A new method for real-time lattice rescoring in speech recognition

Petar Mitankin[1,2] and Stoyan Mihov[1]

[1] Institute of Information and Communication Technologies
Bulgarian Academy of Sciences
[2] Faculty of Mathematics and Informatics
Sofia University
{petar,stoyan}@lml.bas.bg

Abstract. We introduce a novel efficient method, which improves the performance of speech recognition systems by providing the option to partially compile the word lattice into a deterministic finite-state automaton, making it suitable for the rescoring step in the speech recognition process. In contrast to the widely used n-best method our method permits the consideration of significantly larger number of alternatives within the same time-constraint and thus provides better recognition results. In this paper we present a description of the new method and empirical evaluation of its performance in comparison with the n-best method. An important advantage of our method is its applicability for real-time applications.

Most speech recognition systems [1] use a rescoring method. This consists of first using a simple acoustic and grammar model to produce a word lattice, and then to reevaluate these alternative hypotheses with a more sophisticated model on a subset of the lattice. The consideration of hypotheses other than one corresponding to the best path increases the chances of finding the correct transcription. Often, some information source or model not used for building the lattice, such as a more precise grammar or language model, is used to rescore the best hypotheses and determine the best candidate.

Currently, most systems select the n-best paths from the word lattice [2], which are then subject to rescoring using the more sophisticated models. Since the lattice contains significantly more alternative paths, there is the chance that the correct transcription will not be included in the n-best selection. Thus, if we could increase the number of hypotheses without sacrificing processing time we eventually improve the recognition performance. An alternative approach, on-the-fly rescoring, is to rescore the lattice while it is being construced, [6, 5][1].

Here we present a method, which could be considered as a natural development of the method presented in [2] for efficiently generating the n-best hypotheses. We apply essentially a similar procedure, but instead of outputting

[1] Our initial experiments show that the on-the-fly rescoring underperformes the n-best rescoring.

the n-best sentences, we produce within the same time limit a deterministic finite-state automaton, which represents together with the n-best candidates also a significantly larger number of hypotheses. Since the result is in the form of a deterministic automaton, the rescoring of the result by another n-gram language model will not require additional time for finding the best candidate after rescoring.

In the framework of the AComIn project[2] we have implemented our new method for a Large vocabulary continuous speech recognition system (LVCSR), for Bulgarian [3]. The evaluation of the method has been performed over a newly compiled Bulgarian speech corpus using the AComIn equipment.

Method description

We use a beam search based on three features: n-gram language model probabilities, acoustic probabilities and word insertion penalty. The lattice resulted from the beam search represents a nondeterministic acyclic string-to-weight transducer over the tropical semiring, [4]. Each transition of the transducer is labeled with a word and weight which represents a combination of the three features.

We start to determinize the lattice by building the initial state of the deterministic (subsequential) transducer. On each iteration the determinization procedure chooses one unexpanded state and expands it, i. e. we generate all its outgoing transitions along with the states they lead to. Since a real time determinization of the whole lattice is unfeasible, we continue the determinization procedure iteration after iteration until a timing criterion is satisfied[3]. So we do not determinize the whole lattice but only a part of it. In order to determinize the most plausible part of it possibly in the required time on each iteration we choose from all unexpanded states the one with the minimal weight as the weight of a state is the minimum of the weights of all succesful paths through the given state.

The result of the determinization procedure is a subsequential string-to-weight transducer. We apply the final rescoring on all hypotheses it represents. For this purpose this transducer is intersected with two sequential string-to-weight transducers: a transducer which represents the n-gram language model used in the beam search and a transducer which represents the rescoring language model. In the intersection the weights of the former transducer are subtracted while the weights of the latter are added. The best path in the intersection represents the best rescored hypothesis. We found experimentaly that the time needed for the intersection is negligible with respect to the time needed for the determinization procedure.

We compare our algorithm with the determinization algorithm presented in [2] desgined to generate the n-best hypotheses in the lattice. The determinization

[2] AComIn "Advanced Computing for Innovation", grant 316087, funded by the FP7 Capacity Programme (Research Potential of Convergence Regions)

[3] For the experiments presented in this paper the time for the determinization procedure is 10% of the time of the input signal

procedure in [2] is similar to our procedure, but explicitly generates the n-best hypotheses which leads to multiple traversals of one and the same transition of the subsequential transducer during the determinization. In contrast in our procedure each transition is traversed only once during the determinization. This allows us to determinize a bigger part of the lattice and to obtain in the same time limit a subsequential transducer representing more hypotheses.

Evaluation

We have tested the new method using our LVCSR system for Bulgarian [3]. For training and testing the Bulgarian Phonetic Corpus[4] created in the framework of the AComIn project was used. All tests are performed using a speaker independent (SI) acoustic model. The n-gram language models were constructed using a ∼250M words legal corpus for Bulgarian. The test set consists of 9 speakers with 50 long legal utterances each. We have varied the beam width between 1000 and 3000 states by a step of 500. For building the lattice a two-gram language model was used. The rescoring has been performed using a three-gram language model. The rescoring process time for both – the n-best method and the proposed new method – has been constrained to 0.1x of the duration of the utterance.

	Word Accuracy			WER reduction w.r.t. n-best		
	No rescoring	N-best	New method	No rescoring	N-best	New method
Beam=1000	88.77%	92.61%	92.63%	-52.04%	0.00%	0.16%
Beam=1500	89.40%	93.53%	93.57%	-63.87%	0.00%	0.56%
Beam=2000	89.57%	93.64%	93.78%	-64.02%	0.00%	2.27%
Beam=2500	89.76%	93.64%	93.82%	-60.98%	0.00%	2.84%
Beam=3000	89.85%	93.61%	93.85%	-58.87%	0.00%	3.77%
	Number of hypotheses			Time ratio		
	No rescoring	N-best	New method	No rescoring	N-best	New method
Beam=1000	1	1527	3.99E+17	0.31x	0.39x	0.40x
Beam=1500	1	378	3.04E+15	0.45x	0.54x	0.54x
Beam=2000	1	155	3.12E+12	0.65x	0.74x	0.74x
Beam=2500	1	78	1.12E+09	0.92x	1.02x	1.02x
Beam=3000	1	44	1.52E+07	1.28x	1.37x	1.37x

Table 1. Comparison of speech recognition with no rescoring, n-best rescoring and the new method.

Table 1 presents the resulting word accuracy, word error rate reduction with respect to the n-best method, the average number of hypotheses per utterance considered in the rescoring process and the time ratio for the recognition. The measurements are performed on the recognition without rescoring, with rescoring using the n-best method and with rescoring using the newly proposed method by varying the beam width.

[4] Bulgarian Phonetic Corpus: `http://lml.bas.bg/BulPhonC/`

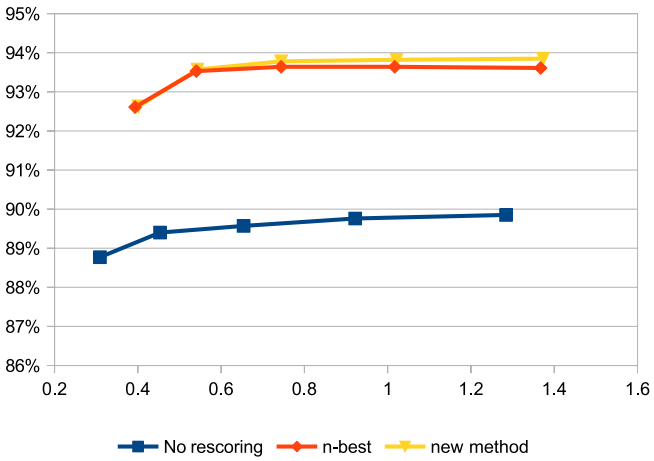

Figure 1. Time to accuracy dependency for the recognition with no rescoring, with rescoring using the n-best method and with rescoring using the new method.

Conclusion

The evaluation clearly shows that considering more hypotheses in the rescoring process within a fixed time constraint reduces the word error rate by up to 3.77% with respect to the n-best rescoring method. The improvement value depends on the particular setup – the beam width, the time constraint for rescoring, the language models used for building the lattice and for rescoring. A question for future investigation is how those factors influence the improvement of the recognition accuracy by the newly proposed method.

References

1. Xuedong Huang, Alex Acero, Hsiao-Wuen Hon. Spoken Language Processing: A Guide to Theory, Algorithm, and System Development (1st ed.). Prentice Hall PTR, Upper Saddle River, NJ, USA. 2001.
2. Mehryar Mohri and Michael Riley. An Efficient Algorithm for the N-Best-Strings Problem. In Proceedings of the International Conference on Spoken Language Processing 2002 (ICSLP 02), Denver, Colorado, September 2002.
3. Petar Mitankin, Stoyan Mihov, Tinko Tinchev: Large vocabulary continuous speech recognition for Bulgarian, Proceedings of the RANLP 2009, Sept. 2009.
4. Mehryar Mohri. Finite-state Transducers in Language and Speech Processing, Computational Linguistics, 23(2), June 1997.
5. Hasim Sak, Murat Saraclar, and Tunga Güngör. On-the-fly lattice rescoring for real-time automatic speech recognition. INTERSPEECH, ISCA, 2010.
6. T. Hori, C. Hori, Y. Minami, and A. Nakamura, Efficient WFST-Based One-Pass Decoding With On-The-Fly Hypothesis Rescoring in Extremely Large Vocabulary Continuous Speech Recognition, Audio, Speech, and Language Processing, IEEE Transactions on, 15, p. 1352-1365, May 2007.