JOINT N-BEST RESCORING FOR REPEATED UTTERANCES IN SPOKEN DIALOG SYSTEMS

Dan Bohus, Geoffrey Zweig, Patrick Nguyen, Xiao Li

Microsoft Research
One Microsoft Way
Redmond, WA, 98033

ABSTRACT

Due to speech recognition errors, repetitions are a frequent phenomenon in spoken dialog systems. In previous work [1] we have proposed a joint decoding model that can leverage structural relationships between repeated utterances for improving recognition performance. In this paper we extend this work in two directions. First, we propose a direct, classification-based model for the same task. The new model can leverage features that were fundamentally hard to capture in the previous framework (e.g. spellings, false-starts, etc.) and leads to an additional performance improvement. Second, we show how both models can be used to perform a combined rescoring of two n-best lists that are part of a repetition pair.

Index Terms: speech recognition, repetitions, rescoring

1. INTRODUCTION

Due to current limitations in speech recognition technology non-understandings and user repetitions are frequent phenomena in spoken language interfaces. A non-understanding occurs when a speech input signal is detected, but the system is unable to construct a valid semantic interpretation for the user's turn. Even when an interpretation of the user's turn can be obtained, if the confidence score is too low, the system might choose to reject (ignore) the recognition hypothesis altogether; in effect, this also leads to a non-understanding. When such non-understandings occur, most systems use simple strategies to recover, such as reissuing the previous prompt, providing some help, or simply asking the user to repeat.

Traditionally, when the user repeats, most systems will try decoding the new utterance using the same recognition process as for the initial one. The two user utterances are however tied by the underlying semantics, i.e. the user tries to express the same concept again. In a large number of cases, they are also strongly tied at the lexical level. For instance, an analysis of a dataset from a commercially deployed directory assistance system [1] indicated that repeated utterance oftentimes follow simple structural changes. In 46% of the cases in which the user had to repeat, the second utterance was lexically identical to the first one. A number of other common structural repetition patterns can be easily identified – see Table 1. For example, in 13.7% of the cases the repeated utterance was a lexical right truncation of the previous utterance, e.g. "Blockbuster Video" → "Blockbuster". As Table 1 illustrates, simple structural patterns like exact repetitions, and (left and right) extensions and truncations account for 70.7% of the data. In principle, we should be able to leverage this information to improve the quality of the recognition for repeated utterances.

In our previous work [1], we have signaled this opportunity and proposed a joint decoding model for this task. Essentially, instead of independently decoding each utterance, we compute the optimal sentence pair $(\boldsymbol{w_1}, \boldsymbol{w_2})$ according to a model for $P(\boldsymbol{w_1}, \boldsymbol{w_2} | \boldsymbol{a_1}, \boldsymbol{a_2})$[1]. Experimentally, we showed that this model led to a 2% absolute improvement in accuracy over the existing baseline in a commercially deployed directory assistance system.

In this paper, we extend our previous work in two different ways. First, we propose and evaluate an alternative approach for improving recognition on repeated utterances based on a direct, classification model. The new model can leverage other features, such as information about spelled words, false-starts, etc. that were fundamentally hard to capture in the previous joint decoding framework. Experimental results confirm that these additional features lead indeed to additional gains in performance. Second, the model previously discussed in [1] computes the optimal sentence pair $(\boldsymbol{w_1}, \boldsymbol{w_2})$ according to $P(\boldsymbol{w_1}, \boldsymbol{w_2} | \boldsymbol{a_1}, \boldsymbol{a_2})$. In a dialog system however we would like to select a single, most likely hypothesis. In this paper we also show how both models can be used to perform a combined rescoring of the two n-best lists.

The idea of integrating information across multiple turns in a conversation appears also in a number of other works. For instance, in [2] a Dynamic Bayesian Net is used to update belief states across multiple utterances over the course of a dialog. Similarly, [3] presents a method for learning belief updating models that scale up in a more complex spoken dialog system. In [4] joint acoustic modeling is used to improve the performance of single-word recognition. [5, 6] study repetition from a descriptive point-of-view (duration, intensity, hyper-articulation, etc.) but do not address automatic speech recognition. The work described in this paper is different from these and other works in that we investigate the particular lexical structure of repeated utterances, and leverage this structure in a classification-based approach for joint n-best list rescoring.

Pattern	(%)	First utterance	Second utterance
Exact Match	46.0	Starbucks	Starbucks
Right Extension	6.6	Starbucks	Starbucks Coffee
Right Truncation	13.7	Blockbuster Video	Blockbuster
Left Extension	1.6	Roma's Pizza	Tony Roma's Pizza
Left Truncation	2.8	The Red Lion Inn	Red Lion Inn

Table 1. Frequencies and examples of structured repetition patterns

[1] we denote by $\boldsymbol{w_1}$ and $\boldsymbol{w_2}$ the word sequences (sentences) spoken by the user the first and second time around in a repetition pair, and by $\boldsymbol{a_1}$ and $\boldsymbol{a_2}$ the corresponding acoustic sequences

2. APPROACH

2.1. A generative joint decoding model

We begin with a brief review of the generative model for joint decoding of repeated utterances that we have proposed and evaluated earlier, in [1].

If we denote by l the underlying concept that the user is trying to convey to the system (*e.g.* a particular business listing in a directory assistance application), the proposed model computes the optimal pair $(\boldsymbol{w_1}, \boldsymbol{w_2})$ as follows:

$$arg \max_{\boldsymbol{w_1}, \boldsymbol{w_2}} P(\boldsymbol{w_1}, \boldsymbol{w_2} | \boldsymbol{a_1}, \boldsymbol{a_2})$$
$$\approx arg \max_{\boldsymbol{w_1}, \boldsymbol{w_2}} \sum_{l} P(l) P(\boldsymbol{w_1}|l) P(\boldsymbol{w_2}|\boldsymbol{w_1}, l) P(\boldsymbol{a_1}|\boldsymbol{w_1}) P(\boldsymbol{a_2}|\boldsymbol{w_2})$$

The model is factored into several components: $P(l)$ captures the prior probability distribution over the concept l. $P(\boldsymbol{w_1}|l)$ can be thought of as a translation model that maps from the canonical form of the concept l to the corresponding spoken form (*e.g.* for the canonical form "Kung Ho Cuisine of China", the spoken form might be "Kung Ho Chinese Restaurant", or "Kung Ho Chinese Food", etc.) The third component in the model, $P(\boldsymbol{w_2}|\boldsymbol{w_1}, l)$, captures how users repeat, at the lexical level. As we have seen earlier, a large proportion of repetitions follow simple structural patterns; in [1] we have discussed a series of incrementally more complex models for $P(\boldsymbol{w_2}|\boldsymbol{w_1}, l)$. Finally, $P(\boldsymbol{a_1}|\boldsymbol{w_1})$ and $P(\boldsymbol{a_2}|\boldsymbol{w_2})$ are the acoustic scores for the corresponding utterances.

The proposed model was therefore used to select the optimal pair $(\boldsymbol{w_1}, \boldsymbol{w_2})$ from the two n-best lists [1]. In addition, the model can also be used to rescore either n-best list individually, in light of information contained in both n-best lists. For instance, to produce a rescoring of the second n-best list using this model, we can simply sum over $\boldsymbol{w_1}$ and compute:

$$P(\boldsymbol{w_2}|\boldsymbol{a_1}, \boldsymbol{a_2}) = \sum_{\boldsymbol{w_1}} P(\boldsymbol{w_1}, \boldsymbol{w_2}|\boldsymbol{a_1}, \boldsymbol{a_2})$$

2.2. A direct model for 2nd n-best list rescoring

We now discuss a direct, classification-based approach for rescoring the second n-best list in light of information from both n-best lists. We formulate a binary classification problem as follows: given the initial n-best list NB_1 and a particular hypothesis $\boldsymbol{w_2}$ which is part of a second n-best list NB_2, compute whether or not this hypothesis $(\boldsymbol{w_2})$ is correct. Specifically, we can use a logistic regression (or maximum-entropy) model, of the form:

$$P(\boldsymbol{w_2}\ is\ correct | \boldsymbol{w_2}, NB_1, NB_2) = \frac{e^{\boldsymbol{\alpha} \cdot \boldsymbol{f}}}{1 + e^{\boldsymbol{\alpha} \cdot \boldsymbol{f}}}$$

where $\boldsymbol{f}$ is a feature vector that characterizes the hypothesis $\boldsymbol{w_2}$, the two n-best lists NB_1 and NB_2, as well as the relationships between these entities. We reserve the discussion of the full set of features used in this model for a later section (4).

In this binary formulation, each hypothesis $\boldsymbol{w_2}$ in the second n-best list is treated independently and provides a data-point for training the model. As we shall see later, to mitigate the potential negative effects of this independence assumption we also design features that capture the relationship between the given hypothesis $\boldsymbol{w_2}$ and the other surrounding hypotheses in the second n-best list NB_2. Finally, note that, given the existing symmetry in the problem, the proposed classification-based approach can also be used to re-score the initial n-best list in light of information from the second n-best list. This can be accomplished by simply reversing the roles of the two n-best lists.

2.3. Determining the single most likely hypothesis

Both models discussed above can be used to improve recognition performance by rescoring the first (or the second) n-best list in light of information from both n-best lists. In addition the first, generative joint model allows us to compute an optimal hypothesis pair $(\boldsymbol{w_1}, \boldsymbol{w_2})$. However, in a spoken dialog system, after the user repeats, we are generally interested in computing the single, most likely hypothesis from all the ones that have been heard so far.

In this paper we propose and experiment with a very simple approach for computing the single most likely hypothesis. First, we re-score the second n-best list; we then reverse the roles of the two n-best lists and perform a rescoring of the first n-best list. Finally, we merge the hypotheses from the two rescored n-best lists, and create a combined list based on the resulting model scores.

3. DATA

The experiments we report in section 5 are based on a corpus of 150,000 orthographically transcribed user utterances from a commercially deployed directory assistance system. In this application, users call in to obtain toll-free numbers for businesses that they are interested in. The system knows of approximately 43,000 businesses. Each business may have several synonyms (*e.g.* "Greater Alarm" and "Greater Alarm Company"), leading to a total of approximately 149,000 canonical listings.

Each use session with the system was assigned a unique identifier at runtime. This allowed us to detect pairs of repeated utterances, consisting of an initial request, followed by a repetition of that request. Such paired utterances account for about half of the total number of utterances. For these experiments, we separated (by random sampling of pairs) a training set of 11,838 utterances, and a test set of 12,650 utterances.

4. FEATURES

The key role in the direct model proposed in subsection 2.2 is played by the features that describe a given recognition hypothesis $\boldsymbol{w_2}$ and its relationship to other hypotheses in the first and second n-best lists (NB_1 and NB_2). In this section, we discuss this set of features in more detail.

$\boldsymbol{w_2}$ basic features. These features capture basic information about the current hypothesis $\boldsymbol{w_2}$. They include the language model score for $\boldsymbol{w_2}$; the absolute- and relative-rank of $\boldsymbol{w_2}$ in NB_2 (Rank and RankRatio); the number of words in $\boldsymbol{w_2}$ (WordNum); whether or not $\boldsymbol{w_2}$ contains any repeated (HasRepeatedWords) words, for instance indicating a false-start – "america america online"; whether or not $\boldsymbol{w_2}$ contains single-letter words (HasOneLetterWord). In addition, we also included two features that capture the size of the two n-best lists (NB_1Size and NB_2Size).

$\boldsymbol{w_2}$ spelling features. An inspection of the dataset revealed that in a significant number of cases users would also spell some of the words in their requests. These spellings might provide additional information about the correctness of a given hypothesis; for instance "Starbucks S T A R B U C K S" is likely to be a correct recognition while "Meridian M E R I T O R" is likely to be incorrect. To capture this information, we designed a set of

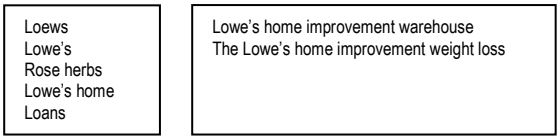


Figure 1. Example pair n-best lists

spelling features as follows: HasSpelledWords indicates if $\boldsymbol{w_2}$ contains a spelled word. SpellingMatches indicates if $\boldsymbol{w_2}$ contains a spelled word that matches another non-spelled word in $\boldsymbol{w_2}$; similarly, NotSpellingMatches indicates if $\boldsymbol{w_2}$ contains a spelled word that does not match any other word in $\boldsymbol{w_2}$ (e.g. "Meridian M E R I T O R"). SpellingInListing (and NotSpellingInListing) indicate if $\boldsymbol{w_2}$ contains a spelled word that appears (or not) in one of the canonical listings.

Listing basic features. These features capture information about how the current hypothesis and the two n-best lists relate to the canonical set of listings. We begin by identifying the subsets L_1 (and L_2) of canonical listings that have at least one word in common with one of the hypotheses from NB_1 (and respectively NB_2). Similarly, we identify the subset L of canonical listsings that have at least one word in common with a hypothesis from either NB_1 or NB_2. We then use the size of these subsets as features in the proposed classification model (NumberOfListingsNB$_1$, NumberOfListingsNB$_2$, NumberOfListings).

Repetition structural features. As we have noted earlier (see Table 1), oftentimes repeated user utterances follow simple structural patterns at the lexical level, like exact repetition, left or right extension, left or right truncation. To leverage this information in the proposed direct model we designed a set of features that capture the structural patterns between the given hypothesis $\boldsymbol{w_2}$ and other hypotheses on the previous n-best list.

We begin by adding to the five patterns illustrated in Table 1 the Other pattern which indicates that a pair of hypotheses does not follow any of these 5 structural patterns. Then, for each of these 6 patterns we define a [Pattern].NB$_1$-w$_2$.Count feature that captures how many hypotheses from the previous n-best list NB_1 are in that particular relationship with the current hypothesis $\boldsymbol{w_2}$. For instance, the LeftTruncation.NB$_1$-w$_2$.Count feature indicates the number of hypotheses contained in NB_1 for which $\boldsymbol{w_2}$ is a left truncation. Similarly, the RightExtension.NB$_1$-w$_2$.Count feature indicates the number of hypotheses contained in NB_1 for which $\boldsymbol{w_2}$ is a right extension, and so on. Concretely, in the example illustrated in Figure 1, considering that we are training or evaluating for $\boldsymbol{w_2}$ = "Lowe's home improvement warehouse", then the RightExtension.NB$_1$-w$_2$.Count feature is 2, since "Lowe's home improvement warehouse" is a right extension for two of the hypotheses from the previous n-best list – "Lowe's" and "Lowe's home." At the same time, LeftTruncation.NB$_1$-w$_2$.Count is 0, etc.

In addition, we defined a set of derived binary features [Pattern].NB$_1$-w$_2$.Count>0. For instance, the RightExtension.NB$_1$-w$_2$.Count>0 feature captures whether or not $\boldsymbol{w_2}$ is a right extension for at least one of the hypotheses in NB_1. Finally, we also computed a set of binary structural features ([Pattern].Top$_1$-w$_2$) based on comparing $\boldsymbol{w_2}$ only against the top hypothesis from the previous n-best list (as opposed to comparing it to all hypotheses from the previous n-best list and counting). For instance RightExtension.Top$_1$-w$_2$ indicates whether or not $\boldsymbol{w_2}$ is a right extension of the hypothesis at the top of NB_1(in the example from Figure 1, RightExtension.Top$_1$-w$_2$ is 0.)

Listing structural features. This set of features captures lexical structural relationships between the current hypothesis $\boldsymbol{w_2}$ and the canonical set of listings – more precisely the subset L of canonical listings that have at least one word in common with a hypothesis from either NB_1 or NB_2. The listing structural features are defined in a similar manner to the repetition structural features. For each of the 6 structural patterns, we define a [Pattern].L-w$_2$.Count feature that captures the number of listings in L that are in that particular structural relationship with $\boldsymbol{w_2}$. Like for the repetition structural features, we also defined a set of derived binary features of the form [Pattern].L-w$_2$.Count>0.

Delta features. As we have mentioned earlier in subsection 2.2., the proposed binary classification approach for rescoring treats each hypothesis in the second n-best list as an independent data-point, and asks the question: is this hypothesis correct? In an effort to mitigate the potential negative effects of this independence assumption, we have expanded the set of features to capture the relationships between the current hypothesis $\boldsymbol{w_2}$ and other hypotheses on the second n-best list.

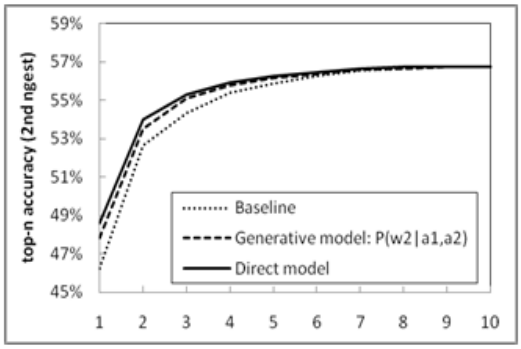
This was accomplished by designing a set of derived features as follows: for all real-valued features f, we introduced additional features f.DiffToMax, f.DiffToMin, f.DiffToMean, which capture the difference between the value of f computed for $\boldsymbol{w_2}$ and the maximum, minimum, and mean value of f across all hypotheses in NB_2. Similarly, for all binary features f, we introduced an additional feature f.IsSingle, which captures whether $\boldsymbol{w_2}$ is the only hypothesis in NB_2 for which f=1. For instance, RightExtension.Top1-w2.IsSingle indicates that the given hypothesis is the only one in NB_2 that's a left extension for the previous top hypothesis.

5. EXPERIMENTS AND RESULTS

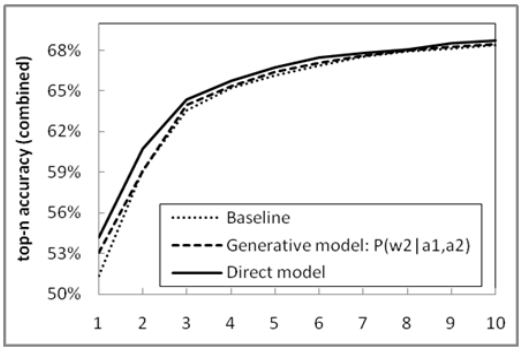
We begin by reporting results obtained when using both models to re-score the second n-best list in each repeat pair.

For the generative joint decoding approach, we computed the posterior $P(\boldsymbol{w_2}|\boldsymbol{a_1}, \boldsymbol{a_2})$, as explained previously in subsection 2.1 (more details about how this model was trained are available in [1].) In a separate approach using the same underlying model, we also computed the optimal pair $(\boldsymbol{w_1}, \boldsymbol{w_2})$, and considered the resulting $\boldsymbol{w_2}$ as the top rescored hypothesis in the second n-best list (note that this does not give us a complete rescoring of the second n-best list, but rather just a recomputed top hypothesis). The direct model (a stepwise logistic regression model) was trained using data from the development set. In each step, the next most informative feature was added to the model as long as the average data likelihood on the training set improved by a statistically significant amount (p-accept=0.05). To avoid over-fitting, we used the Bayesian Information Criterion as a stopping criterion. In addition, in an effort to identify whether (and how much) the spelling features contributed to the model, we trained an additional direct model that excluded these features.

The results, measured in terms of sentence level accuracy on the test set, are shown in Table 2, and illustrated in Figure 2 (at different n-best depths.) As these numbers show, both models improve upon the existing baseline. The generative joint decoding model earlier reported in [1] leads to a 1.7% absolute improvement in top-1 sentence accuracy or a 2.0% absolute improvement when using the top hypothesis from the optimal pair $(\boldsymbol{w_1}, \boldsymbol{w_2})$. The new proposed approach leads to a slightly larger, 2.5% absolute



(a)



(b)

Figure 2. Accuracies for baseline, generative and direct joint rescoring models: (a) 2nd utterance rescoring (b) combined n-best rescoring

	2nd n-best			Combined		
	top-1	top-2	top-3	top-1	top-2	top-3
Baseline	46.2	52.6	54.3	51.3	59.1	63.6
Generative: P(w2|a1,a2)	47.8	53.5	55.1	53.1	59.1	63.9
Generative: (w1,**w2**)	48.1	-	-	-	-	-
Direct (full)	48.6	54.0	55.3	54.2	60.7	64.4
Direct (no spelling)	48.3	53.9	55.2	53.8	60.2	64.0

Table 2. Experimental results for 2nd n-best rescoring and combined n-best rescoring using the generative and direct models

improvement. All improvements are statistically significant: $p < 10^{-5}$ in a two-tailed paired sign-test. The difference between the direct and generative models is also statistically significant ($p = 0.0017$). Finally, as the last line in Table 2 shows, the spelling features provide indeed an additional boost (the difference in top-1 accuracy between the full and no-spelling models is statistically significant $p=0.0078$; at the same time, no statistically significant difference can be detected at $p=0.05$ between the no-spelling and the generative joint decoding approach.)

Next, we used a simple approach to combine the results from the two n-best lists and select a single best hypothesis (or more generally, perform a combined n-best list rescoring.) We exchanged the roles of the two n-best lists, and also rescored the first n-best list. We then merged the hypotheses from the two lists in a combined list and sorted them in decreasing order of their model scores. To evaluate, we considered a hypothesis from the combined n-best list as correct if it matched either one of the transcripts for the first or the second utterance. In this evaluation, we are making the assumption that the 2nd user request expresses the same underlying concept and that if we obtain either one of the correct transcripts the mapping from it to the concept is known or can be easily obtained[2]. As a baseline, we merged the 2 original n-best lists based on the initial scores for each hypothesis.

The results on the test set are shown in Table 2 and illustrated in Figure 2. Under the proposed evaluation criterion, the baseline combined rescoring has a top-sentence accuracy of 51.3%. The earlier proposed generative model improves upon the baseline by 1.8% absolute; the direct model leads to an additional improvement of 1.1%. All these differences (including the one between the direct and generative models) are statistically significant ($p < 10^{-3}$.) Spelling features again provide a significant contribution to the improved performance of the direct model.

[2] unfortunately, we were not able to work directly at the semantic level since we only had orthographic transcripts but not information about the actual semantic intent of the user

7. CONCLUSION

In this paper we have proposed and evaluated a direct, classification-based approach for joint rescoring of repeated utterances in a spoken dialog system. Like the earlier, generative joint decoding model we have proposed in [1], this approach leverages information about the structural relationships between repeated utterances to improve performance. The new model also allows us to incorporate additional features (e.g. spelling, false-starts, etc.) which were fundamentally difficult to capture in the original approach. Experimental results confirm that the proposed model leads to an additional improvement in performance, and that this improvement stems to a large extent from the added features.

In addition, we have also shown how both models can be used to perform a combined rescoring and obtain a single most likely hypothesis from two n-best lists that are part of a repetition pair.

8. ACKNOWLEDGEMENTS

The authors would like to thank Bruce Buntschuh, Kyle Oppenheim, Shawn Chang, Tim Paek, Eric Horvitz and Alex Acero for helpful comments and invaluable assistance.

9. REFERENCES

[1] Zweig, G., Bohus, D., Li, X. and Nguyen, P., 2008 – *Structured Models for Joint Decoding of Repeated Utterances*, in Proceedings of Interspeech'2008, Brisbane, Australia

[2] Paek, T. and Horvitz, E., 2000 - *DeepListener: Harnessing Expected Utility to Guide Clarification Dialog in Spoken Language Systems*, in Proceedings of ICSLP, 2000.

[3] Bohus, D. and Rudnicky, A., 2006 - *A K Hypotheses + Other Belief Updating Model*, in AAAI Workshop on Statistical and Empirical Approaches to Spoken Dialogue Systems, 2006.

[4] Nair, N. U., and Sreenivas, T. V., 2007 - *Joint Decoding of Multiple Speech Patterns for Robust Speech Recognition*, in Proceedings of ASRU-2007, San Juan, Puerto Rico

[5] Oviatt, S., Levow, G.-A., MacEachern, M. and Kuhn, K., 1996 - *Modeling Hyperarticulate Speech During Human-Computer Error Resolution*, in Proceedings ICSLP, 1996.

[6] Bell, L. and Gustafson, J., 1999 - *Repetition and its Phonetic Realizations: Investigating a Swedish Database of Spontaneous Computer-Directed Speech*, in Proceedings ICPhS, 1999.