Cross-Linguistic Analysis of Prosodic Features for Sentence Segmentation

James G Fung[1], *Dilek Hakkani-Tur*[1], *Mathew Magimai Doss*[1], *Liz Shriberg*[1], *Sebastien Cuendet*[1], *Nikki Mirghafori*[1]

[1]International Computer Science Institute, University of California Berkeley, USA

(jgf, dilek, mathew, ees, cuendet, nikki)@eecs.berkeley.edu

10.21437/Interspeech.2007-686

Abstract

In this paper, we perform a cross-linguistic study of prosodic features in sentence segmentation by using two different feature selection approaches: a forward search wrapper and feature filtering. Experiments in Arabic, English, and Mandarin show that prosodic features make significant contributions in all three languages. Feature selection results indicate that feature relevancy can vary greatly depending on the target language, and therefore the optimal feature subset varies considerably between languages. We observe patterns in the feature selection and the affinity of the different languages toward certain feature types, which gives us insight into future feature selection and feature design.

Index Terms: prosodic features, cross lingual, feature selection, sentence segmentation

1. Introduction

The role of sentence segmentation is to break the stream of words provided by automatic speech recognition into sentences for further processing by downstream language processing tasks, such as parsing, machine translation, question answering, etc.

Sentence boundary detection in speech has been studied in an attempt to enrich speech recognition output [1]. In the previous approaches for this task, different classifiers have been evaluated (e.g, HMM, maximum entropy), utilizing both textual and prosodic information. In the DARPA EARS program, special efforts were made for rich transcription of speech with automatically generated structural information, including sentence boundaries, disfluencies, and filler words. For example, [1] 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. A reranking technique [2] further improved sentence boundary detection performance upon the baseline of [1].

State-of-the-art sentence segmentation systems use different kinds of features, including lexical, prosodic, speaker turn-related, and syntactic. Using all of these features usually results in a large dimensional feature vector and can also result in performance degradation. Furthermore, the set of features which work best for different languages or genres can vary.

In this work, we use the results of feature selection experiments to see how well prosodic features originally design for English port to other languages in order to gain insight for future work in these languages and for the design of the next generation of prosodic features.

2. Sentence segmentation system

2.1. Task

The ICSI sentence segmentation system receives, as input, word time alignments from a transcript. The task is to classify each word-final boundary as either a sentence or non-sentence boundary. For conversational speech, sentence boundaries can further be divided into more specific categories that reflect dialog acts, such as statement, question, backchannel, disruption, and floor-grabber / floor-holder. However, in the broadcast news domain, many of these rarely occur. Thus little richness is lost by restricting the task to a binary classification problem.

2.2. Learning algorithm

The learning algorithm used for this study is BoosTexter, a member of the family of boosting algorithms [3]. Boosting algorithms combine multiple weak classifiers into a single strong classifier. The learning algorithm is iterative, and in each iteration, a weak classifier is trained so as to minimize the training error, and in later iterations a different distribution or weighting of training data is used to give emphasis to examples that are often misclassified by the preceding weak classifiers.

2.3. Evaluation measures

We used two standard measures to evaluate system performance: F-measure and NIST error. F-measure, a widely-used metric, is the harmonic mean of recall and precision. NIST error, the official evaluation metric, is the ratio of false positives and false negatives to the number of reference sentences boundaries.

3. Prosodic features

The prosodic features calculated over each word are based off [4]. Currently, there are 84 features: pitch and pitch slope (33), energy and energy slope (33), vowel and rhyme duration (12), pause (2), and speaker turn-related (4).

3.1. Pitch and energy features

Pitch, RMS energy, and probability of voicing are extracted over 10ms frames and fed into Algemy [5] for processing. A piecewise-linear fit is performed while correcting for pitch halving and doubling to create "stylized pitch." Maximum, minimum, mean, and last value statistics are calculated over the word and beginning and ending windows. Combinations of these statistics are used as the basis for all the pitch features.

Pitch slope features are based off of the slopes of the stylized pitch. These include normalized and unnormalized final slope of the word and the difference in slope and rising and

falling pitch patterns across the word boundary.

Energy and energy slope features are calculated in parallel to their pitch counterparts, substituting raw RMS energy for stylized pitch.

3.2. Pitch and energy normalization

To mitigate speaker variation, all of the pitch and energy features are normalized in some manner. One approach is to normalize by some value or statistic from the same or adjacent word. An alternative is to normalize by a speaker statistic, such as speaker baseline pitch or pitch range. These speaker statistics are calculated from the pitch and energy values over speaker regions provided by the ICSI diarization system [6].

3.3. Vowel and rhyme duration features

The vowel duration features consist of Gaussian-normalized and mean-normalized durations of the vowels. The rhyme duration features measure the average phone duration in the rhymes. Two normalizations are used: 1) using the Gaussian-normalized phone duration instead of its unnormalized duration, and 2) normalizing by the speaker's average rhyme phone duration. For both vowel and rhyme duration features, two versions are created: one using the value from the word-final syllable, the other using the maximum value over all syllables.

For Mandarin, phone duration statistics are computed separately for each tone of each vowel. The difference between the mean durations of different tones is usually less than a frame, but since the standard deviation of vowel duration is on the order of frames, this could be significant.

3.4. Pause features

There are two pause features: `PAU_DUR` and `PREV_PAU_DUR` for the pause before and after the current word, respectively. While technically not prosodic features, a long pause after a word is a strong indication that it is at the end of a sentence, making it the best single feature by far.

3.5. Turn features

There are four features related to speaker turns, which are provided by the ICSI diarization system [6]. `TURN_TIME` measures the time that has elapsed since the beginning of the speaker turn. `TURN_TIME_N` is normalized by turn duration, giving the word's position within the turn. A flag, `TURN_F`, is true if and only if the word is the last one in the turn. `TURN_TIME_N` and `TURN_F` are a good example of features with redundant information since `TURN_F` true correlates with `TURN_TIME_N` near 1.

4. Feature selection

In the context of a learning algorithm, a set of features may have two problems: irrelevant features and redundant features. In our set of 84 prosodic features, there is considerable redundancy, especially given the way the pitch and energy features are designed.

The effect of these is to cause the learning algorithm to process more features than necessary, consuming more resources and generating more complicated models. Generally, there is also a decrease in performance as the extra features act like noise, obscuring other features from the learning algorithm.

The goal of feature selection is to remove these extra features from a feature set. More precisely, feature selection is usually performed in the context of a particular learning algorithm, and the goal is to find the feature subset which maximizes its performance. We used two forms of feature selection: filtering and a forward search wrapper.

4.1. Filtering

In filtering, each feature is independently scored according to how related it is to the class feature. An overview of popular scores is given in [7]. Filtering is quick and simple, but it does not take into consideration which learning algorithm will be used or the interaction between features. Thus, it is generally used to estimate feature relevance and filter out irrelevant features.

We used 4 different measures as implemented in the Weka toolkit [8]: Chi-Squared, Gain Ratio (2), Information Gain (1), and Symmetrical Uncertainty (3). The latter three are related information theoretic measures:

$$IG = H(class) - H(class|feat) \tag{1}$$

$$GR = \frac{IG}{H(feat)} \tag{2}$$

$$SU = 2\frac{IG}{H(class) + H(feat)} \tag{3}$$

The Chi-Squared statistic measures how likely the joint $(class, feat)$ probability distribution is assuming a null hypothesis that they are independent.

The scores from these four measures are fairly correlated, so we smoothed the differences by combining them into an overall score. Noting that the distribution of feature scores resemble an exponential distribution, we mean normalized each measure and summed the scores.

4.2. Forward search

A wrapper algorithm is one that iteratively feeds feature subsets to the learning algorithm and uses system performance as the measure of the fitness of the feature subset. A good overview of wrappers is given in [9].

The forward search wrapper starts with an initial feature subset, usually the empty set. At each iteration, a feature is added which improves the feature subset the most. This is a greedy algorithm, and so is suspectible to local maxima. Also, by considering features one at a time, it ignores the relationship between unselected features. Backward search, which begins with the full feature set and iteratively reduces the subset, can see these feature interactions and is generally more effective but was impractical due to the amount of data and number of redundant features. Forward search is computationally less expensive but still very effective.

5. Results and analysis

This study was conducted over three languages: Arabic, English, and Mandarin. The datasets used were a subset of the TDT-4 corpus for Arabic, English and Mandarin (TDT4-ARB, TDT4-ENG, and TDT4-MAN) with word time alignments from ASR output. See Table 1.

To motivate the use of prosodic features, Table 2 compares the result of all prosodic features, a baseline set of only 5 lexical features, and their combination. We see that prosodic features can make significant contributions to the sentence segmentation task. Feature selection can further improve this performance.

Table 1: *Dataset size and average sentence length (in words).*

	ARB	ENG	MAN
Words	185608	931245	459571
Avg. sentence length	21.5	14.7	24.7

Table 2: *Performance of all prosodic features and baseline lexical + speaker features.*

		ARB	ENG	MAN
FM	Baseline	46.4	47.8	43.6
	Prosodic	70.2	67.9	64.2
	Base+Pros	74.9	75.3	68.9
NIST	Baseline	84.0	94.7	87.0
	Prosodic	56.6	62.6	69.7
	Base+Pros	51.8	50.0	61.8

5.1. Forward search results

Table 3 shows the improvement over the first 7 iterations of the forward search wrapper for all languages. We see that with the first 4 to 6 features, we have already exceeded the performance of using all the prosodic features in both NIST error and F-measure, though incremental improvement has already begun to slow and performance soon plateaus.

The first feature selected in each forward search is PAU_DUR. One of either TURN_F or TURN_TIME_N is also selected. Past this, there is little pattern to which features are selected by the forward search wrapper. We conjecture that these features contain most of the relevant information in their respective categories. The algorithm then extracts whatever relevant information it can from the large number of closely-related pitch and energy features. Surprisingly, the rhyme and vowel duration features were not selected.

This suggests a possible heuristic for feature selection: by splitting the large feature set into groups which largely contain the same information, such as pause and turn-related features, a first-pass feature selection performed within each group can reduce the number of candidate features tested in the main feature selection algorithm.

From the results of the forward search and filtering experiments, we wanted to answer 2 questions:

1. Which features have low relevancy and can be removed from the forward search without hurting performace?
2. How do different features perform across different languages?

5.2. Feature relevancy

The best single feature is PAU_DUR. A sizeable pause is a good indication that a sentence has ended and a new one will begin. The other pause feature, PREV_PAU_DUR, the duration of the pause before the word, has mediocre relevancy. For TDT4-ENG, it is ranked 59th out of 84 features but was still selected in the forward search wrapper, improving NIST error by 1.5% absolute. We conjecture that it may help identify single-word sentences.

It is clear that PREV_PAU_DUR contains non-redundant information in TDT4-ENG. A couple inferences can be drawn from this. First, even weakly relevant features can contain useful information which the learning algorithm can exploit. Second, many of the features ranked higher by filtering now contain mostly redundant information by this early stage of the forward search wrapper. Thus, we conclude that many of our features contain redundant information, necessitating feature selection, but our feature set also contains few irrelevant features that can be removed without negatively impacting said feature selection.

Among turn-related features, TURN_F and TURN_TIME_N are ranked very high by filtering, and one of them was selected by the forward search wrapper in each language, at which point the other because almost entirely redundant. All other turn-related features performed very poorly.

Pitch and energy features make up the bulk of the features selected by the forward search wrapper, usually with more pitch features than energy. Filtering results corroborate with this, with pitch features tending to score better than energy features. This may be because pitch inherently has more useful information than energy features, or it may be because the design of our energy features were ill-suited for the task.

For pitch features, the ones that were normalized by speaker pitch range generally performed badly. In comparison, the features which were normalized using speaker baseline pitch or another pitch value from the same speaker tended to do well and were the ones selected by the forward search wrapper.

Pitch slope features tended to perform poorly, though we believe this is due to way they were calculated. Because the segmenter which creates the piecewise-linear fitted stylized pitch operates independently of the transcript word time alignments, the segmenter often does not split a voiced region right on the word boundary. Thus, regions of uniform slope may straddle a word boundary, and so it looks like there's no difference in slope across the boundary. While we compensate for regions up to 30ms, this appears to be insufficient.

5.3. Cross-linguistic analysis

Since the filtering scores varied considerably between languages, to compare feature performance across languages we examined their relative ranking as given by filtering. The 15 features which rose the most and dropped the lowest in comparison to other languages are summarized in Tables 4 and 5, respectively.

Mandarin clearly behaves differently from Arabic and English in terms of pitch features. Pitch slope features perform exceptionally badly, which we attribute to Mandarin being a tonal language, and so the pitch contour imposed by the lexical content obscures intonation which may convey sentence structure. In contrast, other pitch features perform better in Mandarin.

The energy features are more difficult to interpret. For instance, in Arabic, a number of cross-boundary energy features perform considerably better, and a number perform considerably worse. Furthermore, while it appears that word-level energy and energy slope do well in Arabic, more often than not this occurs because the same features scored poorly in the other languages. We partly attribute this behavior to the design of the energy features, which we plan to reexamine in future feature design cycles.

While duration features clearly perform better in English than in the other languages, they were not selected in the forward search wrapper. This leads us to believe that, while there is relevant information in the duration features, the features could be designed better or they are largely redundant in the face of other features.

Table 3: *Iterative improvement in first 7 features in forward search. Type lists the type of feature added that iteration (E* = energy; F* = pitch; *B = cross-boundary; *L = last / word-final; *W = whole word; *S = slope; P = pause; T = turn). For comparison, performance of all prosodic features is also listed.*

	ARB			ENG			MAN		
Iter	Type	NIST	FM	Type	NIST	FM	Type	NIST	FM
1	P	78.6	63.8	P	79.8	63.6	P	77.4	61.3
2	FL	60.3	67.2	FL	67.5	65.5	FW	74.3	62.5
3	FB	57.1	70.8	T	65.1	66.8	T	70.2	63.9
4	EB	56.6	70.2	FB	63.9	67.7	FB	68.6	64.4
5	T	56.8	70.2	FB	63.6	68.0	FB	68.6	64.3
6	FS	55.4	70.6	P	62.1	68.0	EB	68.6	64.9
7	EB	53.8	70.4	ES	61.6	68.3			
All		56.6	70.2		62.6	67.9		69.7	64.2

Table 4: *Tally of top 15 features which ranked well relative to other languages.*

	ARB	ENG	MAN
Pitch (word)			2
Pitch (final)			
Pitch (cross-boundary)			8
Pitch slope	3	3	
Energy (word)	3		
Energy (final)			
Energy (cross-boundary)	4	1	5
Energy slope	3		
Duration (max)		6	
Duration (last)		5	
Pause	1		
Turn-related	1		

Table 5: *Tally of top 15 features which ranked poorly relative to other languages.*

	ARB	ENG	MAN
Pitch (word)	4	4	
Pitch (final)			
Pitch (cross-boundary)	4	4	
Pitch slope			2
Energy (word)		4	1
Energy (final)			
Energy (cross-boundary)	3	3	1
Energy slope			1
Duration (max)	4		5
Duration (last)	4		4
Pause			
Turn-related			1

6. Conclusion

We have shown that prosodic features can produce a gain in sentence segmentation and that feature selection can further improve performance.

Our forward search feature selection results show pause duration and certain turn-related features are excellent features. The remaining features are drawn from a wide-variety of pitch and energy features, though there is no concensus between languages. Unexpectedly, no duration features were selected.

We performed an analysis of the relevance of different features between different languages. As a tonal language, Mandarin pitch features operate in a fundamentally different manner than the other languages. While our energy and duration features appear to work better in Arabic and English, respectively, certain behavior leads us to believe we should also reexamine their design.

7. Acknowledgements

The authors would very much like to thank Harry Bratt at SRI, who provided Algemy and technical support for the implementation of our prosodic features.

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR0011-06-C-0023. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of DARPA.

8. References

[1] Liu, Y., Shriberg, E., et al., "Structural Metadata Research in the EARS Program," *Proc. of ICASSP*, 2005.

[2] Roark, B., Liu, Y., et al., "Reranking for Sentence Boundary Detection in Conversational Speech," *Proc. of ICASSP*, 2006.

[3] Schapire, R. E., Singer, Y., "BoosTexter: A Boosting-Based System for Text Categorization," *Machine Learning*, vol. 39, p 135-168, 2000.

[4] Shriberg, E., Stolcke, A., Hakkani-Tur, D., Tur G., "Prosody-Based Automatic Segmentation of Speech into Sentences and Topics," *Speech Communication*, Vol. 32, No. 1-2, p 127-154, 2000

[5] Bratt, H., "Algemy, a tool for prosodic feature analysis and extraction," *Personal communication*, 2006.

[6] Wooters, C., Fung, J., Peskin, B., Anguera, X., "Towards Robust Speaker Segmentation: the ICSI-SRI Fall 2004 Diarization System," *RT-04F Workshop*, 2004.

[7] Yang, Y., Pederson J. O., "Feature selection in statistical learning of text categorization," *Machine Learning: Proc. of the 14th Intl. Conf.* p 412-420, 1997.

[8] Witten, I. H., Frank E., "Data Mining: Practical machine learning tools and techniques," 2nd Edition, Morgan Kaufmann, San Francisco, 2005.

[9] Kohavi, R., John, G. H., "Wrappers for Feature Subset Selection," *AIJ special issue on relevance*, 1997.