Learning Extraction Patterns for Subjective Expressions

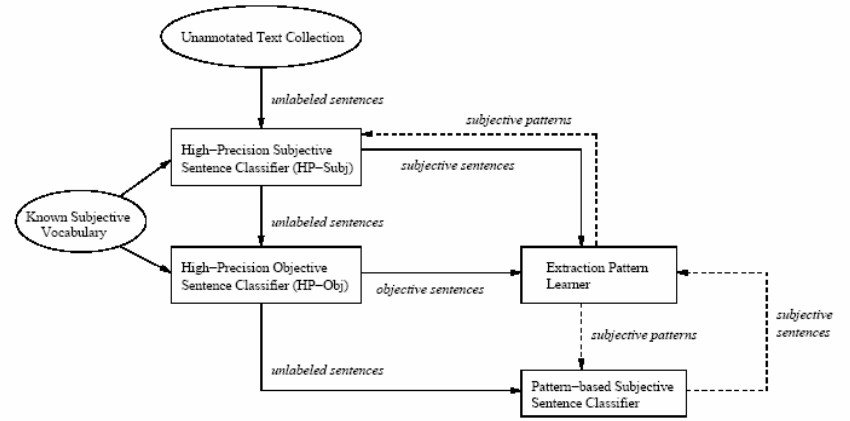
By Ellen Riloff and Janyce Wiebe

Presented by: Joe Dillman

Motivations

- IE systems should be able to discern facts and non-facts
- Annotated data difficult to find
- Believed that much data needed to train subjectivity classifers
- Like a way to automatically generate annotated corpora from small amounts of data using bootstrapped process

Bootstrapping Process



HP-Subj Classifier

- High-precision classifier used to identify subjective sentences
- Subjectivity classes:
  - Strongly subjective - seldom used without subjective meaning
  - Weakly subjective - commonly has subjective and objective use
- Identifies as subjective if 2 or more strongly subjective clues found
- P=91.5%, R=31.9%

HP-Obj Classifier

- High-precision classifier for objective sentences
- Identification heuristic:
  - No strongly subjective clues
  - At most one weakly subjective clue in previous, current and next sentence combined
- P=82.6%, R=16.4%

Learning Subjective Patterns

- Use Autoslog-TS
  - Relevant text: subjective sentences
  - Irrelevant text: objective sentences
- Two phases of processing as before:
  - Phase 1: Apply syntactic patterns to training corpus
  - Phase 2: Apply learned patterns to training corpus and rank

Autoslog-TS Syntactic Templates

SYNTACTIC FORM	EXAMPLE PATTERN
<subj> passive-verb	<subj> was satisfied
<subj> active-verb	<subj> complained
<subj> active-verb dobj	<subj> dealt blow
<subj> verb infinitive	<subj> appear to be
<subj> aux noun	<subj> has position
active-verb <dobj>	endorsed <dobj>
infinitive <dobj>	to condemn <dobj>
verb infinitive <dobj>	get to know <dobj>
noun aux <dobj>	fact is <dobj>
noun prep <np>	opinion on <np>
active-verb prep <np>	agrees with <np>
passive-verb prep <np>	was worried about <np>
infinitive prep <np>	to resort to <np>

- Removal of two rules:
  - passive-verb <dobj>
  - gerund <dobj>

Autoslog-TS: New ranking function

- Conditional probability:

$$Pr(subjective \mid pattern_i) = \frac{subjfreq(pattern_i)}{freq(pattern_i)}$$

- *subjfreq(pattern$_i$)* = frequency of pattern$_i$ in subjective training texts
- *freq(pattern$_i$)* = frequency of pattern$_i$ in all training sentences

Autoslog-TS: New ranking function

- Choose two threshold parameters: θ_1, θ_2
- Choose extraction patterns where:
  - *freq(pattern$_i$)* ≥ θ_1
  - *Pr(subjective | pattern$_i$)* ≥ θ_2

Interesting Generated Patterns

PATTERN	FREQ	%SUBJ
<subj> was asked	11	100%
<subj> asked	128	63%
<subj> is talk	5	100%
talk of <np>	10	90%
<subj> will talk	28	71%
<subj> put an end	10	90%
<subj> put	187	67%
<subj> is going to be	11	82%
<subj> is going	182	67%
was expected from <np>	5	100%
<subj> was expected	45	42%
<subj> is fact	38	100%
fact is <dobj>	12	100%

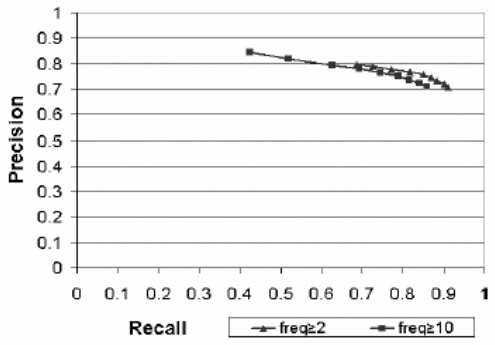
FREQ - # of times pattern appears in training data

%SUBJ – percentage of times seen in subj. sentences

- Presence of noun "*fact*" highly correlated with subjective expression!

Results: Learned Patterns Evaluation

- Precision from 71% to 85%



Results: Modifying HP-Subj

- Modify HP-Subj to use extraction patterns
  - Include originally labeled sentences
  - Add unlabeled sentence if:
    - Contains 2 or more learned patterns
    - Contains 1 original clue and at least 1 pattern
- Results:

HP-Subj		HP-Subj w/Patterns	
Recall	*Precision*	*Recall*	*Precision*
32.9	91.3	40.1	90.2

Concluding Remarks

- Bootstrapping process works to increase amount of labeled data
- Ranking function reduces need for human intervention
- Applications to extraction tasks
- “fact” ≠ fact