Generative and Discriminative Methods using Morphological Information for Sentence Segmentation of Turkish

Guz Umit, Favre Benoit, Hakkani-Tür Dilek, Tur Gokhan

Generative and Discriminative Methods using Morphological Information for Sentence Segmentation of Turkish

Umit Guz, *Member, IEEE*, Benoit Favre, Member, IEEE,
Dilek Hakkani-Tür, *Senior Member, IEEE*, Gokhan Tur, *Senior Member, IEEE*.

***Abstract*—**
This paper presents novel methods for generative, discriminative, and hybrid sequence classification for segmentation of Turkish utterances into sentences. In the literature, this task is generally solved using statistical models that take advantage of lexical information among others. However, Turkish has a productive morphology that generates an exponential vocabulary size, harming language models such as the established hidden event language model (HELM). We extend this model as a factored hidden event language model (fHELM) in order to take advantage of morphologically informed features in addition to the word sequence. Our results indicate that fHELMs result in a 26% reduction in error rate for Turkish broadcast news. Combining lexical, morphological, and prosodic information using these new models and discriminative classifiers (boosting and conditional random fields) results in significant performance improvements over any of the classifiers alone.

I. Introduction

Many useful results have been obtained by applying statistical language modeling techniques to English (and similar languages) – in speech recognition, parsing, word sense disambiguation, part-of-speech (POS) tagging, etc. However, languages that display a substantially different behavior than English, like Turkish, Czech, Hungarian (in that, they have agglutinative or inflective morphology and relatively free constituent order) have not been studied extensively using statistical approaches. In these languages, due to their richer morphology, the vocabulary size for a given corpus size is much larger than other languages [1], [2]. While this causes a data sparseness problem for these languages, the statistical models that look at only words are also blind to the information encoded in the morphology. Usually, the combined effect of these problems is reduction in language processing performance for these languages.

Similarly, in spite of all the advances in discriminative classification techniques in the machine learning community, discriminative sequence classification is still a challenge. Researchers have proposed various techniques such as maximum entropy Markov models [3] or conditional random fields [4], [5]. However these techniques are typically not very successful in handling continuous valued features. On the other hand, for generative sequence modeling, hidden Markov models (HMMs) still dominate the field; however usually only one level of states is employed. For example, for automatic speech recognition (ASR), typically word sequences are modeled for the language model as part of joint modeling [6]. With the advances in graphical models, factored language models (FLMs) handling bundles of features for each sample have been proposed [7]. FLMs have been successfully used for ASR of inflectional languages such as Arabic [8].

In this paper, we address the problem of exploiting morphological information in statistical classification models for sentence segmentation of Turkish speech. Our contributions are four-fold: First, we extend the hidden event language models to factored hidden event language models and combine them with classification models. Second, we introduce a new set of morphological features, extracted from words and their morphological analyses. Third, we extract a set of prosodic features, which are mainly motivated from our previous work for other languages, for the task of Turkish sentence segmentation. Fourth, we propose a discretization method for using continuous-valued features in CRF, that benefits from decision stumps as learned by boosting.

In the next section we briefly summarize the related work on sentence segmentation of speech. Then we present our approach, mainly the generative, discriminative, and hybrid modeling techniques. Then we describe the feature sets for segmenting Turkish speech into sentences. Finally, we provide experimental results showing the effectiveness of the proposed techniques for this morphologically rich language before concluding.

II. Sentence Segmentation

Sentence segmentation for speech aims at finding sentential unit boundaries in a stream of words, output by a speech recognizer. It is a preliminary step for many speech processing applications, such as parsing, machine translation and information extraction, which generally assume the presence of punctuation. One typically leverages the word sequence generated by a speech recognizer and prosodic cues such as pitch, energy and pause duration in order to segment the audio in sentences.

Previous work on sentence segmentation has considered this task as a word boundary classification problem, by determining whether or not two consecutive words are separated by a sentence boundary. The features used are mainly limited to words

Umit Guz, Benoit Favre, and Dilek Hakkani-Tür are with the International Computer Science Institute (ICSI), Berkeley, CA 94704 (email:{umit,favre,dilek}@icsi.berkeley.edu). Umit Guz is also with the Isik University, Istanbul, Turkey. Gokhan Tur is with SRI International, Menlo Park, CA 94025 (email:gokhan@speech.sri.com).

neighboring the boundary [9], [10], [11], with the exception of [12], who included a reranking phase using sentence-level features. [13] showed that for segmentation of speech into sentences, prosodic and lexical cues provide complementary information. [14] evaluated different modeling approaches (HMM, maximum entropy, and conditional random fields) and various prosodic and textual features, in both conversational telephone speech and broadcast news speech.

There is also related work for sentence boundary detection in languages other than English, for example, in Czech [15] where an HMM approach was used, and in Chinese [16], [17] where a maximum entropy classifier was used with mostly textual features. [11] used lexical and prosodic features with several classifiers, including maximum entropy and boosting for English and Mandarin. [18] investigated the use of the same set of prosodic features and feature selection for English, Mandarin, and Arabic. [19] used syntactic dependency structure and support vector machines for sentence boundary detection in Japanese. [20] is the first work that used morphological features for sentence segmentation of Turkish; our work, in a way, extends that work to also include prosodic features and more sophisticated classification models.

Sentence segmentation has also been studied according to various other aspects. [21] showed the benefits of speaker-adapted models and [22] focused on domain adaptation. Sentence segmentation can be optimized to improve downstream tasks, such as speech translation [23], [24] or information extraction [25].

III. Approach

In the literature, typically sentence or dialog act segmentation is treated as a boundary classification problem where the goal is finding the most likely boundary tag sequence, $Y = Y_1 \dots Y_n$ given the features, $\mathcal{X} = \mathcal{X}_1 \dots \mathcal{X}_n$:

$$argmax_Y P(Y|\mathcal{X})$$

To this end mostly generative, discriminative, or hybrid models have been used. Below we summarize these approaches and explain how we extend them to handle the speech input of morphological languages.

A. Factored Hidden Event Language Models

We propose using factored language models with hidden event language models. Below, first we describe HELM and FLM and then describe how we combine them.

1) Hidden Event Language Models: The most popular generative model for sentence segmentation is the hidden event language model (HELM), as introduced by [26]. HELM was originally designed for speech disfluencies, such as deletion (DEL) and repetition (REP). The approach was to treat such events as extra meta-tokens. To ease the computation, an imaginary "no disfluency" (NODF) token is inserted between two words where there is no disfluency between them. The following example is a conceptual representation of a sequence with disfluencies:

... she NODF got REP got NODF real NODF lucky ...

For sentence segmentation, sentence boundaries are simply treated as hidden events, and the word sequence is augmented with fictitious sentence boundary tokens (S for sentence boundary, N for else). So an example would be as follows:

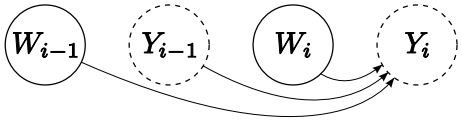

Fig. 1. Conceptual hidden event language model for sentence segmentation.

.. real N lucky S he N was ...

Note that this is different from using an HMM as is typically done in similar tagging tasks, such as POS tagging [27] or named entity extraction [28]. For sentence segmentation, the conceptual model is depicted in Figure 1. In this model one state is reserved for each of the boundary tokens, S and N, and the rest of the states are for generating words. It has been shown that HELM outperforms the conventional HMM approach, and since it allows an explicit point to emit the boundary token, hence can incorporate nonlexical information via combining with other models as presented in the next subsection [13].

The Bayesian optimization is simply done by the Viterbi decoding using only lexical features, i.e., the language model, to model $P(\mathcal{X}, Y)$, where $\mathcal{X}$ and Y represent all the words and boundary tokens.

$$argmax_Y P(Y|\mathcal{X}) = argmax_Y P(\mathcal{X}, Y)$$

2) Factored Language Models: Factored language models aim to model a sequence of feature sets, extending the conventional language modeling. In other words, the goal is building probabilistic language models using the subsets of feature sets (or factors).

Factored language models have been successfully used for ASR [8] of inflectional languages, by defining factors or feature sets consisting of surface forms, stems, morphological analyses, etc. of the words.

More formally, the factored language model aims to estimate the probability of a feature set sequence, $\mathcal{X}_1, ..., \mathcal{X}_n$ instead of a word sequence $W_1, ..., W_n$. Here we consider $\mathcal{X}_t = (W_t, M_t)$ where M_t is a morphological feature for word W_t. An example factored language model can be seen in Figure 2. The current word relies on not only the previous two words but also the current and previous morphological analyses. Therefore, it models:

$$P(W_t|W_{t-1}, W_{t-2}, M_t, M_{t-1})$$

Even with lower-order n-gram approximations, since it may be possible to have unseen n-gram sequences, one important issue with FLMs is how to back off to reliably estimate such probabilities. A new generalized parallel back-off technique was proposed to tackle this problem [7]. Basically, the system is given a back-off graph, which denotes the paths for back-off. Paths in this graph can be chosen manually. In the literature, with complex factors, methods based on genetic algorithms have been proposed to choose the optimal back-off graph [29]. The important point is that many back-off paths can be proposed and the system can process them in parallel.

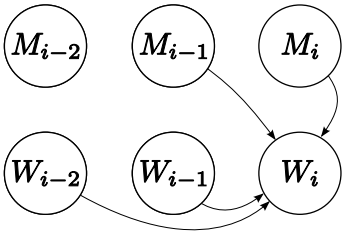

Fig. 2. An example factored language model seen as a directed graphical model over words W and morphological factors, M. The arrows indicate the factors used for estimating the probabilities.

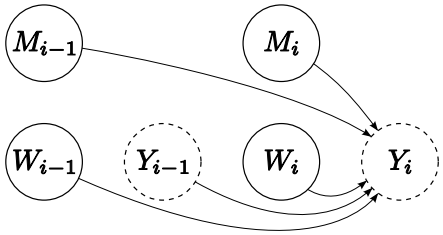

Fig. 3. An example factored language model created for a hidden event language model seen as a directed graphical model over word boundaries, Y, and words, W, and morphological factors, M. The arrows indicate the factors used for estimating the probabilities.

3) From HELM to fHELM: The factored hidden event language models are straightforward extensions of hidden event language models and factored language models. They combine the strength of factored language models for multi-feature sequence modeling with the classification power of hidden event language models. Figure 3 presents the factored hidden event language model topology employed in this paper. The boundary states still exist to potentially build hybrid models (as explained below) and the boundary decision is made according to the formula:

$$P(Y_t|W_t, M_t, Y_{t-1}, W_{t-1}, M_{t-1})$$

where Y_t indicates the boundary decision, S or N after the word W_t with a morphological analysis of M_t.

The next step for building an fHELM is creating a back-off graph indicating the possible back-off paths in case the statistics for the desired n-gram are not reliable. Factored language models are supposed to process them in parallel. In this paper we tried only linear graph back-off (i.e. dropping and forgetting about one factor at a time) and fully connected graph back-off (i.e. backing off to all possible subsets) starting from the most distant feature. More formally, an example back-off dropping the most distant word is defined as follows for factored hidden event language models:

$$P(Y_t|C_t) = \begin{cases} P_{ML}(Y_t|C_t) & \text{if } N(C_t, Y_t) > \tau \\ \alpha(C_t) \times P_{BO}(Y_t|\hat{C}_t) & \text{otherwise.} \end{cases}$$

where $C_t = W_t, M_t, Y_{t-1}, W_{t-1}, M_{t-1}$ is the original context, $\hat{C}_t = W_t, M_t, Y_{t-1}, M_{t-1}$ is the backed off context, P_{ML} is the standard maximum likelihood estimate (with smoothing), $N(\cdot)$ is the number of occurrences, and α is used to ensure that the result is still a probability distribution.

Then the standard Viterbi decoding may be employed to find the most probable state sequence, i.e. the boundary decisions given the words and their other features, such as morphological analysis. This results in a neat method for building a generative classifier when multiple features are used for each sample position. Furthermore, similar to regular HELMs, it is possible to combine the posterior probabilities obtained from other classifiers (preferably discriminative) to improve the performance even more. For example fHELM may exploit the lexical and morphological information and then may be combined with a classifier that uses only prosodic features.

In our experiments, the SRILM [30] toolkit is used for Viterbi decoding and for building the conventional and factored hidden event language models with modified Kneser-Ney smoothing [31].

B. Discriminative Classification Models

One weakness of the hidden event language models is that one can incorporate only a single stream of discrete features such as words. To overcome this obstacle, various classification methods have been used in the literature. In a pioneering study, decision trees were used to build segmentation models to improve the performance also by using additional prosodic features [13]. With the advances in discriminative classification algorithms, researchers tried using CRFs [32] and boosting [33], and hybrid approaches using boosting and maximum entropy classification algorithms [11].

Our system relies on boundary-wise posterior probabilities $P(Y_t|\mathcal{X}_t)$ provided by two classifiers that can be used independently or jointly. The first component is an Adaboost [34] classifier that generates posterior probability estimations out of weighted decision stumps (one-level decision trees):

$$P(Y_t|\mathcal{X}_t) = \left[exp\left(-2m \sum_{i=1}^{m} w_i s_i(\mathcal{X}_t) \right) \right]^{-1}$$

where $s_i(\cdot)$ is a decision stump (presence of a discrete feature or position relative to a threshold of a continuous feature) over a single feature, w_i is the weight given to that decision stump, and m is the number of decision stumps. Adaboost is trained by iterating over the selection of the best decision stump and reweighing of examples where the overall classifier makes mistakes. The implementation used in our experiments is icsiboost.[1] In all our experiments, we used boosting with 1,000 iterations.

The second component of our system uses CRFs as proposed by [4]. We use chain CRFs to estimate the probability of a sequence of boundary events ($Y = Y_1 \dots Y_n$) given a sequence of observations ($\mathcal{X} = \mathcal{X}_1 \dots \mathcal{X}_n$).

$$P(Y|\mathcal{X}) = \frac{1}{Z(\mathcal{X})} exp\left(\sum_{t=1}^{n} \sum_{i=1}^{m} \lambda_i s_i(Y_{t-1}, Y_t, \mathcal{X}_t) \right)$$

$$Z(\mathcal{X}) = \sum_{Y} exp\left(\sum_{t=1}^{n} \sum_{i=1}^{m} \lambda_i s_i(Y_{t-1}, Y_t, \mathcal{X}_t) \right)$$

Here, $s_i(\cdot)$ are decision functions that depend on the examples and a clique of boundaries close to Y_t, λ_i is the weight of s_i estimated on training data, and $Z(\mathcal{X})$ is a normalization factor. Note that CRFs give the probability of the sequence of

[1] http://code.google.com/p/icsiboost

boundary decisions. The forward-backward algorithm can be used to get boundary-level posterior probability estimates.

For our experiments, we use the CRF++ toolkit,[2] which allows binary decision functions dependent on the current boundary and the previous boundary. Features extracted from $\mathcal{X}$ originate from a neighborhood of the boundary and match the features used with Adaboost, though CRF++ does not handle continuous features and requires them to be quantized. After experimenting with different types of quantization, we observed that using thresholds from the decision stumps learned by Adaboost leads to improved performance, probably due to their ability to embed the interaction between features (in Adaboost training, classifiers are chosen in order to correct errors from previous iterations).

C. Hybrid Modeling

One important observation is that nonsequential classification algorithms typically ignore the context, which is critical for the segmentation task. While one may add context as an additional feature, or simply use CRFs, which inherently consider context, these approaches are suboptimal when dealing with real valued features, such as pause duration or pitch range. Most of the previous studies simply tackled this problem by binning the feature space either manually or automatically.

An alternative would be using a hybrid classification approach as suggested by Shriberg *et al.* [13]. The main idea would use the posterior probabilities, P_c, obtained from the other classifiers, such as boosting or CRF, by simply converting them to state observation likelihoods by dividing to their priors following the well-known Bayes rule:

$$argmax_Y \frac{P_c(Y|\mathcal{X})}{P(Y)} = argmax_Y P_c(\mathcal{X}|Y)$$

Applying the Viterbi algorithm to the HMM will then returns the most likely segmentation. In order to handle dynamic ranges of state transition probabilities and observation likelihoods, we apply a weighting scheme as is usually done in the literature

$$argmax_Y P_c(\mathcal{X}|Y)^\alpha \times P(Y)^\beta$$

where $P(Y)$ is estimated by the fHELM, α and β are optimized using a held-out set.

IV. Features

Three types of features - lexical, prosodic and morphological - are used in the classification models.

A. Lexical Features

The lexical features used in this work consist of six word n-gram features for each word boundary that were also used in our previous work for English [35]: three unigrams, two bigrams, and a trigram. Naming the word preceding the word boundary of interest as the *current* word, and the preceding and following words as the *previous* and *next* word respectively, the six lexical features are as follows:

- unigrams: {previous}, {current}, {next},
- bigrams: {current, next}, {previous, current}
- trigram: {previous, current, next}

B. Prosodic Features

The prosodic features are also transferred from the ICSI+ sentence segmentation system [11]. We use about 200 prosodic features, defined for and extracted from the regions around each inter-word boundary. The features include the pause duration at the boundary, normalized phone durations of the word preceding the boundary, and a variety of speaker-normalized pitch features and energy features preceding, following, and across the boundary. These features are an extension of similar features described in [13]. The extraction region around the boundary focuses on either the single words or brief time windows around the boundary. Measures include the maximum, the minimum or the average value in this range. Pitch features are normalized by speaker, using the method to estimate a speaker's baseline pitch values described in [13].

C. Morphological Features

Turkish is also a free-constituent-order language, in which constituents at certain phrase levels can change order rather freely according to the discourse context or text flow. However, the typical order of the constituents, especially for the news genre, is subject-object-verb (SOV).

Let us consider a simple complete sentence, "*çocuk yemek yedi*" in Turkish, which means "*the child ate the meal*" in English. The correct morphological analyses are as follows:

çocuk: Noun+A3sg+Pnon+Nom (the child)
yemek: Noun+A3sg+Pnon+Nom (the meal)
yedi: Verb+Pos(+dH)+Past+A3sg (ate)

Turkish has agglutinative morphology with productive inflectional and derivational suffixations [36]. The number of word forms one can derive from a Turkish root form may be in the millions [37]. For example, [38] shows that one can obtain thousands of new word forms from any noun, a verb, and an adjective root form by suffixing only three morphemes. As an example, let us consider the Turkish word "*yapabileceğim*", which consists of the morphemes "(yap)+(abil)+(ecek)+(im)" which roughly corresponds to "(do)+(able to)+(will)+(I)" in English. It has three potential morphological analyses:

- (yap)yap+Verb+Pos(+yAbil)^DB+Verb+Able(+yAcAk)+Fut(+yHm)+A1sg (I'll be able to do it)
- (yap)yap+Verb+Pos(+yAbil)^DB+Verb+Able(+yAcAk) ^DB+Adj+FutPart(+Hm)+P1sg (The (thing that) I'll be able to do)
- (yap)yap+Verb+Pos(+yAbil)^DB+Verb+Able(+yAcAk) ^DB+Noun+FutPart+A3sg(+Hm)+P1sg+Nom (The one I'll be able to do)

In this representation, the inflectional groups (IGs) denote the derivational boundaries and are marked with "^DB". In this example, the root is a verb but the final IGs have three readings, that are verb, adjective, and noun, respectively.

[2] http://crfpp.sourceforge.net/

Turkish presents an interesting problem for statistical models since the potential POS tag set size (that is, the number of possible morphological parses) is very large because of the productive derivational morphology. Following previous work [39], [2], our approach handles this by breaking up the morphosyntactic tags into inflectional groups, each of which contains the inflectional features for each (intermediate) derived form. To simplify our models further, we only extract morphological features from the final inflectional group of every word, which marks its final category in a sentence.

The morphological features used in this work are obtained using a morphological analyzer for Turkish [36], which outputs all possible morphological parses for all the words. We include the final inflectional group of every word as well as its POS tag, without resolving the ambiguity. For factored HELM, we arbitrarily chose one parse since fHELMs cannot handle multiple parses. With CRF and boosting we used all the possible parses as features. Boosting also exploited parse subsequences as additional features. For the POS tag, we mark the value of the feature as unknown when the word has multiple parses. We also include a single binary feature that checks if any of the possible morphological parses of a word is a Verb according to its final category. We hope, with this, to take advantage of the SOV nature of Turkish. To compare this approach, we also performed experiments with pseudo-morphological features, using the last three letters of each word. Like the "ed" suffix in English, in Turkish certain suffixes may indicate Verb categories.

V. Experiments and Results

A. Data Sets

In our experiments, we use the VOA (Voice of America) Turkish Section[3] part of the Turkish broadcast news (BN) speech corpus collected at the Bogazici University BUSIM Laboratory.[4] The VOA part of the corpus contains approximately 21 hours of single-channel Turkish broadcast news speech data recorded at a 16 bit, 32KHz sampling rate. For sentence segmentation experiments 42 Turkish broadcast news programs (30 minutes each) are used. These 42 files are split into a training set (22 files, 97,330 words), a development set (5 files, 14,897 words), and a test set (5 files, 15,688 words). The development set is used to optimize the parameters, such as probability thresholds and combination weights α and β. The vocabulary size of the training set is 19,328 words, and 33.5% of the words in the development set vocabulary and 35.8% of the test set vocabulary are not observed in the training data (these correspond to 14.8% and 17.3% of the development and test set words, respectively).

There are in total 128,005 words in the training, test, and development sets. 6.76% of these were not parsed by the morphological analyzer, mainly because of foreign person and city names and typos in the data. The remaining words that were parsed have on average 1.95 parse. This drops down to on average 1.83 analyses per word if only the last inflectional group of each word is considered, and to 1.30 if only the POS tag category of the last IG is considered. Table I lists the average number of parses per word as well as the percentage of words that have a single parse in the overall data set with these different conditions.

[3] http://www.voanews.com/turkish/
[4] http://www.busim.ee.boun.edu.tr/

Morphological Feature	Avg. Parse/Word	% of Unamb
Full Morph. Analysis	1.95	37.0
Last IG	1.83	39.5
POS of Last IG	1.30	62.9

TABLE I
AMBIGUITY STATISTICS FOR DIFFERENT LEVELS OF MORPHOLOGICAL FEATURES: AVERAGE NUMBER OF PARSES PER WORD FOR EVERY WORD THAT WAS PARSED BY THE MORPHOLOGICAL ANALYZER AND PERCENTAGE OF WORDS THAT HAVE A SINGLE PARSE (I.E., UNAMBIGUOUS WORDS).

Classifier	F	NIST
Boosting	0.749	44.0%
CRF	0.756	43.3%
HELM	0.782	36.7%

TABLE II
F-MEASURE AND NIST ERROR RATES WITH BOOSTING, CRF, AND HELM USING ONLY LEXICAL FEATURES.

B. Evaluation Methods

For performance evaluation, we report NIST error rate and F-measure on forced alignment output of an automatic speech recognizer [40]. The NIST error rate is the number of misclassified word boundaries divided by the number of reference sentence boundaries. F-measure is the harmonic mean of precision and recall. The NIST error rate is explained in detail with examples in [41].

C. Experiments with Lexical and Morphological Features

We compare our results with a baseline of using only lexical features for all classification methods. Table II presents results using boosting, CRF, and HELM with only lexical features. HELM outperforms other methods probably because of the large number of lexical features they must tackle due to the agglutinative nature of Turkish.

When we add morphological and pseudo-morphological (last three letters of words) to the feature sets, we observe significant improvements in the performance with all classifiers. This is intuitive because of the morphological characteristics and SOV sentence order of Turkish. One interesting observation is that with boosting the performance degrades when both morphological and pseudo-morphological features are employed instead of only one of them. CRF consistently performs a little better than boosting. The error rate of fHELM is reduced by 26% relative compared to HELM when only lexical features are used. This shows the effectiveness of factored hidden event language models for generative sequence classification. Furthermore, the

Features	L+M		L+PM		L+M+PM	
	F	NIST	F	NIST	F	NIST
Boosting	0.884	24.7%	0.853	30.0%	0.869	26.5%
CRF	0.887	24.0%	0.864	26.0%	**0.891**	**21.7%**
fHELM	0.865	25.9%	0.862	27.1%	-	-

TABLE III
F-MEASURE AND NIST ERROR RATES WITH BOOSTING, CRF AND HELM USING LEXICAL (L), MORPHOLOGICAL (M), AND/OR PSEUDO-MORPHOLOGICAL FEATURES.

Classifier	F	NIST
Boosting(L+M)+fHELM(L+M)	0.879	23.8%
CRF(L+M)+fHELM(L+M)	0.890	21.5%

TABLE IV
F-MEASURE AND NIST ERROR RATES WHEN COMBINING BOOSTING AND CRF WITH FHELM WITH LEXICAL (L) AND MORPHOLOGICAL (M) FEATURES.

Classifier	F	NIST
Boosting(P)	0.862	27.2%
Boosting(P)+fHELM(L+M)	**0.919**	**15.8%**

TABLE V
F-MEASURE AND NIST ERROR RATES WHEN USING ONLY PROSODIC (P) INFORMATION WITH BOOSTING AND COMBINING WITH FHELM USING LEXICAL (L) AND MORPHOLOGICAL (M) INFORMATION.

relative NIST error rate reductions are even more with boosting (44%) and CRF (50%) with morphological features. These results are shown in Table III.

Table IV presents results with the combination of discriminative and generative sequence classification methods when both lexical and morphological features are used. The performance is more or less the same as using only the discriminative classifiers, suggesting that they already incorporate the information coming with hidden event models.

D. Experiments with Prosodic Features

Since we expect the prosody to provide orthogonal information for sentence segmentation, we first combined boosting trained with only prosodic features with factored HELMs. Table V presents these results. Note that, before combination, boosting and fHELMs have comparable performance. This shows the utility of the prosodic features that were originally designed for English. Furthermore, this hybrid model reduces the NIST error rate by 39% relative. This demonstrates the power of the model combination with complementary information provided by two different sets.

Then we exploited the prosodic features along with lexical and morphological information with boosting and CRF. Table VI presents these results. As seen, for both classifiers, performance improved significantly. This is in part due to the nature of the data, i.e., broadcast news, in which the reporters and anchor people explicitly mark sentence boundaries with prosody.

As the final set of experiments, we tried combining fHELM with boosting and CRF using all the features. Table VII presents these results. With this final combination, the model including boosting did not improve. The CRF model improved, however only slightly.

VI. DISCUSSION AND CONCLUSIONS

We have presented generative, discriminative, and hybrid classification methods using lexical, morphological, and prosodic information for Turkish sentence segmentation. We have shown significant improvements over a lexical baseline.

While CRF results in better performance with prosodic and lexical features only, boosting benefits more from the morphological features. This is probably due to the ability of boosting to handle unknown feature values. For example, one of the morphological features is set to unknown in case the word is morphologically ambiguous. This requires further investigation, but a prior morphological disambiguation step may provide benefits.

The prosodic features are mainly transferred from English and model only word-level phenomena. They can also be improved by modeling at subword level. For example, the morphological ambiguity for the sentence final words may be resolved using morpheme-level prosodic features.

One significant benefit of using fHELMs is that they can be trained using millions of examples, also benefiting from the textual data that can be found easily (such as from the WWW), whereas the discriminative models are more limited for that case. In this work, we have used the same data for training all models, and investigating the use of more data for fHELMs is part of our future work, in addition to experimenting with real ASR output.

fHELMs can be used for other language processing tasks requiring sequence classification such as POS tagging and named entity extraction and can easily be combined with state-of-the-art discriminative models.

VII. ACKNOWLEDGMENTS

We thank Kemal Oflazer, for providing us with his morphological analyzer, Siddika Parlak and Murat Saraclar for providing us with the data sets and the forced alignments, and Andreas Stolcke and Dimitra Vergyri for many helpful discussions.

Features	L+P		L+M+P		L+PM+P		L+M+PM+P	
	F	NIST	F	NIST	F	NIST	F	NIST
Boosting	0.894	20.4%	0.922	16.5%	0.918	15.8%	**0.927**	14.7%
CRF	0.895	20.2%	0.921	14.6%	**0.916**	16.9%	0.923	15.3%

TABLE VI

F-MEASURE AND NIST ERROR RATES WITH BOOSTING AND CRF USING LEXICAL (L), PROSODIC (P), MORPHOLOGICAL (M), AND/OR PSEUDO-MORPHOLOGICAL (PM) FEATURES.

Classifier	F	NIST
Boosting(L+P+M+PM)+fHELM(L+M)	0.925	14.8%
CRF(L+P+M+PM)+fHELM(L+M)	0.926	14.9%

TABLE VII

F-MEASURE AND NIST ERROR RATES WHEN COMBINING FHELM WITH BOOSTING AND CRF USING LEXICAL (L), MORPHOLOGICAL (M+PM), AND PROSODIC (P) INFORMATION.

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) CALO (NBCHD-030010) program, the DARPA GALE (HR0011-06-C-0023) program, the Scientific and Technological Research Council of Turkey (TUBITAK) fundings at SRI and ICSI, (TUBITAK CAREER Project No: 107E182, Extracting and Using Prosodic Information for Turkish Spoken Language), and the Isik University Research Fund (Project No:05B304). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding agencies.

REFERENCES

[1] A. Waibel, P. Geutner, L. Mayfield-Tomokiyo, T. Schultz, and M. Woszczyna, "Multilinguality in speech and spoken language systems," *Proceedings of the IEEE, Special Issue on Spoken Language Processing*, vol. 88, no. 8, pp. 1297–1313, August 2000.

[2] D. Hakkani-Tür, K. Oflazer, and G. Tur, "Statistical morphological disambiguation for agglutinative languages," in *Proceedings of the 18th International Conference on Computational Linguistics (COLING)*, Saarbruecken, Germany, August 2000.

[3] A. McCallum, D. Freitag, and F. Pereira, "Maximum entropy markov models for information extraction and segmentation," in *Proceedings of the International Conference on Machine Learning (ICML)*, Palo Alto, CA, June 2000.

[4] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proceedings of the Eighteenth International Conference on Machine Learning (ICML)*. 2001, pp. 282–289, Morgan Kaufmann Publishers Inc. San Francisco, CA, USA.

[5] Y. Altun, *Discriminative Methods for Label Sequence Learning*, Ph.D. thesis, Brown University, Department of Computer Science, Providence, RI, 2005.

[6] C. D. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*, MIT Press, 1999.

[7] J. A. Bilmes and K. Kirchhoff, "Factored language models and generalized parallel backoff," in *Proceedings of the Human Language Technology Conference (HLT)-Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, Edmonton, Canada, May 2003.

[8] D. Vergyri, K. Kirchhoff, K. Duh, and A. Stolcke, "Morphology-based language modeling for Arabic speech recognition," in *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, Jeju-Island, Korea, October 2004.

[9] A. Srivastava and F. Kubala, "Sentence boundary detection in Arabic speech," in *Eighth European Conference on Speech Communication and Technology*. ISCA, 2003.

[10] J. Huang and G. Zweig, "Maximum entropy model for punctuation annotation from speech," in *Proceedings of the Seventh International Conference on Spoken Language Processing (ICSLP)*, Denver, CO, 2002.

[11] M. Zimmermann, D. Hakkani-Tür, J. Fung, N. Mirghafori, L. Gottlieb, Y. Liu, and E. Shriberg, "The ICSI+ multi-lingual sentence segmentation system," in *Proceedings of International Conference on Spoken Language Processing (Interspeech)*, Pittsburgh, PA, September 2006.

[12] B. Roark, Y. Liu, M. Harper, R. Stewart, M. Lease, M. Snover, I. Shafran, B. Dorr, J. Hale, A. Krasnyanskaya, and L. Yung, "Reranking for sentence boundary detection in conversational speech," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toulouse, France, 2006.

[13] E. Shriberg, A. Stolcke, D. Hakkani-Tür, and G. Tur, "Prosody-based automatic segmentation of speech into sentences and topics," *Speech Communication*, vol. 32, no. 1-2, pp. 127–154, 2000.

[14] Y. Liu, E. Shriberg, A. Stolcke, B. Peskin, J. Ang, D. Hillard, M. Ostendorf, M. Tomalin, P. Woodland, and M. Harper, "Structural metadata research in the EARS program," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2005.

[15] J. Kolar, J. Svec, and J. Psutka, "Automatic punctuation annotation in Czech broadcast news speech," in *Proceedings of the 9th Conference Speech and Computer*, 2004.

[16] C. Zong and F. Ren, "Chinese utterance segmentation in spoken language translation," in *The 4th International Conference on Computational Linguistics and Intelligent Text Processing*, 2003, pp. 516–525.

[17] D. Liu and C. Zong, "Utterance segmentation using combined approach based on bi-directional n-gram and maximum entropy," in *Proceedings of the ACL Workshop: The Second SIGHAN Workshop on Chinese Language Processing*, 2003, pp. Pages 16–23.

[18] J. Fung, D. Hakkani-Tür, M. Magimai Doss, E. Shriberg, S. Cuendet, and N. Mirghafori, "Prosodic features and feature selection for multi-lingual sentence segmentation," in *Proceedings of International Conference on Spoken Language Processing (Interspeech-Eurospeech)*, Antwerp, Belgium, 2007.

[19] T. Kawahara, M. Saikou, and K. Takanashi, "Automatic detection of sentence and clause units using local syntactic dependency," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toulouse, France, April 2007.

[20] G. Tur, *A Statistical Information Extraction System for Turkish*, Ph.D. thesis, Bilkent University, Department of Computer Science, Ankara, Turkey, 2000.

[21] J. Kolár, E. Shriberg, and Y. Liu, "On speaker-specific prosodic models for automatic dialog act segmentation of multi-party meetings," in *Proceedings of the Ninth International Conference on Spoken Language Processing*. ISCA, 2006.

[22] S. Cuendet, D. Hakkani-Tur, and G. Tur, "Model adaptation for sentence unit segmentation from speech," *Proceedings of SLT, Aruba*, 2006.

[23] E. Matusov, A. Mauser, and H. Ney, "Automatic sentence segmentation and punctuation prediction for spoken language translation," in *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, 2006, pp. 158–165.

[24] S. Rao, I. Lane, and T. Schultz, "Optimizing sentence segmentation for spoken language translation," in *Proceedings of International Conference on Spoken Language Processing (Interspeech-Eurospeech)*, Antwerp, Belgium, 2007.

[25] B. Favre, R. Grishman, D. Hillard, H. Ji, D. Hakkani-Tur, and M. Ostendorf, "Punctuating speech for information extraction," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Las Vegas, NV, 2008.

[26] A. Stolcke and E. Shriberg, "Statistical language modeling for speech dis-

fluencies," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Atlanta, GA, May 1996.

[27] K. W. Church, "A stochastic parts program and noun phrase parser for unrestricted text," in *Proceedings of the Conference on Applied Natural Language Processing (ANLP)*, Austin, Texas, 1988, pp. 136–143.

[28] D. M. Bikel, R. Schwartz, and R. M. Weischedel, "An algorithm that learns what's in a name," *Machine Learning Journal Special Issue on Natural Language Learning*, vol. 34, no. 1-3, pp. 211–231, 1999.

[29] K. Duh and K. Kirchhoff, "Automatic learning of language model structure," in *Proceedings of the International Conference on Computational Linguistics (COLING)*, Geneva, Switzerland, 2004.

[30] A. Stolcke, "SRILM–An extensible language modeling toolkit," in *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, Denver, CO, September 2002.

[31] R. Kneser and H. Ney, "Improved clustering techniques for class-based statistical language modeling," in *Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH)*, Berlin, Germany, 1993.

[32] Y. Liu, A. Stolcke, E. Shriberg, and M. Harper, "Using conditional random fields for sentence boundary detection in speech," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, Ann Arbor, MI, 2005.

[33] S. Cuendet, D. Hakkani-Tür, and G. Tur, "Model adaptation for sentence segmentation from speech," in *Proceedings of the IEEE/ACL Spoken Language Technologies (SLT) Workshop*, Aruba, 2006.

[34] R. E. Schapire and Y. Singer, "Boostexter: A boosting-based system for text categorization," *Machine Learning*, vol. 39, no. 2/3, pp. 135–168, 2000.

[35] S. Cuendet, D. Hakkani-Tür, E. Shriberg, J. Fung, and B. Favre, "Cross-genre feature comparisons for spoken sentence segmentation," in *Proceedings of the IEEE International Conference on Semantic Computing (ICSC)*, Irvine, California, 2007.

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

[37] J. Hankamer, In W. Marslen-Wilson, editor, *Lexical Representation and Process*, chapter Morphological Parsing and the Lexicon, The MIT Press, 1989.

[38] D. Hakkani-Tür, K. Oflazer, and G. Tur, "Statistical morphological disambiguation for agglutinative languages," *Computers and the Humanities*, vol. 36, no. 4, 2002.

[39] K. Oflazer, "Dependency parsing with an extended finite state approach," in *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL)*, Maryland, USA, 1999.

[40] E. Arisoy, H. Sak, and M. Saraclar, "Language modeling for automatic Turkish broadcast news transcription," in *Proceedings of International Conference on Spoken Language Processing (Interspeech-Eurospeech)*, Antwerp, Belgium, 2007.

[41] Y. Liu, E. Shriberg, A. Stolcke, D. Hillard, M. Ostendorf, and M. Harper, "Enriching speech recognition with automatic detection of sentence boundaries and disfluencies," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1526–1540, September 2006.

PLACE PHOTO HERE

Umit Guz graduated from the Department of Computer Programming, Yildiz Technical University, Istanbul, Turkey in 1990. He received the B.S. degree with high honors from the Department of Electronics Engineering, College of Engineering, Istanbul University, Istanbul, Turkey in 1994. He received M.S. and Ph.D. degrees in Electronics Engineering with high honors from the Institute of Science, Istanbul University in 1997 and 2002, respectively. From 1995 to 1998 he was a research and teaching assistant in the Department of Electronics Engineering, Istanbul University. He has been an instructor in the Department of Electronics Engineering, Engineering Faculty, Isik University, Istanbul, Turkey since 1998. He was awarded a Post-Doctoral Research Fellowship by The Scientific and Technical Research Council of Turkey (TUBITAK) in 2006. He was accepted as an International Fellow by the SRI International Speech Technology and Research (STAR) Laboratory in 2006. He was awarded a J. William Fulbright Post-Doctoral Research Fellowship for 2007. He was accepted as an International Fellow by the International Computer Science Institute (ICSI) Speech Group at the University of California at Berkeley for 2007. His research interest covers speech processing, speech modeling, speech coding, speech compression, automatic speech recognition, natural language processing, and biomedical signal processing.

PLACE PHOTO HERE

Benoit Favre received his B.S., M.S., and Ph.D. degrees from the University of Avignon, France, in 2001, 2003, and 2007. He was a teaching assistant at the University of Avignon from 2003 to 2007. He was also a research engineer at Thales Land&Joint Systems in Paris, France from 2004 to 2007. He currently holds a post-doc position at the International Computer Institute (ICSI), Berkeley, California. His Ph.D. thesis explores interactive speech summarization of broadcast news archives; his research interests include Natural Language Processing, Speech Understanding, and Text and Speech Summarization. He is also interested in Machine Learning on structured outputs and global inference. He is a member of ISCA and IEEE, and was a member and webmaster of AFCP.

PLACE PHOTO HERE

Dilek Hakkani-Tür is a senior researcher at ICSI. Prior to joining ICSI, she was a senior technical staff member in the Voice Enabled Services Research Department at AT&T Labs-Research in Florham Park, New Jersey. She received her B.Sc. degree from Middle East Technical University in 1994, and M.Sc. and Ph.D. degrees from Bilkent University, Department of Computer Engineering in 1996 and 2000, respectively. Her Ph.D. thesis is on statistical language modeling for agglutinative languages. She worked on machine translation during her visit to Carnegie Mellon University, Language Technologies Institute in 1997, and her visit to Johns Hopkins University, Computer Science Department, in 1998. In 1998 and 1999, she visited SRI International, Speech Technology and Research Laboratory, and worked on using lexical and prosodic information for information extraction from speech. In 2000, she worked in the Natural Sciences and Engineering Faculty of Sabanci University, Turkey. Her research interests include natural language and speech processing, spoken dialog systems, and active and unsupervised learning for language processing. She has co-authored several papers in natural language and speech processing. She is a member of ISCA, IEEE, Association for Computational Linguistics and was an associate editor of *IEEE Transactions on Audio, Speech and Language Processing* from 2005 to 2008.

PLACE PHOTO HERE

Gokhan Tur received his B.S., M.S., and Ph.D. degrees from the Department of Computer Science, Bilkent University, Turkey in 1994, 1996, and 2000, respectively. From 1997 to 1999, he visited the Center for Machine Translation of CMU, then the Department of Computer Science of Johns Hopkins University, and then the Speech Technology and Research Laboratory of SRI International. He worked at AT&T Labs - Research from 2001 to 2006. He is currently with the Speech Technology and Research Lab of SRI International. His research interests include spoken language understanding (SLU), speech and language processing, machine learning, and information retrieval and extraction. He co-authored more than 60 papers published in refereed journals and presented at international conferences. Dr. Tur is also the recipient of the Speech Communication Journal Best Paper awards by ISCA for 2004-2006 and by EURASIP for 2005-2006. Dr. Tur is the organizer of the HLT-NAACL 2007 Workshop on Spoken Dialog Technologies, and the HLT-NAACL 2004 and AAAI 2005 Workshops on SLU, and the editor of the *Speech Communication Journal* Special Issue on SLU in 2006. He is also the spoken language processing area chair for IEEE ICASSP 2007 and IEEE ICASSP 2008 conferences, spoken dialog area chair for HLT-NAACL 2007 conference, finance chair for IEEE/ACL SLT 2006 workshop, and SLU area chair for IEEE ASRU 2005 workshop. Dr. Tur is a senior member of IEEE, ACL, and ISCA, and a member of the IEEE Signal Processing Society (SPS), Speech and Language Technical Committee (SLTC) for 2006-2008.