Discriminative N-gram Language Modeling for Turkish

Ebru Arısoy[1]*, Brian Roark*[2]*, Izhak Shafran*[2] *and Murat Saraçlar*[1]

[1]Electrical and Electronics Engineering Department, Boğaziçi University, Istanbul, Turkey
[2]Center for Spoken Language Understanding, OGI/OHSU, Beaverton, OR, USA
{arisoyeb, murat.saraclar}@boun.edu.tr, roark@cslu.ogi.edu, zakshafran@gmail.com

Abstract

In this paper Discriminative Language Models (DLMs) are applied to the Turkish Broadcast News transcription task. Turkish presents a challenge to Automatic Speech Recognition (ASR) systems due to its rich morphology. Therefore, in addition to word n-gram features, morphology based features like root n-grams and inflectional group n-grams are incorporated into DLMs in order to improve the language models. Various feature sets provide reductions in the word error rate (WER). Our best result is obtained with the inflectional group n-gram features. 1.0% absolute improvement is achieved over the baseline model and this improvement is statistically significant at p<0.001 as measured by the NIST MAPSSWE significance test.

Index Terms: discriminative language modeling, speech recognition, agglutinative languages

1. Introduction

Discriminatively trained conditional models, such as Conditional Random Fields (CRF) [1] and Maximum Mutual Information Estimation (MMIE) [2], have been applied successfully to both acoustic modeling [3] and language modeling [4, 5, 6] for ASR. These sequence modeling approaches have been demonstrated to consistently outperform generative modeling approaches, partly due to improved parameter estimation and partly due to the ease with which very many overlapping features can be included in the models. Feature based language models allow for easy integration of relevant knowledge sources such as syntactic and semantic dependencies [5, 7, 8].

In this paper DLMs are applied to the Turkish Broadcast News (BN) transcription task. Turkish presents a challenge to ASR systems due to its rich morphology. Previous work on Turkish ASR has focused on using sub-word units for language modeling [9, 10, 11] to alleviate the data sparseness and out-of-vocabulary (OOV) problems that plague word-based systems in agglutinative languages. In this work our aim is to incorporate morphological information to the word-based language model and also to take advantage of discriminative training. Therefore, we use word based generative language models in the first pass and attempt to incorporate morphological information via features in DLMs in the second pass. Morphological features have been incorporated into discriminative framework for Czech ASR [12]. Turkish, being an agglutinative language, contains variable number of morphemes, typically 3-4 including the stem [13], which leads to new methods for defining features from morphological glosses, as outlined in this work.

The paper is organized as follows. In Section 2 we present the baseline ASR system. Section 3 gives a summary of DLMs. Experimental setup and results are presented in Sections 4 and 5 respectively. Finally, Section 6 discusses the results.

2. Baseline ASR System

In this study approximately 107 hours of Turkish Broadcast news data is used [14]. This data is partitioned into training (100.4 hours), held-out (3.2 hours) and test (3.1 hours) sets. The training data, held-out and test sets are disjoint in terms of the dates of the selected shows.

For the acoustic models, 13 dimensional Perceptual Linear Prediction (PLP) features are extracted using 16kHz acoustic data. Cepstral mean normalization for each speaker is computed and every normalized 9 consecutive cepstral frames are spliced together. Then, those frames are projected to 40 dimensional features by Linear Discriminant Analysis (LDA). Since in Turkish, each letter approximately corresponds to a phoneme, the acoustic models are based directly on letters (29 letters and 1 silence model). Context Dependent (CD) pentaphone models are trained with Maximum Likelihood (ML) training with embedded Semi-Tied Covariance (STC) re-estimation. The total number of HMM states and gaussians are set as 2800 and 60000 respectively. Vocal Tract Length Normalization (VTLN) and Speaker Adaptive Training (SAT) are applied to compensate for speaker variations. For SAT, feature space Maximum Likelihood Linear Regression (FMLLR) transforms are computed for each speaker. Then, these transforms are used in ML training of the acoustic models. In decoding FMLLR-adapted features and Maximum Likelihood Linear Regression (MLLR) adapted models are used. Acoustic models are trained with IBM's software tools [15].

For baseline generative language modeling, we have used text corpora of 96.4M words collected from the web and the reference transcriptions of the acoustic data (500K words). The most frequent 50K words from the corpora are selected as the vocabulary items and this gives 8.7% OOV rate over the test data. Trigram word-based language models with modified Kneser-Ney smoothing are generated using the SRILM toolkit [16]. Language models are linearly interpolated in order to reduce the effect of out-of-domain data. Interpolation constants are optimized over the held-out set Word Error Rate (WER).

The recognition task is also performed with IBM's software tools. The recognition results after various speaker compensation transforms are given in Table 1. In the DLM experiments, we only used the output of the CD decoding since it will be computationally expensive to generate the DLM training data for all the acoustic models. Our baseline system with CD acoustic models yields 33.3% WER over the held-out set and 31.1% WER over the test data.

Table 1: Recognition Results in terms of WERs

	Acoustic Models		
	CD	VTLN	SAT
Heldout Data	33.3	32.4	29.5
Test Data	31.1	29.4	26.9

3. Discriminative Language Modeling

DLM is a complementary approach to the existing baseline language model. In contrast to the generative language model, it is trained on acoustic sequences with their transcripts to optimize discriminative objectives using both positive (correct utterances) and negative (recognition errors) examples. DLM training data consists of lattices or N-best lists. The feature vector, $\Phi(x, y)$, is defined as a function of the acoustic input, x, and the candidate hypothesis, y. $\bar{\alpha}$ is the vector of parameters associated with the features. The candidate hypothesis which maximizes this inner product, $\Phi(x, y) \cdot \bar{\alpha}$, is the best scoring hypothesis under this model. The parameter values are estimated during training and y that maximizes the inner product is searched for in the decoding.

The DLM training data (lattices or N-best lists) is generated by breaking the training data into k folds, and recognizing the utterances in each fold using the baseline acoustic model (trained on all of the utterances) and an n-gram language model trained on the other $k-1$ folds to alleviate over-training of the language models. Acoustic model training data is not typically controlled in the same way since baseline acoustic model training is far more expensive and less prone to over-training than standard n-gram language model training [6].

The first element of the feature vector, $\Phi_0(x, y)$, is defined as the “log-probability of x, y in the lattice produced by the baseline recognizer”. This feature will present the effect of our baseline acoustic and language models to the DLM training. Other features are defined as the number of times a particular n-gram is seen in the candidate hypothesis.

The feature parameters, $\bar{\alpha}$, are learned discriminatively from the labeled training examples. Either the perceptron algorithm or Global Conditional Log-Linear Models (GCLMs) can be used for learning these parameters. The perceptron algorithm penalizes features associated with the current 1-best hypothesis, and rewards features associated with the gold-standard hypothesis (reference or lowest-WER hypothesis). GCLMs, in contrast, use the ASR system output to calculate feature gradients so as to iteratively maximize the conditional likelihood of the gold-standard hypotheses. See [6] for an overview of the approaches.

4. Experimental Setup

In our experiments, first DLM training data is generated. Then features are extracted from the training data and the parameters of these features are updated using the perceptron algorithm [17] and an n-best reranker with a regularized conditional likelihood objective[1] [18]. Our experimental set-ups are based on the findings of the previous research on DLMs [6]: (i) k-fold cross validation is applied in DLM training data generation; (ii) the oracle hypothesis is selected as the gold standard instead of the reference; (iii) N-best lists are used instead of lattices.

4.1. DLM Training Data

In this research, only the language model training is controlled via 24-fold cross validation. The same vocabulary and the interpolation constant optimized for the baseline ASR system are used for those language models. In our experiments, N-best lists (50-best and 1000-best) are utilized instead of lattices as the DLM training data. The lattice, 1000-best and 50-best oracle error rates reported as 17.4%, 21.2% and 24.5% respectively.

4.2. Feature sets for Turkish

Turkish has an agglutinative morphology where many new words can be derived from a single stem by addition of several suffixes. Thus, in our experiments, we have also used feature sets that take the morphological characteristics of Turkish into account. In the morphological analysis [19], we have a rich set of morphological features such as major root Parts of Speech (PoS), minor PoS, derivation boundaries and morphological tags. The ambiguity in the words with multiple morphological analysis is solved using a perceptron-based morphological disambiguation tool for Turkish [20].

Phrase: sunulacak bildiridekiler *(those in the paper that will be presented)*

Morphological Analysis:

sun*(present)*+Verb+ ˆDB +Verb+Pass+Pos

+ ˆDB +Adj+FutPart+Pnon

bildiri*(paper)*+Noun+A3sg+Pnon+Loc+ ˆDB +Adj+Rel

+ ˆDB +Noun+Zero+A3pl+Pnon+Nom

Inflectional Groups:

root(sunulacak)=sun*(present)*

IG1(sunulacak)=+Verb

IG2(sunulacak)=+Verb+Pass+Pos

IG3(sunulacak)=+Adj+FutPart+Pnon

root(bildiridekiler)=bildiri*(paper)*

IG1(bildiridekiler)=+Noun+A3sg+Pnon+Nom

IG2(bildiridekiler)=+Adj+Rel

IG3(bildiridekiler)=+Noun+Zero+A3pl+Pnon+Nom

Figure 1: Example morphological analysis for features

In the example in Figure 1, morphological feature sequences are separated by ˆDB symbol which denotes the derivation boundaries. These sub-lexical units are called the Inflectional Groups (IGs). The words in the example are analyzed as a root and a sequence of 3 IGs. In this paper, we have used word, root and IG-based n-grams as our feature sets. Special emphasis is given to IG-based features.

4.2.1. Word n-gram features

All the word unigrams and bigrams seen in the N-best list are used as word n-gram features. For instance:

$\Phi_j(x, y)$ = Number of times “*sunulacak bildiridekiler*” is seen in y

4.2.2. Root n-gram features

All the root n-grams, up to 2, seen in the N-best list are used as root n-gram features. For instance:

$\Phi_j(x, y)$ = Number of times “*sun bildiri*” is seen in y

4.2.3. IG-based n-gram features

As was shown in Figure 1, a Turkish word can be represented as a root and a sequence of IGs. In dependency parsing and morphological disambiguation tasks of Turkish, IGs have a common usage to handle the challenges introduced by the agglutinative nature of Turkish. It has been found that the last IG of a word determines its role as a dependent and the syntactic relations are between the last IG of the current word and one of the IGs of the word on the right [21]. In addition, using IG-based features in a discriminative framework gives the best accuracy for the Turkish morphological disambiguation and PoS tagging tasks [20].

In this research, IGs are also used in DLMs to optimize the WER of the decoder. Each word in the DLM training data is

[1] Reranker [18] is used as a tool for GCLM training.

Table 2: n-gram features used in the DLM experiments. These features are extracted from an N-best utterance composed of m words.

	n-grams	Features	index ranges
Word n-grams	unigrams	(w_i)	$1 \leq i \leq m$
	bigrams	(w_{i-1}, w_i)	$2 \leq i \leq m$
Root n-grams	unigrams	(r_i)	$1 \leq i \leq m$
	bigrams	(r_{i-1}, r_i)	$2 \leq i \leq m$
IG-based (SET1)	unigrams	(r_i), $(IG_{i,j})$	$1 \leq i \leq m$ and $1 \leq j \leq n_i$
n-grams	bigrams	(r_{i-1}, r_i), $(r_{i-1}, IG_{i,j})$, $(IG_{i-1,k}, r_i)$, $(IG_{i-1,k}, IG_{i,j})$	$2 \leq i \leq m$, $1 \leq k \leq n_{i-1}$ and $1 \leq j \leq n_i$
IG-based (SET2)	unigrams	(r_i), $(IG_{i,j})$	$1 \leq i \leq m$ and $1 \leq j \leq n_i$
n-grams	bigrams	$(IG_{i-1,n_{i-1}}, IG_{i,j})$	$2 \leq i \leq m$ and $1 \leq j \leq n_i$

represented in terms of roots and IGs and the n-gram features are extracted from this training data. The i'th word in an N-best utterance is represented as

$$r_i + IG_{i,1} + \hat{}DB + ... + IG_{i,j} + \hat{}DB + ... + IG_{i,n_i}$$

where r_i is the root of the i'th word (w_i), n_i is the number of inflectional groups in w_i and IG_{i,n_i} is the last inflectional group of w_i. $IG_{i,j}$ represents the j'th inflectional group of w_i where $1 \leq j \leq n_i$. If there are m words in an utterance, the possible IG-based unigrams and bigrams from that utterance are

1. Unigrams in a word: (r_i), ($IG_{i,j}$) where $1 \leq i \leq m$ and $1 \leq j \leq n_i$.
2. Bigrams in consecutive words: (r_{i-1}, r_i), (r_{i-1}, $IG_{i,j}$), ($IG_{i-1,k}$, r_i), ($IG_{i-1,k}$, $IG_{i,j}$) where $2 \leq i \leq m$, $1 \leq k \leq n_{i-1}$ and $1 \leq j \leq n_i$.

In the example of Figure 1, there are one (r_{i-1}, r_i), three $(r_{i-1}, IG_{i,j})$, three $(IG_{i-1,k}, r_i)$, and nine $(IG_{i-1,k}, IG_{i,j})$ bigram features. For instance:

$\Phi_j(x, y)$ = Number of times "*IG3(sunulacak) IG2(bildiridekiler)*" is seen in y

$\Phi_k(x, y)$ = Number of times "*root(sunulacak) IG3(bildiridekiler)*" is seen in y

The above features, all the possible unigrams and bigrams seen in the N-best list, are used as the IG-based (SET1) n-gram features in our experiments. We also try out reducing the number of bigram features by using a subset of the IG-based bigrams that may convey syntactic relations between the words. This feature set is called the IG-based (SET2) n-gram features and it contains all the unigrams and only the bigrams from the last inflectional group of the previous word to all the inflectional groups of the current word, $(IG_{i-1,n_{i-1}}, IG_{i,j})$, seen in the N-best list. The word, root and IG-based n-gram features used in our experiments are summarized in Table 2.

4.3. Parameter Estimation

In our experiments oracle best path is used as the gold standard. The perceptron algorithm (as presented in [6]) and the conditional likelihood reranker (as presented in [18]) are used for training the feature parameters. Due to computational limitations, the reranker is only applied to the 50-best lists.

In the perceptron model, the parameter for the baseline model, α_0 (the weight associated with $\Phi_0(x, y)$), is set to a fixed number and is not updated during training. We perform two different training scenarios. In one scenario, 8 different models are trained with α_0 from 0 to 16. The best model is chosen using the held-out set WER. In an alternative training scenario, α_0 is set to 0 during the training and then an optimal mixing parameter between the baseline score and the DLM score is chosen using the held-out set WER. Averaged perceptron parameters are used in the evaluation of the held-out and test sets. The number of iterations over the training data are also optimized on the held-out set.

5. Experimental Results

The DLM results for various unigram and bigram feature sets are given in Table 3. The results are reported for the best iterations and α_0 constants that yield the lowest WERs over the held-out set. All the feature sets converged in 1-4 iterations of the perceptron algorithm. In addition to the N-best perceptron results, WERs with the reranker are also given.

Although, oracle WERs differ significantly, 21.2% for 1000-best and 24.5% for 50-best, increasing the N-best size from 50 to 1000 does not give any significant difference on the DLM performance, an observation noticed in other work on discriminative training as well [22]. The number of active features (features with non-zero weights) are comparably smaller than the number of the initial features for all the feature sets. Even though, 1000-best lists introduce approximately 5 times more initial features than 50-best lists, the number of active features are almost the same.

The perceptron model gives 0.5-0.9% WER reductions on the test set and these reductions are statistically significant at p<0.001 as measured by the NIST MAPSSWE significance test. We also experimented with unigrams, bigrams and trigrams for the root, word and IG-based (SET1) n-gram features. These new feature sets also improved the baseline performance, however no significant performance improvement is achieved compared to not using trigram features.

IG-based (SET2) n-grams reduced the number of IG-based (SET1) initial features to 0.1M from 1.96M (94.9%) in 50-best list and to 0.17M from 4.78M (96.4%) in 1000-best list without effecting the performance of the perceptron model. In the IG-based (SET1) experiments, we do not know the position of the IGs in a word in the feature sets, since a particular IG may occur at the end in one word and at the middle in an other word. However, we found out that 15% of the ($IG_{i-1,k}$, $IG_{i,j}$) active bigram features of the best IG-based (SET1) model overlapped with IG-based (SET2) bigrams, ($IG_{i-1,n_{i-1}}$, $IG_{i,j}$). Although this finding does not lead us to the conclusion that ($IG_{i-1,n_{i-1}}$, $IG_{i,j}$) is the most salient feature among all the IG-based (SET1) bigrams, the importance of this feature is pronounced in the IG-based (SET2) experiments which give the same improvements with the IG-based (SET1) experiments, however, with a comparably smaller set of features.

GCLM trained with unigram and bigram features on the 50-best list gives 0.5-1.0% improvement over the test baseline. For the IG-based features, it performs slightly better than the perceptron training. These improvements are also statistically significant at p<0.001 as measured by the NIST MAPSSWE significance test.

6. Discussion

As seen in Table 3, we do not observe any significant performance difference between word and morphological features. However, IG-based features seem more robust than the other

Table 3: DLM results. Feats represents the number of features extracted from the N-best lists. Feats (Act.) represents the percentage of the features with non-zero weights after the parameter training.

	Feats (millions)	Feats (Act.) (%)	WER held-out	WER test
Baseline	-	-	33.3	31.1
	Word n-gram features			
Perceptron 50-best	2.08	13.3	32.6	30.5
Reranker 50-best	2.08	100.0	32.4	30.6
Perceptron 1000-best	7.18	4.2	32.5	30.5
	Root n-gram features			
Perceptron 50-best	1.04	14.4	32.6	30.6
Reranker 50-best	1.04	100.0	32.4	30.6
Perceptron 1000-best	2.83	6.7	32.5	30.5
	IG-based (SET1) n-gram features			
Perceptron 50-best	1.96	31.2	32.4	30.3
Reranker 50-best	1.96	100.0	32.3	30.1
Perceptron 1000-best	4.78	11.2	32.4	30.5
	IG-based (SET2) n-gram features			
Perceptron 50-best	0.10	43.0	32.4	30.4
Reranker 50-best	0.10	100.0	32.5	30.2
Perceptron 1000-best	0.17	27.4	32.4	30.2

features explored as the gains observed in the held-out set were preserved in the test set. In order to understand whether there is any real complementary information in the word and morphological feature sets, we also trained models using “word n-gram+root n-gram” and “word n-gram+IG (SET2) n-gram” features together. However, for the first combined feature set, we obtained the same test set error rate as when the feature sets were used separately. For the second feature set, the test set WER decreased to 30.2% for the 50-best list. To judge whether this lack of observed additivity is an indication that no real additive benefit can be expected, we performed a small cheating experiment. Among the best scoring hypotheses of each feature set associated with the same utterance, we chose the one with the lowest WER. This reduced the WERs to 29.3% and 28.9% for the above combined feature sets respectively. These cheating experiments show that the effect of each feature set can be additive. This additivity can be revealed if only the non-overlapping discriminative features are trained together, for example, using different factorizations of the tags, as in [12].

Future directions will include a key question raised by these results. How can we effectively combine complementary feature sets to achieve the additive results that we should expect based on our small cheating experiment? The current feature sets represent one possible way of defining features from the morphological analysis. We will continue to explore a range of feature definitions that may achieve further improvements. Moreover, sub-word-based language models are more appropriate for the agglutinative languages and we will investigate those units in DLMs.

While the current results are preliminary, they do present a promising direction for including morphological information within a Turkish ASR system.

7. Acknowledgements

This research was done when the first author was visiting OGI/OHSU. Her visit was supported by TUBITAK BDP (Unified Doctorate Program of the Scientific and Technological Research Council of Turkey). This research is partially supported by TUBITAK (Project code: 105E102) and Bogazici University Research Fund (Project code: 07HA201D).

8. References

[1] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, “Conditional random fields: Probabilistic models for segmenting and labeling sequence data,” in *Proc. ICML 2001*, 2001, pp. 282–289.

[2] L. Bahl, P. Brown, P. deSouza, and R. Mercer, “Maximum mutual information estimation of hidden markov model parameters for speech recognition,” in *Proc. ICASSP 1986*, 1986, pp. 49–52.

[3] D. Povey and P. C. Woodland, “Large-scale MMIE training for conversational telephone speech recognition,” in *Proc. NIST Speech Transcription Workshop*, College Park, MD, 2000.

[4] B. Roark, M. Saraclar, M. J. Collins, and M. Johnson, “Discriminative language modeling with conditional random fields and the perceptron algorithm,” in *Proc. ACL 2004*, 2004, pp. 47–54.

[5] M. Collins, M. Saraclar, and B. Roark, “Discriminative syntactic language modeling for speech recognition,” in *Proc. ACL 2005*, 2005, pp. 507–514.

[6] B. Roark, M. Saraclar, and M. Collins, “Discriminative n-gram language modeling,” *Computer Speech and Language*, vol. 21, no. 2, pp. 373–392, 2007.

[7] S. Khudanpur and J. Wu, “Maximum entropy techniques for exploiting syntactic, semantic and collocational dependencies in language modeling,” *Computer Speech and Language*, 2000.

[8] H. Erdogan, R. Sarikaya, S. F. Chen, Y. Gao, and M. Picheny, “Using semantic analysis to improve speech recognition performance,” *Computer Speech and Language*, vol. 19, pp. 321–343, 2005.

[9] E. Arisoy, H. Dutagaci, and L. M. Arslan, “A unified language model for large vocabulary continuous speech recognition of Turkish,” *Signal Processing*, vol. 86, pp. 2844–2862, 2006.

[10] K. Hacioglu, B. Pellom, T. Ciloglu, O. Ozturk, M. Kurimo, and M. Creutz, “On lexicon creation for Turkish LVCSR,” in *Proc. EUROSPEECH 2003*, Geneva, Switzerland, 2003, pp. 1165–1168.

[11] M. Kurimo, A. Puurula, E. Arisoy, V. Siivola, T. Hirsimäki, J. Pylkkönen, T. Alumäe, and M. Saraclar, “Unlimited vocabulary speech recognition for agglutinative languages,” in *Proc. HLT-NAACL 2006*, New York, USA, 2006.

[12] I. Shafran and K. Hall, “Corrective models for speech recognition of inflected languages,” in *Proc. EMNLP 2006*, Sydney, Australia, 2006.

[13] O. Cetinoglu and K. Oflazer, “Morphology-syntax interface for Turkish LFG,” in *Proc. COLING/ACL 2006*, Sydney, Australia, 2006.

[14] E. Arisoy, H. Sak, and M. Saraclar, “Language modeling for automatic Turkish broadcast news transcription,” in *Proc. INTERSPEECH-EUROSPEECH 2007*, Antwerp, Belgium, 2007.

[15] H. Soltau, G. Saon, D. Povey, L. Mangu, J. Kuo, M. Omar, and G. Zweig, “The IBM 2006 GALE Arabic ASR system,” in *Proc. ICASSP 2007*, Honolulu, HI, USA, 2007.

[16] A. Stolcke, “Srilm – An extensible language modeling toolkit,” in *Proc. ICSLP 2002*, vol. 2, Denver, 2002, pp. 901–904.

[17] M. J. Collins, “Discriminative training methods for Hidden Markov Models: Theory and experiments with perceptron algorithms,” in *Proc. EMNLP 2002*, 2002, pp. 1–8.

[18] E. Charniak and M. Johnson, “Coarse-to-fine n-best parsing and maxent discriminative reranking,” in *Proc. ACL 2005*, 2005, pp. 173–180.

[19] K. Oflazer, “Two-level description of Turkish morphology,” *Literary and Linguistic Computing*, vol. 9, pp. 137–148, 1994.

[20] H. Sak, T. Gungor, and M. Saraclar, “Morphological disambiguation of Turkish text with perceptron algorithm,” in *Proc. CICLing 2007*, Mexico City, Mexico, 2007.

[21] G. Eryiğit and K. Oflazer, “Statistical dependency parsing of Turkish,” in *Proc. EACL 2006*, Trento, 3-7 April 2006, pp. 89–96.

[22] I. Shafran and W. Byrne, “Task-specific minimum bayes-risk decoding using learned edit distance,” in *Proc. of INTERSPEECH*, 2004, pp. 1945–48.