MT-BASED ARTIFICIAL HYPOTHESIS GENERATION FOR UNSUPERVISED DISCRIMINATIVE LANGUAGE MODELING

Erinç Dikici, Murat Saraçlar

Bogazici University, Department of Electrical and Electronics Engineering
34342, Bebek, Istanbul, Turkey

ABSTRACT

Discriminative language modeling (DLM) is used as a post-processing step to correct automatic speech recognition (ASR) errors. Traditional DLM training requires a large number of ASR N-best lists together with their reference transcriptions. It is possible to incorporate additional text data into training via artificial hypothesis generation through confusion modeling. A weighted finite-state transducer (WFST) or a machine translation (MT) system can be used to generate the artificial hypotheses. When the reference transcriptions are not available, training can be done in an unsupervised way via a target output selection scheme. In this paper we adapt the MT-based artificial hypothesis generation approach to unsupervised discriminative language modeling, and compare it with the WFST-based setting. We achieve improvements in word error rate of up to 0.7% over the generative baseline, which is significant at $p < 0.001$.

Index Terms— Discriminative language model, confusion model, machine translation, unsupervised training

1. INTRODUCTION

An automatic speech recognition (ASR) system outputs possible transcriptions of a given input speech utterance. These transcriptions, also called the *hypotheses*, are generally arranged in an *N-best list* accompanied by their recognition scores (posterior probabilities assigned by the recognizer). In a typical example, one can observe that the hypothesis with the highest recognition score, called the *1-best*, is not necessarily the most accurate transcription among the N-best. Here the accuracy is measured by aligning the hypothesis to its *reference* (i.e., the manual transcription of that utterance) and counting the number of word errors.

Discriminative language modeling (DLM) techniques are applied as a post-processing step for ASR to reorganize the N-best list such that more accurate hypotheses occur at the top [1]. Training a DLM requires knowing the reference of the utterances beforehand, which takes too much time and effort especially if the number of data is large. In fact, in some applications, there may not even be any permission to listen to the recordings due to privacy issues.

One way to facilitate DLM training when the data is inadequate is to make use of a separate text source through confusion modeling. Most of the time, finding such a corpus is easier than transcribing all of the utterances. In this scenario, a small number of transcribed speech data is used to build a *confusion model* (CM) which captures the confusions (errors) made by the ASR system. This CM is then used to transform the source text into artificial hypotheses which look like the real ASR hypotheses. The accuracy of the artificial hypotheses can easily be determined since their reference, the source text, is already known. This process is sometimes referred to as *semi-supervised* training in the literature, because a small number of transcribed data is used to "label" the rest of the training examples [2–4].

It is also possible to perform discriminative language modeling even when no manual transcriptions are available at all. This process, called *unsupervised* training, can be done in a variety of ways including extracting phrasal cohorts [5], retraining with a weaker acoustic model [6], or relying on a confidence score rather than the reference [7]. In this study we follow the work of Kuo et al. [8] by utilizing the Minimum Bayes Risk (MBR) score to replace the missing reference with a selected hypothesis called the *target output*. Once the target ranks of the N-best hypotheses are determined, one can use them to train the DLM directly or via confusion modeling. We will refer to these cases as *unsupervised-DLM* and *unsupervised-CM*, respectively.

There exists several approaches to build a CM. One approach forms a weighted finite-state transducer (WFST) to store the confusions made by the recognizer along with their occurrence probabilities. The artificial hypothesis generation process then becomes a simple composition operation of the source text with the CM. Another approach includes confusion modeling as part of a machine translation (MT) system in which artificial hypotheses make up the possible target translations of the source sentence.

The authors would like to thank Arda Çelebi for the WFST-based artificial hypothesis generation framework. This research is supported in part by the TUBITAK project 109E142, Turkish State Planning Organization (DPT) under the TAM project 2007K120610 and the Bogazici University Research Fund (BU-BAP) project D-7948 (14A02D3). The numerical calculations reported in this paper were performed at TUBITAK ULAKBIM, High Performance and Grid Computing Center (TRUBA Resources).

978-0-9928626-3-3/15/$31.00 ©2015 IEEE

The study [9] compares these two approaches for semi-supervised training where the references of examples for training the CM are known, and reports that artificial hypotheses generated by the MT system are more effective than those of the WFST approach. The WFST approach has also been applied to the unsupervised-CM case in [10] where the authors achieve a performance comparable to the semi-supervised case. The main purpose of this paper is to present the use of MT-based confusion modeling in an unsupervised-CM scenario, and to compare its performance to its WFST-based counterpart. To the authors' knowledge, this is the first paper to feature an MT-based confusion model in an unsupervised setting.

This paper is organized as follows: In Section 2, we present the methods required to train the DLM. In Section 3 we explain the two artificial hypothesis generation approaches in detail. Section 4 includes our experimental setup and results. Section 5 concludes this paper with a summary and discussion.

2. DISCRIMINATIVE LANGUAGE MODEL TRAINING

This section deals with the basics of discriminative language model training. We choose the linear model and use the ranking perceptron algorithm to train its parameters. To select the target ranks of the hypotheses for unsupervised training, we employ the minimum Bayes risk criterion.

2.1. Linear model

In this study we adapt the linear model of [11] for unsupervised modeling. In this setting, x represents the spoken utterance that is input to the recognizer, and y represents its written counterpart. In a supervised scenario, y would stand for the reference (manual transcription) of x. Throughout this paper, since we do not have any reference, we will use y to refer to the target output, which is the hypothesis selected to replace the reference.

$\tilde{y} \in \tilde{\mathcal{Y}}$ are the N-best hypotheses that serve as training examples for discriminative modeling. These may be either real hypotheses output by the ASR system, or artificial hypotheses generated by the CM, to be explained in Section 3.

The symbol $\mathbf{\Phi}(\tilde{y})$ denotes the feature vector which represents a hypothesis in a d-dimensional modeling space. In our implementation, this vector contains the unigram frequencies of the hypothesis. $\mathbf{w}$ is the model vector that is estimated by discriminative training. Each element of $\mathbf{w}$ is the weight associated with the corresponding feature of $\mathbf{\Phi}$.

2.2. Ranking perceptron

We use a variant of the ranking perceptron algorithm as in [12] to train the parameters of the linear model. The WPerRank, whose pseudocode is given in Figure 1, considers the N-best hypotheses in pairs (a, b), and aims to reorganize the list such that if a has fewer word errors (ranked higher) than b, a's score (its inner product with the current model) must be significantly greater than b's. The score difference threshold is adjusted by a margin multiplier denoted by $\tau\Delta(a, b)$ where $\Delta(a, b)$ is the Levenshtein (edit) distance between the hypotheses. This multiplier also occurs in the model update to take into account the total number of word errors [13]. Learning rate (η) and decay rate (γ) multipliers are included to facilitate the convergence of the iterative optimization procedure. The WPerRank makes several passes over the data and in the end, the model weights obtained at each update step are averaged for robustness.

```
input number of training examples $I$,
number of iterations $T$, margin multiplier $\tau > 0$,
learning rate $\eta > 0$, decay rate $\gamma > 0$
$\mathbf{w} = 0, \mathbf{w}_{sum} = 0$
for $t = 1 \dots T$ do
  for $i = 1 \dots I$ do
    for $(a, b) \in \tilde{\mathcal{Y}}$ do
      if $r_a \succ r_b$ & $\langle \mathbf{w}, \mathbf{\Phi}(a) - \mathbf{\Phi}(b) \rangle < \tau\Delta(a, b)$ then
        $\mathbf{w} = \mathbf{w} + \eta\Delta(a, b)(\mathbf{\Phi}(a) - \mathbf{\Phi}(b))$
    $\mathbf{w}_{sum} = \mathbf{w}_{sum} + \mathbf{w}$
  $\eta = \eta \cdot \gamma$
return $\mathbf{w}_{avg} = \mathbf{w}_{sum}/(IT)$
```

Fig. 1. The WPerRank algorithm.

In the testing phase, the estimated model vector $\mathbf{w}_{avg}$ is used to reweight the N-best hypotheses of an ASR output. The final result is the hypothesis which gives the highest inner product score with the estimated model:

$$y^* = \underset{\tilde{y} \in \tilde{\mathcal{Y}}}{\operatorname{argmax}} \Big\{ w_0 \log P(\tilde{y}|x) + \langle \mathbf{w}_{avg} \mathbf{\Phi}(\tilde{y}) \rangle \Big\}. \quad (1)$$

Here, $\log P(\tilde{y}|x)$ is the recognition score assigned to $\tilde{y}$ by the baseline recognizer for the given utterance x, and w_0 is a scaling factor which is optimized on a held-out set. The overall system performance is computed by considering all y^* and represented in word error rate (WER).

2.3. Using MBR hypothesis as the target output

The rank of a hypothesis is an essential element in discriminative language modeling as it provides the supervision for training. If the reference transcriptions are not available, the accuracy, thus the ranks of the hypotheses cannot be determined. One way to overcome this is to utilize a computed score as an indicator of their target ranks. In this study we follow the same method as in [8] to select a target output hypothesis in place of the missing reference, using the Minimum Bayes Risk (MBR) scores. The MBR score for a target output candidate $\hat{y}$ is defined as:

$$\begin{aligned} MBR(\hat{y}|x) =& E_{\tilde{y}|x}[\Delta(\tilde{y},\hat{y})] \\ =& \sum_{\tilde{y}\in\tilde{\mathcal{Y}}} \Delta(\tilde{y},\hat{y})p(\tilde{y}|x), \end{aligned} \tag{2}$$

where $\Delta(\tilde{y},\hat{y})$ denotes the edit distance of the other hypotheses $\tilde{y}$ aligned to the candidate, and $p(\tilde{y}|x)$ denotes the recognition score assigned to $\tilde{y}$ by the ASR system. The MBR target output is the hypothesis which yields the lowest MBR score:

$$y = \underset{\hat{y}\in\tilde{\mathcal{Y}}}{\operatorname{argmin}}\, MBR(\hat{y}|x). \tag{3}$$

3. GENERATING ARTIFICIAL HYPOTHESES

It is possible to include additional text data (that is not accompanied by any acoustic recording) into DLM training by generating artificial hypotheses via confusion modeling. The CM is built using the real N-best hypotheses and is supposed to capture the inherent variability in the ASR output. In this section we present two hypothesis generation approaches that we use in this study.

3.1. WFST-based confusion modeling

The first approach uses a context independent CM similar to that presented in [4]. We first align each hypothesis to the MBR target output using the edit distance. This alignment yields a list of matching language unit pairs that are confused by the ASR, and the frequency of their match-ups gives the probability of their confusion. The CM is represented by a single-node WFST having these pairs as input-output values and the confusion probabilities as weights.

In order to generate artificial hypotheses, the source text is composed with the CM. This yields alternative hypotheses together with their occurrence probabilities in the form of a graph which resembles the lattice output of an ASR system that processed a spoken version of that source text. In the end, the most probable N paths are selected and listed.

3.2. MT-based confusion modeling

We compare the WFST-based artificial hypothesis generation approach with a statistical phrase-based machine translation (MT) framework. An MT system typically tries to match the words or phrases of a source language to those of the target language, and requires a bilingual parallel corpus. In our implementation, we treat the MBR target outputs as the source language text and the N-best ASR hypotheses as their translations in the target language. This way, the translation alternatives learned by the MT system will yield a CM which is similar to that obtained in the WFST-based approach. However, this time the CM is context dependent as the MT system is phrase-based.

The steps of the MT-based system are as follows: First, morph alignment is performed between the parallel text. Unlike traditional MT, we use the Levenshtein algorithm as in the WFST setup rather than a more complicated word alignment package such as Giza++ [14], since there is no variation in the order of morphs in our data. Using these alignments, the system computes the maximum likelihood of the lexical translations, extracts phrases and tunes the weights of the feature functions for the phrase translation rules. Finally, the source text is decoded into artificial hypotheses using these translation probabilities. In order to preserve the alignment structure, a phrase reordering model is not built and distortions are not allowed during decoding.

4. EXPERIMENTS

In this study we apply discriminative language modeling for Turkish large vocabulary continuous speech recognition. We first introduce our experimental setup and then present the experimental results.

4.1. Experimental setup

In our experiments, we use parts of our Turkish Broadcast News Speech and Transcripts Database [15], which is a collection of Turkish TV and radio channel recordings that are manually transcribed. The dataset is divided into two non-overlapping pieces as follows:

The first piece consists of 60-hours of acoustic data, which are passed through the ASR system to obtain the real hypotheses. These hypotheses are used either to train the DLM directly, or to build the CM. Due to its acoustical nature, we will refer to this piece as $\mathcal{A}$ throughout this section.

The second piece consists of manual transcriptions of 34K utterances, which roughly corresponds to the same duration of speech as the first piece. We use the second piece as the source text upon which artificial hypotheses are generated. We intentionally select these transcriptions but not some other text so that we can compare the artificial hypotheses with the real hypotheses of the same source. Following a similar notation, we will refer this piece as $\mathcal{T}$.

Turkish is an agglutinative language of the Altaic family. It has a highly inflectional morphology which causes high out-of-vocabulary rates in ASR. In order to compensate for this, we use morphs instead of words to as the language unit. Morphs are statistically derived pieces of a word, similar to morphemes. We choose morphs because previous experiments have shown that they suit better to the agglutinative nature of Turkish [16], and because experiments provide higher accuracies with this setup [4].

Our baseline ASR system is composed of a triphone acoustic model and a morph trigram generative language model [16], and is prepared using the AT&T library [17] and the SRILM [18] toolkit. For artificial hypothesis generation, the WFST-based system is implemented by the OpenFST library [19] whereas the MT-based system is implemented by the Moses toolkit [20]. The Morfessor tool is used to generate the morphs [21]. Significance tests are done using the NIST MAPSSWE tool [22].

The ASR N-best lists include 50 hypotheses. In order to obtain equivalent results, we limit the number of artificial hypotheses to 50. The feature vector of the linear model, $\mathbf{\Phi}$, consists of morph unigram frequencies. There are 34K unique morphs in the real N-bests of set $\mathcal{A}$ and 19K unique morphs in the source text of set $\mathcal{T}$.

The held-out (parameter optimization) set contains 3.1 hours and the test (evaluation) set contains 3.3 hours of speech, both of which correspond to around 2K utterances. Prior to discriminative modeling, the generative baseline WER is 22.9% for the held-out set and 22.4% for the test set. The oracle rates (the lowest WER that can be obtained on the N-best lists) for the same sets are 14.2% and 13.9%, respectively.

4.2. Experimental results

We first investigate the performance of three different discriminative language modeling scenarios. Table 1 presents the WERs obtained by the WPerRank algorithm on the held-out and test subsets. The first row represents the unsupervised-DLM scenario, where the DLM is trained directly using the real ASR hypotheses ($\mathcal{A}$). In the second row, $\mathcal{A}$ is used to build a WFST-based CM, which in turn generates the artificial hypotheses $\mathcal{T}_{WFST}$ through unsupervised-CM. The third row is similar to the second, this time using the MT-based approach to generate $\mathcal{T}_{MT}$. For all experiments, the WPerRank is allowed to make at most 50 iterations over the training data, and algorithmic parameters are optimized on the held-out set.

Training Data	Held-out	Test
$\mathcal{A}$	22.5	22.0
$\mathcal{T}_{WFST}$	22.5	22.3
$\mathcal{T}_{MT}$	22.4	22.0

Table 1. WPerRank WER (%) for different training data types.

We see from Table 1 that all three experiments provide lower WER than the held-out baseline of 22.9%. Training the DLM using $\mathcal{T}_{MT}$ improves the held-out accuracy by 0.5%, which is statistically significant at $p < 0.001$. The test set performance is also better than using $\mathcal{T}_{WFST}$, which shows that the model obtained by the MT-based approach is more generalizable.

The unsupervised-DLM experiment which uses real hypotheses $\mathcal{A}$ to train the DLM shares the best test set WER with $\mathcal{T}_{MT}$. Note that in this experiment, the manual transcriptions of the acoustic data are not known, and the MBR target outputs are used instead. For comparison, if the manual reference transcriptions of $\mathcal{A}$ were available, the rate on the same set would be 21.6% (not shown on the table). This suggests that unsupervised training is able to provide half of the gains that could be obtained with the supervised technique, without altering the test set accuracy.

Training Data	Held-out	Test
$\mathcal{A}$ + $\mathcal{T}_{WFST}$	22.2	22.0
$\mathcal{A}$ + $\mathcal{T}_{MT}$	22.3	22.1
$\mathcal{A}$ + $\mathcal{T}_{WFST}$ + $\mathcal{T}_{MT}$	22.2	22.0

Table 2. N-best combination WPerRank WER (%).

It is also possible to combine the unsupervised-DLM and unsupervised-CM approaches by combining the real hypotheses of set $\mathcal{A}$ with the artificial hypotheses of set $\mathcal{T}$. Table 2 shows the performances for possible combinations of different data types. We see that combining all three sources decreases the held-out WER by an additional 0.2%, down to 22.2%.

5. CONCLUSION

In this study we compare WFST- and MT-based artificial hypothesis generation approaches for unsupervised discriminative language modeling. These techniques allow us to make use of acoustic and text data that are coming from different sources to train the discriminative language model, with no supervision at all.

Experiments have shown that the MT-based approach is able to yield to slightly better WER than the WFST-based approach, although the superiority of MT-generated hypotheses are not as apparent as reported in [9] where a similar analysis was carried out for semi-supervised training. Combining the real and artificial examples also creates a positive effect on the system accuracy.

In order to achieve good discriminative language performance without supervision, the generated hypotheses must be sufficiently diverse, yet as similar to the real hypotheses as possible. This in turn requires a well-trained confusion model. In the future we would like to extend our CM coverage by aligning the N-best hypotheses to themselves instead of the MBR-hypothesis. We also would like to improve hypothesis diversity by trying efficient data sampling strategies on the generated N-bests.

REFERENCES

[1] Brian Roark, Murat Saraçlar, Michael Collins, and Mark Johnson, "Discriminative language modeling with conditional random fields and the perceptron algorithm," in *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, Stroudsburg, PA, USA, 2004, ACL, Association for Computational Linguistics.

[2] Gakuto Kurata, Abhinav Sethy, Bhuvana Ramabhadran, Ariya Rastrow, Nobuyasu Itoh, and Masafumi Nishimura, "Acoustically discriminative language model training with pseudo-hypothesis," *Speech Communication*, vol. 54, no. 2, pp. 219 – 228, 2012.

[3] Q.F. Tan, K. Audhkhasi, P.G. Georgiou, E. Ettelaie, and S. Narayanan, "Automatic speech recognition system channel modeling," in *Proc. Interspeech*, 2010, pp. 2442–2445.

[4] Arda Çelebi, Hasim Sak, Erinç Dikici, Murat Saraçlar, M. Lehr, E. Prud'hommeaux, P. Xu, N. Glenn, D. Karakos, S. Khudanpur, B. Roark, K. Sagae, I. Shafran, D. Bikel, C. Callison-Burch, Y. Cao, K. Hall, E. Hasler, P. Koehn, A. Lopez, M. Post, and D. Riley, "Semi-supervised discriminative language modeling for Turkish ASR," in *Proc. ICASSP*, 2012, pp. 5025–5028.

[5] P. Xu, B. Roark, and S. Khudanpur, "Phrasal cohort based unsupervised discriminative language modeling," in *Proc. Interspeech*, Portland, Oregon, Sept 2012.

[6] P. Jyothi, L. Johnson, C. Chelba, and B. Strope, "Distributed discriminative language models for Google voice search," in *Proc. ICASSP*, 2012, pp. 5017–5021.

[7] Takanobu Oba, Atsunori Ogawa, Takaaki Hori, Hirokazu Masataki, and Atsushi Nakamura, "Unsupervised discriminative language modeling using error rate estimator.," in *Proc. Interspeech*, 2013, pp. 1223–1227.

[8] H-K. J. Kuo, E. Arisoy, L. Mangu, and G. Saon, "Minimum Bayes risk discriminative language models for Arabic speech recognition," in *Proc. ASRU*, 2011, pp. 208–213.

[9] Erinç Dikici, Emily Prud'hommeaux, Brian Roark, and Murat Saraçlar, "Investigation of MT-based ASR confusion models for semi-supervised discriminative language modeling," in *Proc. Interspeech*, Lyon, France, August 2013.

[10] Erinç Dikici and Murat Saraçlar, "Unsupervised training methods for discriminative language modeling," in *Proc. Interspeech*, Singapore, September 2014, pp. 2857 – 2861.

[11] Michael Collins and Nigel Duffy, "New ranking algorithms for parsing and tagging: Kernels over discrete structures, and the voted perceptron," in *ACL*, 2002, pp. 263–270.

[12] Erinç Dikici, Murat Semerci, Murat Saraçlar, and Ethem Alpaydın, "Classification and ranking approaches to discriminative language modeling for ASR," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 2, pp. 291–300, 2013.

[13] Hasim Sak, Murat Saraçlar, and Tunga Gungor, "Morpholexical and discriminative language models for Turkish automatic speech recognition," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 20, no. 8, pp. 2341–2351, 2012.

[14] Franz Josef Och and Hermann Ney, "A comparison of alignment models for statistical machine translation," in *Proceedings of the 18th Conference on Computational Linguistics*, 2000, pp. 1086–1090.

[15] Murat Saraçlar, "Turkish broadcast news speech and transcripts LDC2012S06," 2012, Philadelphia: Linguistic Data Consortium. Web Download.

[16] Ebru Arısoy, Doğan Can, Sıddıka Parlak, Haşim Sak, and Murat Saraçlar, "Turkish broadcast news transcription and retrieval," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 5, pp. 874–883, 2009.

[17] Mehryar Mohri, "Finite-state transducers in language and speech processing," *Computational Linguistics*, vol. 23, no. 2, pp. 269–311, 1997.

[18] Andreas Stolcke, "SRILM – an extensible language modeling toolkit," in *Proceedings of ICSLP*, Denver, 2002, vol. 2, pp. 901–904, http://www.speech.sri.com/projects/srilm/.

[19] Cyril Allauzen, Michael Riley, Johan Schalkwyk, Wojciech Skut, and Mehryar Mohri, "OpenFst: A general and efficient weighted finite-state transducer library," in *CIAA 2007*. 2007, vol. 4783 of *LNCS*, pp. 11–23, Springer, http://www.openfst.org.

[20] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbsts, "Moses: Open source toolkit for statistical machine translation," in *Proceedings of the 45th Annual Meeting of the ACL Interactive Poster and Demonstration Sessions*, 2007, pp. 177–180.

[21] Mathias Creutz and Krista Lagus, "Unsupervised discovery of morphemes," in *Proceedings of the ACL-02 workshop on Morphological and phonological learning - Volume 6*, Stroudsburg, PA, USA, 2002, MPL '02, pp. 21–30, Association for Computational Linguistics.

[22] D. Pallett, William M. Fisher, and Jonathan G. Fiscus, "Tools for the analysis of benchmark speech recognition tests," in *Proc. ICASSP*, 1990, vol. 1, pp. 97 – 100.