Statistical Modeling of Turkish for Automatic Topic Segmentation

Gökhan Tür Dilek Z. Hakkani-Tür Kemal Oflazer

Department of Computer Engineering
Bilkent University, Bilkent, Ankara, TR-06533, TURKEY
{tur,hakkani,ko}@cs.bilkent.edu.tr

January 5, 2000

Abstract

This paper presents the firstever study on statistical segmentation of Turkish text into topics. Our results indicate that it is possible to overcome the problems due to the highly agglutinative nature of Turkish by building a statistical model based on the morphological analyses of the words instead of the surface forms. We have achieved 10.90% segmentation error rate on our test set according to the weighted TDT2 segmentation cost metric. This is 32% better than the word-based baseline model.

1 Introduction

Topic segmentation is the task of automatically dividing a stream of text or speech into topically homogeneous blocks. Given a sequence of (written or spoken) words, the aim of topic segmentation is to find the boundaries where topics change.

Topic segmentation is an important task for various language understanding applications, such as information extraction and retrieval (IR). An application may be as follows: Given a corpus of newspaper articles strung together, and a user query, return a collection of coherent segments matching the query. Lacking a tool for detecting topic breaks, an IR application may be able to locate positions in its database, but be unable to determine how much of the surrounding data to provide to the user. Similarly in video-on-demand applications, there is no mark-up to indicate the topic boundaries. Also, segmenting text along topic boundaries may be useful for text summarization and anaphora resolution [Kozima, 1993]. Figure 1 gives an example of a topic change boundary from a broadcast news transcript.

There has recently been increased interest in segmenting such information streams into topics. In 1997, the U.S. Defense Advanced Research Projects Agency (DARPA) initiated the Topic Detection and Tracking (TDT) Program [Allan *et al.*, 1998]. The purpose of this effort is to advance and accurately measure the state of the

> ... tens of thousands of people are homeless in northern china tonight after a powerful earthquake hit an earthquake registering six point two on the richter scale at least forty seven people are dead few pictures available from the region but we do know temperatures there will be very cold tonight minus seven degrees **<TOPIC_CHANGE>** peace talks expected to resume on monday in belfast northern ireland former u. s. senator george mitchell is representing u. s. interests in the talks but it is another american center senator rather who was the focus of attention in northern ireland today here's a. b. c.'s richard gizbert ...

Figure 1: An example of a topic boundary in a broadcast news word transcript.

art in TDT and to assess the technical challenges to be overcome. In the framework of this program, topic segmentation is an *enabling* technology for other applications, such as topic tracking and new event detection.

Past topic segmentation systems have depended mostly on the lexical information, i.e., they all have used surface forms of the words. However, in morphologically rich languages, like Turkish, a sequence of inflectional and derivational morphemes can be added to a word root [Oflazer, 1993]. The number of word forms one can derive from a root form may be in the millions [Hankamer, 1989]. Thus the number of distinct word forms is much larger than that of languages like English For instance, the derived modifier *sağlamlaştırdığımızdaki*[1] would be morphologically decomposed as:

sağlam+laş+tır+dı+ğı+mız+da+ki

and morphologically analyzed as:[2]

```
strong+Adj^DB
+Verb+Become^DB
+Verb+Caus+Pos^DB
+Adj+PastPart+P1sg^DB
+Noun+Zero+A3sg+Pnon+Loc^DB
+Adj
```

When employing statistical techniques, using the surface forms of the words would result in data sparseness in agglutinative languages, such as Turkish. Words with different inflectional and derivational suffixes must not be considered different, when trying to segment Turkish text. Furthermore, roots are more informative than the full words. Note that, stemming words is a common practice in other fields, such as information retrieval [Porter, 1980].

Although in recent years there have been numerous studies in processing Turkish text, we are not aware of any studies of developing a topic segmentation system for Turkish. Besides this, corpus-based supervised statistical methods have never been studied on Turkish text processing except a forthcoming Ph.D. thesis on statistical language modeling of Turkish [Hakkani-Tür, 2000 forthcoming].

In the next section, we review previous work on topic segmentation. In Section 3, we describe our methodology of incorporating morphological analyses of the words in processing Turkish. Section 4 reports on experimental procedures and results.

2 Previous Work

Most automatic topic segmentation work based on text sources has explored topical word usage cues in one form or other. Kozima [1993] uses mutual similarity of words in a sequence of text as an indicator of text structure. Reynar [1994] presents a method which finds topically similar regions in the text by graphically modeling the distribution of word repetitions. Hearst [1994] uses cosine similarity in a word vector space as an indicator of topic similarity.

Several participating systems in the TDT-Pilot Study rely essentially on word usage: Ponte and Croft [1997] extract related word sets for topic segments with the information retrieval technique of local context analysis (LCA), and then compare the expanded word sets. Yamron *et al.* [1998] model topics with unigram language models (LMs) and their sequential structure with hidden Markov models (HMMs). The overall structure of the

[1] Literally, "(the thing existing) at the time we caused (something) to become strong".

[2] The unobvious morphological features used in this paper are: `^DB`: Derivation boundary, `Adj`: Adjective, `Det`: Determiner, `Become`: Become verb, `Caus`: Causative verb, `Past`: Past tense marker, `PastPart`: Derived past participle, `Pnon`: No possessive agreement, `P1sg`: First person singular possessive agreement, `P2sg`: Second person singular possessive agreement, `P3sg`: Third person singular possessive agreement, `P3pl`: Third person plural possessive agreement, `A3sg`: Third person singular agreement, `A3pl`: Third person plural agreement, `Zero`: Zero derivation with no overt derivation, `Nom`: Nominative case, `Loc`: Locative case, `Acc`: Accusative case, `Dat`: Dative case, `Abl`: Ablative case, `Gen`: Genitive case, `Ins`: Instrumantive case, `Pos`: Positive polarity, `Prop`: Proper Name, `Agt`: Agent.

model is that of an HMM [Rabiner and Juang, 1986] in which the states correspond to topic clusters T_j, and the observations are sentences (or chopped units) $W_1, \ldots, W_N$. The resulting HMM, depicted in Figure 2, forms a complete graph, allowing for transitions between any two topic clusters. The observation likelihoods for the HMM states, $P(W_i|T_j)$, represent the probability of generating a given sentence W_i in a particular topic cluster T_j. 100 topic cluster LMs are automatically constructed, using the multi-pass k-means algorithm [Hartigan and Wong, 1979]. Since HMM emissions are meant to model the topical usage of words, but not topic-specific syntactic structures, the LMs consist of unigram distributions that exclude stop words (high-frequency function and closed-class words). To account for unobserved words they interpolate the topic cluster-specific LMs with the global unigram LM obtained from the entire training data. The observation likelihoods of the HMM states are then computed from these smoothed unigram LMs. All HMM transitions within the same topic cluster are given probability one, whereas all transitions between topics are set to a global *topic switch penalty* (TSP) which is optimized on the development data. The TSP parameter allows trading off between false alarms and misses. Once the HMM is trained, the Viterbi algorithm [Viterbi, 1967] is used to search for the best state sequence and corresponding segmentation.[3]

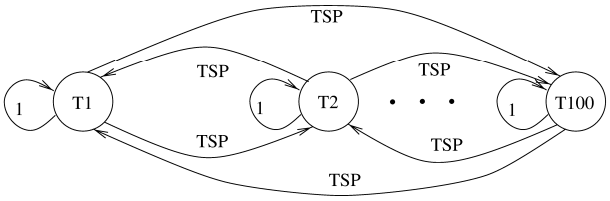

Figure 2: Structure of the basic HMM developed by Dragon for the TDT Pilot Study. The labels on the arrows indicate the transition probabilities. TSP represents the topic switch penalty.

Previous work on both text and speech has found that cue phrases or discourse particles (items such as "now" or "by the way"), as well as other lexical cues, can provide valuable indicators of structural units in discourse. The University of Massachusetts "HMM" approach described in the TDT Pilot Study Report [1998] uses an HMM that models the initial, middle, and final sentences of a topic segment. At CMU, Beeferman *et al.* [1999] combined a large set of automatically selected lexical discourse cues in a maximum-entropy model.

In our previous work, we have successfully combined lexical and prosodic cues for automatic topic segmentation of speech [Stolcke *et al.*, 1999; Tür *et al.*, 1999; Shriberg *et al.*, 2000; Hakkani-Tür *et al.*, 1999]. Topical word usage and lexical discourse cues are modeled by language models embedded in a hidden Markov model. Prosodic discourse cues, such as pause durations and pitch resets, are modeled by a decision tree based on automatically extracted acoustic features and alignments. Lexical and prosodic features can be combined either in the HMM or in the decision tree framework.

3 The Approach

Assuming that topics do not change in mid-sentence, our aim can be stated as grouping sentences into contiguous stretches belonging to one topic, i.e., sentence boundaries are classified into "topic boundaries" and "nontopic boundaries". Topic segmentation is thus reduced to a boundary classification problem. In order to do this, we used an extension of the Dragon's approach, explained in Section 2. This approach is based purely on topical word distributions. We extended it to also handle morphological complications of Turkish, using stems of the words and then using only nouns in forming the topic clusters.

[3]Note that the transition probabilities in the model are not normalized to sum to one; this is convenient and permissible since the output of the Viterbi algorithm depends only on the relative weight of the transition weights.

Word	Freq	Meaning
gol	1222	goal
ikinci	912	second
Beşiktaş	867	Beşiktaş
teknik	781	technical
Galatasaray	773	Galatasaray
Fenerbahçe	699	Fenerbahçe
orta	678	middle
takım	665	team
dk.	655	min.
sarı	622	yellow
maç	592	match
yarıda	575	half (Loc)
top	521	ball
Trabzonspor	479	Trabzonspor
yaptı	473	did
Mehmet	471	Mehmet
Hakan	462	Hakan
dakikada	450	minute (Loc)
maçı	449	match (Acc)
futbol	445	football
Fatih	413	Fatih
yarı	412	half
oyun	406	game
Ali	384	Ali

Table 1: The most frequent words in one of the clusters, containing mostly football news articles

3.1 Word-based Modeling

In order to gauge our baseline performance, we constructed a word-based model, and then automatically constructed 100 topic cluster LMs, using the multi-pass k-means algorithm. This model enables us to evaluate our extended models.

Table 1 gives a list of the most frequent words in one of the clusters, containing mostly football news articles. *Beşiktaş*, *Galatasaray*, *Fenerbahçe*, and *Trabzonspor* are top Turkish football teams, *Hakan*, *Mehmet*, and *Ali* are the top players, and *Fatih Terim* is the trainer of *Galatasaray*.

3.2 Stem-based Modeling

Word-based modeling works well in languages in which there is very little or no morphology, such as English. On the other hand, morphologically rich languages, like Turkish, suffer from the fact that the number of word forms one can derive from a Turkish root form may be in the millions [Hankamer, 1989]. Because of this reason, the number of distinct word forms is much larger than that of English. Table 2 gives a list of 26 different forms of the stem word *gol* (goal), in the cluster mentioned in Table 1.

More specifically, word-based approach suffers from this characteristic of Turkish in two ways:

1. Using the surface forms of the words results in data sparseness in the training data. This sparseness badly damages the performance of the clustering algorithm, hence the quality of the language models.

2. The second drawback of using a word-based model is that, while segmenting, using the surface forms of the words leads to a lower performance. For example, a word with an unseen inflectional or derivational form would not contribute to the statistical computation, even if its stem is in the vocabulary.

Word	Freq	Morphological Analysis
gol	1222	goal+Noun+A3sg+Pnon+Nom
golü	350	goal+Noun+A3sg+Pnon+Acc or goal+Noun+A3sg+P3sg+Nom
gole	150	goal+Noun+A3sg+Pnon+Dat
golle	138	goal+Noun+A3sg+Pnon+Ins
goller	126	goal+Noun+A3pl+Pnon+Nom
golde	85	goal+Noun+A3sg+Pnon+Loc
golün	75	goal+Noun+A3sg+Pnon+Gen or goal+Noun+A3sg+P2sg+Nom
golünü	63	goal+Noun+A3sg+P3sg+Acc or goal+Noun+A3sg+P2sg+Acc
golüyle	62	goal+Noun+A3sg+P3sg+Ins
golcü	59	goal+Noun+A3sg+Pnon+Nom^DB +Adj+Agt
golleri	48	goal+Noun+A3pl+P3sg+Nom or goal+Noun+A3pl+Pnon+Acc or goal+Noun+A3pl+P3pl+Nom or goal+Noun+A3sg+P3pl+Nom
golden	45	goal+Noun+A3sg+Pnon+Abl
gollerle	40	goal+Noun+A3pl+Pnon+Ins
gollük	37	goal+Noun+A3sg+Pnon+Nom^DB +Adj+FitFor
gollü	26	goal+Noun+A3sg+Pnon+Nom^DB +Adj+With
golüne	24	goal+Noun+A3sg+P3sg+Dat or goal+Noun+A3sg+P2sg+Dat
golleriyle	20	goal+Noun+A3pl+P3sg+Ins or goal+Noun+A3pl+P3pl+Ins or goal+Noun+A3sg+P3pl+Ins
golsüz	18	goal+Noun+A3sg+Pnon+Nom^DB +Adj+Without
golcüsü	18	goal+Noun+A3sg+Pnon+Nom^DB +Noun+Agt+A3sg+P3sg+Nom
golünde	16	goal+Noun+A3sg+P3sg+Loc or goal+Noun+A3sg+P2sg+Loc
gollerde	15	goal+Noun+A3pl+Pnon+Loc
goldeki	15	goal+Noun+A3sg+Pnon+Loc^DB +Det
gollerin	12	goal+Noun+A3pl+Pnon+Gen or goal+Noun+A3pl+P2sg+Nom
golünden	10	goal+Noun+A3sg+P3sg+Abl or goal+Noun+A3sg+P2sg+Abl
gollerini	9	goal+Noun+A3pl+P3sg+Acc or goal+Noun+A3pl+P2sg+Acc or goal+Noun+A3pl+P3pl+Acc or goal+Noun+A3sg+P3pl+Acc
gollere	8	goal+Noun+A3pl+Pnon+Dat

Table 2: The frequency table for the root word *gol* (goal) in the cluster mentioned in Table 1.

Word	Freq	Meaning
gol	2271	goal
maç	2048	match
oyun	1781	game
takım	1382	team
ol	1317	be
oy[4]	1273	vote
al	1264	take
top	1228	ball
futbol	1227	football
oyna	1224	play
yap	1219	do or make
yarı	1101	half
Galatasaray	1018	Galatasaray
sahıa	996	field
Hakan	986	Hakan
Beşiktaş	974	Beşiktaş
at	948	throw
dakika	892	minute
Fenerbahçe	872	Fenerbahçe
rakip	866	opponent
çık	826	exit
orta	785	middle
et	755	do or make
ikinci	734	second

Table 3: The most frequent stems in a cluster, containing mostly football news articles.

It is clear that, removing the suffixes of the words, and using the roots of the words would reduce the data sparseness significantly, and the unigram language models obtained from the topic clusters would be more effective. So we decided to use the root words instead of the surface forms of the words, and build stem-based language models, instead of word-based language models.

In order to do this, we used a preprocessing module, developed by Hakkani-Tür [2000 forthcoming], which tokenizes the training data, analyzes the tokens using the morphological analyzer developed by Oflazer [1993], groups the collocations, and finally removes some obviously improbable morphological parses in order to reduce the morphological ambiguity. Then, we extracted the roots of the words, and rebuilt the training corpus using only these roots. When there were more than one root for a word, we used all of the roots. However, this ambiguity was not a real problem as there were only 1.15 distinct roots per word on the average.

As expected, we obtained clusters with smaller number of words, and with higher frequencies. Table 3 lists the most frequent words in a corresponding cluster, containing mostly football news articles.

3.3 Noun-based Modeling

When we analyzed Table 3, and other clusters, we saw that in order to model the topical usage of words, more than just stopwords could have been excluded. In fact, only nouns would be sufficient to model the topics. Since we had the morphological analyses of the words, it was straightforward for us to test this hypothesis.

Instead of using the stems of words, we only used the stems of the morphological parses that have a noun root form. After using the same clustering algorithm, we ended up with new clusters. The most frequent nouns for the cluster containing mostly football related articles is listed in Table 4. Common verbs such as, *ol* (be) , *al* (take), *yap* (make), and *et* (do) and somewhat football related verbs, such as *oyna* (play), *çık* (exit), and *at* (score) disappeared in Table 4 when we compare with Table 3.

[4]Note that the frequent word *oyun* (game) has another morphological parse, meaning "your vote", hence the appearance of the

Word	Freq	Meaning
gol	2562	goal
maç	2412	match
oyun	2071	game
takım	1659	team
futbol	1492	football
oy	1429	vote
yarı	1275	half
top	1257	ball
Galatasaray	1230	Galatasaray
Beşiktaş	1201	Beşiktaş
saha	1189	field
Fenerbahçe	1162	Fenerbahçe
dakika	1029	minute
orta	868	middle
rakip	852	opponent
lig	695	league
kale	657	goal
dk.	642	min.
pozisyon	638	position
hata	606	error
teknik	594	technical
Hakan	579	Hakan
hakem	543	referee
alan	541	space or field

Table 4: The most frequent nouns in a cluster, containing mostly football news articles.

4 Experiments and Results

To evaluate our models we carried out experiments in the TDT paradigm. We first describe our training and test data, then give results obtained with the baseline word-based, stem-based, and noun-based language models. We used SRILM toolkit for language modeling and decoding [Stolcke, 1999]. In our work, we assumed that each news piece contains only one topic, and tried to find out article boundaries. Hand-checking of a subset of articles showed that this assumption was true except for a few cases.

4.1 Training Data

Topic unigram language models were trained from the web resources of Milliyet newspaper articles, covering the period from January 1, 1997 through September 12, 1998. For training the language models we removed stories with fewer than 300 and more than 3,000 words, leaving 14,495 stories with an average length of 432 words, 500 stems, or 310 nouns, excluding stop words. Accordingly, in this training data, there are 376,371 distinct words, 128,125 distinct stems, or 119,475 distinct nouns.

4.2 Test Data

We evaluated our system on a test set of 100 news articles, covering the period from September 12, 1998 through September 14, 1998, comprising 2,803 sentences, 32,772 words, 38,329 stems, or 24,807 nouns, excluding stopwords. The topic switch penalty was optimized on the development set of 99 news articles from the same newspaper, between September 14, 1998 and September 16, 1998, including 3,180 sentences, 33,728 words, 39,106 stems, or 25,615 nouns, excluding stopwords.

root *oy* (vote).

Model	Development Set			Test Set		
	P_{Miss}	$P_{FalseAlarm}$	C_{Seg}	P_{Miss}	$P_{FalseAlarm}$	C_{Seg}
Chance	1.0000	0.0000	0.3000	1.0000	0.0000	0.3000
Human Performance	0.2093	0.0176	0.0742	N/A	N/A	N/A
Word-based	0.4394	0.0658	0.1779	0.3560	0.0752	0.1594
Word-based (Random)	0.3412	0.0286	0.1224	0.3840	0.0427	0.1451
Stem-based	0.2704	0.0655	0.1270	0.2552	0.0708	0.1261
Noun-based	**0.2627**	**0.0413**	**0.1077**	**0.2487**	**0.0492**	**0.1090**

Table 5: Summary of error rates with different language models. A "chance" classifier that labels all potential boundaries as non-topic would achieve 0.3 weighted segmentation cost. "Random" indicates that the articles are shuffled.

4.3 Evaluation metrics

We have adopted the evaluation paradigm used by the TDT2—Topic Detection and Tracking Phase 2 [Doddington, 1998] program, allowing fair comparisons of various approaches both within this study and with respect to other recent work. Segmentation accuracy was measured using TDT evaluation software from NIST, which implements a variant of an evaluation metric suggested by Beeferman *et al.* [1999]. The evaluation metric reflects the probability that two positions in the corpus probed at random and separated by a distance of k words are correctly classified as belonging to the same story or not. If the two words belong to the same topic segment, but are erroneously claimed to be in different topic segments by the segmenter, then this will increase the system's *false alarm* probability. Conversely, if the two words are in different topic segments, but are erroneously marked to be in the same segment, this will contribute to the *miss* probability. The false alarm and miss rates are defined as averages over all possible probe positions with distance k. In the TDT-2 program, k is a constant and equals 50. Miss and false alarm probabilities are combined into a single *segmentation cost* metric

$$C_{Seg} = P_{Miss} \times P_{seg} + P_{FalseAlarm} \times (1 - P_{seg})$$

where $P_{Seg} = 0.3$ is the *a priori* probability of a segment being within an interval of k words on the TDT2 training corpus.

4.4 Segmentation Results

Table 5 shows the results of the Turkish topic segmenter, using word-based, stem-based, and noun-based approaches.

These results are consistent with the rationale given in the previous section. As expected, the word-based model suffered from data sparseness, and 28.61% improvement is achieved for the development set when we use the stems of the words. Furthermore, it is possible to obtain 15.19% more improvement using only nouns, achieving a total of 39.46% improvement over our baseline word-based model. For the test set, the results are also similar, and we achieve 20.89% improvement when we have used the stem-based approach, and our results are 31.61% better when we have used the noun-based approach.

Comparing these three modeling approaches, we observe that stem-based and noun-based models have a 38%-40% lower miss probability than the word-based model in the development data. This rate is 28%-30% in the test set. This enormous decrease in the miss probability is the main reason of the final improvement. We would say that, using stems of the words or nouns, we have obtained more discriminative topic unigram language models in the clustering phase, hence we have missed fewer topic boundaries. Additionally, when we have used the noun-based models, we see that there is a 31%-37% improvement over the stem-based models in the false

Word	Morphological Analysis	Word-based Probability	Stem-based Probability	Noun-based Probability
Son	`Last+Adj`	0	0	0
dakikalarda	`minute+Noun+A3pl+Pnon+Loc`	0.000337	0.004930	0.007296
Galatasaray'ın	`Galatasaray+Noun+Prop+A3sg+Pnon+Gen`	0.001433	0.005598	0.008679
atakları	`attack+Noun+A3pl+P3sg+Nom`	0.000072	0.001192	0.001600
sıklaştı	`frequent+Adj^DB+Verb+Become+Pos+Past+A3sg`	0	0.000557	0
Hakan	`Hakan+Noun+Prop+A3sg+Pnon+Nom`	0.002556	0.005422	0.004087
attığı	`score+Verb+Pos^DB+Adj+PastPart+P3sg`	0.001232	0.005458	0
golle	`goal+Noun+A3sg+Pnon+Ins`	0.000760	0.012454	0.018019
ağları	`net+Noun+A3pl+Pnon+Acc`	0.000138	0.000428	0.000595
sarstı	`shake+Verb+Pos+Past+A3sg`	0.000001	0.000127	0

Table 6: The unigram probabilities of the words in the example sentence. Note that the word *son* (last) is a stopword, hence gets 0 probability.

Corpus	P_{Miss}	$P_{FalseAlarm}$	C_{Seg}
Turkish	0.4394	0.0658	0.1779
English	0.4685	0.0817	0.1978

Table 7: Word-based segmentation error rates for English and Turkish corpora.

alarm probabilities.

Let's analyze these results using a concrete example. Consider the following sentence from an article on football: *Son dakikalarda Galatasaray'ın atakları sıklaştı, Hakan attığı golle ağları sarstı.*[5] Table 6 shows the individual unigram probabilities of the words in a cluster including mainly football news articles for both word-based and stem-based approaches. Note that, due to data sparseness, all of these words, though related with football have less probability when compared to stem-based and noun-based models. Furthermore, the word *sıklaştı* (became frequent) received 0 probability, since its surface form is unseen in the training data, although its stem *sık* (frequent) gets some probability.

When we analyze our errors, we see that errors are made when there are topically very similar news articles in a sequence, or when an article contains more than one topic, though this second case is less likely. This is why we obtained better performance on the test set than the development set for both word-based and stem-based models, although we set the topic switch penalty on the development set. When we analyzed this, we see that development set is *harder* to segment than the test set, in the sense that it includes articles with very similar consecutive topics. Note that because of this, the miss probability of a human annotator is about 20%. When we ordered the articles randomly, this difference disappeared.

It would be useful to provide word-based segmentation error rates obtained from a recent work [Tür *et al.*, 1999] for English Broadcast News corpus. As shown in Table 7, the two test sets have comparable behavior. Stem-based and noun-based models are not available for English. It would be interesting to try these approaches for English, too.

5 Conclusions

We have presented a probabilistic model for automatically segmenting Turkish text into topically homogeneous blocks. We tried three different approaches to model topics so that we can overcome the problems arising from

[5] Literally, "In the last minutes, Galatasaray's attacks became more frequent, Hakan shook the net with his goal."

the agglutinative nature of Turkish. First, we tried a baseline model, using only the surface forms of the words, then we have modeled the stems of the words, and obtain a huge win. Finally we modeled only the stems of the nouns, and reached 10.90% segmentation error rate according to the weighted TDT2 segmentation cost metric on our test set, which was 32% better than the baseline model.

These results are important in the sense that, statistical methods have been largely ignored for processing Turkish. Mainly due to the agglutinative nature of Turkish words and the structure of Turkish sentences, the construction of a language model for Turkish can not be directly adapted from English. It is necessary to incorporate some other techniques. This work is a preliminary step in the application of corpus-based statistical methods to Turkish text processing.

6 Acknowledgments

This work was begun while the first two authors were visiting Speech Technology and Research Laboratory, SRI International, with support from DARPA under contract no. N66001-97-C-8544 and from NSF under grant IRI-9619921. We thank Andreas Stolcke and Elizabeth Shriberg for many helpful discussions.

References

[Allan *et al.*, 1998] Allan, J.; Carbonell, J.; Doddington, G.; Yamron, J.; and Yang, Y. 1998. Topic detection and tracking pilot study: Final report. In *Proceedings DARPA Broadcast News Transcription and Understanding Workshop*, Lansdowne, VA. 194–218.

[Beeferman *et al.*, 1999] Beeferman, Doug; Berger, Adam; and Lafferty, John 1999. Statistical models for text segmentation. *Machine Learning* 34(1-3):177–210. Special Issue on Natural Language Learning.

[Doddington, 1998] Doddington, George 1998. The Topic Detection and Tracking Phase 2 (TDT2) evaluation plan. In *Proceedings DARPA Broadcast News Transcription and Understanding Workshop*, Lansdowne, VA. 223–229.

[Hakkani-Tür *et al.*, 1999] Hakkani-Tür, Dilek; Tür, Gökhan; Stolcke, Andreas; and Shriberg, Elizabeth 1999. Combining words and prosody for information extraction from speech. In *Proceedings of the 6th European Conference on Speech Communication and Technology*, volume 5, Budapest. 1991–1994.

[Hakkani-Tür, 2000 forthcoming] Hakkani-Tür, Dilek Z. oming. *Statistical Language Modeling of Turkish.* Ph.D. Dissertation, Department of Computer Engineering, Bilkent University, Ankara, Turkey.

[Hankamer, 1989] Hankamer, Jorge 1989. Lexical representation and process. In Marslen-Wilson, W., editor 1989, *Morphological Parsing and the Lexicon.* The MIT Press.

[Hartigan and Wong, 1979] Hartigan, J. A. and Wong, M. A. 1979. A k-means clustering algorithm. *Applied Statistics* 28:100–108.

[Hearst, 1994] Hearst, M. A. 1994. Multi-paragraph segmentation of expository text. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, New Mexico State University, Las Cruces, NM. 9–16.

[Kozima, 1993] Kozima, H. 1993. Text segmentation based on similarity between words. In *Proceedings of the 31rd Annual Meeting of the Association for Computational Linguistics*, Ohio State University, Columbus, Ohio. 286–288.

[Oflazer, 1993] Oflazer, Kemal 1993. Two-level Description of Turkish Morphology. *Literary and Linguistic Computing* 8(3).

[Ponte and Croft, 1997] Ponte, J. M. and Croft, W. B. 1997. Text segmentation by topic. In *Proceedings of the First European Conference on Research and Advanced Technology for Digital Libraries*, Pisa, Italy.

[Porter, 1980] Porter, M. F. 1980. An algorithm for suffix stripping. *Program* 14:130–137.

[Rabiner and Juang, 1986] Rabiner, L. R. and Juang, B. H. 1986. An introduction to hidden Markov models. *IEEE ASSP Magazine* 3(1):4–16.

[Reynar, 1994] Reynar, Jeffrey C. 1994. An automatic method of finding topic boundaries. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, New Mexico State University, Las Cruces, NM. 331–333.

[Shriberg *et al.*, 2000] Shriberg, Elizabeth; Stolcke, Andreas; Hakkani-Tür, Dilek; and Tür, Gökhan 2000. Prosody-based automatic segmentation of speech into sentences and topics. *Speech Communication.* to appear.

[Stolcke *et al.*, 1999] Stolcke, Andreas; Shriberg, Elizabeth; Hakkani-Tür, Dilek; Tür, Gökhan; Rivlin, Ze'ev; and Sönmez, Kemal 1999. Combining words and speech prosody for automatic topic segmentation. In *Proceedings DARPA Broadcast News Workshop*, Herndon, VA. 61–64.

[Stolcke, 1999] Stolcke, Andreas 1999. SRILM—the SRI language modeling toolkit. http://www.speech.sri.com/projects/srilm/.

[Tür *et al.*, 1999] Tür, Gökhan; Hakkani-Tür, Dilek; Stolcke, Andreas; and Shriberg, Elizabeth 1999. Integrating prosodic and lexical cues for automatic topic segmentation. Submitted.

[Viterbi, 1967] Viterbi, A. 1967. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory* 13:260–269.

[Yamron *et al.*, 1998] Yamron, J.P.; Carp, I.; Gillick, L.; Lowe, S.; and Mulbregt, P.van 1998. A hidden Markov model approach to text segmentation and event tracking. In *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing*, volume 1, Seattle, WA. 333–336.