# Extracting Phrase Patterns with Minimum Redundancy for Unsupervised Speaker Role Classification

**Bin Zhang, Brian Hutchinson, Wei Wu and Mari Ostendorf***
University of Washington, Seattle, WA 98125

## Abstract

This paper addresses the problem of learning phrase patterns for unsupervised speaker role classification. Phrase patterns are automatically extracted from large corpora, and redundant patterns are removed via a graph pruning algorithm. In experiments on English and Mandarin talk shows, the use of phrase patterns results in an increase of role classification accuracy over n-gram lexical features, and more compact phrase pattern lists are obtained due to the redundancy removal.

## 1 Introduction

The identification of speaker roles is fundamental to the analysis of social content and information reliability in conversational speech. Previous work has primarily used supervised learning in automatic role classification. Barzilay et al. (2000) described a speaker role classification system for English broadcast news (BN), where the speakers were categorized into three types: anchor, journalist, and guest. The authors used supervised learning to discover n-gram signature phrases for speaker introduction and structural features such as duration, achieving an accuracy of 80% on ASR derived transcripts. Liu et al. (2006) studied speaker role classification on TDT-4 Mandarin BN data. Hidden Markov and maximum entropy models were used to label the sequence of speaker turns with the roles anchor, reporter, and other, based on n-gram features, which yielded 80% classification accuracy on human transcripts.

Hutchinson et al. (2010) extend previous work to the case of unsupervised learning, with the goal of portability across languages. That work explored speaker role classification using structural and n-gram features on talk show (or broadcast conversation (BC)) data. In this paper, we address a limitation of n-grams as features by proposing a method for learning phrases with gaps, which is particularly important for conversational speech, since there are more disfluencies that can cause failure of n-gram matching. In addition, we want to avoid topic words (e.g., proper nouns) in order to model speaker roles rather than topics. For example, for identifying the host, the phrase pattern "*We'll be back with * in a minute*" is more general than the n-grams "*We'll be back with John Smith in a minute*." To prevent these problems with n-grams, one must limit the length of learned n-grams, making them less discriminative.

Phrase patterns have been used in other NLP applications such as (Sun et al., 2007). To remove the redundancies in the automatically extracted phrase patterns, we propose a redundancy removal algorithm based on graph pruning that does not require role-labeled data. The resulting set of patterns is then used to extract lists of signature and conversational phrases, from which features are derived that are used to distinguish between the different roles. Using the phrase pattern-based lexical features in clustering, we obtain 82-89% speaker role classification accuracy on human transcripts of BC shows.

## 2 Method

Phrase patterns are generalizations of n-grams. A phrase pattern $p$ of length $n$ is an ordered list of words $(w_1, w_2, \ldots, w_n)$. It is matched by a word sequence of length $m \geq n$ if the sequence contains the words in the order defined by the pattern. Because the words in a phrase pattern need not appear contiguously in the sequence, phrase matching is less sensitive to disfluencies and topic words.

### 2.1 Phrase Pattern Extraction

Phrase patterns can be extracted by the sequential pattern mining algorithm PrefixSpan (Pei et al.,

* This research was funded by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA). All statements of fact, opinion or conclusions contained herein are those of the authors and should not be construed as representing the official views or policies of IARPA, the ODNI or the U.S. Government.



*Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL*, pages 717–720,
Los Angeles, California, June 2010. ©2010 Association for Computational Linguistics

2001). This algorithm efficiently extracts frequent phrase patterns in a corpus (i.e., relative frequency greater than a given threshold). Prior to the extraction, we perform text preprocessing including splitting the text into lines based on commas and periods to limit the pattern span, followed by case and punctuation removal. The extracted phrase patterns have variable length. As a result, longer patterns may contain shorter patterns. Phrase patterns with the same length may also be overlapped. These redundancies should be removed; otherwise, the same phrase may match several patterns, resulting in biased counts.

### 2.2 Phrase Pattern Redundancy Removal

Define a phrase pattern $p$ as contained in another phrase pattern $q$ if $q$ contains all the words in $p$ in the same order. $p$ is called a parent pattern and $q$ is the corresponding child pattern. Instead of constructing a tree as in (Siu et al., 2000) for variable length n-grams, we create a graph of phrase patterns based on containment, because a pattern can contain and be contained by multiple patterns. Our redundancy removal algorithm involves pruning this graph. With the nodes being the phrase patterns, the edges of the phrase pattern graph are constructed by connecting length-$n$ phrase pattern $p$ to length-$(n+1)$ phrase pattern $q$ for all $n$, if $p$ is contained in $q$. We connect only phrase patterns that differ by one word in length for computational efficiency, and this results in a multi-layer structure: the phrase patterns in each layer have the same length. For the convenience of pruning, a virtual node $T$ is created as the "zeroth"-layer, and it is directly connected to all the nodes in the layer with the shortest pattern length.

Once a phrase pattern graph has been created, we prune the graph in order to remove the redundant nodes. First, we remove edges based on the ratio of counts $c(q)/c(p)$ between child node $q$ and parent node $p$. A large ratio implies that the child appears in most of the cases where the parent appears. Hence, we keep the edge to indicate that the child can be used as a preferred substitute for the parent. On the other hand, the edge is removed if the ratio is small (less than a threshold $t$, see Fig. 1).

After this procedure is performed on all the edges in the graph, we determine whether a node is pruned based on its connectivity to parents and children. We

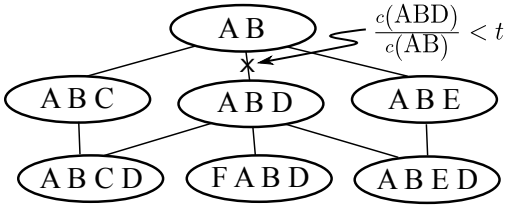


Figure 1: A fragment of an example phrase pattern graph. The letters represent words. The edge between "AB" and "ABD" is removed because the ratio of counts is less than the threshold.

define two levels of pruning, which differ in whether a node can be preserved even if its connections to parents are removed:

**Conservative pruning** A node is pruned if it has at least one child.

**Aggressive pruning** A node is pruned if it has at least one child or is not on a path connected to $T$.

Both methods were investigated, in case some useful phrase patterns ended up being pruned with the more aggressive approach.

### 2.3 Features Based on Phase Patterns

Although (Hutchinson et al., 2010) uses both lexical and structural features, here we use only lexical features to better assess impact. Once the graph pruning has provided a list of phrase patterns (eliminating phrases of length one because of low reliability), two subsets are extracted to represent signature phrases as might be used by a host and conversational phrases as might occur more frequently in live interviews. The signature statistic

$$\theta_1 = \frac{DF}{SF} + \alpha \log(f_{\text{BC}}). \quad (1)$$

is based on the speaker frequency ($SF$, # speakers whose utterances match $p$), document frequency ($DF$, # shows that match $p$), and genre-dependent frequency $f_{\text{BC}}$ (# occurrences of $p$ in BC), all computed on the training data. The ratio $\frac{DF}{SF}$ favors phrases that occur in many documents but few speakers, e.g. one per show, as for a host. The log BC frequency term is a penalty to eliminate infrequent patterns. The conversation statistic

$$\theta_2 = \frac{f_{\text{BC}}}{f_{\text{BN}}+1}\mathbf{1}_{SF>\beta}. \quad (2)$$

uses frequency $f_{\text{BN}}$ (# occurrences of $p$ in BN), to look for phrases that are more frequent in BC data than BN, ideally live discussion phenomena. The indicator function $\mathbf{1}_{SF>\beta}$ eliminates phrases used by a small number of speakers to avoid topic-related phrases. Hyper-parameters $\alpha$ and $\beta$ are tuned by inspecting the top phrase patterns after ranking. We use $\alpha = 10^{-3}, \beta = 500$ for English and $\alpha = 10^{-4}, \beta = 1000$ for Mandarin. Phrase patterns are ranked by the two statistics to generate lists of *signature* and *conversational* patterns, respectively.

During speaker-level feature extraction in role detection, each phrase pattern in the lists is matched against a speaker's utterances. The lexical features have two dimensions: the count of matches using the signature and conversational patterns, each normalized by the total number of patterns matched in the show to account for differences between shows.

## 3 Experiments

### 3.1 Task and Data

In the absence of speaker-role-labeled conversational speech training data, we perform unsupervised speaker role classification with three classes: host, expert guest, and soundbite. We evaluate on two human-labeled evaluation sets (English and Mandarin). The English eval set contains nine BC shows (150 speakers), while the Mandarin eval set contains 14 shows (140 speakers). There is an additional labeled Mandarin development set composed of ten shows (71 speakers). There are on average 7.6k words and 7.5k characters per show for English and Mandarin, respectively. The phrase patterns are learned from much larger corpora with speaker labels but without speaker role labels, including web transcripts for 310 English shows and quick rich transcripts for 4395 Mandarin shows. Because of the larger amount of Mandarin data, we use a lower threshold ($5 \times 10^{-5}$) for phrase pattern extraction than for English ($10^{-4}$).

### 3.2 Classification

Spectral clustering (Shi and Malik, 2000) is used in this work, since we found it to outperform other clustering approaches such as $k$-means and Gaussian mixture models. Given a two-dimensional feature vector for each speaker in a show, we construct a speaker graph with edge weights defined by Gaussian similarity $\exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right)$. The spectral clustering is non-deterministic, because it uses $k$-means as its final step ($k = 3$), which is initialized by randomly choosing $k$ samples as initial centroids. We vary $\sigma$ as an integer from 1 to 100 in combination with different random initializations to generate multiple clustering alternatives, and then use a partition selection algorithm to pick the most common clustering among the candidates. We use domain knowledge to map speaker clusters into speaker roles: the cluster whose members have the largest average number of speaker turns is the host cluster, that with the smallest average number of turns is the soundbite cluster, and the remaining cluster contains the expert guests.

### 3.3 Results

The phase pattern pruning threshold $t$ was tuned on the Mandarin dev set. We varied $t$ from 0.1 to 0.9, and measured the classification accuracy. $t = 0.8$ was found to be optimal (Fig. 2).

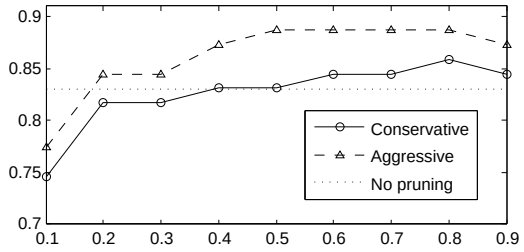


Figure 2: Accuracy on Man dev vs. pruning threshold $t$

The list of classification results on all the data sets is shown in Tab. 1. Aggressive pruning yields the best classification performance on all the data sets. It is also better than using n-gram matching for feature extraction (the last row of the table).

| | Man dev | Man eval | Eng eval |
|---|---|---|---|
| No pruning | 0.83 | 0.86 | 0.81 |
| Cons. pruning | 0.86 | 0.83 | 0.81 |
| Aggr. pruning | **0.89** | **0.89** | **0.82** |
| N-gram | 0.86 | 0.86 | 0.77 |

Table 1: Classification results

The size of phrase pattern lists is given in Tab. 2, and the number of redundant phrase patterns (the patterns that are contained in other patterns) is in Tab. 3 for different pruning levels. Using aggressive pruning, the list size and number of redun-

dant phrase patterns are greatly reduced. However, the classification accuracy does not decrease. This demonstrates that the redundant phrase patterns are not helpful and can be harmful for this task.

| Pruning level | Signature ptn. | | Conv. ptn. | |
|---|---|---|---|---|
| | Eng | Man | Eng | Man |
| No pruning | 2000 | 2000 | 1000 | 1000 |
| Cons. pruning | 1605 | 946 | 928 | 998 |
| Aggr. pruning | 244 | 370 | 465 | 835 |

Table 2: Phrase pattern list size

| Pruning level | Signature ptn. | | Conv. ptn. | |
|---|---|---|---|---|
| | Eng | Man | Eng | Man |
| No pruning | 396 | 1331 | 337 | 142 |
| Cons. pruning | 35 | 307 | 334 | 142 |
| Aggr. pruning | 6 | 59 | 8 | 0 |

Table 3: Number of redundant phrase patterns in the list

The unsupervised speaker role classification system in (Hutchinson et al., 2010) uses both n-gram and structural features, giving classification accuracy on English and Mandarin eval sets of 0.86 and 0.84, respectively. Adding structural features to phrase-pattern-based lexical features improves the performance on English but not Mandarin, perhaps because soundbites in English tend to be much shorter than those in Mandarin.

### 3.4 Discussion

The experiments reflect differences between the two languages. We observe that the main gain in Mandarin comes from improved classification of hosts, due to the signature phrase patterns. In English, the improvement is attributed to improved classification of expert guests and soundbites, suggesting an improved conversational dimension of the lexical features. The performance difference of the two languages seems more related to the languages themselves, rather than the size of data sets on which phrase patterns are learned, because we were able to obtain similar performance on Mandarin even when the training set size is reduced.

Anecdotal inspection of the phrase patterns learned for the signature phrases suggests that the combination of redundancy pruning and the heuristic signature statistic is quite effective. For example, we observed English signature patterns such as "*back with after this*" and "*let's take a look at*." The former pattern can be matched by phrases with names or topics inserted, and the latter can be matched by "*let's just take a look at*" or "*let's take a brief look at*." In the Mandarin signature patterns, we also found patterns such as "今天 请到 演播室 的 嘉宾 是 * 的 * 教授" (*today the guest invited to the studio is Professor from*) and "谢谢 来自 * 的 报道" (*thanks to the report from*). These patterns can be considered to be templates for hosts, where the named-entities are skipped.

## 4 Conclusions

We have presented a method for extracting phrase patterns with minimum redundancy for speaker role classification. The proposed algorithm removes most of the redundancies in the phrase patterns, leading to more compact pattern lists and improved classification accuracy over n-gram lexical features. We can apply the algorithm to other applications such as text classification, where phrase patterns can be used in place of n-grams. One way to extend this work is to use the automatically extracted phrase patterns as initial features, and then employ supervised or semi-supervised learning techniques to learn a more discriminative feature set.

## References


R. Barzilay et al. 2000. The Rules Behind Roles: Identifying Speaker Role in Radio Broadcasts *Proc. AAAI*, pp. 679–684.

Y. Liu. 2006. Initial Study on Automatic Identification of Speaker Role in Broadcast News Speech. *Proc. HLT*, pp. 81–84.

B. Hutchinson et al. 2010. Unsupervised Broadcast Conversation Speaker Role Labeling *Proc. ICASSP*, pp. 5322–5325.

G. Sun et al. 2007. Detecting Erroneous Sentences Using Automatically Mined Sequential Patterns. *Proc. ACL*, pp. 81–88.

J. Pei et al. 2001. PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-projected Pattern Growth. *Proc. ICDE*, pp. 215–224.

M. Siu and M. Ostendorf. 2000. Variable N-grams and Extensions for Conversational Speech Language Modeling. *IEEE Transactions on Speech and Audio Processing*, 8(1):63–75.

J. Shi and J. Malik. 2000. Normalized Cuts and Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905.