Using Symbolic Prominence to Help Design Feature Subsets for Topic Classification and Clustering of Natural Human-Human Conversations

Constantinos Boulis, Mari Ostendorf

Department of Electrical Engineering
University of Washington, Seattle, USA
boulis,mo@ssli.ee.washington.edu

Abstract

In this work, we use the output of a symbolic prominence classifier rather than acoustic cues of prominence, to improve the tasks of clustering and classification of spontaneous conversations to topics. In our experiments, we combine the output of a prominence classifier with lexical feature selection and combination methods to build improved feature subsets. Evaluated for the task of topic classification on a subset of Switchboard-I, the combination method offered a 11% relative reduction of classification error compared to using lexical-only feature selection methods; similar gains are reported for clustering.

1. Introduction

Various aspects of prosody have been successfully integrated in a number of spoken language understanding (SLU) tasks such as dialog act classification in conversational speech [1] and dialogue systems [2], discourse segmentation [3], error detection in dialogue systems [4] and voicemail summarization [5]. A less explored avenue to improve such tasks is prominence. Prominence, also referred to as pitch accent in English, is phrase-level emphasis given to one or more syllables of a word that goes beyond word-level strong/weak syllable differences associated with lexical stress. Prominence can be in combination but is not the same as intonation marking phrase or sentence boundaries. Some of the acoustic correlates of prominence are the F0 range, duration and energy of a syllable. Prominence has long been linked with information structure in numerous ways such as to contrast new vs. old information [6, 7] and to give local focus. Despite the wealth of literature on the role of prominence for human understanding, few attempts have been made to integrate prominence in SLU tasks. In [2], stress is used to disambiguate between words. In [8] a spoken document retrieval system is augmented with acoustic features such as duration and magnitude but no F0 information was used, which is an important feature for detecting prominence. Links between prominence and simple measures of word saliency have been established in [9, 10]. In [5], a number of acoustic and lexical features have been used to learn which words to extract for voicemail summarization. Acoustic cues that correlate with prominence have been found to be important. Most related with the current work is [11], where prosody is used to discriminate content from function words, but the approach was not integrated in a SLU system.

In this work, we use the output of a prominence detection system to facilitate classification and clustering of topics in natural human-human conversations. Semantic characterization of conversations can be valuable in a number of applications, such as analyzing customer-support call-center conversations or business meetings. Incorporating prominence into the semantic characterization of natural conversations has the major advantage that it can be equally useful in supervised and unsupervised cases. Traditional lexical measures of word saliency rely on annotated data in terms of topics. These measures become less reliable as the number of annotated examples decreases and do not apply at all in unsupervised cases. In contrast, prominence may be useful in cases of total lack of supervision, such as absence of the correct transcript of a conversation as well as topic-conversation pairs, although in this work we have experimented only with the true transcripts. Importantly, training a baseline prominence detection system with accuracy around 80% does not require a large amount of annotated data. The prominence classifier used in this work was trained on 124 Switchboard-I conversations, where each word was hand-annotated with a binary value (prominent or not).

2. Leveraging Prominence for Topic Detection

Spoken language classification tasks, such as call routing or characterizing human-human conversations, have a number of unique issues that distinguish them from text classification tasks. Handling ASR errors has been the one that most researchers have focused on. Other less explored issues are the handling of disfluencies and prominence. In this work we show that prominence can be integrated with standard techniques to design better feature subsets. The problem of feature selection and combination is at the core of text and spoken language classification. Typical vocabulary sizes are on the order of tens of thousands and only some of these features are deemed relevant for classification. Traditional techniques for determining which words should be removed rely on lexical information only. For example, one of the best performing feature selection measures is the Information Gain (IG) [12] which is given by:

$$IG(w) = H(\mathbf{C}) - p(w)H(\mathbf{C}|w) - p(\bar{w})H(\mathbf{C}|\bar{w}) \quad (1)$$

where $H(\mathbf{C}) = -\sum_{c=1}^{C} p(c)\log p(c)$ denotes the entropy of the discrete topic category random variable $\mathbf{C}$. Each document is represented with the Bernoulli model, i.e. a vector of 1 or 0 depending if the word appears or not in the document. All words in the vocabulary are ranked according to IG and the top N words are selected.

Prominence can be used to complement measures such as IG. First, the prominence classifier can be used to produce a score for every word occurrence in the dataset. The score will be the probability of each occurrence being prominent. Words can then be ranked according to their average prominence, i.e.

the average value of the prominence scores of all occurrences of a word. Having two alternative ranked lists - one from IG and another from prominence - the objective is to merge them in an optimal way. One way of combining the lists is to cascade them. First, the N_p words with the lowest average prominence are eliminated, the remaining words are ranked according to IG and the top N_l words selected. This scheme will be most successful when the two lists produce their best results in different regions. As shown in section 3 their effects are complementary, prominence can robustly identify irrelevant words but not the most relevant, while IG can identify quite well the most relevant words but not the most irrelevant words.

A second way of leveraging prominence is to use the prominence scores of each occurrence rather than their average. An appealing characteristic of prominence scoring is that different occurrences of a word will have different prominence scores, whereas the common bag-of-words representation treats all occurrences of a word uniformly. Conceptually, this scheme is a generalization of the bag-of-words representation in that a word counter is conditionally incremented based on the prominence value. One combination method is to only count the word occurrences above a certain threshold and then calculate IG.

3. Experiments

Experiments were performed on 648 conversations or 1296 conversation sides from the Switchboard-I corpus [13]. A single topic from a list of 64 is associated with each conversation. The true transcripts of the conversations were used for all experiments. Each word of the transcript was automatically annotated with a score between 0 and 1, indicating the probability that the word is prominent. The total number of word occurrences was approximately 800K, and keeping words with 5 or more occurrences, resulted in a vocabulary of 5211.

The prominence classifier described in [14] was used in all our experiments. All words of 124 conversation sides were annotated with binary values (prominent or not) and C4.5 decision trees were used for training. The prominence model predicts the posterior probability of prominence for each word using a decision tree with a combination of prosodic and text features. The prosodic features include: various F0 statistics within and across words (normalizing for speaker mean and variance), normalized duration statistics, silence duration, and energy statistics over a word. The text features included part-of-speech labels of the target word and its neighbors. The model was trained using a weakly supervised approach described in [12], where an initial model is trained on a small set of labeled data (roughly 4 hrs from 124 conversations), and then the EM algorithm is used to incorporate additional data that does not have prosodic mark-up but does have syntactic parses that can be used as additional features. For reference, the error rate of the classifier when used in predicting prominence was 21.3%, though hard decisions were not used in this work.

3.1. Classification experiments

The Bow toolkit [15] was used for all the classification experiments. A number of popular text classification algorithms are implemented in the Bow toolkit and it has been extensively used for research purposes. The 1296 conversation sides were equally split in train and test datasets and a 10-fold cross validation methodology is used. The standard deviation of all classification experiments are also reported. In the first experiment, we remove words from the vocabulary according to their average prominence. Three methods are compared; removing words with the highest average prominence, lowest average prominence and removing words at random. In Figure 1, the three different ways are shown. In all cases the Naive Bayes method is used as the learning method.

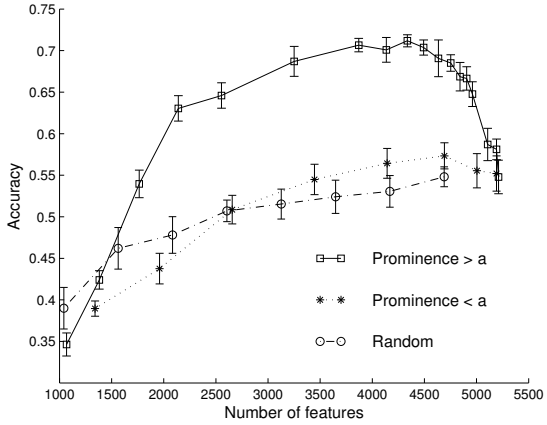

Figure 1: *Eliminating words according to their average prominence.*

Removing words with average prominence probability 0.45 or lower resulted in a very significant gain in performance over using all features; the classification accuracy rose from .548 when using all 5211 words to .712 when using words with average prominence 0.45 or higher (4337 words). Removing words at random consistently degraded the performance, while removing words with the highest prominence was not much different from random. These results suggest that prominence can quite robustly identify the least important words but not the most important words. If the most important words could be identified with prominence then removing words with the maximum average prominence should have resulted in worse results than removing words at random.

In the second experiment, we combine the information gain (IG) measure with prominence. We compare two methods to select words. The first method is to rank all 5211 words according to IG and select the N highest and the second method is to remove words with average prominence 0.45 or higher, rank the remaining 4337 words according to IG and select the N highest. The results are shown in Figure 2 where it can be seen that using only IG improves classification accuracy substantially compared to using all 5211 features, but combining prominence and IG offers additive gains, for every number of final features.

The results in Figures 1,2 were obtained with the Naive Bayes model. To make sure that prominence can help build better feature spaces, independent of the choice of classifier, we repeated the experiments of Figure 2 for a number of successful learning methods for text classification. The results are shown in Table 1, where we compared the lowest achievable classification errors using IG only and prominence+IG for a variety of learning methods.

The four different classifiers used are Naive Bayes with Laplace smoothing(NB), Rocchio (Rocchio)[16], Probabilistic indexing (Prind)[17] and Support Vector Machines (SVMs)[18]. The results of Table 1 show consistent gains in performance using the combined feature selection method, re-

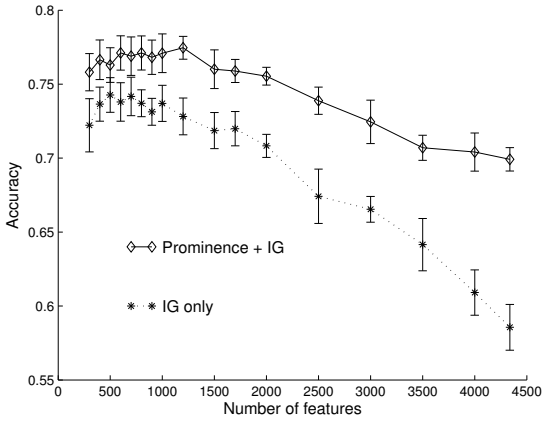

Figure 2: *Feature subsets determined by IG only and prominence combined with IG.*

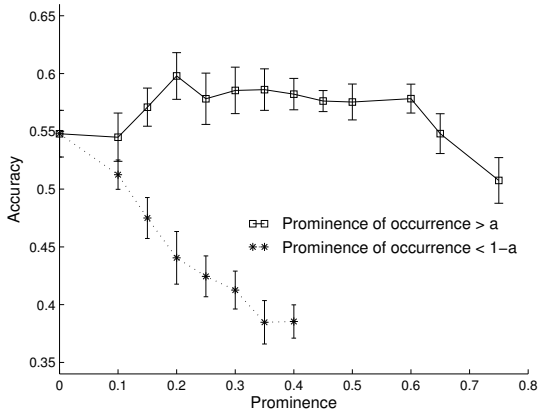

Figure 3: *Eliminating word occurrences according to their prominence.*

Table 1: *Relative reduction of classification error using prominence+IG compared to IG only for various learning methods.*

NB	Rocchio	Prind	SVM
11%	12%	15%	7%

inforcing our hypothesis that prominence helps design better feature spaces irrespective of the classifier. Even for SVMs, that are known to provide state-of-the-art results in text classification and are considered robust to irrelevant features, the proposed method still offers gains.

Up to this point we have only experimented with the average prominence of a word. In Figure 3 we select word occurrences rather than words, according to their prominence. The results are shown in Figure 3 and show that this is not as good of a selection mechanism as with average prominence. This can be due to two main reasons. First, because using the average prominence reduces the variability of the prominence prediction, in other words using the average reduces the "noise" in the data. Second, if we remove a high percentage of an irrelevant word, but not every occurrence then the remaining few occurrences may be biased towards a specific topic, therefore the classifier will mistakenly train this word as relevant.

It is also instructive to see some of the words that are removed using prominence. In Table 2 we see the 16 words with the least average prominence and their corresponding ranking in the IG list. We observe that for many common words, IG fails to identify them as irrelevant, as intuition would suggest. Using a default stopword list would capture some of these words, but not all. For example, the word *bye-bye* is specific to the task at hand and would not be included in a default stopword list.

Table 2: *The 16 words with the least average prominence and their position in the IG list (out of 5211 words, smaller numbers are less important).*

1-8		9-16	
thinner	349	shall	946
to	116	of	56
bye-bye	4368	from	4642
than	3413	at	4214
an	4686	with	4232
till	3477	within	4254
a	4	and	900
the	300	should've	1735

3.2. Clustering experiments

An important characteristic of utilizing symbolic prominence is that it does not require any supervised data in terms of topics, while all lexical word saliency measures do. IG is not applicable in an unsupervised scenario. Designing feature subsets for text clustering is usually done with feature correlation methods, such as Latent Semantic Analysis (LSA) [19]. In this respect, using prominence to select irrelevant features is complimentary to feature combination/correlation approaches such as LSA. A natural way to combine LSA and prominence is to first remove words according to prominence and then combine the remaining ones with LSA. We used the CLUTO toolkit[1][20], a software package for clustering in high-dimensional spaces, to cluster the 1296 conversations with the number of topics (64) assumed to be known a priori. We used the default values of CLUTO to perform clustering. The objective function to maximize is intra-cluster cosine similarity, ten random restarts are performed and the one with the highest objective function is retained. Since the final result depends on the initial conditions we performed 10 trials and also report standard deviations. Note that the standard deviations here do not have the same interpretation as in classification, since here the same dataset is used for clustering. The results were evaluated using the adjusted Rand index [21] a common measure to evaluate clustering solutions. The adjusted Rand index is the fraction of pairs of points that were correctly clustered together and correctly clustered in different classes, adjusted to zero for chance results. Figure 4 shows the clustering results, where we see that for a variety of features combining prominence with LSA is better than using LSA only. Words with average prominence 0.4 or higher were selected. LSA is performed on the tf·idf conversation-side word matrix.

Using all 5211 features resulted in an adjusted Rand index of .667 with standard deviation of .016. Using LSA alone did not offer gains than the baseline, while using the combined scheme resulted in an adjusted Rand index of .694.

[1] http://www-users.cs.umn.edu/~karypis/cluto/

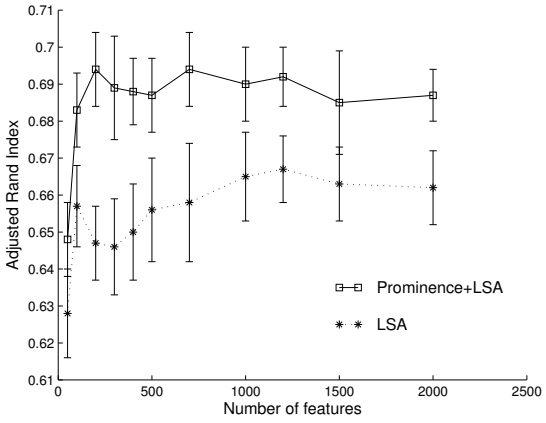

Figure 4: *Effect of LSA only and prominence combined with LSA on clustering performance.*

4. Conclusions

We have demonstrated how automatically detected symbolic prominence can be used to design better feature subsets for semantic characterization of natural human-human conversations. Despite the fact that the prominence classifier had an error rate of about 21.3% it was shown to be useful for the tasks of topic classification and clustering. Specifically, words with very low and high average prominence were shown to be mostly irrelevant for classification purposes. This suggests that prominence may have a bigger role in SLU systems by filtering out uninteresting areas rather than detecting areas of high content. Further, prominence was shown to lift the gains from other common feature selection and combination methods, such as IG and LSA. In the future, it will be interesting to examine the words with high prominence and low IG, since this may be a way to detect discourse markers. In addition, prominence can be useful when using the ASR transcriptions as well. Using an ASR system adds "noise" to the word sequence and therefore lexical measures of word saliency may be affected.

5. Acknowledgments

We would like to acknowledge the help of Darby Wong and Jeremy Kahn in providing us with the prominence classifier. This work has been supported by NSF grant IIS-0121396.

6. References

[1] E. Shriberg and et al, "Can prosody aid the automatic classification of dialog acts in conversational speech?" *Language and Speech*, vol. 41, no. 3-4, pp. 439–487, 1998.

[2] E. Nöth, A. Batliner, A. Kießling, R. Kompe, and H. Niemann, "VERBMOBIL: The use of prosody in the linguistic components of a speech understanding system," *IEEE Trans. on Speech and Audio Proc.*, vol. 8, no. 5, pp. 519–532, 2000.

[3] E. Shriberg, A. Stolcke, D. Hakkani-Tür, and G. Tür, "Prosody-based automatic segmentation of speech into sentences and topics," *Speech Communication*, vol. 32, pp. 127–154, 2000.

[4] J. Ang, R. Dhillon, A. Krupski, E. Shriberg, and A. Stolcke, "Prosody-based automatic detection of annoyance and frustration in human-computer dialog," in *Proc. of ICSLP*, 2002.

[5] K. Koumpis and S. Renals, "The role of prosody in a voicemail summarization system," in *Proc. of ISCA Workshop on Prosody in Speech Recognition and Understanding*, 2001, pp. 87–92.

[6] D. Bolinger, "Accent is predictable (if you're a mind-reader)," *Language*, vol. 48, pp. 633–644, 1972.

[7] M. Horne, P. Hansson, G. Bruce, and J. Frid, "Prosodic correlates of information structure in Swedish human-human dialogues," in *Proc. of Eurospeech*, 1999, pp. 29–32.

[8] B. Chen, H.-M. Wang, and L.-S. Lee, "Improved spoken document retrieval by exploring extra acoustic and linguistic cues," in *Proc. of Eurospeech*, vol. 1, 2001, pp. 299–302.

[9] S. Pan and K. R. McKeown, "Word informativeness and automatic pitch accent modeling," in *Proc. of the Joint SIGDAT Conference on EMNLP and VLC*, 1999, pp. 148–157.

[10] R. Silipo and F. Crestani, "Prosodic stress and topic detection in spoken sentences," in *Proc. of SPIRE*, 2000, pp. 242–252.

[11] J.-M. Blanc and P. Dominey, "Using prosody to discriminate between function and content words," in *Proc. of International Conference on Speech Prosody*, 2004.

[12] G. Forman, "An extensive empirical study of feature selection metrics for text classification," *Journal of Machine Learning Research*, pp. 1289–1305, 2003.

[13] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "Switchboard: Telephone speech corpus for research development," in *Proc. of ICASSP*, 1992, pp. 517–520.

[14] D. Wong, M. Ostendorf, and J. Kahn, "Using weakly supervised learning to improve prosody labeling," University of Washington, Electrical Engineering Department, Tech. Rep., 2005.

[15] A. McCallum, "Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering," 1996, http://www.cs.cmu.edu/ mccallum/bow.

[16] T. Joachims, "A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization," in *Proc. of ICML*, 1997, pp. 143–151.

[17] N. Fuhr, "Models for retrieval with probabilistic indexing," *Inf. Process. Manage.*, vol. 25, no. 1, pp. 55–72, 1989.

[18] T. Joachims, *Making large-Scale SVM Learning Practical*, B. Schölkopf, C. Burges, and A. Smola, Eds. MIT-Press, 1999.

[19] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American Society for Information Science*, vol. 41, pp. 391–407, 1990.

[20] Y. Zhao and G. Karypis, "Empirical and theoretical comparisons of selected criterion functions for document clustering," *Machine Learning*, vol. 55, pp. 311–331, June 2004.

[21] L. Hubert and P. Arabie, "Comparing partitions," *Journal of Classification*, pp. 193–218, 1985.