Feature Subsumption for Opinion Analysis

Ellen Riloff[1], Siddharth Patwardhan[1] and Janyce Wiebe[2]

[1]School of Computing, University of Utah

[2]Department of Computer Science, University of Pittsburgh

Introduction

- Lexical features are commonly used in approaches to sentiment, subjectivity and opinion recognition.
- Lexico-syntactic features capture subtleties and non-compositional meanings that simpler features do not.
- Our goal is to automatically identify features that representationally overlap with simpler features, but that are better opinion indicators.
  - Features that are identified as strong opinion indicators can be added to a subjectivity lexicon.
- A subsumption hierarchy can be used to:
  - gain an understanding of features.
  - automatically eliminate unnecessary features and improve classification performance.

Text Representations

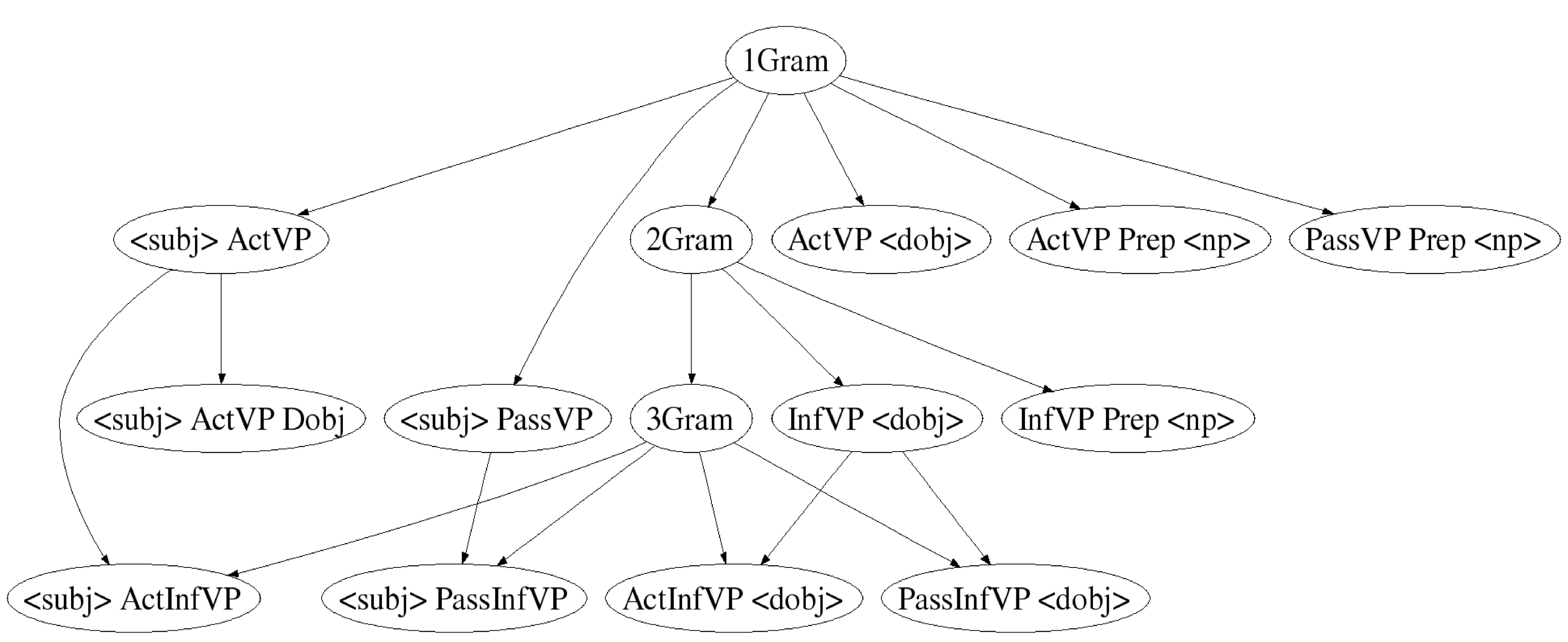
- Ngrams: unigrams and bigrams.
- Extraction Patterns: can represent flexible subjective expressions.

dealt the company a decisive blow
ActVP(dealt) Dobj(blow)

no benefits or advantages to the economy
NP(benefits) Prep(to)

The Subsumption Hierarchy

- A manually created hierarchy defines the representational scope of different types of features.
- The subsumption rules detect representational overlap between features and their relative value for classification.



- Each node in the hierarchy defines the representation of one type of feature based on words, sequential, and syntactic dependencies.
- Two sample node definitions:

Name = **2Gram**
Constituent[0] = WORD1
Constituent[1] = WORD2
Dependency = Sequential(0, 1)

Name = **NP Prep <np>**
Constituent[0] = NP
Constituent[1] = PREP
Constituent[2] = NP_EXTRACTION
Dependency = Syntactic(0, 1)
Dependency = Syntactic(1, 2)

Data Sets Used in the Experiments

- **OP data set (Wiebe et al., 2004)**: document-level opinion classification. 2,452 documents (9% opinion + 91% non-opinion).
- **Polarity data set (Pang and Lee, 2004)**: document-level polarity classification. 1,400 documents (50% positive + 50% negative).
- **MPQA data set (Wiebe et al., 2005)**: sentence-level opinion classification. 9,732 sentences (55% opinion + 45% non-opinion).

Using Subsumption for Analysis

- Opinion vs. non-opinion features:

Feature	IGain	Example
line	.0016	… notes backed by credit *line* receivables
the line	.0075	… lays it on *the line*; … steps across *the line*
benefits	.0040	… included $235,000 in tax *benefits*
NP(benefits) Prep(to)	.0090	... with no proven *benefits to* the economy
due	.0001	… estimated $1.23 bn in debt *due* next week
ActVp(due) Prep(to)	.0038	… it's all *due to* the intense scrutiny

- Polarity features:

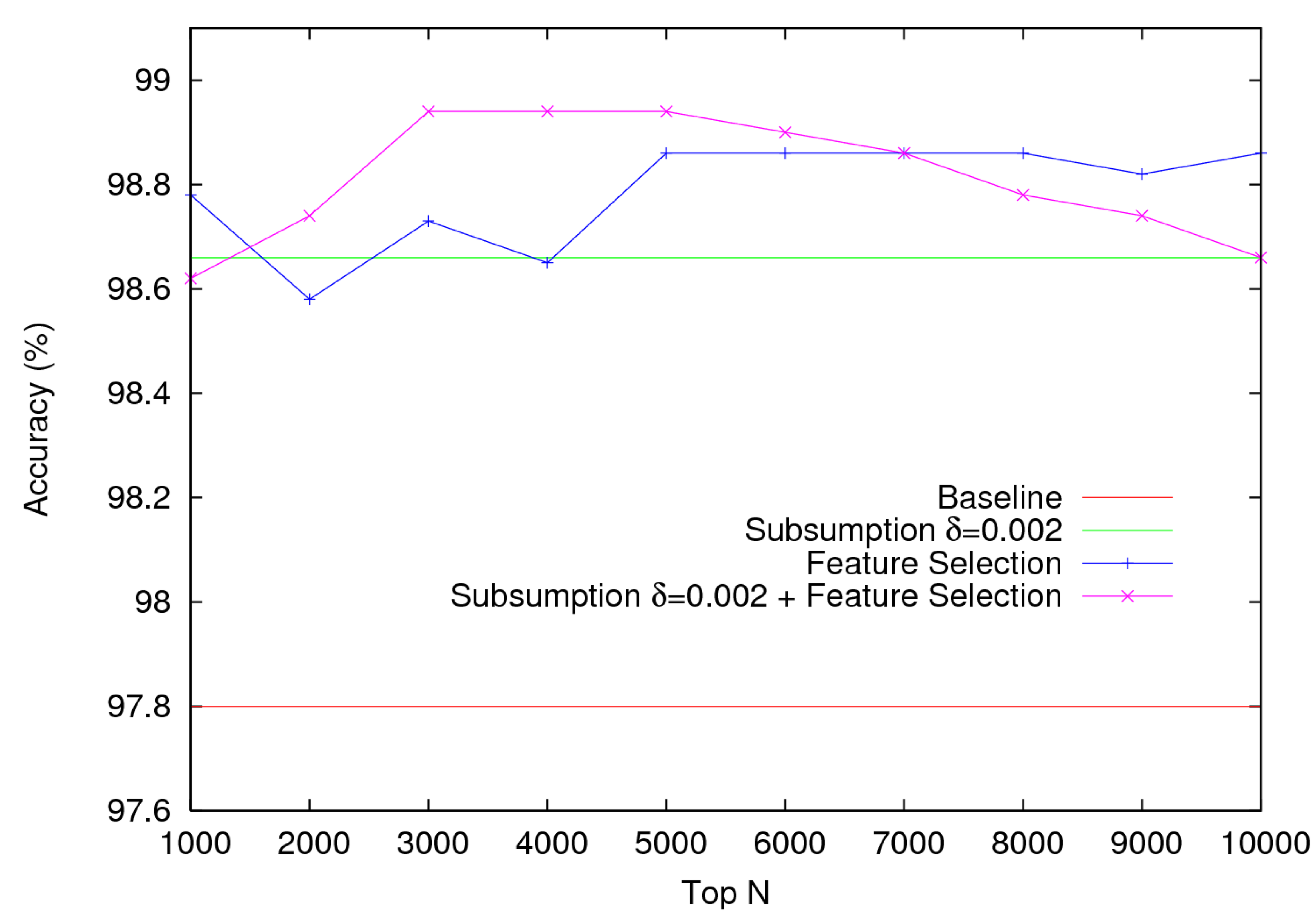
Feature	IGain	Example
short	.0014	… to make a long story *short*
nothing short	.0039	… *nothing short* of spectacular
disaster	.0010	… rated PG-13 for *disaster* related elements
AuxVp(be) Dobj(disaster)	.0048	... this *is* such a confused *disaster* of a film
work	.0001	… the next day during drive to *work*
ActVp(work)	.0038	… the film *will work* just as well

Subsumption for Feature Set Reduction

- The subsumption hierarchy was used to eliminate unnecessary features in three opinion-related classification tasks.
- The results on the OP data:

Features	Accuracy		F-measure	
	Baseline	δ = 0.002	Baseline	δ = 0.002
1Gram	97.5	-	84.5	-
1+2Gram	98.0	98.7	88.0	92.3
1Gram+EP	97.0	97.9	82.4	87.4
1+2Gram+EP	97.8	98.7	86.7	92.3

- Subsumption provided additional performance benefits over traditional Feature Selection as well:



Conclusions

- Our research investigates the use of a subsumption hierarchy to detect representational overlap in features and their relative performance value.
- Can be used:
  - as an analytic tool to compare features of different complexities.
  - to automatically remove unnecessary features, improving classification accuracy.