321–331, 2001.

[25] R. Rubino, S. Huet, F. Lefèvre, and G. Linarès. Statistical post-editing of machine translation for domain adaptation. In *EAMT*, 2012.

[26] H. Schmid. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*, 1994.

[27] C. Servan, C. Raymond, F. Béchet, and P. Nocera. Conceptual decoding from word lattices: application to the spoken dialogue corpus MEDIA. In *ISCA INTERSPEECH*, 2006.

[28] M. Simard, C. Goutte, and P. Isabelle. Statistical phrase-based post-editing. In *HLTNAACL*, 2007.

[29] A. Stolcke. Srilm-an extensible language modeling toolkit. In *IEEE ICASSP*, 2002.

[30] G. Tür, J. H. Wright, A. L. Gorin, G. Riccardi, and D. Z. Hakkani-Tür. Improving spoken language understanding using word confusion networks. In *ISCA INTERSPEECH*, 2002.