# Semi-Supervised Discriminative Language Modeling with Out-of-Domain Text Data

**Arda Çelebi**[1] **and Murat Saraçlar**[2]
[1]Department of Computer Engineering
[2]Department of Electrical and Electronics Engineering
Boğaziçi University, Istanbul, Turkey
{arda.celebi, murat.saraclar}@boun.edu.tr

## Abstract

One way to improve the accuracy of automatic speech recognition (ASR) is to use discriminative language modeling (DLM), which enhances discrimination by learning where the ASR hypotheses deviate from the uttered sentences. However, DLM requires large amounts of ASR output to train. Instead, we can simulate the output of an ASR system, in which case the training becomes semi-supervised. The advantage of using simulated hypotheses is that we can generate as many hypotheses as we want provided that we have enough text material. In typical scenarios, transcribed in-domain data is limited but large amounts of out-of-domain (OOD) data is available. In this study, we investigate how semi-supervised training performs with OOD data. We find out that OOD data can yield improvements comparable to in-domain data.

## 1 Introduction

Discriminative language modeling (DLM) helps ASR systems to discriminate between acoustically similar word sequences in the process of choosing the most accurate transcription of an utterance. DLM characterizes and learns from ASR errors by comparing the reference transcription of the utterance and the candidate hypotheses generated by the ASR system. Although previous studies based on this supervised setting have been successful (Roark et al., 2007; Arısoy et al., 2009; Arısoy et al., 2012; Sak et al., 2012), they require large amounts of transcribed speech data and a well-trained in-domain ASR system, both of which are hard to obtain. To overcome this difficulty, instead of training with the real ASR output, we can use simulated output, in which case the training becomes semi-supervised.

Semi-supervised training for discriminative language modeling has been shown to achieve as good word error rate (WER) reduction as the training done with real ASR output (Sagae et al., 2012; Çelebi et al., 2012). In this approach, first a confusion model (CM) is estimated from supervised data. This CM contains all seen confusions and their occurrence probabilities in hypotheses generated by an ASR system. Then, the CM is used to generate a number of alternative-but-incorrect hypotheses, or simulated hypotheses, for a given sentence. Since the CM characterizes the errors that the ASR system makes, simulated hypotheses carry these characteristics. At the end, the DLM is trained on the reference sentences and their simulated hypotheses. Although being able to simulate the output of the ASR system allows us to generate as much output as we need for the DLM training, there is not always enough text data that is in the same domain as the ASR system. Yet, it is easier to find large amounts of out-of-domain (OOD) text data. In this study, we extend the previous studies where in-domain text data was used for hypothesis simulation. Instead of using limited in-domain data, we experiment with larger amounts of OOD data for hypothesis simulation.

The rest of the paper is organized as follows. In Section 2, we summarize the related work. In Section 3, we explain the methods to simulate the hypotheses and to train the DLM. We give the experimental results in Section 4 before concluding with Section 5.

*Proceedings of NAACL-HLT 2013*, pages 727–732,
Atlanta, Georgia, 9–14 June 2013.

## 2 Related Work

The earliest work on hypothesis simulation for DLM was done by Kurata et al. (2009; 2012). They generate the probable n-best lists that an ASR system may output for a hypothetical input utterance given a word sequence. In another study, Tan et al. (2010) propose a system for channel modeling of ASR for simulating the ASR corruption using a phrase-based machine translation system trained between the reference and output phoneme sequences from a phoneme recognizer. Jyothi and Fosler-Lussier (2010) also model the phonetic confusions using a confusion matrix that takes into account word-based phone confusion log likelihoods and distances between the phonetic acoustic models. This model is then used to generate confusable word graphs for training a DLM using the perceptron algorithm. Xu et al. (2009) propose the concept of cohorts and report significant WER improvement for self-supervised DLM. Similarly, Sagae et al. (2012) use phrasal cohorts to simulate ASR output and the perceptron algorithm for training. They observe half of the WER reduction that the fully supervised methods achieve. In another parallel study, Çelebi et al. (2012) work on a Turkish ASR system and consider various confusion models at four different granularities (word, morph, syllable, and phone) and different sampling methods to choose from a large list of simulated hypotheses. They observe that the strategy that matches the word error (WE) distribution of the simulated hypotheses to the WE distribution of the ASR outputs yields the best WER reduction.

While the previous studies use in-domain data sets for simulation, it is quite common to collect large amounts of OOD text data from the web. However, given the nature of web data, some kind of selection mechanism is needed to ensure quality. Bulyko et al. (2007) use perplexity-based filtering to select a relevant subset from vast amounts of web data in order to increase the training data of the generative LM used by the ASR system. There are also studies that use a relative-entropy based selection mechanism in order to match the n-gram distribution of the selected data against the in-domain data by Sethy et al. (2006; 2009). In this study, we consider the perplexity-based selection method for a start.

## 3 Method

### 3.1 Sentence Selection from OOD Data

In order to select sentences from the OOD data, we use three methods in addition to random selection. We calculate the perplexity of each sentence with SRILM toolkit, which gives normalized scores with respect to the length of the sentence. Then, we order sentences based on their perplexity scores in increasing order. Perplexity is calculated by a LM trained on in-domain data. After ordering, the top of the list contains those sentences that resemble the in-domain data the most whereas the sentences at the bottom resemble the in-domain data the least. We apply the three methods on this ordered list of sentences. The first two methods, TOP-$N$ and BOTTOM-$N$, simply get the top and bottom $N$ sentences, respectively. The third method, RC-$N$x$M$, picks uniformly separated $N$ clusters of $M$ consecutive sentences, while making sure that top and bottom $M$ sentences are among the selected ones.

### 3.2 Hypothesis Simulation

Semi-supervised DLM training uses artificially generated hypotheses which mimic the ASR system output. To generate the hypotheses, we follow the three-step finite state transducer based pipeline given in Çelebi et al. (2012) and summarized by the following composition sequence:

$$\mathrm{sample}(N\text{-best}(\mathrm{prune}(\mathcal{W}\circ\mathcal{L}_{\mathcal{W}}\circ\mathcal{CM})\circ{\mathcal{L}_{\mathcal{M}}}^{-1}\circ\mathcal{G}_{\mathcal{M}}))$$

In the first step of the pipeline, we use the confusion model transducer ($\mathcal{CM}$) to generate all possible confusions that the ASR system can make for a given reference sentence $\mathcal{W}$. We consider syllable, morph and word based confusion models, and convert $\mathcal{W}$ to these units using the lexicon $\mathcal{L}_{\mathcal{W}}$. The generated alternatives are pruned for efficiency reasons.

As the output of the first step may include many implausible sequences, the second step converts them to morphs using ${\mathcal{L}_{\mathcal{M}}}^{-1}$ and reweights them with a morph-based language model $\mathcal{G}_{\mathcal{M}}$ to favor the meaningful sequences. For this, we use three approaches. The first approach is to use the LM that is used by the ASR system, called GEN-LM. The second LM called ASR-LM is trained from the output of the ASR system, whereas the third approach is not to use any language model, denoted by NO-LM, in which case we just use the scores coming from the confusion model in the first step. A large list of of $N$-best ($N = 1000$) hypotheses are produced at this stage.

The third step, called sampling, involves picking a subset of the hypotheses from a larger set with broad variety. This step is done in order to pick samples so as to make sure that they include error variety instead of just high scoring hypotheses. As done by Çelebi et al. (2012), we use four sampling methods to pick 50 hypotheses out of the highest scoring 1000 hypotheses. The simplest of them is Top50, where we select the highest scoring 50 hypotheses. Another method is Uniform Sampling (US) which selects instances from the WER-ordered list in uniform intervals. Third method, called RC5x10, forms 5 clusters separated uniformly, each containing 10 hypotheses. Lastly, ASRdist-50 selects 50 hypotheses in such a way that the WE distribution of selected hypotheses resembles the WE distribution of the real ASR output as much as it can. We accomplish this by filling the WE bins with the hypotheses having required number of WEs.

### 3.3 DLM Estimation

The training of the DLM involves representing the training data as feature vectors and processing via a discriminative learning algorithm. We represent the simulated $N$-best lists using unigram features as described by Dikici et al. (2012). As the learning algorithm, we apply the WER-sensitive perceptron algorithm proposed by Sak et al. (2011b), which has been shown to perform better for reranking ASR hypotheses as it minimizes an objective function based on the WER rather than the number of misclassifications.

## 4 Experiments

### 4.1 Experimental Setup

We employ DLM on a Turkish broadcast news transcription data set (Arısoy et al., 2009), which comprises disjoint training (105356 sentences), held-out (1947 sentences) and test (1784 sentences) subsets consisting of ASR outputs represented as $N$-best lists. We use Morfessor (Creutz and Lagus, 2005) to obtain the morph level word segmentations from which we build the LMs. For semi-supervised experiments, we use the first half of the training subset ($t_1$: 53992 sentences, 965K morphs) to learn the confusion models, and the reference transcriptions of the second half ($t_2$: 51364 sentences, 935K morphs) to generate in-domain simulated n-best lists to be compared against OOD simulated ones. For this setup, the generative baseline WER and oracle WER on the held-out set are 22.9% and 14.2% and on the test set are 22.4% and 13.9%, respectively. When we use ASR 50-best from $t_1$ for DLM training, WERs drop to 22.2% and 21.8% on the held-out and the test sets, respectively.

For OOD data, we use a data set of 10.8M sentences (140M morphs) from newspaper articles downloaded from the Internet (Sak et al., 2011a). To calculate the perplexity of OOD sentences for selection, we use a language model trained over the reference transcripts and 50-best lists of $t_1$ and $t_2$.

### 4.2 Results on Out-of-Domain Data

We start our experiments with 500K randomly selected OOD sentences, or RAND-500K. We run the simulation pipeline with four sampling methods, three confusion and three language models, giving 36 experiments in total. We choose among the proposed sampling approaches and confusion models using a rank-based comparison as done by Dikici et al. (2012).

We look at which sampling method performs the best by first dividing experiments into 9 groups, each having 4 results from all sampling methods. Within each group, we rank the sampling methods based on the WER they achieve in increasing order and take the average of assigned ranks. ASRdist-50 gets the lowest average rank of 1.8, while RC5x10, US-50, and TOP-50 come after with the averages of 2.1, 2.4, and 3.4, respectively. This shows that ASRdist-50 gives the best WER reduction on OOD data, which is also true for in-domain data (Çelebi et al., 2012).

Doing the same rank-based comparison for the CMs this time, we observe that the syllable and morph-based models have the same average rank of 1.5, whereas the word-based model has 2.8. However, a closer look reveals that the syllable-based CM paired with NO-LM is an outlier because NO-LM approach allows variety at the output but when the unit of the confusion model is as small as syllables, it produces too much variety that deteriorates the discriminative model. If we don't consider the ranks coming from NO-LM, the average rank of syllable- and morph-based models become 1.1 and 1.8, respectively. Thus, we use syllable-based models over the others for the rest of the experiments.

Knowing that the ASRdist-50 sampling method and syllable-based CM together give the best results for RAND-500K, we experiment with three more sentence selection methods described in Section 3.1. Table 1 shows all the results obtained from four 500K OOD data sets.

| OOD Data sets | GEN-LM | ASR-LM | NO-LM |
|---|---|---|---|
| TOP-500K | 22.6 | 22.6 | 22.6 |
| BOTTOM-500K | 22.4 | **22.2** | 22.5 |
| RAND-500K | **22.2** | 22.5 | 22.6 |
| RC-5x100K | 22.4 | 22.6 | 22.5 |

Table 1: WER (%) on held-out set obtained with syllable-based CMs and ASRdist-50 sampling method

According to Table 1, the highest WER reduction is achieved with BOTTOM-500K+ASR-LM and RAND-500K+GEN-LM combinations. While ASR-LM exceeds the other two LMs only in the case of BOTTOM-500K, for other three OOD data sets GEN-LM gives the best results. More interestingly, using OOD sentences resembling in-domain data (or TOP-500K) is outperformed in all cases, especially by BOTTOM-500K. To understand this, we look at the number of morphs in each data set given in Table 2. Even though each OOD data set has 500K sentences, BOTTOM-500K has the highest number of morphs (∼6.5M) and TOP-500K had the lowest (∼3.5M), while the other two have around 5.5M morphs. We also look at the morph unigram distribution (M) of all four data sets and calculating the KL divergence KL(M || U)[1] of each M to uniform distribution (U). We observe that the unigram morph distribution of the TOP-500K data set is the least uniform with KL distance of 6.6, whereas BOTTOM-500K has KL distance of 2.7 and the other two have KL distances of around 4.3. In other words, this shows that TOP-500K has the lowest content variation, especially when compared to BOTTOM-500K. Note also the slightly high value of KL distance for $t_2$, which can be attributed to the relatively low number of unique morphs (types).

| Data set | KLD | Types | Tokens |
|---|---|---|---|
| $t_2$ (50K) | 4.65 | 22,107 | 935,137 |
| TOP-500K | 6.63 | 20,689 | 3,519,012 |
| BOTTOM-500K | 2.71 | 54,458 | 6,474,385 |
| RAND-500K | 4.36 | 50,422 | 5,559,763 |
| RC-5x100K | 4.35 | 50,561 | 5,343,342 |

Table 2: KL distance, KL(M || U), between uniform distribution (U) and unigram morph distribution (M); number of unique morphs and tokens.

### 4.3 Out-of-Domain vs In-Domain Data

In this section, we compare the results for in-domain data with the results for four OOD data sets in Table 3. In order to see how the size of OOD data set affects the WER reduction, we start with 50K sentences and increase the size gradually up to 500K. The first row of Table 3 shows the WER obtained with the in-domain data $t_2$, containing approximately 50K sentences.

| Data | 50K | 100K | 200K | 500K |
|---|---|---|---|---|
| $t_2$ | 22.4 | - | - | - |
| TOP | 22.8 | 22.7 | 22.7 | 22.6 |
| BOTTOM | 22.6 | 22.4 | 22.3 | 22.2 |
| RAND | 22.5 | 22.3 | 22.3 | 22.2 |
| RC-5 | 22.5 | 22.5 | 22.3 | 22.4 |

Table 3: WER (%) on held-out set for in-domain (Syllable+ASR-LM+ASRdist-50) and four OOD data sets in increasing sizes

According to Table 3, even though 50K OOD sentences yield worse results than the same amount of in-domain sentences, as the size of OOD data set increases, the amount of WER reduction increases and surpasses the level obtained by using in-domain data. What is more interesting is that RAND outperforms in-domain data starting from 100K, whereas BOTTOM starts at a higher WER but drops relatively fast, leveling with RAND starting at 200K. Note that the best WER achieved with the simulated data matches the supervised DLM performance using ASR 50-best from $t_1$, reported in Section 4.1.

Then we go one step further and expand the BOTTOM data set to 1M sentences and we observe WER of 22.1% on the held-out set. This further supports

[1] KL(M || U) = $\sum_i p_i log(\frac{p_i}{1/V}) = log(V) - H(p)$, where $V = 61294$ and $H(p)$ is the entropy of $p$.

the observation that the more OOD data we use, the lower WER we can achieve.

As a side observation, when we calculate the WER of five 100K-blocks from the RAND-500K set, we find that the standard deviation of WER is 0.06%, which gives and idea about the significance level of the WER differences.

### 4.4 Merging Real and Simulated Hypotheses

We also evaluate whether merging simulated hypotheses with real ASR hypotheses yields further WER reductions. The result of merging the real hypotheses from $t_1$ with the simulated ones from in-domain and OOD data are shown in Table 4. The first row shows the WER of the combination with the simulated hypotheses from in-domain data $t_2$.

| Real | Simulated | WER (%) |
|---|---|---|
| $t_1$ | $t_2$ (50K) | 22.0 |
| $t_1$ | TOP-500K | 22.3 |
| $t_1$ | BOTTOM-500K | 22.1 |
| $t_1$ | RAND-500K | 22.0 |
| $t_1$ | RC-5x100K | 22.1 |
| $t_1$ | BOTTOM-1M | 21.9 |

Table 4: WER (%) on held-out set obtained by merging real and simulated hypotheses

When combined with the real hypotheses from $t_1$, RAND500K achieves the same level of WER reduction as the simulated hypotheses from $t_2$ on the heldout set. The results on the test set are also similar. On the test set, the combination of the real hypotheses from $t_1$ and the simulated hypotheses from $t_2$ achieve 21.5% WER, whereas the WER is 21.6% when the simulated hypotheses from $t_2$ are replaced by those from RAND500K. This indicates that enough OOD data can replace the in-domain data and yield similar performance, even in combination with in-domain real data.

Moreover, we further expand the OOD data to 1M for BOTTOM, however even though it reduces the WER on the heldout set, it achieves slightly higher WER on the test set (21.7%).

Next, we combine the in-domain real hypotheses from $t_1$, simulated hypotheses from $t_2$ and simulated ones from the OOD data sets. However, compared to the combination of $t_1$ and $t_2$, adding extra 500K OOD hypotheses on top of those two gives similar WERs on the held-out set while WERs on the test set increases slightly. From another point of view, adding in-domain simulated hypotheses from $t_2$ on top of real ones from $t_1$ and 500K OOD data (rows 2-5 in Table 4) provides slight WER improvement on the held-out set but not on the test set.

## 5 Conclusion

In this study, we investigate whether we can achieve the same level of WER reduction for semi-supervised DLM with the large amounts of OOD data instead of in-domain data. We observe that ASRdist-50 sampling method and syllable-based CMs yield the best results with the OOD data. Moreover, selecting OOD sentences randomly rather than using perplexity-based methods is enough to achieve the best WER reduction. We also observe that simulated hypotheses from the OOD data is almost as good as in-domain simulated hypotheses or even real ones. As a future work, we will increase the size of the OOD data and examine other methods like relative entropy based OOD selection.

## Acknowledgments

This research is supported in part by TÜBİTAK under the project number 109E142.

## References

Ebru Arısoy, Doğan Can, Sıddıka Parlak, Haşim Sak, and Murat Saraçlar. 2009. Turkish broadcast news transcription and retrieval. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(5):874–883, July.

Ebru Arısoy, Murat Saraçlar, Brian Roark, and Izhak Shafran. 2012. Discriminative language modeling with linguistic and statistically derived features. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(2):540–550, February.

Ivan Bulyko, Mari Ostendorf, Man-Hung Siu, Tim Ng, Andreas Stolcke, and Özgür Çetin. 2007. Web resources for language modeling in conversational speech recognition. *ACM Transactions on Speech and Language Processing*, 5(1):1–25, December.

Arda Çelebi, Haşim Sak, Erinç Dikici, Murat Saraçlar, Maider Lehr, Emily T. Prud'hommeaux, Puyang Xu, Nathan Glenn, Damianos Karakos, Sanjeev Khudanpur, Brian Roark, Kenji Sagae, Izhak Shafran, Daniel Bikel, Chris Callison-Burch, Yuan Cao, Keith Hall, Eva Hasler, Philipp Koehn, Adam Lopez, Matt Post, and Darcey Riley. 2012. Semi-supervised discriminative language modeling for Turkish ASR. In *Proc. ICASSP*, pages 5025–5028.

Mathias Creutz and Krista Lagus. 2005. Unsupervised morpheme segmentation and morphology induction from text corpora using morfessor 1.0. Technical report, Helsinki University of Technology. Publications in Computer and Information Science Report A81.

Erinç Dikici, Arda Çelebi, and Murat Saraçlar. 2012. Performance comparison of training algorithms for semi-supervised discriminative language modeling. In *Proc. Interspeech*, Oregon, Portland, September.

Preethi Jyothi and Eric Fosler-Lussier. 2010. Discriminative language modeling using simulated ASR errors. In *Proc. Interspeech*, pages 1049–1052.

Gakuto Kurata, Nobuyasu Itoh, and Masafumi Nishimura. 2009. Acoustically discriminative training for language models. In *Proc. ICASSP*, pages 4717–4720.

Gakuto Kurata, Abhinav Sethy, Bhuvana Ramabhadran, Ariya Rastrow, Nobuyasu Itoh, and Masafumi Nishimura. 2012. Acoustically discriminative language model training with pseudo-hypothesis. *Speech Communication*, 54(2):219–228.

Brian Roark, Murat Saraçlar, and Michael Collins. 2007. Discriminative n-gram language modeling. *Computer Speech and Language*, 21(2):373–392, April.

Kenji Sagae, Maider Lehr, Emily T. Prud'hommeaux, Puyang Xu, Nathan Glenn, Damianos Karakos, Sanjeev Khudanpur, Brian Roark, Murat Saraçlar, Izhak Shafran, Daniel Bikel, Chris Callison-Burch, Yuan Cao, Keith Hall, Eva Hasler, Philipp Koehn, Adam Lopez, Matt Post, and Darcey Riley. 2012. Hallucinated n-best lists for discriminative language modeling. In *Proc. ICASSP*.

Haşim Sak, Tunga Güngör, and Murat Saraçlar. 2011a. Resources for turkish morphological processing. *Language Resources and Evaluation*, 45(2):249–261.

Haşim Sak, Murat Saraçlar, and Tunga Güngör. 2011b. Discriminative reranking of ASR hypotheses with morpholexical and n-best-list features. In *Proc. ASRU*, pages 202–207.

Haşim Sak, Murat Saraçlar, and Tunga Güngör. 2012. Morpholexical and discriminative language models for Turkish automatic speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(8):2341–2351, October.

Abhinav Sethy, Panayiotis G. Georgiou, and Shrikanth Narayanan. 2006. Text data acquisition for domain-specific language models. In *EMNLP '06 Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 382–389.

Abhinav Sethy, Panayiotis G. Georgiou, Bhuvana Ramabhadran, and Shrikanth Narayanan. 2009. An iterative relative entropy minimization-based data selection approach for n-gram model adaptation. *IEEE Transactions on Audio, Speech and Language Processing*, 17(1):13–23, January.

Qun Feng Tan, Kartik Audhkhasi, Panayiotis G. Georgiou, Emil Ettelaie, and Shrikanth Narayanan. 2010. Automatic speech recognition system channel modeling. In *Proc. Interspeech*, pages 2442–2445.

Puyang Xu, Damianos Karakos, and Sanjeev Khudanpur. 2009. Self-supervised discriminative training of statistical language models. In *Proc. ASRU*, pages 317–322.