

# Detecting Overlap in Features Using a Subsumption Hierarchy

Siddharth Patwardhan and Ellen Riloff

## Introduction

- Features used in Natural Language Processing tasks vary in complexity.
- Some Natural Language Processing tasks do extremely well by learning statistical properties of sequences of words, called *n*-grams.
- For example, some of the best *Text Categorization* systems use *n*-gram models based on statistical distribution of word sequences.
- However, many *Information Extraction* systems do a much deeper linguistic analysis of text.
- For instance, the *Autoslog* system uses lexico-syntactic “Extraction Patterns” to identify and extract relevant information from free text.
- In this research, we describe an analysis tool called the *Subsumption Hierarchy* which can be used to:
  - ✧ identify overlap in features
  - ✧ select the best features among the overlapping features for an NLP task.

## Background

- N-grams are defined as:

a subsequence of *n* items from a given sequence of items.

- In the context of Natural Language Processing these are usually words, part of speech tags or words with their part of speech tags.

For example, in the sentence “the cow jumped over the moon”, some of the n-grams (among others) present:

cow, jumped, the cow, cow jumped, jumped over the prep article, noun verb prep  
cow(noun) jumped(verb), moon(noun)

- Lexico-syntactic patterns provide a deeper and richer analysis of text.
- These patterns can identify the syntactic roles and constituents within the text.

Mr. Scindia immediately announced his resignation  
<subj>\_ActVp (ANNOUNCED) \_<doobj>

Jared was momentarily shocked by the news  
<subj>\_PassVp (SHOCKED)

The natives frequently tried to scare the unsuspecting tourists  
<subj>\_ActInfVp (TRIED\_SCARE) \_<doobj>

They arrived home by midnight  
<subj>\_ActVp (ARRIVED) \_PP (BY)

## Motivation

- For many tasks, surprisingly, simpler representations like 1-grams as features outperform the richer representations.
- Simply adding all the complex features to the mix produces only a small improvement in performance.
- For example, in a *Text Classification* task of classifying a set of movie reviews as positive or negative:

Features	Accuracy	Precision	Recall	F-Measure
1-grams	81.84%	83.71%	79.06%	81.32%
2-grams	77.14%	80.38%	71.79%	75.84%
Lexico-Syntactic Patterns	72.86%	76.10%	66.67%	71.07%
1+2-grams	82.91%	85.00%	79.91%	82.38%
LSPatterns+1+2-grams	82.91%	87.38%	76.92%	81.82%

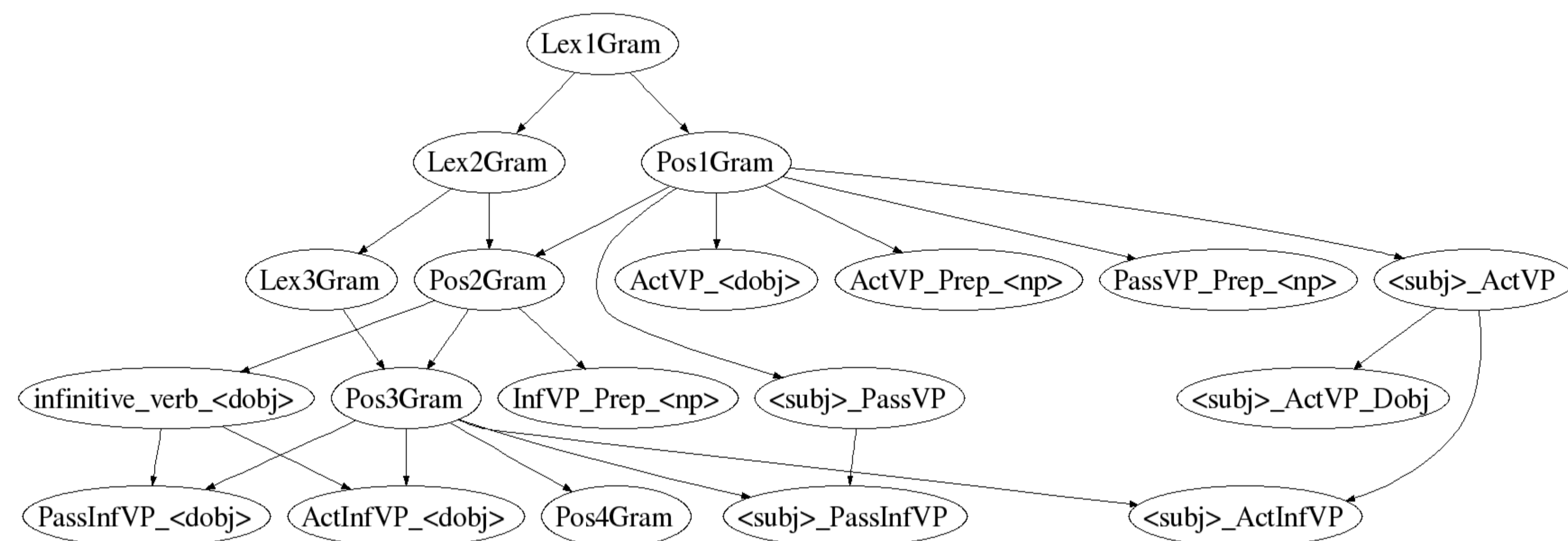
- Observe that the more complex features (2-grams and lexico-syntactic patterns) perform worse than the 1-grams.
- Adding lexico-syntactic patterns and 2-grams to the 1-grams improves the results only slightly.
- The more complex features are sparser than the simpler features.

- Hypotheses:

- ✧ The complex features “overlap” with the simpler features, and the classifier then tends to choose the more prevalent simpler features.
- ✧ Detecting overlapping features, and selectively adding complex features to the classifier should improve performance.

## The Subsumption Hierarchy

- Subsumption occurs if the presence of one feature implies the presence of another feature in a majority of the cases (second subsumes the first).
- For example, if the 2-gram “CALL FOR” occurs in text, it implies that the 1-gram “CALL” also occurs, and hence “CALL” subsumes “CALL FOR”.
- Manually created hierarchy that determines overlap between two features by defining rules for subsumption.



- Every feature is assigned to a node in the hierarchy.
- For example, the 2-gram “CALLED FOR” gets assigned to the Lex2Gram node, and the feature <subj>\_ActVp(CALLED) gets assigned to the <subj>\_ActVp node.
- Every feature is converted into a sequence of words.
- *Sequential* or *syntactic* dependencies are defined for pairs of words in the sequences.

*Sequential Dependency* defined between two words implies that the two words must occur consecutively in text.

jumped over the

*Syntactic Dependencies* are defined only in the lexico-syntactic features to capture the attachments between syntactic elements.

Bob wanted to buy a small present for Cathy  
InfVp (BUY) \_PP (FOR)

- A feature *x* is subsumed by another feature *y* if:
  - ✧ Feature *y* is an ancestor of *x*.
  - ✧ Words in feature *y* are a subsequence of words in feature *x*.
  - ✧ Feature *x* has at least all the sequential and syntactic dependencies of feature *y*.

## Feature Selection

- Subsumption hierarchy can be used to eliminate overlap in features before passing them on to a Natural Language Processing task.
- First, statistical properties (indicating the usefulness of a feature) are computed for each feature, using training data.
- The features are then pushed through the subsumption hierarchy.
- For every pair of features *x* and *y*, if *x* subsumes *y*, and if they are statistically similar, either *x* or *y* may be deleted from the set.
- We use the notion of the *delta* of a statistical property of two features to determine if they are statistically similar.
- Evaluation will be performed by using this tool for feature selection in a Text Categorization task.