SYNTACTIC AND SUB-LEXICAL FEATURES FOR TURKISH DISCRIMINATIVE LANGUAGE MODELS

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

[1]Electrical and Electronics Engineering Department, Boğaziçi University, Istanbul, Turkey
[2]Center for Spoken Language Understanding, OGI/OHSU, Beaverton, OR, USA

ABSTRACT

This paper investigates syntactic and sub-lexical features in Turkish discriminative language models (DLMs). DLM is a feature-based language modeling approach. It reranks the ASR output with discriminatively trained feature parameters. Syntactic information is incorporated into DLM as part-of-speech (PoS) tag n-gram features and head-to-head dependency relations. Sub-lexical units are first utilized as language modeling units in the baseline recognizer. Then, sub-lexical features are used to rerank the sub-lexical hypotheses. We explore features, similar to syntactic features, on sub-lexical units to reveal the implicit morpho-syntactic information conveyed by these units. We find out that DLM yields more improvement for sub-lexical units than for words. Basic sub-lexical n-gram features result in 0.6% reduction over the baseline and morpho-syntactic features yield an additional 0.4% reduction on the test set.

Index Terms— language modeling, automatic speech recognition, discriminative training, sub-lexical recognition units

1. INTRODUCTION

Turkish, being an agglutinative language with rich morphology, presents a challenge for automatic speech recognition (ASR) systems. Its agglutinative nature leads to a high number of out-of-vocabulary (OOV) words which degrade the ASR accuracy. To handle the OOV problem, vocabularies composed of sub-lexical units have been proposed for agglutinative languages. Sub-lexical vocabularies result in higher recognition accuracies than word vocabularies for Turkish [1].

In state-of-the-art ASR systems, language model parameters are estimated from a text corpus with maximum likelihood estimation. DLMs, trained on ASR transcriptions, are proposed as a complementary approach to this baseline model [2]. Discriminative training of language models has been shown to improve the performance of ASR systems partly due to optimizing the model parameters with respect to the objection function directly related to word error rate (WER) and partly due to incorporating relevant information sources such as morphology and syntax into the language modeling [3, 4, 5].

The authors would like to thank AT&T Labs – Research for the software, Kemal Oflazer for the morphological parser and Gülşen Cebiroğlu Eryiğit for the dependency parser. This research is supported in part by Turkish State Planning Organization under the project number DPT2007K120610, TUBITAK under the project number 105E102 and BU Research Fund under the project number 07HA201D. Ebru Arısoy was supported by TUBITAK BDP. Murat Saraçlar is supported by TUBA GEBIP award. This research was also supported in part by NSF 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.

In this paper we make use of both sub-lexical recognition units and discriminative training in Turkish language models. In our DLM approach, first relevant information obtained from the baseline ASR system is encoded as features. Then, feature parameters are discriminatively trained using training samples and utilized to rerank N-best hypotheses. The basic features in the models are n-grams. Morphology and syntax are also important information sources for language modeling. Therefore better improvements can be achieved with the addition of morphological and syntactic features to basic n-grams. Our previous research [3] has shown the effectiveness of morphological features on Turkish DLM and we investigate syntactic features in this paper. Syntactic information is obtained from the ASR hypotheses by the help of morphological and syntactic tools [6, 7, 8].

In our sub-lexical DLM approach, first sub-lexical units are utilized as language modeling items in the baseline recognizer. Then, features obtained from the output of this recognizer are incorporated into DLM to rerank the N-best sub-lexical hypotheses. Generating the training data for DLM makes use of the baseline language model. Therefore, there can be an interaction between language modeling units and discriminative training. Using word features on word hypotheses and using sub-lexical features on sub-lexical hypotheses allow us to investigate this interaction.

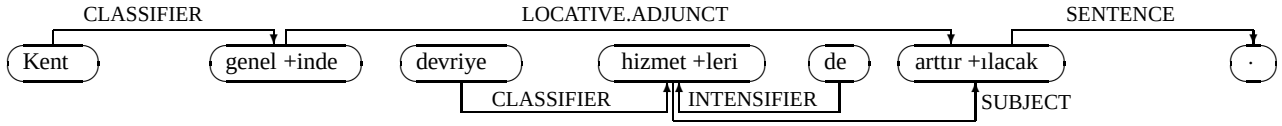
Linguistic information has been shown to be useful in feature-based language models [3, 4, 5, 9]. However, obtaining linguistic information from sub-lexical hypotheses is not trivial, since the morphological and syntactic tools can not be directly applied to sub-lexical units. Therefore we explore features, similar to syntactic word features, on sub-lexical hypotheses to reveal the morpho-syntactic information conveyed by sub-lexical units.

The main contributions of this paper are; (i) syntactic information is incorporated into Turkish DLM; (ii) effect of language modeling units on DLM is investigated; (iii) morpho-syntactic information is explored when using sub-lexical units. The next section summarizes the language models utilized in our research. Section 3 explains the syntactic and sub-lexical DLM feature sets. Experiments and results are presented in Section 4 and Section 5 concludes the paper.

2. LANGUAGE MODELS

2.1. Sub-lexical language models

In this approach, the recognition lexicon is composed of sub-lexical units instead of words. Sub-words in the lexicon are capable of covering most of the words of a language, thus addressing the OOV problem and leading to a decrease in WER. Grammatically-derived units, stems, affixes or their groupings, and statistically-derived units, morphs, have both been proposed as lexical items for Turkish ASR [1]. Morphs are learned statistically from words by the Morfessor algorithm [10]. Morfessor uses a Minimum Description Length



MORPH:	Kent	genel +inde	devriye	hizmet +leri	de	arttır +ılacak	.
WORD:	Kent	genelinde	devriye	hizmetleri	de	arttırılacak	.
GLOSS:	The city	throughout	patrol	services	also	will be increased	.
ROOT:	[kent]	[genel]	[devriye]	[hizmet]	[de]	[art]	.
PoS:	[Noun]	[Noun]	[Noun]	[Noun]	[Conj]	[Verb]	[Punc]
IG:	[A3sg+Pnon+Nom]	[A3sg+P3sg+Loc]	[A3sg+Pnon+Nom]	[A3pl+P3sg+Nom]	[-]	[Caus]$\hat{}DB$ [Pass+Pos+Fut+A3sg]	[-]

Fig. 1. Dependency tree of a morph hypothesis sentence. It means "Patrol services will also be increased throughout the city".

principle to learn a sub-word lexicon in an unsupervised manner. Morphs have been shown to outperform words and grammatically-derived units [1], therefore they are used as the sub-lexical approach in this paper.

In morph-based language models, all the words in the text corpus are split into morphs and generative language models are trained as if the morphs are words. After decoding, morph sequences are converted to words. A marker, "+", is attached to the non-initial morphs to facilitate this conversion (See Figure 1). Note that morph sequences do not always yield grammatically correct Turkish words.

2.2. Discriminative language models (DLMs)

DLM is a complementary approach to the baseline language model. In contrast to the generative language model, it is trained on acoustic sequences with their transcripts to optimize discriminative objective functions using both positive (reference transcriptions) and negative (recognition errors) examples.

The first step in DLM is to generate the training data which consists of lattices or N-best lists. DLM is a feature-based language modeling approach. Therefore, each candidate hypothesis in DLM training data is represented as a feature vector of the acoustic input, x, and the candidate hypothesis, y. The first element of this vector, $\Phi_0(x, y)$, is defined as the "log-probability of y in the lattice produced by the baseline recognizer for utterance x". The word n-grams [2], morphological relations [3, 4] and syntactic dependencies [5] can be used as the other features. These features are defined as the number of times a particular n-gram, relation or dependency is seen in the candidate hypothesis. Example word and morph bigram features from Figure 1 are as follows.

$\Phi_i(x, y)$ = Number of times "*hizmetleri de*" is seen in y

$\Phi_j(x, y)$ = Number of times "*hizmet +leri*" is seen in y

$\bar{\alpha}$ is the vector of parameters associated with the features. The parameters are discriminatively estimated using the perceptron algorithm (See [2]). Under this model, the best hypothesis maximizes the inner product of the feature and the parameter vectors ($\Phi(x, y) \cdot \bar{\alpha}$).

3. FEATURE SETS FOR DLM

We use word and sub-lexical n-grams as the basic features and investigate morphological, syntactic features on words, and morpho-syntactic features on sub-lexical units. These feature sets will be explained using Figure 1 and Table 1. Dependency relations between words and English gloss, PoS tag, root and Inflectional Groups (IGs) of each word are given in Figure 1.

3.1. Morphological Features

Morphological features are obtained with a Turkish morphological analyzer [6]. As was given in the last line of Figure 1, morphological feature sequences are separated by DB symbol which denotes the derivation boundaries. These sequences are called the IGs. The morphological analyzer [6] analyzes the i'th word, $\mathtt{w_i}$, as a root, $\mathtt{r_i}$, and a sequence of IGs, given as

$\mathtt{w_i} = \mathtt{r_i} + \mathtt{IG_{i,1}} + \mathtt{DB} + ... + \mathtt{IG_{i,j}} + \mathtt{DB} + ... + \mathtt{IG_{i,n_i}}$ where $\mathtt{IG_{i,j}}$ and $\mathtt{IG_{i,n_i}}$ are the j'th and the last IGs of $\mathtt{w_i}$ respectively. Morphological features for Turkish have already been investigated in [3]. In this paper we experiment with the same set of features to compare their performance with syntactic features[1]. Morphological features are root n-grams and IG pairs from the last IG of a word and any IGs of the next word. These features are illustrated in Table 1.

3.2. Syntactic Features

Syntax is an important information source for language modeling due to its role in sentence formation. Syntactic information has been incorporated into generative [11] and feature-based [5, 9] language models. The success of these approaches leads us to investigate syntactic features for Turkish DLM.

To obtain syntactic features, first all the words in the N-best lists are analyzed with the morphological analyzer [6]. Then, multiple analyses are disambiguated [7] and dependency trees of hypothesis sentences are derived using the dependency parser [8]. This is a classifier-based deterministic parser utilizing IGs as the parsing units to obtain the word dependencies. Here, it is important to note that hypothesis sentences contain recognition errors. However, the parser generates the best possible relations even for incorrect hypotheses.

We investigate similar feature definitions with [5]. We use PoS tag (denoted by $\mathtt{t_i}$ for $\mathtt{w_i}$) n-grams and head-to-head (H2H) dependency relations between lexical items or their PoS tags as the syntactic features. PoS tag features are utilized in an effort to obtain class-based generalizations that may capture well-formedness tendencies. H2H dependency relations are utilized 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. The syntactic features are illustrated in Table 1. The dependency relations between i'th and k'th words and their PoS tags are denoted by $DR(w_i, w_k)$ and $DR(t_i, t_k)$ respectively. For instance the feature notation H2H(tw) corresponds to the dependency relation between the PoS tag of one of words and the other word, $DR(t_i, w_k)$, and the PoS tag and word pair, *i.e.*, (INTENSIFIER [Conj] hizmetleri) for the phrase "hizmetleri de".

3.3. Sub-lexical 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. In Section 3.2, syntactic generalizations are considered with PoS

[1]The baseline system and amount of the DLM training data in this paper are different than the ones in [3].

Table 1. Feature sets utilized in the experiments. Features are defined as $\Phi_i(x, y)$ = Number of times a "feature template" is seen in y

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), (himetleri de)
Morph unigrams, bigrams	M(1), M(2)	$(m_i), (m_{i-1}m_i)$	(hizmet), (himet +leri)
Morphological features			
Root unigrams, bigrams	R(1), R(2)	$(r_i), (r_{i-1}r_i)$	(hizmet), (hizmet de)
IG unigrams, pairs	IG(1), IG(2)	$(IG_{i,j}), (IG_{i-1,last}IG_{i,j})$	([A3pl+P3sg+Nom]), ([A3pl+P3sg+Nom][-])
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			
Clustering (Brown et al.) unigrams, bigrams	$C_B(1)$, $C_B(2)$	$(c_i), (c_{i-1}c_i)$	
Clustering (MED) unigrams, bigrams	$C_{MED}(1)$, $C_{MED}(2)$	$(c_i), (c_{i-1}c_i)$	
Long distance triggers	$M_{LD}(2)$	$(m_{i,k}m_{j,l})$ where $j > i$	(hizmet de), (hizmet +ılacak)

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 and between non-initial morphs and suffixes.

3.3.1. Clustering of sub-lexical units

One way of information extraction from morphs is to convert them into words and to apply the same procedure with words. However, this indirect approach tends to fail when concatenation of morph sequences does not generate grammatically correct words. In addition, it contradicts with the main idea of statistical morphs – obtaining sub-lexical units without any linguistic tools. Therefore, this section focuses on obtaining syntactic information, similar to PoS tags of words, directly from morph sequences. This is achieved via grouping morphs that have similar functions. We apply two hierarchical clustering approaches on morphs to obtain meaningful groupings. The first one is Brown et al.'s algorithm [12] which aims to cluster words into semantically-based or syntactically-based groupings by maximizing the average mutual information of adjacent classes. The second approach utilizes minimum edit distance (MED) as the similarity function in bottom-up clustering. In our application we modify MED to softly penalize the substitutions and deletions due to vowel and consonant harmony, and consonant drop rules of Turkish. The motivation in this modification is to group morphs that are similar in lexical form in addition to surface form. This clustering is only meaningful for non-initial morphs since graphemic similarity of initial morphs does not reveal any linguistic information. Features are illustrated in Table 1 ($\texttt{c}_\texttt{i}$ represents the cluster of the i'th morph, $\texttt{m}_\texttt{i}$).

3.3.2. Long distance triggers

We also investigate long distance triggers, similar to the trigger features in [13, 14], between morph pairs. 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. In a word hypothesis, the dependency relations can be good syntactic trigger pairs. Since these relations are not directly extracted from morph sequences, we extract all the morph pairs between the morphs of any two words in a sentence as the candidate morph triggers. 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 [15]. The pairs with higher likelihood ratios ($\log \frac{L(H_1)}{L(H_0)}$) are assumed to be the morph triggers and utilized as features. Feature template is given in Table 1 where $m_{i,k}$ represents the k'th morph of the i'th word.

4. EXPERIMENTS

4.1. Baseline ASR system

The baseline ASR system was developed with AT&T tools for automatic transcription of Turkish broadcast news (BN) [1]. We used a 200K word vocabulary, resulting in a 2% OOV rate, and a 76K morph vocabulary, resulting in full coverage[2] on the test data. Turkish web corpus [7] (182.3M words) and the reference transcriptions of the acoustic training data (1.8M words) were utilized in language modeling. Generative language models were built and linearly interpolated using the SRILM toolkit[3]. The best results were obtained with 3-gram word and 4-gram morph language models. Results for the held-out set (3.1 hours) are given in Tables 2 and 3. A separate test set (3.3 hours) is used for the final evaluations.

4.2. DLM Results

DLM training data, 50-best lists, for words and morphs were generated by decoding the acoustic model training data (188 hours) with the baseline ASR systems. Language model over-training was controlled via 12-fold cross validation. Feature parameters were trained with the perceptron algorithm in 1-3 iterations. α_0 parameter and the number of iterations were optimized on the held-out set.

The results of DLM experiments on word hypotheses for the held-out set are given in Table 2. 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 and roots. However bigrams and trigrams do not give any improvements over unigrams. Word, root and IG features give gains over the baseline error. These gains are consistent with the ones reported in [3]. Then, we utilize syntactic features together with word features. PoS tags yield additive 0.4% improvement on top of the gain obtained with word unigrams. Dependency features, H2H(all)[4], are incorporated into word and PoS tag features. However, the performance is degraded most probably due to the data sparsity problems with 5.8M 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. The best performing feature set on the held-out word hypotheses, R(1) + IG(1,2), reduces the test set error from 23.4% to 22.7% (significant at $p < 0.001$ as measured by the NIST MAPSSWE test).

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

[3] http://www.speech.sri.com/projects/srilm/

[4] H2H(all) = H2H(ww) + H2H(wt) + H2H(tw) + H2H(tt)

Table 2. Results on word 50-best hypotheses for the held-out set.

	Feats(x10^3)	ActFeats(x10^3)	WER(%)	ΔWER
Baseline word-based ASR	-	-	24.1	-
W(1)	154.9	51.2	23.8	0.3
R(1)	34.3	12.9	23.4	0.7
R(1) + IG(1,2)	167.9	57.0	23.3	0.8
W(1)+PoS(1,2)	155.7	60.6	23.4	0.7
W(1)+PoS(1,2)+H2H(tt)	157.5	66.7	23.4	0.7
W(1)+PoS(1,2)+H2H(all)	5783.6	1152.5	23.6	0.5

The results of DLM experiments on sub-lexical hypotheses are given in Table 3. As with the word and root features, bigrams and trigrams do not introduce any gains over unigrams, moreover they degrade the performance of the unigram features. DLM yields more improvement for morphs than for words with basic n-gram features (See Tables 2 and 3). Then, we incorporate syntactic features into morph unigrams. We both experiment with PoS tags of concatenated morph sequences and automatically derived morpho-syntactic clusters. SRILM toolkit is used for clustering with Brown et al.'s algorithm and 50 classes are generated from the morph corpora. Clustering with MED similarity is only applied to non-initial morphs and all initial morphs are assigned to the same cluster, resulting in totally 5186 clusters. The n-gram features obtained from the PoS tags of morph sequences and automatic clustering approaches give additional 0.4% and 0.3% significant improvements on top of the gain obtained with morph unigram features. Long distance trigger features are extracted from the oracle and 1-best morph hypotheses and incorporated into morph unigrams. However no additional gain is achieved. The best performing feature set on the held-out morph hypotheses, M(1)+PoS(1,2), reduces the test set error from 22.4% to 21.4% (significant at $p < 0.001$).

5. CONCLUSION AND DISCUSSION

This paper reveals important and interesting results on Turkish DLM. First, unigrams are shown to be effective for obtaining significant gains on the baseline and increasing the n-gram context does not introduce any further gain. Considering that the perceptron training penalizes features associated with the current 1-best and rewards features associated with the oracle, the success of word unigram features can be explained as adaptation of the language model with the perceptron algorithm [16]. The language model interpolation constant is optimized on the held-out set before lattice generation. However, DLM provides a better optimization than the linear interpolation. Second, it is shown that DLM with basic features yields more improvement for morphs than for words. This demonstrates the superiority of sub-lexical units also in DLM. Third, morpho-syntactic clusters yield 1.1% significant gain over the baseline. It is interesting that this gain is obtained independent of the clustering approach, showing the effectiveness of the n-gram features that capture the generalizations of the training data.

Neither H2H nor morph dependencies give any improvements. Morph triggers are just a brute-force attempt to incorporate longer distance morph relations. However, H2H dependencies are linguistically motivated, therefore they are expected to be more effective in DLM. One possible problem is that the parser generates the best possible dependency relations even for incorrect hypotheses, as a result, it may not provide good negative examples for discrimination. Therefore, we will investigate a better way of incorporating longer distance information into DLM in our future research.

Our final observation is that the high number of features are masking the expected gains of the proposed features, mostly due to the sparseness of the observations per parameter. This will make feature selection a crucial issue for our future research.

Table 3. Results on morph 50-best hypotheses for the held-out set.

	Feats(x10^3)	ActFeats(x10^3)	WER(%)	ΔWER
Baseline morph-based ASR	-	-	22.9	-
M(1)	45.9	20.1	22.1	0.8
M(1,2)	2272.7	241.3	22.4	0.5
M(1)+PoS(1,2)	46.7	22.2	21.7	1.2
M(1)+C_B(1,2)	48.5	24.2	21.8	1.1
M(1)+C_{MED}(1,2)	95.1	28.5	21.8	1.1
M(1)+M_{LD}(2)	46.9	21.5	22.2	0.7

6. REFERENCES

[1] E. Arısoy, D. Can, S. Parlak, H. Sak, and M. Saraçlar, "Turkish Broadcast News Transcription and Retrieval," *IEEE Trans Audio, Speech and Lang. Process.*, vol. 17, no. 5, pp. 874–883, 2009.

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

[3] 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.

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

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

[6] K. Oflazer, "Two-level description of Turkish morphology," *Literary and Linguistic Computing*, vol. 9, no. 2, 1994.

[7] H. Sak, T. Güngör, and M. Saraçlar, "Turkish language resources: Morphological parser, morphological disambiguator and web corpus," in *Proc. GoTAL, LNAI 5221*, 2008.

[8] G. Eryiğit, J. Nivre, and K. Oflazer, "Dependency parsing of Turkish," *Computational Linguistics*, vol. 34, no. 3, 2008.

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

[10] M. Creutz and K. Lagus, "Unsupervised morpheme segmentation and morphology induction from text corpora using Morfessor 1.0," Publications in Computer and Information Science Report A81, Helsinki University of Technology, March 2005.

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

[12] P. F. Brown, V. J. Della Pietra, P. V. deSouza, J. C. Lai, and R. L. Mercer, "Class-based n-gram models of natural language," *Computational Linguistics*, vol. 18, no. 4, 1990.

[13] R. Rosenfeld, *Adaptive Statistical Language Modeling: A Maximum Entropy Approach*, Ph.D. thesis, Carnegie Mellon University, Pittsburgh, PA, USA, 1994.

[14] N. Singh-Miller and M. Collins, "Trigger-based language modeling using a loss-sensitive perceptron algorithm," in *Proc. ICASSP*, Honolulu, Hawaii, USA, 2007.

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

[16] M. Bacchiani, B. Roark, and M. Saraçlar, "Language model adaptation with MAP estimation and the perceptron algorithm," in *Proc. HLT-NAACL*, Boston, MA, USA, 2004.