Discriminative Language Modeling Research at Boğaziçi University

Erinç Dikici, Murat Saraçlar

Department of Electrical and Electronics Engineering, Boğaziçi University, Istanbul, Turkey

Introduction

- The ASR output hypothesis having the highest recognition score is not necessarily the most accurate transcription. DLM aims to correct this.
- We suggest approaching the DLM task as a reranking problem rather than the common structured prediction technique.
- Depending on the availability of acoustic data and their manual transcriptions, we investigate supervised, semi-supervised and unsupervised training settings.

Discriminative Language Model Training

Elements	Acoustic input x	N-best list ..... $\tilde{\mathcal{Y}}$
	Reference transcription ..... y	Feature vector ..... $\mathbf{\Phi}(x, \tilde{y})$
Linear Model	$\langle \mathbf{w}, \mathbf{\Phi}(x, \tilde{y}) \rangle$	
Objective	Learn $\mathbf{w}$, the model parameters	

Structured Perceptron (Per)

Picks the hypothesis with the least word errors by rewarding features of the *gold-standard* (y_i) and penalizing features of the *current best* (z_i)

input set of training examples $\{1 \leq i \leq I\}$,
number of iterations T
$\mathbf{w} = 0, \mathbf{w}_{sum} = 0$
for $t = 1 \dots T, i = 1 \dots I$ **do**
  $z_i = \text{argmax}_{z \in \tilde{\mathcal{Y}}} \langle \mathbf{w}, \mathbf{\Phi}(z) \rangle$
  $\mathbf{w} = \mathbf{w} + g(y_i, z_i)(\mathbf{\Phi}(y_i) - \mathbf{\Phi}(z_i))$
  $\mathbf{w}_{sum} = \mathbf{w}_{sum} + \mathbf{w}$
return $\mathbf{w}_{avg} = \mathbf{w}_{sum}/(IT)$

Ranking Perceptron (PerRank)

Considers all hypotheses in $\tilde{\mathcal{Y}}$ and makes sure the one with fewer word errors (higher rank r) has also a higher score

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

The margin function g( ) determines the loss function to be optimized and the effect of favoring and penalizing of the model update.

Canonical	WER-sensitive *(W)*	Reciprocal *(R)*
$g(a, b) = 1$	$g(a, b) = r_a - r_b$	$g(a, b) = \frac{1}{r_a} - \frac{1}{r_b}$

Testing Choose the highest scoring hypothesis: $y^* = \underset{\tilde{y} \in \tilde{\mathcal{Y}}}{\text{argmax}} \{ w_0 \log P(\tilde{y}|x) + \langle \mathbf{w}_{avg}, \mathbf{\Phi}(\tilde{y}) \rangle \}$

Generating Artificial Hypotheses

- When transcribed acoustic data are not sufficient, we simulate ASR output by generating hypotheses based on a separate text corpus to use in DLM training.

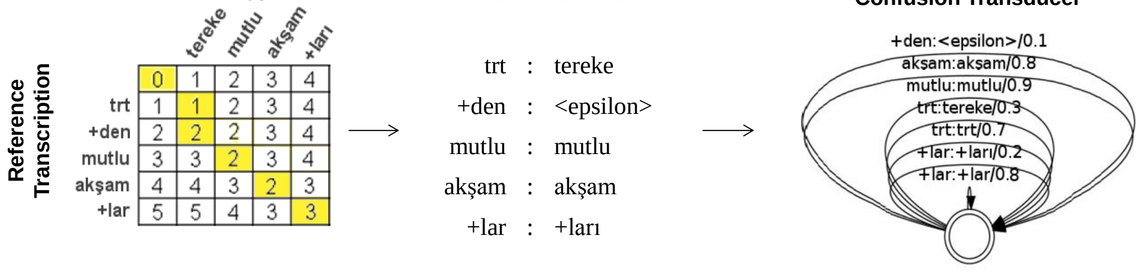
$sample\,(\mathcal{N}\text{-}best(\,prune(\mathcal{W} \circ \mathcal{L}_{\mathcal{W}} \circ \mathcal{CM}) \circ \mathcal{L}_{\mathcal{M}}^{-1} \circ \mathcal{G}_{\mathcal{M}}))$

ASR Confusion Modeling by WFSTs

- Modeling the way ASR confuses acoustically similar language units by aligning the reference transcript with the ASR output hypotheses and computing confusion statistics

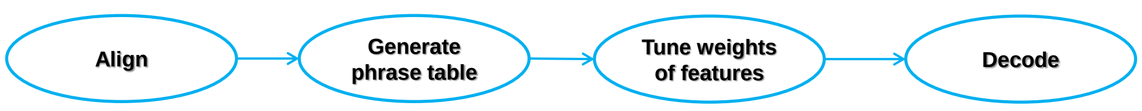
Confusion Models	Phone	Syllable	Morph	Word
Model Size	135	43 k	137 k	362 k

ASR N-best Hypothesis **Extracted Confusions** **Confusion Transducer**



ASR Confusion Modeling by MT

- Modeling ASR confusions through a statistical phrase-based MT system
- No reordering during translation, no distortions are allowed during decoding



Language Model Reweighting

- Not all hypotheses generated by the confusion model are linguistically plausible. These are reweighted using a generative language model to favor the meaningful sequences.

Language Models	ASR-LM	GEN-LM	NO-LM
Source	ASR N-bests	Newspaper websites	-

Hypothesis Sampling

- We apply several data sampling schemes to increase the accuracy, robustness and computational efficiency of the discriminative model.
- With real data, reducing the number of hypotheses during training does not alter system accuracy, though decreasing CPU times drastically.
- With simulated data, we would like to obtain a sufficiently errorful subset of artificial hypotheses with broader variety.

Methods	Top50	US50	RC5x10	ASRdist50
How it selects	First 50 hyp. with the highest score	Uniformly distributed 50 hyp. in terms of word error	5 clusters separated uniformly, each having 10 hyp.	50 hyp. having a WE distribution similar to that of ASR

Reference Selection for Unsupervised Learning

- In the unsupervised setting, no manual transcriptions are available at all. We investigate three ways to find a sequence that could serve as the missing reference and use these to determine the hypothesis ranks or to build a confusion model.

1. 1-best
2. Minimum Bayes Risk (MBR): $l(y|x) = E_{\tilde{y}|x}[L(\tilde{y}, y)] = \sum_{\tilde{y} \in \mathbf{GEN}(x)} L(\tilde{y}, y) p(\tilde{y}|x) \quad \hat{y} = \underset{y \in \mathbf{GEN}(x)}{\text{argmin}}\ l(y|x)$
3. Segmental MBR

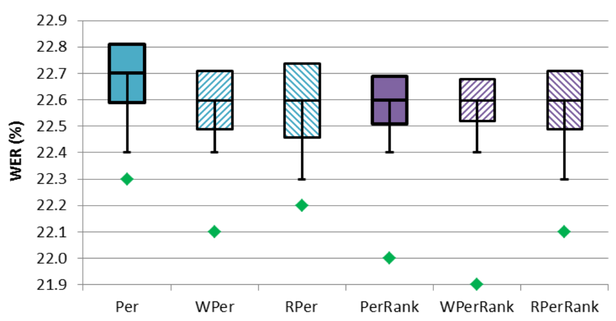
Experimental Setup

Task Turkish broadcast news transcription

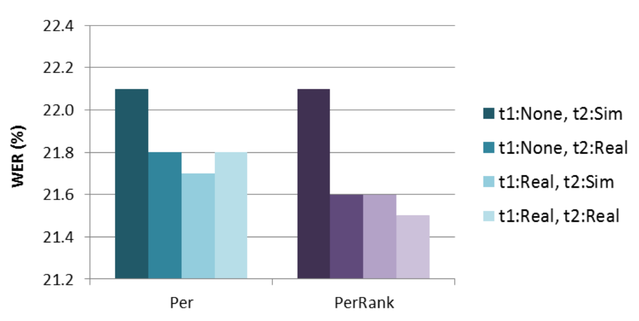
Setup 50-best list of hypotheses, Morph unigram counts (46k dimensional), highly sparse

Dataset Partitions & Usage	# of utterances	# of words	Time
t_1: Confusion model construction ASR output and reference transcriptions	53,992	686,551	188 h
t_2: Simulated hypothesis generation Reference transcriptions only	51,364	666,462	
Held-out	1,947	23,199	3.1 h
Test	1,784	23,410	3.3 h

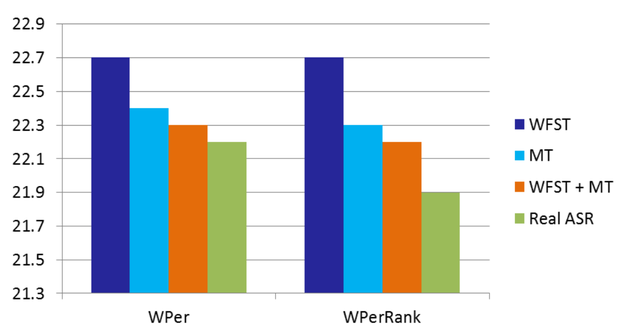
Experimental Results



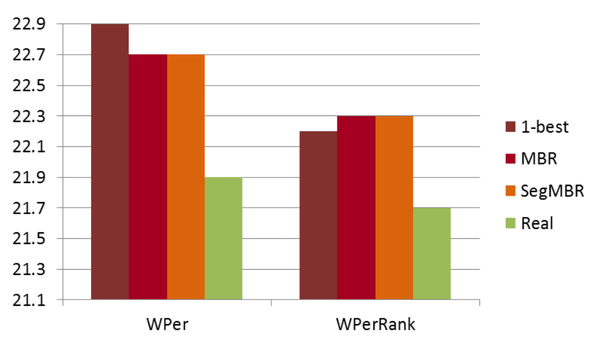
t_2 simulated (WFST CM), Held-out (◆: t_2 real)



WFST CM, Test



t_2 simulated, WFST and MT CM, GEN-LM, Held-out



t_1+t_2, unsupervised, Held-out

Discussion

- Ranking perceptrons provide accuracies similar or better than those of structured perceptrons for all experiments.
- All algorithms show a similar overall performance with simulated data, contrary to real data where ranking algorithms outperform.
- Non-constant g(.) has a positive effect on system accuracy. *W* is slighty better for the structured perceptron whereas *R* takes the lead in the ranking variant.
- Syllable- and morph-based CMs give better results. Phone-based confusions are so local that they cannot generate enough variety.
- None of the LM approaches seems dominant. Mimicking ASR WER distribution with ASRdist sampling gives better results.
- Real and simulated N-bests combined performs as good as the real ASR N-bests.
- MT CM yields more suitable simulated hypotheses for the semi-supervised setup.
- With the unsupervised training setting, improvements in WER up to half of those of the supervised case can be obtained.

Acknowledgments

This research is supported in part by TUBITAK under project number 109E142 and by the Turkish State Planning Organization (DPT) under the TAM Project number 2007K120610. The authors would like to thank Ethem Alpaydın, Ebru Arısoy, Arda Çelebi, Brian Roark, Haşim Sak, Murat Semerci and Izhak Shafran for their contributions.