Discriminative Language Modeling With Linguistic and Statistically Derived Features

Ebru Arısoy, *Member, IEEE*, Murat Saraçlar, *Member, IEEE*, Brian Roark, *Senior Member, IEEE*, and Izhak Shafran, *Member, IEEE*

***Abstract*—This paper focuses on integrating linguistically motivated and statistically derived information into language modeling. We use discriminative language models (DLMs) as a complementary approach to the conventional n-gram language models to benefit from discriminatively trained parameter estimates for overlapping features. In our DLM approach, relevant information is encoded as features. Feature weights are discriminatively trained using training examples and used to re-rank the N-best hypotheses of the baseline automatic speech recognition (ASR) system. In addition to presenting a more complete picture of previously proposed feature sets that extract implicit information available at lexical and sub-lexical levels using both linguistic and statistical approaches, this paper attempts to incorporate semantic information in the form of topic sensitive features. We explore linguistic features to incorporate complex morphological and syntactic language characteristics of Turkish, an agglutinative language with rich morphology, into language modeling. We also apply DLMs to our sub-lexical-based ASR system where the vocabulary is composed of sub-lexical units. Obtaining implicit linguistic information from sub-lexical hypotheses is not as straightforward as word hypotheses, so we use statistical methods to derive useful information from sub-lexical units. DLMs with linguistic and statistical features yield significant, 0.8%–1.1% absolute, improvements over our baseline word-based and sub-word-based ASR systems. The explored features can be easily extended to DLM for other languages.**

***Index Terms*—Discriminative training, language modeling, morphologically rich languages, speech recognition.**

I. Introduction

A Statistical language model assigns a probability distribution over all possible word strings in a language. The most common language modeling approach in state-of-the-art ASR technology is n-gram modeling, due to model simplicity and strong recognition performance. In n-gram models, the probability of a word string $P(W)$ is approximated with the product of conditional word probabilities, conditioned on $n-1$ previous words. These probabilities are estimated using a large text corpus with maximum likelihood estimation (MLE).

Discriminative training of language models has been recently introduced to obtain improved parameter estimates for language models (see Section II). The advantage of discriminative parameter estimation to MLE is that discriminative training takes negative examples into account as well as the positive examples and therefore results in a better discrimination between alternative classes. Positive examples are the correct transcriptions and negative examples are erroneous candidate transcriptions. DLM parameters are trained by optimizing an objective function that is directly related to the system performance, word error rate (WER) in ASR systems. Estimation of the DLM parameters is relatively straightforward for linear models. In addition to the improved parameter estimates with discriminative training, another advantage of this DLM approach to the conventional n-gram language model is that it is a featured-based approach. Consequently, it allows for easy integration of many relevant knowledge sources, such as morphology, syntax and semantics, into language modeling.

In this paper, we explore DLMs in the context of Turkish Large Vocabulary Continuous Speech Recognition (LVCSR). Turkish is a particularly good test case for this sort of approach. As an agglutinative language with rich morphology, straightforwardly applied n-gram models suffer from very serious out-of-vocabulary (OOV) rates (about 8% for 50 K, 2% for 200 K, and 1% for 500 K vocabularies). Sub-lexical-based ASR systems, where the vocabulary is composed of sub-lexical units, have been proposed for agglutinative languages, e.g., Turkish, Finnish, Estonian, Hungarian to handle the OOV problem caused by moderate size (50 K) word vocabularies and non-robust language model estimates caused by huge (500 K) word vocabularies. The state-of-the-art Turkish LVCSR is based on automatically obtained sub-lexical units, called "morphs." Having a rich morphology makes Turkish interesting for DLM research since implicit morphological constraints seem likely to be useful as DLM features. However, deriving and exploiting

Manuscript received January 31, 2011; revised May 31, 2011; accepted June 26, 2011. Date of publication July 18, 2011; date of current version nulldate. This work was supported in part by the Scientific and Technological Research Council of Turkey (TUBITAK) under Project 105E102, Boğaziçi University Research Fund (BAP) under Project 07HA201D, and the Turkish State Planning Organization under Project DPT2007K120610. The work of M. Saraçlar was supported by the TUBA GEBIP award. This work was also supported in part by the National Science Foundation (NSF) under Grant IIS-0447214. Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the NSF. Parts of this study were presented in conferences [1] and [2]. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Gokhan Tur.

E. Arısoy was with Boğaziçi University, Bebek, 34342, Istanbul, Turkey. She is now with the Speech Technologies for Media and ACCES Department, IBM T. J. Watson Research Center, Yorktown Heights, NY 10598 USA (e-mail: earisoy@us.ibm.com).

M. Saraçlar is with the Electrical and Electronics Engineering Department, Boğaziçi University, Bebek, 34342, Istanbul, Turkey (e-mail: murat.saraclar@boun.edu.tr).

B. Roark and I. Shafran are with the Center for Spoken Language Understanding, Oregon Health and Science University, Beaverton, OR 97006 USA (e-mail: roark@cslu.ogi.edu; zak@cslu.ogi.edu).

Digital Object Identifier 10.1109/TASL.2011.2162323

such information from sequences of sub-word units, automatically learned from the data in an unsupervised manner, is a challenge.

In our approach, DLMs are applied as a complementary approach to the generative n-gram language models, in either the word-based or morph-based systems. Following prior approaches, we first explore basic n-grams of both words and morphs as DLM features. We then explore novel feature sets that extract implicit information available at lexical and sub-lexical levels using linguistic and statistical approaches. Linguistic features are investigated in the word-based ASR system to integrate complex morphological and syntactic language characteristics of Turkish into language modeling. The linguistic information is extracted from the word hypothesis sentences with the help of morphological and dependency parsers. Obtaining linguistic information from morph hypotheses, however, is not as straightforward as word hypotheses, since the linguistic tools can not be directly applied to sub-word units. Therefore we use statistical approaches to derive useful implicit information from morph hypotheses. DLMs with these linguistic and statistical features yield significant improvements over our baseline word-based and morph-based ASR systems.

The paper is organized as follows. Previous related work is presented in Section II. We explain a general framework for DLMs in Section III. We present novel linguistically and statistically motivated feature sets in Section IV and give the experimental results in Section V. Discussion and conclusions are given in Section VI

II. Previous Work

N-gram language models utilize word probabilities conditioned on only the $n-1$ previous words in calculating the probability of the word strings. Language models have been proposed that try to include deeper language characteristics than these local collocations, e.g., morphology, syntax, and semantics. A way of integrating relevant information into language models is to represent words or sentences as a bundle of arbitrary features and to build language models based on these features either with generative or discriminative approaches.

Factored language models [3] generalize word n-grams by representing words as a set of factors (features). Conditional exponential language models, i.e., maximum entropy (ME) language models [4], utilize features, which are functions of the predicted word and its history, to estimate word n-gram probabilities. Within this ME paradigm, various feature sets were explored, including: trigger features, to adapt the model expectations to the topic of discourse [4]; semantic dependencies and syntactic structure for ASR [5]; and semantic analysis for spoken dialogue systems [6]. A joint morphological-lexical language model which is also based on ME modeling was proposed in [7] for morphologically complex Arabic language. The conditional structure in ME language models is exchanged for a global, sentence-level structure using whole sentence exponential language models [8].

DLMs have been demonstrated to consistently outperform generative modeling approaches, partly due to improved parameter estimates with discriminative training and partly due to the ease with which many overlapping features can be included into the same model. Linear models have been successfully applied to discriminative language modeling for speech recognition [9]–[12] and utterance classification [13]. In DLMs based on linear models, model parameters are used to define a cost, $F(W, A)$, on the word sequence which also includes the likelihood from the baseline recognizer, $P(W, A)$:

$$F(W,A) = \alpha_0 \log P(W,A) + \sum_i \alpha_i \Phi_i(W) \tag{1}$$

Here $\Phi_i(W)$ represents sentence level features and α_i's are the parameters associated with the features. α_i's can be learned discriminatively using the perceptron algorithm [9], [14] or using methods based on maximizing the conditional log-likelihood [9] or minimizing a discriminative loss function [12]. A comparison of different loss functions for DLMs is given in [15]. Within the DLM paradigm, utilizing only the word n-grams as features outperformed generative word n-grams [9], and incorporating syntactic [11] and morphological [16] features yielded additional improvements. Trigger-based features were also incorporated in order to use the conversation context as an additional information source [17]. The discriminative model has also been extended beyond the language model to jointly estimate the acoustic state transitions, duration model and language model, and thus learn the weights of the finite-state transducer for a speech recognizer [18]. These methods are similar to approaches for reranking in the syntactic parsing literature [19], [20], where instead of speech inputs being paired with word string outputs (as in this paper), the problem has word string inputs paired with syntactic parse tree outputs.

In addition to discriminative linear models, adjusting the parameters of n-gram language models trained with MLE was proposed to achieve minimum sentence error for discriminative training of language models [21], [22]. Minimum classification error framework was also used to discriminatively train language and semantic classifier models for spoken utterance classification task [23].

In this paper, we apply DLMs based on linear models to Turkish which is an agglutinative language with rich morphology. Agglutinative languages suffer from a large number of OOV words in LVCSR. This problem has been mostly handled by using meaningful segmentations of words, called sub-lexical units, in n-gram language modeling. In sub-lexical language modeling, the recognition vocabulary is composed of sub-lexical units instead of words. Sub-words in the vocabulary are capable of covering most of the words of a language, thus addressing the OOV problem and leading to a decrease in WER. Consequently, various sub-lexical units obtained with grammatical or unsupervised data-driven approaches have been proposed for these languages. Previous studies for ASR in Estonian [24], Finnish [25], Hungarian [26], and Turkish [27] show that sub-lexical units improve the ASR accuracy over words. A comparison among morphologically rich languages Estonian, Finnish and Turkish, as well as Arabic, is given in [28]. Previous work on Turkish language modeling focused on sub-lexical n-gram language models. Grammatically derived sub-lexical units, e.g., morphemes [29], inflectional groups [30], stem+endings [31] and their combinations [32], as well as automatically derived sub-lexical units, called "morphs" [27],

[33], have been used for Turkish ASR. Morphs are learned from the data in an unsupervised manner by the Morfessor algorithm [34] which uses the minimum description length principle, and they have been shown to outperform words and grammatically derived units in Turkish LVCSR [27]. Therefore, morphs are selected as the sub-lexical units in this paper. The most comprehensive sub-lexical language modeling experiments on Turkish LVCSR were reported in [27], [35].

III. Discriminative Language Modeling

This section describes a general framework for DLMs based on linear models. We will follow the definitions and notations given in [9]. The main components of DLMs are as follows:

1) **Training Examples**: These are the input:output pairs (x_i, y_i) for $i = 1 \dots N$. Inputs, $x \in \mathcal{X}$, are the utterances and outputs, $y \in \mathcal{Y}$, are the corresponding reference transcriptions. Here $\mathcal{X}$ is the set of all possible inputs and $\mathcal{Y}$ is the set of all possible outputs.
2) $\boldsymbol{GEN(x)}$: This function enumerates a finite set of candidates for the inputs, x, where $GEN(x) \subseteq \mathcal{Y}$. $GEN(x)$ function can be the lattice or the N-best list output of the baseline ASR system for the utterance x.
3) $\boldsymbol{\Phi(x,y)}$: A d-dimensional real-valued feature vector $(\Phi(x,y) \in \Re^d)$. The representation $\boldsymbol{\Phi}$ defines the mapping from the (x_i, y_i) pair to the feature vector $\Phi(x_i, y_i)$.
4) $\boldsymbol{\bar{\alpha}}$: A vector of discriminatively learned feature parameters $(\bar{\alpha} \in \Re^d)$

A. Training Data Generation

Like many other supervised learning approaches, DLM requires labeled input:output pairs as the training examples. Utterances with the reference transcriptions, $(x_1, y_1) \dots (x_N, y_N)$, are used in DLM training. The utterances are decoded with the baseline ASR system to obtain the candidate outputs for the $GEN(x)$ function. Since speech data with transcriptions are limited compared to the text data, DLM training data is generated by breaking the acoustic 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 manner since baseline acoustic model training is more expensive and less prone to over-training than n-gram language model training [9].

B. Feature Representations

Discriminative language modeling is a feature-based sequence modeling approach. Therefore, each (x_i, y_i) pair is mapped to a d-dimensional real-valued feature vector $\Phi(x_i, y_i)$. This feature vector is defined as a function of the acoustic input, x, and the candidate hypothesis, y. Each element of the feature vector, $\Phi_0(x_i, y_i) \dots \Phi_{d-1}(x_i, y_i)$, corresponds to a different feature. The zeroth element of the feature vector, $\Phi_0(x,y)$,[1] is the contribution of the baseline ASR system to DLM and defined as the "log-probability of y in the lattice produced by the baseline recognizer for utterance x." The basic approach for the other DLM features is to use word n-grams in defining features. The word n-gram features are defined as the number of times a particular n-gram is seen in the candidate hypothesis.

```
Inputs: Training examples $(x_i, y_i)$ for $i = 1 \dots N$
Initilization: $\bar{\alpha}_0^N = (\alpha_0, 0, \dots, 0)$
Algorithm:
  For $t = 1 \dots T$
    $\bar{\alpha}_t^0 = \bar{\alpha}_{t-1}^N$
    For $i = 1 \dots N$
      $z_i = \underset{z \in GEN(x_i)}{\operatorname{argmax}} \langle \Phi(x_i, z),\ \bar{\alpha}_t^{i-1} \rangle$
      $\bar{\alpha}_t^i = \bar{\alpha}_t^{i-1} + \Phi(x_i, y_i) - \Phi(x_i, z_i)$
Output: Averaged parameters $\bar{\alpha}_{AVG} = \sum_{i,t} \bar{\alpha}_t^i / NT$
```

Fig. 1. Variant of the perceptron algorithm [9]. $\bar{\alpha}_t^i$ represents the feature parameters after the tth pass on the ith example.

C. Parameter Estimation

Each DLM feature has an associated parameter, i.e., α_i for $\Phi_i(x,y)$. In this paper, a variant of the perceptron algorithm is used for parameter estimation (shown in Fig. 1). This algorithm is the same as the one given in [9]. The main idea in this algorithm is to penalize features associated with the current 1-best hypothesis, and to reward features associated with the gold-standard hypothesis (reference or lowest-WER hypothesis). It has been found that the perceptron model trained with the reference transcription as the gold-standard hypothesis is much more sensitive to the value of the α_0 constant [9]. Therefore, we use the lowest-WER hypothesis (oracle) as the gold-standard hypothesis. Averaged parameters, $\bar{\alpha}_{AVG}$, are used in decoding held-out and test sets, since averaged parameters have been shown to outperform regular perceptron parameters in tagging tasks [14].

IV. DLM Features

This section describes the feature sets used in Turkish DLMs. Features are extracted from the N-best list output of the training utterances decoded by the word-based and morph-based ASR systems. We investigate word n-gram features on word hypotheses and morph n-gram features on morph hypotheses as the basic features. Morphological and syntactic information is extracted from word hypothesis sentences using linguistic tools. Implicit morpho-syntactic information is explored on morph hypothesis sentences with statistical approaches and it is used to define statistically motivated features. We also explore statistical approaches on word hypotheses to derive the topic of the utterances and topic sensitive features are incorporated into DLMs. Beginning of sentence symbol, "$\langle \texttt{s} \rangle$," and end of sentence symbol, "$\langle / \texttt{s} \rangle$," are added to the hypothesis sentences before feature extraction.

A. Basic n-Gram Features

We use n-grams as the basic DLM features as used in [9], [10]. n-gram features are defined as the number of times an n-gram is seen in the candidate hypothesis. Word unigram and bigram feature templates are given in Table I where w_i represents the ith word in the hypothesis sentence. A Turkish phrase with English glosses is given in Fig. 2. An example word bigram feature is as follows:

$$\Phi_i(x,y) = \text{number of times ``}\langle \texttt{s} \rangle\ \texttt{hizmetleri}\text{'' is seen in } y.$$

[1] $\Phi_0(x,y) = \log P_l(y) + (1/\beta) \log P_a(x|y)$ where $\log P_l(y)$ and $\log P_a(x|y)$ are the baseline language model and acoustic model scores respectively and β is the language model weight.

TABLE I
FEATURE SETS USED IN THE EXPERIMENTS

Descriptions	Notations	Feature Templates	Examples of the Feature Templates
Basic word and sub-lexical features			
Word unigrams, bigrams	W(1), W(2)	$(w_i), (w_{i-1}w_i)$	(hizmetleri), (hizmetleri de)
Morph unigrams, bigrams	M(1), M(2)	$(m_i), (m_{i-1}m_i)$	(hizmet), (hizmet +leri)
Morphological features			
Root unigrams, bigrams	R(1), R(2)	$(r_i), (r_{i-1}r_i)$	(hizmet), (hizmet de)
Stem+ending unigrams, bigrams	SE(1), SE(2)	$(s_i), (e_i), (s_ie_i), (e_{i-1}s_i)$	(+leri), (+leri de)
IG unigrams, pairs	IG(1), IG(2)	$(IG_{i,j}), (IG_{i-1,last}IG_{i,j})$	([Noun+A3pl+P3sg+Nom]), ([Noun+A3pl+P3sg+Nom][Conj])
Syntactic features			
PoS unigrams, bigrams	PoS(1), PoS(2)	$(t_i), (t_{i-1}t_i)$	([Noun]), ([Noun] [Conj])
H2H between words	H2H(ww)	$(DR(w_i, w_k)w_iw_k)$	(INTENSIFIER de hizmetleri)
H2H between PoS tags and words	H2H(tw)	$(DR(t_i, w_k)t_iw_k)$	(INTENSIFIER [Conj] hizmetleri)
H2H between words and PoS tags	H2H(wt)	$(DR(w_i, t_k)w_it_k)$	(INTENSIFIER de [Noun])
H2H between tags	H2H(tt)	$(DR(t_i, t_k)t_it_k)$	(INTENSIFIER [Conj] [Noun])
Sub-lexical (morpho-syntactic) features			
Cluster unigrams, bigrams	C(1), C(2)	$(c_i), (c_{i-1}c_i)$	((CLUSTER-I), (CLUSTER-I CLUSTER-5)
Long distance triggers	$M_{LD}(2)$	$(m_{i,k} \rightarrow m_{j,l})$ where $j > i$	(hizmet → +ılacak)
Topic sensitive features			
Topic unigrams, bigrams	T(1), T(2)	$(w_i : T(W)), (w_{i-1}w_i : T(W))$	(hizmetleri:TOPIC-I), (hizmetleri de:TOPIC-I)

Turkish phrase: <s> hizmetleri de arttırılacak </s> *(services will also be increased)*

Turkish phrase as a morph sequence: <s> hizmet +leri de arttır +ılacak </s>

Morphological analyses of words in phrase:
hizmetleri: hizmet+Noun+A3pl+P3sg+Nom
de: de+Conj
arttırılacak: art+Verb+ ˆDB +Verb+Caus+ ˆDB +Verb+Pass+Pos+Fut+A3sg

Roots+Endings in words:
hizmetleri: hizmet*(service)* +leri
de: de*(also)*
arttırılacak: art*(increase)* +tırılacak

Roots+Inflectional Groups (IGs) in words:
hizmetleri: hizmet*(service)* +Noun+A3pl+P3sg+Nom
de: de*(also)* +Conj
arttırılacak: art*(increase)* +Verb +Verb+Caus +Verb+Pass+Pos+Fut+A3sg

Fig. 2. Turkish phrase with morphological analysis. English glosses are given in parentheses. Endings and IGs are the groupings of the morphological tags in surface and lexical forms, respectively.

Morph features are extracted from the N-best hypotheses obtained from the morph-based ASR system. In morph-based ASR, the recognition vocabulary is composed the automatically obtained sub-lexical units called morphs [34] and the generative language model is built with morph n-grams. Therefore, the recognition output is the most likely morph sequence corresponding to the input utterance. Note that concatenation of morphs in ASR transcriptions can generate any sequence including non-word items and correcting the ungrammatical items has been shown to improve the recognition accuracy [36]. We do not deal with this problem in this paper and to facilitate conversion of morph sequences to word sequences after decoding, a marker, "+", is attached to the non-initial morphs. Fig. 2 shows an example morph phrase. Morph n-gram features are defined in the same way as the word n-gram features. Morph unigram and bigram feature templates are given in Table I where m_i represents the ith morph in the morph hypothesis sentence.

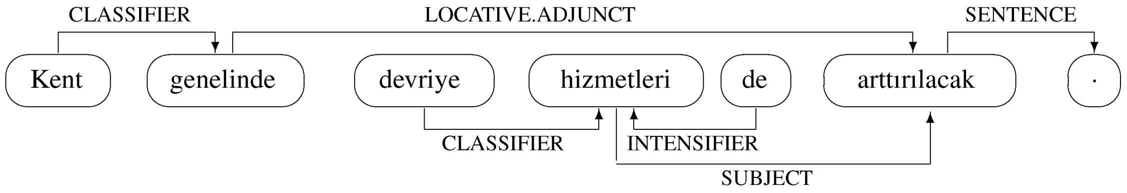
B. Linguistically Motivated Features

This section focuses on extracting linguistic information from word hypothesis sentences and utilizing this information as DLM features.

1) Morphological Features: Morphology is an important information source for feature-based language models, especially for morphologically rich languages [7], [16]. While introducing the morphological features, we aim to turn the challenging morphological structure of Turkish into a useful information source in DLMs.

Morphological features will be explained using the example Turkish word phrase and the corresponding morphological analysis given in Fig. 2. The decompositions are given in terms of the roots followed by morphological tag sequences. Morphological tag sequences are separated by the $\hat{}DB$ symbol which denotes the derivation boundaries. The morphological analysis is obtained with Oflazer's morphological analyzer [37]. The complex morphological structure of Turkish can lead to several different morphological analyses for the same word. We found that in a vocabulary of 1.6 M morphologically decomposed words, each word can have on average 2.4 different morphological analyses. The maximum number of morphological analyses per word is 68. Therefore, the ambiguity in the words with multiple morphological analyses was resolved using Sak *et al.*'s perceptron-based morphological disambiguation tool for Turkish [38].

We define the morphological features using roots, stem+endings and inflectional groups (IGs). The morphological tag sequences separated by derivation boundaries are called the IGs. All morphological features are defined in a similar manner as the basic n-gram features.



WORD:	Kent	genelinde	devriye	hizmetleri	de	arttırılacak	.
GLOSS:	The city	throughout	patrol	services	also	will be increased	.
PoS:	[Noun]	[Noun]	[Noun]	[Noun]	[Conj]	[Verb]	[Punc]

Fig. 3. Example dependency analysis for syntactic.

In **root n-gram features**, first the words in the hypothesis sentences are represented only with their roots using the morphological decompositions. Then the n-gram features are generated in the same way as words, as if roots are words. The root unigram and bigram feature templates are given in Table I where r_i represents the root of the ith word.

In **stem+ending n-gram features**, the stem is extracted from the morphological decomposition and the remaining part of the word is taken as the ending. If there is no ending in the word, a special symbol is inserted to represent the empty ending. Hypothesis sentences are converted to stem and ending sequences and the n-gram features are generated in the same way as words, as if stems and endings are words. The stem+ending unigram and bigram feature templates are given in Table I where s_i and e_i represent the stem and the ending of the ith word, respectively.

IG-based features are commonly used to handle the challenges introduced by the agglutinative nature in dependency parsing and morphological disambiguation tasks of Turkish [38]–[40]. In previous studies, it has been shown that the dependency relations between words are determined by the relations between the last IG of a word and any IGs of the word on the right [40]. Moreover, IG-based features in a discriminative framework gives the best accuracy for the Turkish morphological disambiguation and PoS tagging tasks [38]. Therefore, IGs are also used in defining the morphological features in this research. According to the IG definition a word can be represented as a root and a sequence of IGs, i.e., $r_i + IG_{i,1} + \hat{}DB + \ldots + IG_{i,j} + \hat{}DB + \ldots + IG_{i,n_i}$. Here, r_i, $IG_{i,j}$ and IG_{i,n_i} represent the root, jth IG and the last IG of the ith word, respectively. Note that the morphological tags following the root of the word until the derivation boundary constitute the first IG of this word. The words in the hypothesis sentences are represented as a sequence of IGs and the IG-based n-gram features are extracted from these sequences. IG-based unigram features are extracted by treating IGs as words. IG-based bigram features are obtained by using only the IG pairs between the last IG of the previous word and all the IGs of the current word. The main motivation in using these pairs as the bigram features is to decrease the number of features using the knowledge that $(IG_{i-1,n_{i-1}} IG_{i,j})$ pairs are more likely to contain linguistic information, i.e., potential dependency relations between words. IG-based unigram and bigram feature templates are given in Table I.

2) Syntactic Features: Syntax is an important information source for language modeling due to its role in sentence formation. Syntactic information have been incorporated into conventional generative language models to capture the long distance dependencies [41], [42]. Syntactic information has also been used in feature-based reranking approaches [5], [8], [11]. The success of these approaches lead us to investigate syntactic features for Turkish DLMs.

The syntactic features will be explained with the dependency analysis given in Fig. 3. Eryiğit *et al.*'s dependency parser [40] is used for the analysis. This is a classifier-based deterministic parser utilizing IGs as the parsing units to find the dependency links between words. The incoming and outgoing arrows in the figure show the dependency relations between the head and the dependent words with the type of the dependency. The words with English glosses, PoS tags associated with the words are also given in the example.

For the syntactic DLM features, we explore similar feature definitions as [11]. Features are extracted from the dependency analysis of the hypothesis sentences. Here, it is important to note that hypothesis sentences contain recognition errors and the parser generates the best possible dependency relation even for incorrect hypotheses. We use PoS tag n-grams and head-to-head (H2H) dependency relations as the syntactic features. PoS tag features are used in an effort to obtain class-based generalizations that may capture well-formedness tendencies. PoS tag unigram and bigram features are extracted in a manner similar to that of words but treating PoS tags as words. PoS tag feature templates are given in Table I where t_i represents the PoS tag of the ith word. H2H dependency relations are used since presence of a word or morpheme can depend on the presence of another word or morpheme in the same sentence and this information is represented in the dependency relations. H2H features are defined using the dependency relations and lexical items or their PoS tags. The examples for the H2H feature templates are given in Table I. The dependency relations between ith and kth words are denoted by $DR(w_i, w_k)$ where $DR(\cdot,\cdot)$ represents the type of the dependency relation between the input pairs. For instance in the feature notations given in Table I, H2H(tw) corresponds to the feature defined in terms of the dependency relation between two words, $DR(w_i, w_k)$, and the PoS tag of w_i, denoted by t_i, and word w_k, i.e., (INTENSIFIER [Conj] hizmetleri) for the phrase "hizmetleri de," from Fig. 3.

C. Statistically Derived Features

The advantage of the statistical morphs compared to their grammatical counterparts is that they do not require linguistic knowledge for segmenting words into sub-lexical units. As a result morphs do not convey explicit linguistic information like grammatical morphemes and obtaining linguistic information from morph sequences is not obvious. In Section IV-B2, syntactic generalizations are considered with PoS tag and H2H features. Since this information is not directly acquired from morphs, we focus on exploring representative features of morpho- syntactic information using data driven approaches. Since Turkish is a prefix-free language, we make an analogy between initial morphs and roots as well as between non-initial morphs and suffixes in exploring morpho-syntactic features.

1) Clustering of Sub-Lexical Units: The feature set proposed in this section aims to obtain syntactic information, similar to PoS tags, directly from the morphs in the hypothesis sentences. In order to obtain categories for morphs, like PoS tags of words, morphs in the training corpora are automatically classified with clustering algorithms. The category associated with a particular morph is considered as the tag of that morph and utilized in defining the morpho-syntactic features.

We apply two hierarchical clustering approaches. The first one is Brown *et al.*'s algorithm [43] and the second one uses minimum edit distance (MED) between strings as the similarity function in bottom-up clustering.

Brown *et al.*'s algorithm was proposed for class-based n-gram language models. This algorithm is based on maximizing the average mutual information of adjacent classes (see [43] for details). Utilizing n-gram features in DLMs makes this clustering an attractive approach for our research.

The motivation in the second clustering approach is to capture the syntactic similarity of morphs using their graphemic similarities, since a non-initial morph can contain a grammatical morpheme, a group of grammatical morphemes or pieces of grammatical morphemes. Recall that we treat initial morphs and roots in a similar manner. Therefore, this clustering is only meaningful for non-initial morphs since graphemic similarity of initial morphs does not reveal any linguistic information. So, we only cluster the non-initial morphs. As for the initial morphs, we try both assigning all of them into the same cluster and clustering initial morphs with Brown *et al.*'s algorithm.

We define $d(s, s')$ as the MED between two non-initial morphs s and s'. To make the distance independent of the length of the non-initial morphs, we normalize the MED as $(|s| + |s'|/|s|*|s'|)d(s, s')$ where $|s|$ is the length of s. In defining the similarity of the clusters, we prefer complete-link similarity to obtain tight clusters. We choose the non-initial morph pair yielding the highest distance for defining the similarity of the clusters because this pair gives the most dissimilar members of the clusters for the complete-link similarity function. The most similar clusters are merged in every iteration of the bottom-up clustering. As a stopping criterion, we impose an empirically determined threshold on the similarity of the clusters in order to alleviate merging of clusters to generate a huge cluster containing most of the non-initial morphs.

In Turkish, the same lexical form morpheme can correspond to different surface form morphemes due to vowel harmony, consonant harmony, and consonant deletion/insertion rules. For instance, the lexical plural suffix "lAr"[2] can correspond to two different surface form suffixes, i.e., "ler" and "lar," due to the vowel harmony. Therefore, considering the phonological variations in surface and lexical form morphemes in clustering will help grouping morphs with similar syntactic functions into the same cluster. In our application, we modify MED to softly penalize the phonological variations in the surface and lexical forms of the same suffix. For example, the substitution operation for the vowel-vowel and consonant-consonant pairs that are more likely to occur due to the vowel harmony and consonant harmony rules and the deletion and insertion operations for the consonants that are more likely to occur during suffixation are considered to introduce a cost of 0.5 instead of 1. Note that the cost of all the operations are 1 in the MED calculation.

After clustering morphs with one of the hierarchical clustering algorithms, the cluster of each morph is taken as the tag of that morph. The n-gram morph cluster features are extracted on the morph tag sequences similar to PoS tag features of words. The feature templates are given in Table I where c_i represents the cluster of the ith morph, m_i.

2) Long Distance Triggers: In addition to morph clusters, we also propose long distance triggers as morpho-syntactic features for DLMs with sub-lexical units. These features are motivated by the H2H dependency features in words. Considering initial morphs as stems and non-initial morphs as suffixes, we assume that the existence of a morph can trigger another morph in the same sentence. The morphs in trigger pairs are believed to co-occur for a syntactic function, like the syntactic dependencies of words, and these pairs are used to define the long distance morph trigger features.

Long distance morph trigger features are similar to the word trigger features proposed in [4], [17]. Previously proposed approaches use the document history to find the word trigger features. In our research we only consider sentence level trigger pairs to capture the morpho-syntactic level dependencies instead of discourse level information. We extract all the morph pairs between the morphs of any two words in a sentence as the candidate morph triggers. The candidate morph trigger pairs are extracted from the hypothesis sentences (1-best and oracle) to obtain also the negative examples for DLMs. Among the possible candidates, we try to select only the pairs where morphs are occurring together for a special function. This is formulated with hypothesis testing where null hypothesis (H_0) represents the independence and the alternative hypothesis (H_1) represents the dependence assumptions of morphs in the pairs [44]. The pairs with higher likelihood ratios $(\log(L(H_1)/L(H_0)))$ are assumed to be the morph triggers and used as features. The feature templates are given in Table I where $(m_i \rightarrow m_j)$ represents the trigger pair.

D. Topic Sensitive Features

The main motivation in topic sensitive features is to incorporate semantic information into language modeling. We apply statistical approaches also to word hypotheses to automatically assign topics to the utterances. Here we extend the work given in [5] for ME language models to DLMs.

We only use the reference transcriptions of acoustic model training data for topic clustering. This data contains BN recordings collected from 375 different shows. First we apply the TextTiling algorithm [45] on the reference transcriptions of the consecutive utterances of each show in order to split the entire data

[2] "A" is the lexical symbol realized as /a/ or /e/ in surface form.

into multi paragraphs. In the TextTiling algorithm, the text data is split into fixed length blocks, pseudo sentences, and lexical similarity scores are assigned to adjacent blocks. Lexical similarity scores between adjacent blocks are calculated by the normalized inner product of the word count vectors of these blocks. Thus more common words in adjacent blocks result in higher similarity scores. Then paragraph boundaries are determined based on the lexical similarity scores. See [45] for details of the boundary identification.

In this paper, we use the TextTiling algorithm implementation in Natural Language Toolkit (NLTK).[3] TextTiling is performed over the stems in order to increase the common tokens in pseudo sentences. We only keep nouns and verbs and remove everything else in the sentences before TextTiling since these constituents are considered to carry more useful information than the other constituents for topic clustering. After splitting the reference transcriptions into multi paragraphs, each paragraph is represented as a $tf.idf$ vector using stems as the terms. The paragraphs are clustered into topics using the hierarchical bottom-up clustering algorithm explained in Section IV-C. We apply average link similarity function in clustering and the similarity of the paragraphs are defined with the cosine distance of their $tf.idf$ vectors.

After assigning topic IDs to paragraphs, the training utterances corresponding to the reference transcriptions in the paragraphs get the same topic IDs with the paragraphs. Topic sensitive features are defined using the associated topic IDs of the utterances together with the word n-gram features. Example feature templates are given in Table I where $T(W)$ represents the topic ID of the utterance W.

An important issue arises in assigning topics to the held-out and the test set utterances where the correct transcriptions are not available. Here we use the ASR hypotheses instead of the reference transcriptions in topic assignments. We apply two different scenarios to find the most likely topic of a given utterance. In the first scenario, an utterance is assigned to the most similar topic of its best ASR hypothesis. In the second scenario, TextTiling is performed over the best ASR hypotheses of consecutive utterances and utterances in each paragraph are assigned to most similar topic of the paragraph. The second scenario is the same approach with assigning topics to training utterances except that best ASR hypotheses are used instead of reference transcriptions. Again the similarity is defined with the cosine similarity of the $tf.idf$ vectors of the best ASR hypotheses or $tf.idf$ vectors of the paragraphs containing the best ASR hypotheses with the $tf.idf$ vectors of the documents of the topics.

V. Experiments

This section explains the experimental setups for the baseline ASR systems and DLMs and gives the experimental results for words and sub-lexical units.

A. ASR System Description for Turkish

In our research, we use word-based and morph-based LVCSR systems developed for automatic transcription of Turkish Broadcast News (BN). The acoustic model of the BN transcription system is built with AT&T tools[4] using the Turkish BN corpus containing 184 hours of BN recordings. Acoustic models are speaker-independent and they do not utilize discriminative training and speaker adaptation. Separate held-out (3.1 hours) and test (3.3 hours) data are used to evaluate the system performance. More detailed descriptions of the corpora and the acoustic models are given in [27].

The Turkish web corpus [46] (182.3 M words) collected from major news portals and the reference transcriptions of the acoustic model training data (1.3 M words) were used in building generic and in-domain language models respectively for the BN transcription system. Language models with interpolated Kneser–Ney smoothing as well as entropy-based pruning were built using SRILM toolkit [47]. Generic and in-domain language models were linearly interpolated in order to reduce the effect of out-of-domain data. The interpolation constant was chosen to minimize the heldout set perplexity. We used a 200 K word vocabulary, resulting in a 2% OOV rate, for the word-based system and a 76 K morph vocabulary, resulting in full coverage,[5] for the morph-based system. The word vocabulary was chosen as the most frequent 200 K words and morph vocabulary contains all the morphs in the generic and in-domain text data. The best results were obtained with 3-gram word and 4-gram morph language models. In the morph-based language models, all the words in the text corpus were split into morphs and generative n-gram language models were trained as if the morphs were words. Each word in the text data was segmented on average into 1.4 morphs.

We applied linguistic tools on the ASR transcriptions for DLM feature extraction, and being provided with transcriptions of complete sentences is important for the morphological disambiguation tool and especially important for the dependency parser. Therefore, we converted segments manually labeled by the annotators to complete sentences. Since some sentences can be quite long for Turkish, we generated lattices based on manually labeled segments with the pruned language models in the first pass recognition. Then the lattices belonging to the same sentence were concatenated and re-scored with unpruned language models. If the manually labeled segments contained utterances belonging to more than one sentence, we split these utterances into smaller chunks from the sentence boundaries before first-pass recognition. Sentence boundaries were obtained from the reference transcriptions.

B. Experimental Setup for DLMs

DLM training data for words and morphs were generated by decoding the acoustic model training data (184 hours) with the word-based and morph-based ASR systems, respectively. Language model over-training was controlled via 12-fold cross validation. Utterances in each fold were decoded with the baseline acoustic model trained on all the utterances and the fold-specific language model. A fold-specific language model was generated by interpolating the generic language model with the in-domain

[3] http://www.nltk.org/.

[4] http://www.research.att.com/simfsmtools/{fsm,dcd}

[5] A word is considered as OOV if it cannot be generated by any combination of the morphs in the vocabulary.

language model built from the reference transcriptions of the utterances in the other 11 folds. We did not determine a new interpolation constant for each fold-specific language model. Instead we used the same interpolation constant utilized in the baseline language model. 50-best lists were generated from the lattices based on sentence-based segmentations. We used the scenario explained in Section IV-A to obtain sentence-based segmentations. We chose to work on 50-best hypotheses for DLMs since our previous research [2] showed that increasing the N-best size from 50 to 1000 does not give any significant difference on the DLM performance, an observation also noticed in other work on discriminative training as well [16].

DLM features were extracted from the 50-best hypotheses and feature parameters were trained with the perceptron algorithm given in Fig. 1. In the perceptron model the parameter for the baseline model, α_0 (the weight associated with $\Phi_0(x, y)$), was set to a fixed number and was not updated during training. We trained several models by changing the α_0 constant from 0 to 16. We decided on the α_0 constant and the number of iterations by looking at the held-out set WER. The best held-out set WER was obtained in 1–3 iterations of the perceptron algorithm. Averaged perceptron parameters were used in the evaluation of the held-out and test sets. 50-best word and morph hypotheses were used in discriminative reranking as well. 50-best lists for the held-out and test data for word and morph DLMs are generated by decoding the held-out and test utterances with the baseline word-based and morph-based ASR systems explained in Section V-A.

C. *Experimental Results*

In the experiments we investigated the combinations of different n-gram orders (unigrams, unigrams+bigrams, unigrams+bigrams+trigrams) for the proposed feature set. The n-gram combination yielding the lowest WER on the held-out set was chosen as the best combination. If different combinations yielded the same amount of improvements, then the one with the lowest number of features was chosen as the best combination. Features from other information sources were incorporated into the best n-gram combination of the current feature set. We performed many experiments with different feature combinations but we only report the best scoring ones.

The results of DLM experiments on word hypotheses are given in Table II. Here "Feats" represents the number of features extracted from the 50-best lists. The number of features with non-zero weights after the parameter training is denoted by "ActFeats." n-gram features up to trigrams are tried for words, roots and stem+endings. The best results are obtained with unigrams both in word and morphological feature sets. Word unigram features result in 0.4% improvement over the test set error (significant at $p = 0.009$ as measured by the NIST MAPSSWE test). Morphological feature sets, stem+endings and IGs, yield higher gains than word features. However, root unigram features seem to overtrain on the held-out data since the gain obtained on the held-out set is not preserved on the test set. The best performing morphological feature set, $\mathrm{R}(1) + \mathrm{IG}(1)$, reduces the test set error from 23.4% to 22.6% (significant at $p < 0.001$).

TABLE II
RESULTS ON 50-BEST WORD HYPOTHESES

	Feats ($\mathrm{x}10^3$)	ActFeats ($\mathrm{x}10^3$)	WER(%) held-out	Δ	WER(%) test	Δ
Baseline word-based ASR	-	-	24.1	-	23.4	-
W(1)	154.9	51.2	23.8	0.3	23.0	0.4
R(1)	34.3	12.9	23.4	0.7	23.0	0.4
SE(1)	51.0	21.7	23.5	0.6	22.6	0.8
R(1) + IG(1)	36.0	16.9	23.2	0.9	22.6	0.8
W(1)+PoS(1,2)	155.7	60.6	23.4	0.7	22.5	0.9
W(1)+$\mathrm{C_B}$(1,2)	155.7	70.5	23.4	0.7	22.8	0.6
W(1)+PoS(1,2)+H2H(tt)	157.5	66.7	23.4	0.7	22.6	0.8
W(1)+PoS(1,2)+H2H(all)	5783.6	1152.5	23.6	0.5	22.9	0.5
W(1)+T(1)	627.5	127.4	23.7	0.4	22.9	0.5

Then, we utilize syntactic features together with word features. PoS tag unigram and bigram features yield additive 0.5% (significant at $p < 0.001$) improvement on top of the gain obtained with word unigrams on the test set. Inspired by class-based features for morphs, we also investigated automatically induced word class features in word DLM experiments [denoted by $\mathrm{C_B}$ (1,2)]. We clustered words in the text corpora, web corpus and reference transcriptions of acoustic model training data, into 26 clusters using Brown *et al.*'s approach. The reason of choosing 26 automatically induced classes is to provide a fair comparison with PoS tags since we use 26 different PoS tags in our experiments. Even though the held-out set WERs are the same for PoS tag and class-based features, automatically induced class features yield 0.3% less improvement compared to PoS tag features on the test set. PoS tags may provide a more meaningful word classing with small number of clusters than automatic word clustering. Dependency features, H2H(all),[6] are incorporated into word and PoS tag features. However, the performance is degraded most probably due to data sparsity with 5.8 M features. In addition, only H2H(tt) features are incorporated into the same set to see the effect of PoS tag dependencies together with PoS tag n-grams. However, no improvement is achieved.

Finally, we perform experiments with topic sensitive features. We cluster the training data utterances into 40, 100, and 200 topics using the hierarchical clustering algorithm and assign these topics to held-out and test set utterances as explained in Section IV-D. Topic sensitive features obtained with 40 topics work slightly better than the features obtained with 100 and 200 topics. Additionally, assigning topics to held-out and test sets over the paragraphs, obtained with TextTiling, instead of sentences yield slightly better results. The best gains with topic sensitive features are obtained when topic unigram features are added to word unigram features. However, the improvement on top of word unigram features is small (0.1%).

The results of DLM experiments on morph hypotheses are given in Table III. As with the word unigram features, morph bigrams and trigrams do not introduce any gains over morph unigrams, moreover they degrade the performance of the unigram features. Morph unigram features yield 0.6% (significant at $p < 0.001$) improvement over the baseline.

Then, we incorporate automatically derived morpho-syntactic features into morph unigrams. Automatically induced classes are extracted from the morph corpora which contain web corpus and reference transcriptions of acoustic model

[6] $\mathrm{H2H(all)} = \mathrm{H2H(ww)} + \mathrm{H2H(wt)} + \mathrm{H2H(tw)} + \mathrm{H2H(tt)}$

TABLE III
RESULTS ON 50-BEST MORPH HYPOTHESES

	Feats ($x10^3$)	ActFeats ($x10^3$)	WER(%) held-out	Δ	WER(%) test	Δ
Baseline morph-based ASR	-	-	22.9	-	22.4	-
M(1)	45.9	20.1	22.1	0.8	21.8	0.6
M(1)+PoS(1,2)	46.7	22.2	21.7	1.2	21.4	1.0
M(1)+C_B(1,2)	48.5	24.2	21.8	1.1	21.3	1.1
M(1)+C_{MED}(1,2)	95.1	28.5	21.8	1.1	21.3	1.1
M(1)+M_{LD}(2)	46.9	21.3	22.2	0.7	21.7	0.7
M(1)+C_B(1,2)+M_{LD}(2)	49.5	27.0	21.8	1.1	21.2	1.2
M(1)+C_{MED}(1,2)+M_{LD}(2)	96.1	28.0	21.8	1.1	21.3	1.1
M(1)+T(1)	247.5	78.0	22.2	0.7	21.7	0.7

training data words segmented into morphs. SRILM toolkit is used for clustering with Brown *et al.*'s algorithm and 50, 100, and 200 classes are generated from the morph corpora. Each cluster set gives almost the same result in the experiment. Therefore we only report the result with 50 clusters in Table III. The feature templates with this clustering are denoted by $\mathrm{C_B}(.)$. In clustering with MED similarity, the imposed threshold constraint results in 5186 clusters. The feature templates with this clustering are denoted by $\mathrm{C_{MED}}(.)$ in Table III. Note that MED clustering is only applied to non-initial morphs. For the initial morphs we both try assigning all of them into the same cluster and clustering these morphs with Brown *et al.*'s algorithm. These two approaches give almost the same result and we only report the result where all initial morphs are assigned to the same cluster in Table III. Both of the Brown and MED clustering approaches give additional 0.5% significant ($p < 0.001$) improvements on top of the gain obtained with morph unigram features on the test data.

In addition to morpho-syntactic clusters, we also experiment with PoS tags of concatenated morph sequences. PoS tags of morph sequences are obtained using the morphological parser and disambiguation tool by assuming that concatenation of morph sequences result in grammatically correct Turkish words. The words that are not parsed by the parser are automatically labeled with the "Noun" tag. This is the default labeling in the dependency parser and we used the same tag to be consistent with the word PoS tag experiments. Morph unigrams together with PoS tag unigrams and bigrams of morph sequences are the best performing feature set on the held-out morph hypotheses and reduce the test set error from 22.4% to 21.4% (significant at $p < 0.001$).

Long distance trigger features are extracted from the oracle and 1-best morph hypotheses in the N-best lists. Incorporating long distance morph trigger features into either morph unigrams or into morph unigram and cluster features yields no additional improvements. We also tried topic features on morphs using the same topic clustering that we used for words. The topic features are defined in the same way with words using morph:topic pairs. Topic features for morphs do not give significant gains on top of morph unigram features.

VI. DISCUSSION AND CONCLUSION

In this paper, we investigate linguistically and syntactically motivated features in addition to the basic word and morph n-gram features in Turkish DLMs. The linguistically motivated features, morphological, and syntactic features, are investigated on word hypothesis sentences. Statistically derived features, morph clusters and long distance morph triggers, are investigated on morph hypothesis sentences. We also incorporate topic information into DLMs by obtaining topics using statistical approaches. The proposed feature sets yield significant improvements in DLMs for word-based and morph-based Turkish BN transcription systems.

The DLM experiments on word and morph hypotheses reveal some important and interesting results. First, basic unigrams are shown to be effective for obtaining significant gains on the baseline and increasing the n-gram context does not introduce any further gains, also observed in [18]. One possible explanation is that unigrams reduce the data sparsity compared to the higher order n-grams since the DLM training data is limited. Another explanation can be the adaptation of the language model with the perceptron algorithm [48]. In order to alleviate the effect of out-of-domain data, generic and in-domain language models are linearly interpolated in the ASR systems. However, DLM with unigram features can provide an extra adaptation by biasing the words or morphs occurring more frequently in the in-domain data. In order to investigate if the gains of the word and morph unigram features are really coming from language model adaptation, we investigate the correlation between the active unigram feature weights and the difference of the unigram log probabilities between the in-domain and the baseline language models both for words and morphs. Observing a higher correlation between these two components would be a strong evidence of the language model adaptation. However, we obtain rather low correlations, 0.09 for words and 0.16 for morphs, which means that the gain obtained with the word and morph unigram features is not only coming from language model adaptation.

Second, we demonstrate the superiority of sub-lexical units in DLMs in addition to generative language models. When we compare reranking word hypotheses with basic word features and reranking morph hypotheses with basic morph features, morphs yield more improvement than words. The gain obtained with morph unigrams is higher than the gain obtained with word unigrams (See Tables II and III. Additionally, sub-lexical units, i.e., IGs and stem+endings, also outperform words when reranking word hypotheses. As suggested in [16] for Czech, factorizations of morphological tags can be useful as DLM features. However, for Turkish the number of morphological tags per word can be quite high. We have found that in a vocabulary of 1.6 M morphologically decomposed words each word gets on average 6.1 morphological tags excluding the root. Therefore, meaningful groupings of lexical and surface form morphological tags, i.e., IGs and endings, provide a good compromise between composite and factored representations of these tags.

Third, it is shown that the n-gram features that capture the generalizations of the training data, i.e., PoS tags and morpho-syntactic clusters, are effective in DLM experiments. These features achieve significant additive gains on top of the gains obtained with word and morph unigram features. However, the topic-sensitive features which localize the data by classifying same words/morphs occurring in different topics as different features do not give any significant gain on top of word/morph unigrams. Since our DLM training data is limited, topic-sensitive features may introduce sparsity to the data.

Fourth, neither H2H dependency relation nor morph trigger features give any improvements. Morph triggers are just a brute-force attempt to incorporate longer distance morph relations into DLMs and they fail as DLM features. However, H2H dependencies are linguistically motivated, therefore they are expected to be more effective in DLMs. H2H(all) features even degraded the performance of the existing feature set when they are incorporated into the combination of word and PoS tag n-gram features. One possible problem is that the hypothesis sentences in the DLM training data contain recognition errors and the parser generates the best possible dependency relation even for incorrect hypotheses. As a result, the dependency analysis of the incorrect hypotheses may not provide good negative examples for discrimination of the correct and incorrect hypotheses. Another possible problem is the sparseness of the observations per parameters. If a feature is not seen frequently enough in the training data, we cannot estimate a robust parameter for this feature with the perceptron training. As a result, incorporating sparse features can degrade the performance of the existing model. We believe that the high number of features are masking the expected gains of the proposed features. This will make feature selection a crucial issue in our future research.

Finally, DLM training data is limited and as a result the improvements attained with unigram features (not higher orders) and with class-based, morph, and morphological features may be due to reducing the effect of data sparsity. Therefore, the ultimate importance of "real" linguistic and statistical features is not completely clear, but incorporating linguistic and statistically obtained information into DLMs seems to be the right direction for morphologically rich languages. Using more DLM training data may help to reveal further gains with linguistic and statistically obtained information.

Acknowledgment

The authors would like to thank AT&T Labs—Research for the software, K. Oflazer for the morphological parser, G. Cebiroğlu Eryiğit for the dependency parser, and H. Sak for the morphological disambiguation tool and anonymous reviewers for useful comments.

References

[1] E. Arısoy, M. Saraçlar, B. Roark, and I. Shafran, "Syntactic and sublexical features for Turkish discriminative language models," in *Proc. ICASSP*, Dallas, TX, 2010, pp. 5538–5541.

[2] E. Arısoy, B. Roark, I. Shafran, and M. Saraçlar, "Discriminative n-gram language modeling for Turkish," in *Proc. Interspeech*, Brisbane, Australia, 2008, pp. 825–828.

[3] J. A. Bilmes and K. Kirchhoff, "Factored language models and generalized parallel backoff," in *Proc. NAACL*, Edmonton, AB, Canada, 2003, pp. 4–6.

[4] R. Rosenfeld, "Adaptive statistical language modeling: A maximum entropy approach," Ph.D. dissertation, Carnegie Mellon Univ., Pittsburgh, PA, 1994.

[5] S. Khudanpur and J. Wu, "Maximum entropy techniques for exploiting syntactic, semantic and collocational dependencies in language modeling," *Comput. Speech Lang.*, vol. 14, pp. 355–372, 2000.

[6] H. Erdoğan, R. Sarikaya, S. F. Chen, Y. Gao, and M. Picheny, "Using semantic analysis to improve speech recognition performance," *Comput. Speech Lang.*, vol. 19, pp. 321–343, 2005.

[7] R. Sarıkaya, M. Afify, D. Yonggang, H. Erdoğan, and G. Yuqing, "Joint morphological-lexical language modeling for processing morphologically rich languages with application to dialectal Arabic," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 7, pp. 1330–1339, Sep. 2008.

[8] R. Rosenfeld, S. F. Chen, and X. Zhu, "Whole-sentence exponential language models: A vehicle for linguistic-statistical integration," *Comput. Speech Lang.*, vol. 15, no. 1, pp. 55–73, 2001.

[9] B. Roark, M. Saraçlar, and M. Collins, "Discriminative n-gram language modeling," *Comput. Speech Lang.*, vol. 21, no. 2, pp. 373–392, 2007.

[10] B. Roark, M. Saraçlar, M. J. Collins, and M. Johnson, "Discriminative language modeling with conditional random fields and the perceptron algorithm," in *Proc. ACL*, Barcelona, Spain, 2004, pp. 47–54.

[11] M. Collins, M. Saraçlar, and B. Roark, "Discriminative syntactic language modeling for speech recognition," in *Proc. ACL*, Ann Arbor, MI, 2005, pp. 507–514.

[12] T. Oba, T. Hori, and A. Nakamura, "Round-robin discrimination model for reranking ASR hypotheses," in *Proc. Interspeech*, Makuhari, Japan, 2010, pp. 5126–5129.

[13] M. Saraclar and B. Roark, "Utterance classification with discriminative language modeling," *Speech Commun.*, vol. 48, no. 3–4, pp. 276–287, Mar.–Apr. 2006.

[14] M. Collins, "Discriminative training methods for hidden Markov models: Theory and experiments with perceptron algorithms," in *Proc. EMNLP*, Philadelphia, PA, 2002, pp. 1–8.

[15] T. Oba, T. Hori, and A. Nakamura, "A comparative study on methods of weighted language model training for reranking LVCSR n-best hypotheses," in *Proc. ICASSP*, Dallas, TX, 2010, pp. 2446–2449.

[16] I. Shafran and K. Hall, "Corrective models for speech recognition of inflected languages," in *Proc. EMNLP*, Sydney, Australia, 2006, pp. 390–398.

[17] N. Singh-Miller and M. Collins, "Trigger-based language modeling using a loss-sensitive perceptron algorithm," in *Proc. ICASSP*, Honolulu, HI, 2007, pp. 25–28.

[18] M. Lehr and I. Shafran, "Discriminatively estimated joint acoustic, duration and language model for speech recognition," in *Proc. ICASSP*, Dallas, TX, 2010, pp. 5542–5545.

[19] E. Charniak and M. Johnson, "Coarse-to-fine n-best parsing and maxent discriminative reranking," in *Proc. ACL*, Ann Arbor, MI, 2005, pp. 173–180.

[20] M. Colins and T. Koo, "Discriminative reranking for natural language parsing," *Comput. Linguist.*, vol. 31, no. 1, pp. 25–70, 2005.

[21] H.-K. J. Kuo, E. Fosler-Lussier, H. Jiang, and C.-H. Lee, "Discriminative training of language models for speech recognition," in *Proc. ICASSP*, Orlando, FL, 2002, pp. 325–328.

[22] A. Rastrow, A. Sethy, and B. Ramabhadran, "Constrained discriminative training of n-gram language models," in *Proc. IEEE Workshop ASRU*, Merano, Italy, 2009, pp. 311–316.

[23] S. Yaman, L. Deng, D. Yu, Y.-Y. Wang, and A. Acero, "An integrative and discriminative technique for spoken utterance classification," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 6, pp. 1207–1214, 2008.

[24] T. Alumäe, "Methods for Estonian large vocabulary speech recognition," Ph.D. dissertation, Tallinn Univ. of Technol., Tallinn, Estonia, 2006.

[25] T. Hirsimäki, M. Creutz, V. Siivola, M. Kurimo, S. Virpioja, and J. Pylkkönen, "Unlimited vocabulary speech recognition with morph language models applied to Finnish," *Comput. Speech Lang.*, vol. 20, no. 4, pp. 515–541, 2006.

[26] P. Mihajlik, T. Fegyò, Z. Täske, and P. Ircing, "A morpho-graphemic approach for the recognition of spontaneous speech in agglutinative languages—Like Hungarian," in *Proc. Interspeech*, Antwerp, Belgium, Aug. 2007, pp. 1497–1500.

[27] E. Arısoy, D. Can, S. Parlak, H. Sak, and M. Saraçlar, "Turkish broadcast news transcription and retrieval," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 5, pp. 874–883, 2009.

[28] M. Creutz, T. Hirsimäki, M. Kurimo, A. Puurula, J. Pylkkönen, V. Siivola, M. Varjokallio, E. Arısoy, M. Saraçlar, and A. Stolcke, "Morph-based speech recognition and modeling of out-of-vocabulary words across languages," *ACM Trans. Speech Lang. Process.*, vol. 5, no. 1, pp. 1–29, 2007.

[29] K. Çarkı, P. Geutner, and T. Schultz, "Turkish LVCSR: Towards better speech recognition for agglutinative languages," in *Proc. ICASSP*, İstanbul, Turkey, 2000, pp. 1563–1566.

[30] D. Z. Hakkani-Tür, "Statistical language modeling for agglutinative languages," Ph.D. dissertation, Bilkent Univ., Ankara, Turkey, 2000.

[31] E. Mengüşoğlu and O. Deroo, “Turkish LVCSR: Database preparation and language modeling for an agglutinative language,” in *Proc. ICASSP, Student Forum*, Salt Lake City, UT, 2001.

[32] E. Arısoy, H. Dutağacı, and L. M. Arslan, “A unified language model for large vocabulary continuous speech recognition of Turkish,” *Signal Process.*, vol. 86, no. 10, pp. 2844–2862, 2006.

[33] K. Hacıoğlu, B. Pellom, T. Çiloğlu, Ö. Öztürk, M. Kurimo, and M. Creutz, “On lexicon creation for Turkish LVCSR,” in *Proc. Eurospeech*, Geneva, Switzerland, 2003, pp. 1165–1168.

[34] M. Creutz and K. Lagus, “Unsupervised morpheme segmentation and morphology induction from text corpora using Morfessor 1.0,” Helsinki Univ. of Technol., Publications in Computer and Information Sci. Rep. A81, Mar. 2005.

[35] H. Erdoğan, O. Büyük, and K. Oflazer, “Incorporating language constraints in sub-word based speech recognition,” in *Proc. ASRU*, San Juan, Puerto Rico, 2005, pp. 98–103.

[36] E. Arısoy and M. Saraçlar, “Lattice extension and vocabulary adaptation for Turkish LVCSR,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 1, pp. 163–173, Jan. 2009.

[37] K. Oflazer, “Two-level description of Turkish morphology,” *Literary Linguist. Comput.*, vol. 9, no. 2, pp. 137–148, 1994.

[38] H. Sak, T. Güngör, and M. Saraçlar, “Morphological disambiguation of Turkish text with perceptron algorithm,” in *Proc. CICLing LNCS 4394*, 2007, pp. 107–118.

[39] D. Hakkani-Tür, K. Oflazer, and G. Tür, “Statistical morphological disambiguation for agglutinative languages,” *J. Comput. Humanities*, vol. 36, no. 4, pp. 381–410, 2002.

[40] G. Eryiğit, J. Nivre, and K. Oflazer, “Dependency parsing of Turkish,” *Comput. Linguist.*, vol. 34, no. 3, pp. 357–389, 2008.

[41] C. Chelba and F. Jelinek, “Structured language modeling,” *Comput. Speech Lang.*, vol. 14, no. 4, pp. 283–332, 2000.

[42] B. Roark, “Probabilistic top-down parsing and language modeling,” *Computational Linguistics*, vol. 27, no. 2, pp. 249–276, 2001.

[43] P. F. Brown, V. J. D. Pietra, P. V. deSouza, J. C. Lai, and R. L. Mercer, “Class-based n-gram models of natural language,” *Comput. Linguist.*, vol. 18, no. 4, pp. 467–479, Dec. 1992.

[44] C. Manning and H. Schütze*, Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press, 1999.

[45] M. Hearst, “Texttiling: Segmenting text into multi-paragraph subtopic passages,” *Comput. Linguist.*, vol. 23, no. 1, pp. 33–64, Mar. 1997.

[46] H. Sak, T. Gungor, and M. Saraclar, “Resources for Turkish morphological processing,” *Lang. Resources Eval.*, vol. 45, no. 2, pp. 249–261, 2011.

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

[48] M. Bacchiani, B. Roark, and M. Saraçlar, “Language model adaptation with MAP estimation and the perceptron algorithm,” in *Proc. HLT-NAACL*, Boston, MA, 2004, pp. 21–24.

Ebru Arısoy (M’09) received the B.S., M.S., and Ph.D. degrees from the Electrical and Electronics Engineering Department, Boğaziçi University, Istanbul, Turkey, in 2002, 2004, and 2009, respectively.

She is currently working as a Post-Doctoral Researcher at the Speech Technologies for Media and ACCES Department, IBM T. J. Watson Research Center, Yorktown Heights, NY. Her main research interests include automatic speech recognition and statistical language modeling.

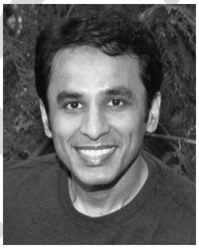
Murat Saraçlar (M’00) received the B.S. degree from the Electrical and Electronics Engineering Department, Bilkent University, Ankara, Turkey, in 1994 and the M.S.E. degree and Ph.D. degree from the Electrical and Computer Engineering Department, The Johns Hopkins University, Baltimore, MD, in 1997 and 2001, respectively.

He is currently an Associate Professor at the Electrical and Electronic Engineering Department, Boğaziçi University, Istanbul, Turkey. From 2000 to 2005, he was with the Multimedia Services Department, AT&T Labs—Research. His main research interests include all aspects of speech recognition, its applications, as well as related fields such as speech and language processing, human–computer interaction, and machine learning.

Dr. Saraçlar was a member of the IEEE Signal Processing Society Speech and Language Technical Committee (2007–2009). He is currently serving as an associate editor for IEEE SIGNAL PROCESSING LETTERS and he is on the editorial board of *Computer Speech and Language*.

Brian Roark (M’04–SM’11) received the B.A. degree from the Departments of Mathematics and Philosophy, University of California, Berkeley, in 1989, the M.S. degree in information science from the Claremont Graduate School, Claremont, CA, in 1997, the M.S. degree in applied mathematics from Brown University, Providence, RI, in 2001, and the Ph.D. degree from the Department of Cognitive and Linguistic Sciences, Brown University, in 2001.

From 2001 to 2004, he was a Senior Technical Staff Member in the Speech Algorithms Department, AT&T Labs—Research. He is currently an Associate Professor in the Center for Spoken Language Understanding and Department of Biomedical Engineering, Oregon Health and Science University, Beaverton.



Izhak Shafran (M’01) received the Ph.D. degree electrical engineering from University of Washington, Seattle.

He is currently an Assistant Professor in the Center for Spoken Language Understanding and Department of Biomedical Engineering, Oregon Health and Science University (OHSU), Beaverton. After the Ph.D. degree, he joined AT&T Labs Research in the Speech Algorithms Group. Before joining OHSU, he was a Visiting Professor at the University of Paris-South and then a Research Faculty in the Center for Language and Speech Processing (CLSP), The Johns Hopkins University. His research interests are in modeling speech and analyzing conversational speech.