



Extracting Sources of Opinions from the World News

Siddharth Patwardhan
Ellen Riloff

[Opinions in the News]

- Newspaper articles are a rich source of opinions expressed by people.

“Mr Tsvangirai endorsed a United States proposal for targeted sanctions against President Mugabe.”

- It would be useful for a QA system to separate opinions from truly factual data.

Why Detect Opinions

- To answer questions, such as,

What is Utah's attitude towards President Bush's social security reform?

would require a subjective analysis of documents.

- Identifying subjectivity could improve the accuracy of a QA system in answering factual questions.

[Identifying Sources]

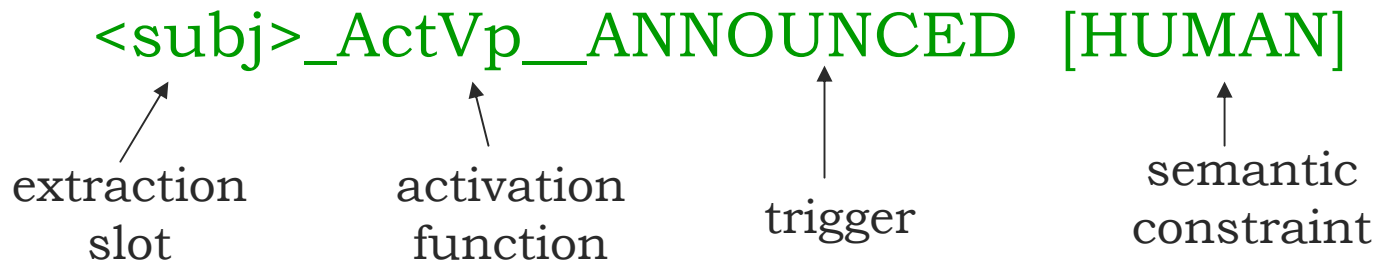
- Our focus is on the *sources* of opinions in text.
 - *Beijing* expressed its disapproval of the Report on Human Rights.
 - *Deseret News* reported that *Utah* will be supporting the INDUCE act.
 - *The Washington Post* reported *Mr. Blair's* view on the oil crisis.

Identifying Sources is *Information Extraction*

- Source Extraction naturally lends itself to being an IE problem.
- As an IE problem, our goal is to learn a set of patterns that, with good accuracy, identify pieces of text as sources.
- Ideally, we want a large set of patterns to achieve maximum coverage.

Case frames: Syntactic Patterns for IE

- Case frames consist of
 - Trigger word(s).
 - Activation functions.
 - Extraction “slot”(s).
 - Optional Semantic Constraints.



Case Frame Internals

- The Sundance (NLP) system parses a sentence.
- The system then searches for any trigger word(s) in the sentence.

Mr. Scindia **announced** his resignation last Monday.

<subj>_ActVp_**ANNOUNCED** [HUMAN]

Case Frame Internals

- If a trigger word is found, the activation functions are evaluated on the phrase containing the trigger.

Active Verb Phrase

Mr. Scindia **announced** his resignation last Monday.

<subj>_ActVp__ANNOUNCED [HUMAN]

Case Frame Internals

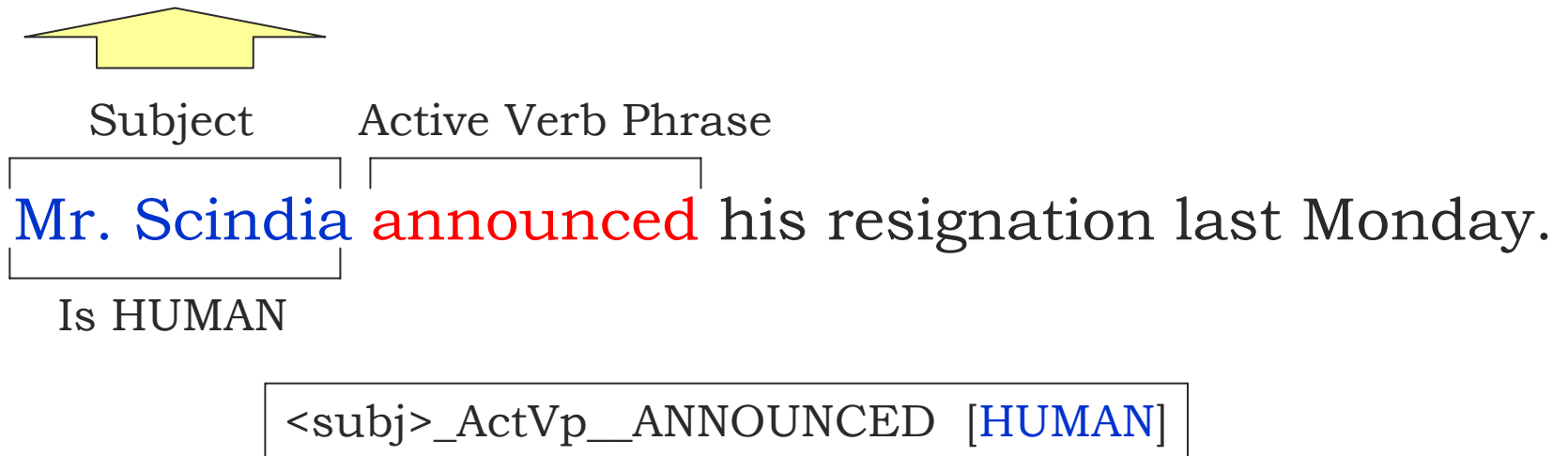
- If all activation functions test Ok, the system searches for available slots.

Subject Active Verb Phrase
┌──────────┬──────────┐
Mr. Scindia announced his resignation last Monday.

<subj>_ActVp__ANNOUNCED [HUMAN]

Case Frame Internals

- If any of the slots are present, and meet the semantic constraints (if any), they are extracted!



[AutoSlog]

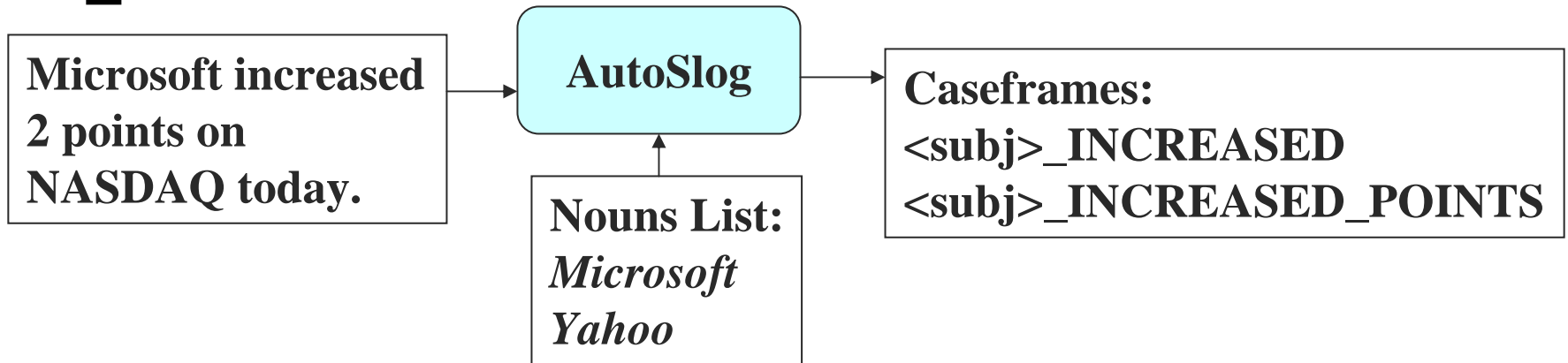
- AutoSlog is an IE system that learns syntactic patterns (*case frames*) from annotated data.

<subj>_BELIEVES

STATED_<dobj>

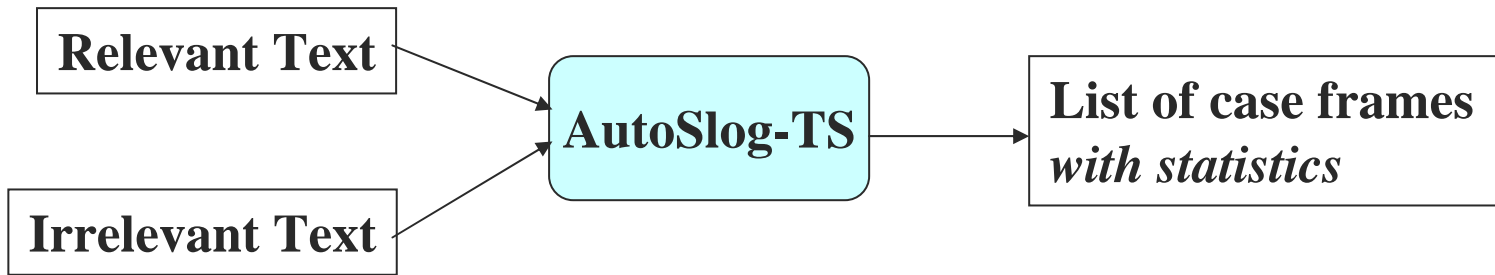
<subj>_MADE_STATEMENT

[AutoSlog]



- Given text and a list of desired extractions, AutoSlog generates case frames that extract those nouns.
- AutoSlog is not perfect. A human must review the generated case frames.

AutoSlog-TS



- Uses sets of *relevant* and *irrelevant* texts to create case frames, and sort them by probability and frequency.
- Human must review top-ranked case frames and assign role labels.

Source Extraction with AutoSlog-SE

- A statistical variation of the supervised AutoSlog learning algorithm.
- In the first pass, AutoSlog runs “exhaustively” on training texts and generates all possible case frames.
- In the second pass, the case frame extractions are scored against the annotated training data -- *probability* and *frequency* are measured.

AutoSlog-SE: First Pass

CNN informed watchers of the massive epidemic of ebola in Tanzania.

Reuters reports the stabbing of a 29-year old man in Santa Clara.

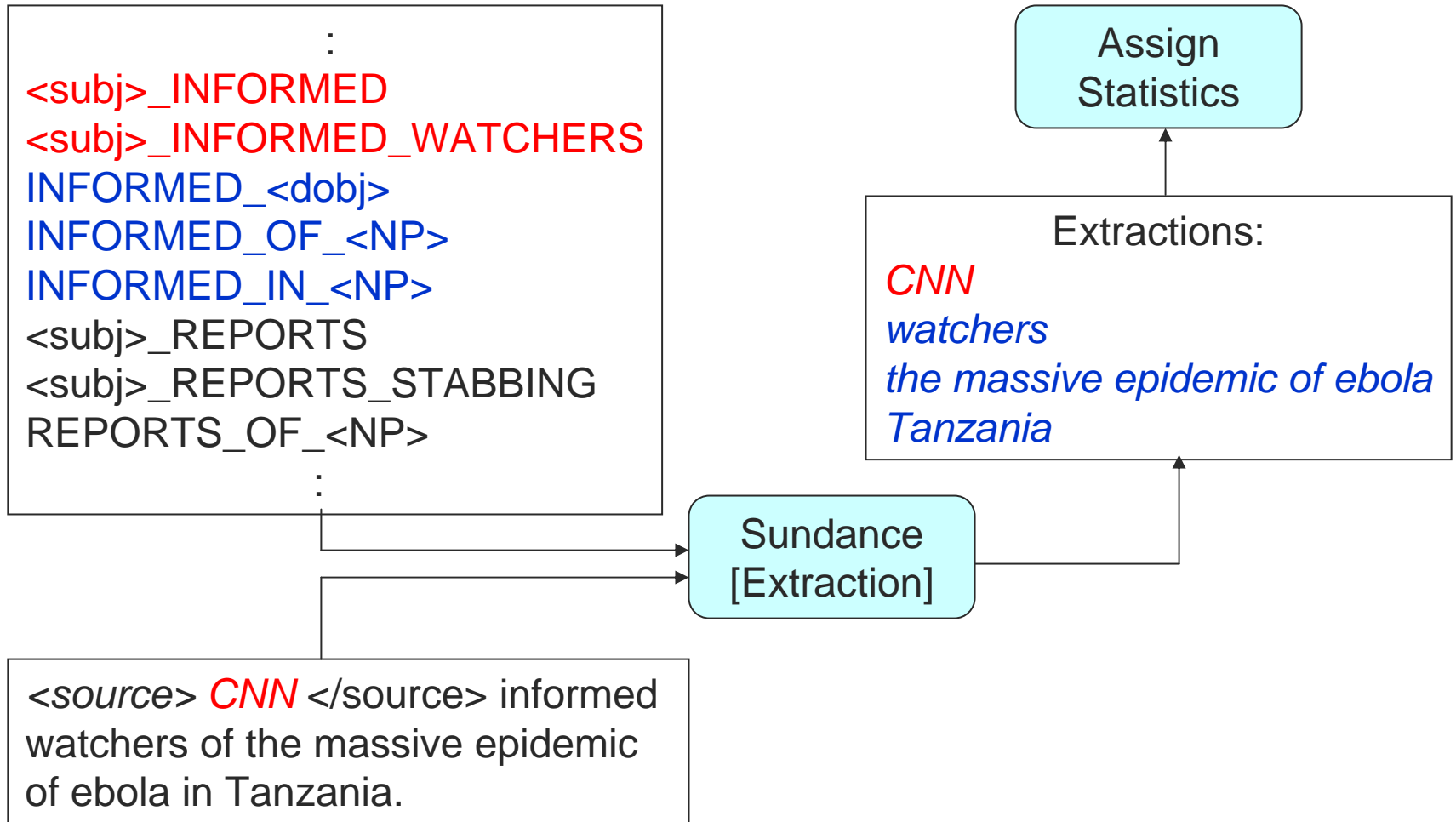
President Bush informed the Pentagon of his intention to withdraw troops.

Training text

AutoSlog
[Exhaustive]

<subj>_INFORMED
<subj>_INFORMED_WATCHERS
INFORMED_<dobj>
EPIDEMIC_OF_<NP>
INFORMED_IN_<NP>
<subj>_REPORTS
<subj>_REPORTS_STABBING
REPORTS_<dobj>
REPORTS_IN_<NP>
STABBING_OF_<NP>
<subj>_INFORMED_PENTAGON
PENTAGON_OF_<NP>

AutoSlog-SE: Second Pass



Semantic Tags

- Sundance assigns *Semantic Tags* to words, using its dictionary. E.g.
 - POLICEMAN: PERSON
 - BEIJING: CITY
- Semantic Tags can be used to apply *Selectional Restrictions* to extractions.

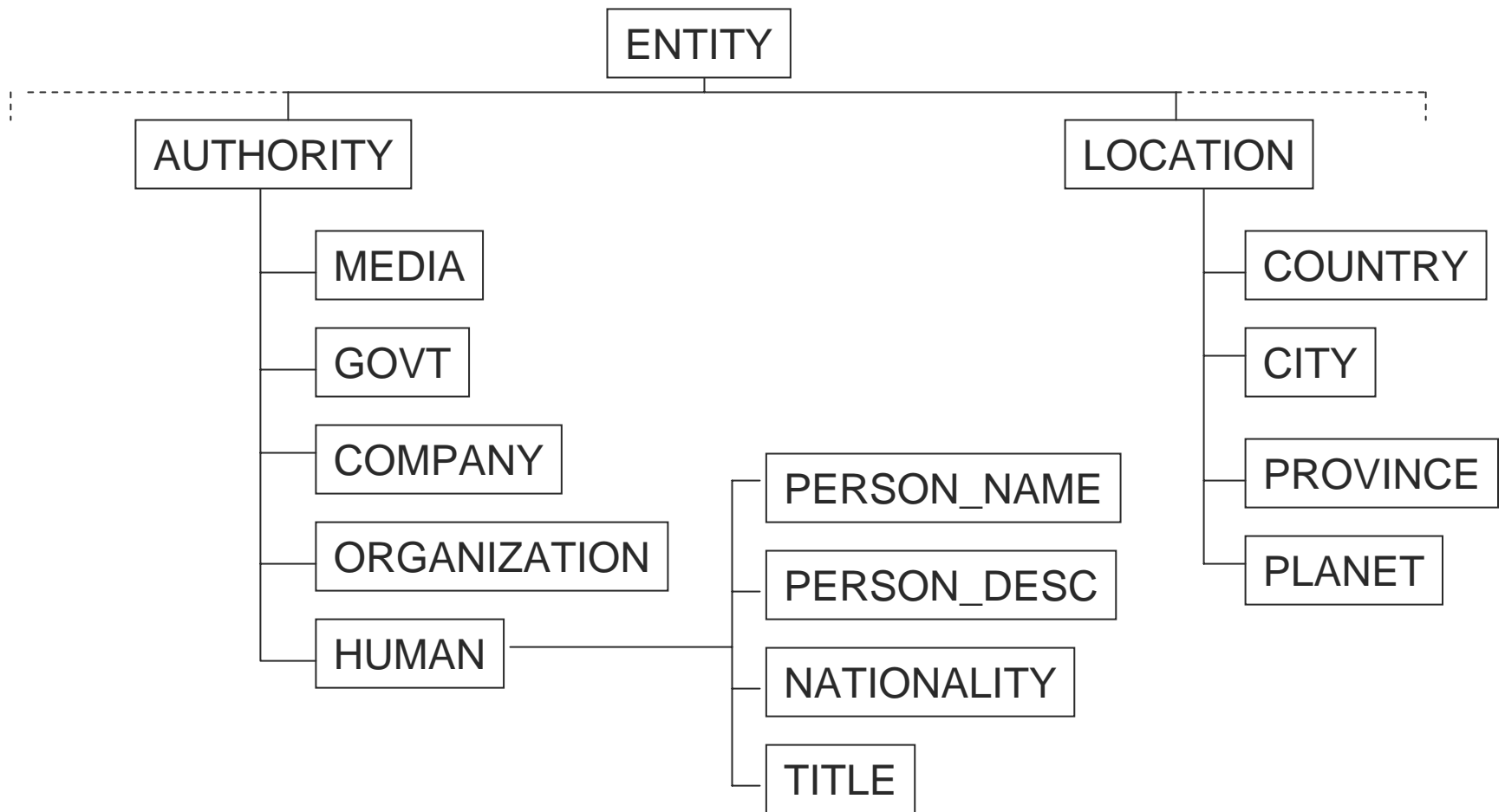
Selectional Restrictions

- To apply *selectional restrictions* to extractions, semantic constraints are attached to case frames. E.g.

<subj>_REPORTED [HUMAN]

which reads as “the subject of REPORTED should be extracted only if it is HUMAN (semantic tag)”.

Hierarchy in the Semantic Tags



Named Entity Patterns

- Patterns that capture certain special cases.

```
<NP-ANYWORD:sem=TITLE  
& ALLWORDS:case=Capitalized>
```

E.g. Mr. Zakhir Hussein

```
<NP-ANYWORD:sem=PERSON_NAME  
& ANYWORD:sem=PERSON_DESC>
```

E.g. dispatcher Shirley Iker

Named Entity Patterns

- Patterns that combine larger pieces of text into a single phrase.

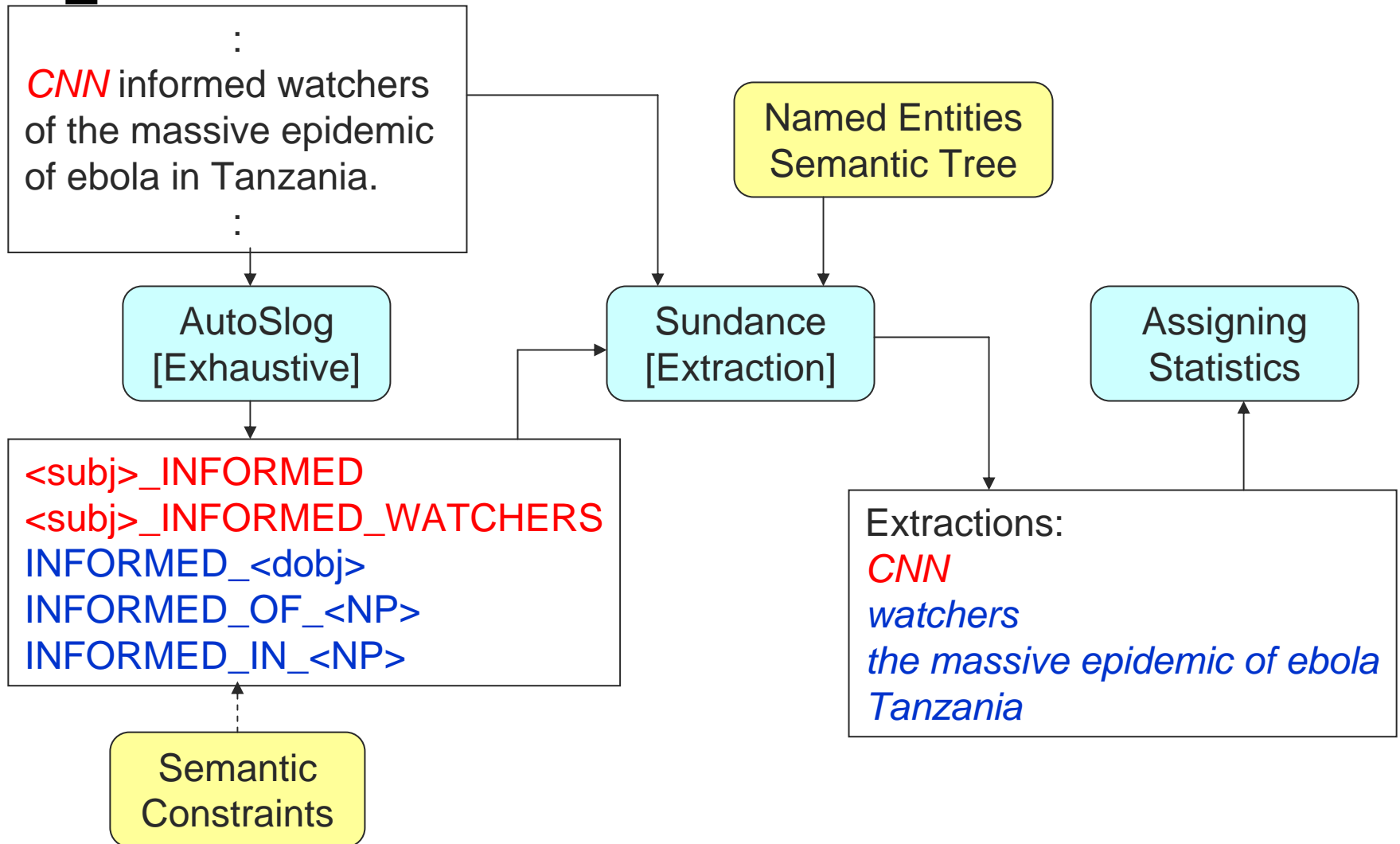
<NP-HEAD:sem=PERSON_DESC>
<WORD:word=including>
<NP-HEAD:sem=PERSON_DESC>

E.g. administration officials including the secretary

<NP-HEAD:word=republic>
<WORD:word=of>
<NP>

E.g. Islamic Republic of Mauritania

AutoSlog-SE: Summary



[Applying Cutoffs]

Total case frames after the 2 passes	~27,000
Case frames with $F > 2$	~1,950
Case frames with $F > 2, P > 0.1$	~750
Case frames with $F > 2, P > 0.3$	~500
Case frames with $F > 2, P > 0.5$	~265
Case frames with $F > 2, P > 0.7$	~170
Case frames with $F > 2, P > 0.9$	~75

[Annotated Data]

- 400 MPQA documents manually annotated for subjectivity.
- ~500 words per document.
- We are evaluating against the agent tags (indicating *sources* of opinions).
- For evaluation purposes, we flag an extraction as correct if its head noun is the same as that of an annotated source in that sentence.

10-Fold Cross-validation

- 10 folds of the data set were created.
- 9 folds were used for training (creating the set of case frames).
- The sets of case frames with *frequency* values greater than 2, and different *probability* cutoffs, were evaluated on the 10th test fold.

(This was repeated 10 times and results averaged)

Results

F>2:

Cutoff	Precision	Recall	CFCount
P>0.1	0.536	0.488	741
P>0.3	0.693	0.451	503
P>0.5	0.797	0.397	267
P>0.7	0.840	0.327	169
P>0.9	0.841	0.057	76

Augmenting Case Frames

- Good precision at high cutoffs.
- But only moderate recall.
- Due to the small training set and the small number of case frames (providing a smaller coverage).
- We try to infer new case frames from the existing set to improve coverage.

[Syntactic Variations]

- Intuition: If `<subj>_REPORTED` is a good case frame, then most likely, so is `WAS_REPORTED_BY_<NP>`
- We automatically generate syntactic variations of the “good” case frames (i.e. those above a certain cutoff), and add them to the existing set.

[Syntactic Variations]

PassiveVP by <NP> → <subj> ActiveVP

E.g. was reported by CNN → CNN reported

<subj> ActiveVP → PassiveVP by <NP>

E.g. Reuters reported → was reported by Reuters

ActiveVP <dobj> → <subj> PassiveVP

E.g. worried John → John was worried

[Syntactic Variations]

<subj> PassiveVP → InfinitiveVerb <NP>

E.g. Mugabe was quoted → to quote Mugabe

ActiveVP <dobj> → InfinitiveVerb <NP>

E.g. quoted Mugabe → to quote Mugabe

NP(Singular) Prep <NP> ↔ NP(Plural) Prep <NP>

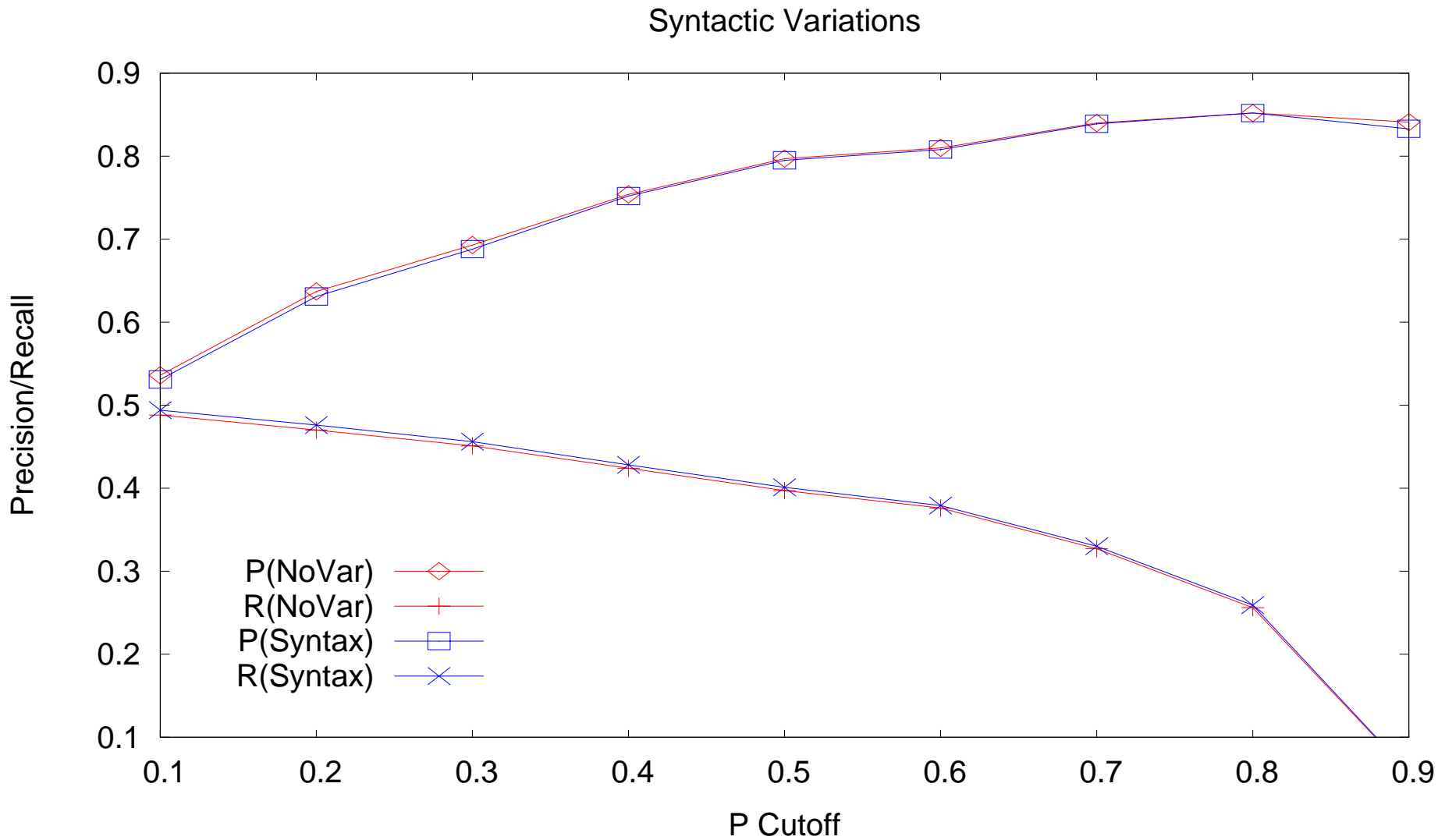
E.g. statement by the Pope ↔ statements by the Pope

Results

F>2, Syntactic Variations:

Cutoff	Prec'n	P-	Recall	R+	CFs
P>0.1	0.531	0.004	0.494	0.006	1036
P>0.3	0.688	0.005	0.456	0.005	716
P>0.5	0.795	0.002	0.401	0.004	388
P>0.7	0.839	0.001	0.330	0.003	248
P>0.9	0.833	0.008	0.058	0.001	109

[Results



Semantic Variations

- Intuition: If `<subj>_STATES` is a good case frame, then most likely, `<subj>_SAYS` is also a good case frame.
- We use *synonyms* and *hyponyms* of verbs (from *WordNet*) to create semantic variants of existing case frames.

[WordNet [Fellbaum98]]

- Semantic network.
- Nodes represents real world concepts.
- Rich network of relationships between these concepts.
- Relationships such as “*car is a kind of vehicle*”, “*high is opposite of low*”, etc. exist.
- Node = Synonym Set (and a definition)

WordNet Structure

{armament}
*Weaponry used by
military or naval force*

{battery}
*Group of guns or missile launchers
operated together at one place*

Is-a

Has-part

{artillery, heavy weapon, gun, ordinance}
Large, but transportable, armament

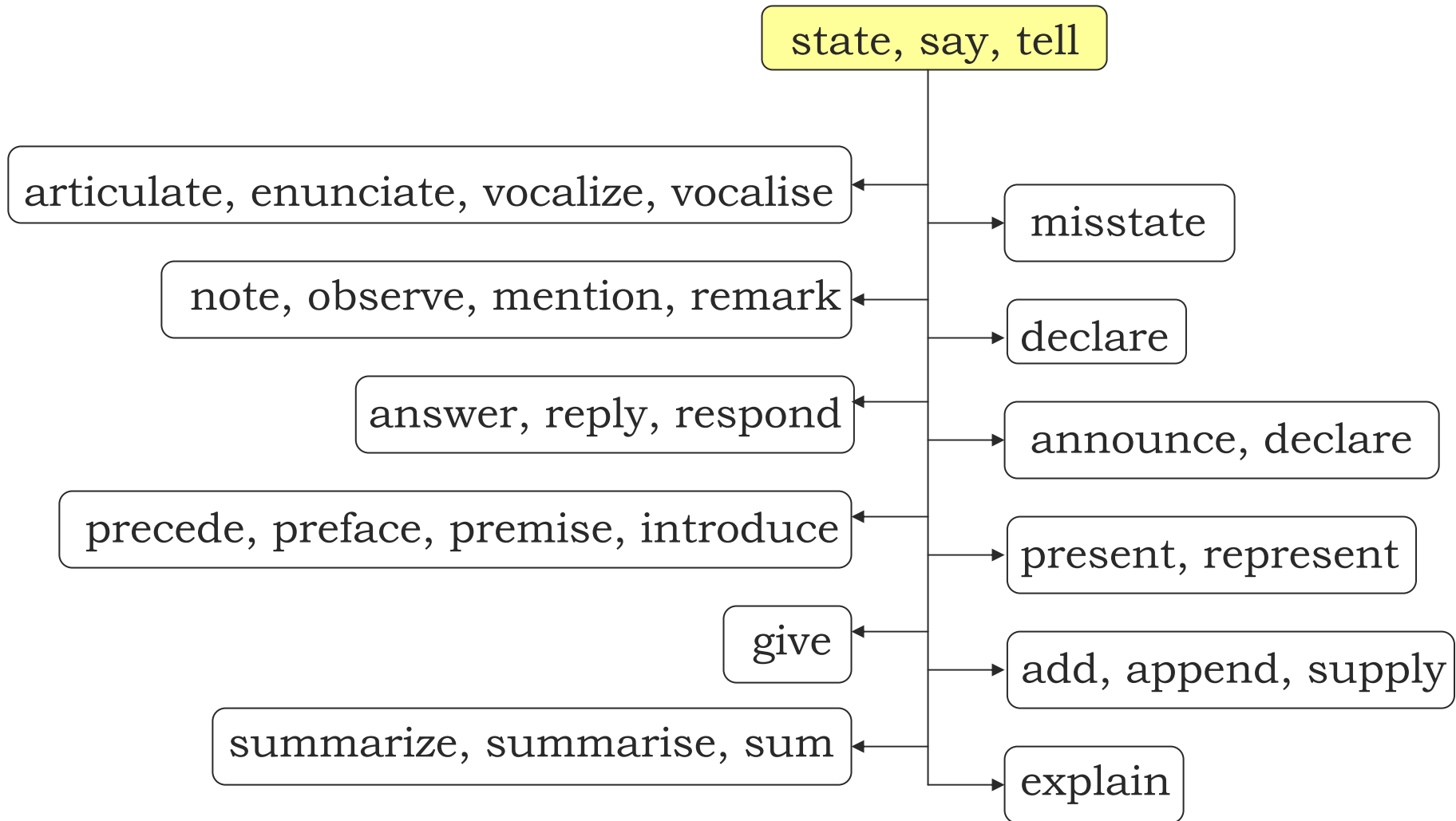
Is-a

Has-part

{cannon}
*A large artillery gun
that is usually on wheels*

{stock, gunstock}
*The handle of a handgun or the butt end
of a rifle or shotgun*

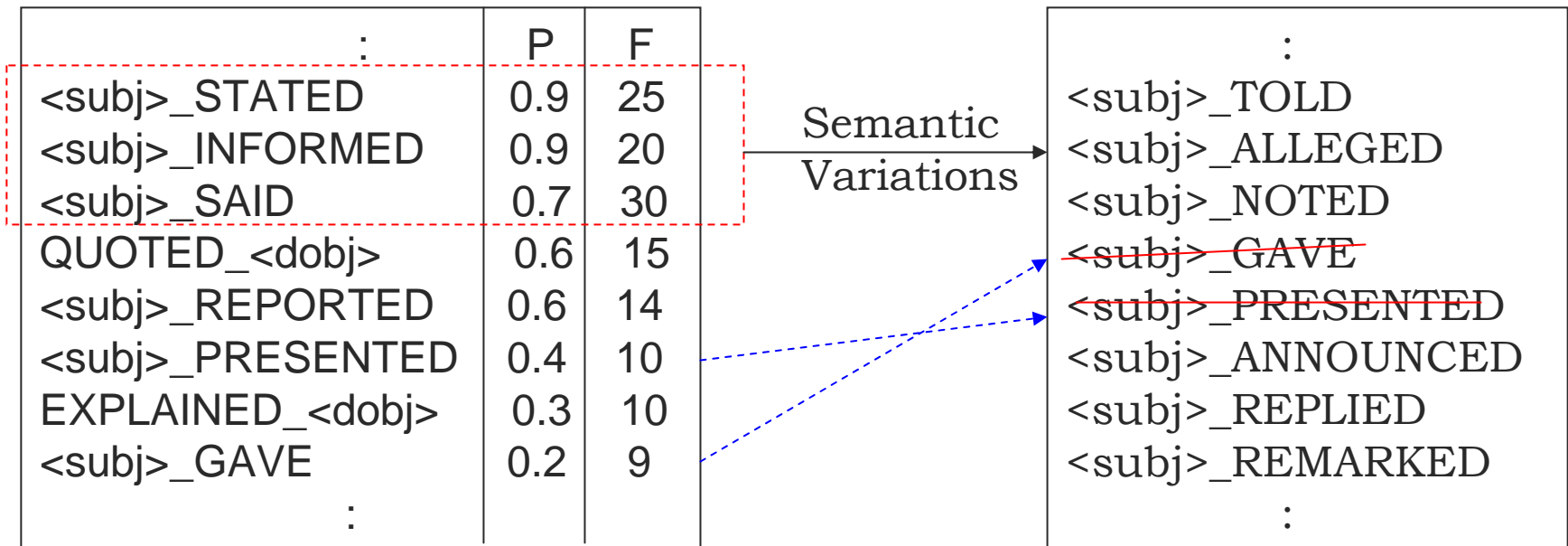
[Synsets and Hyponyms]



[One More Filter]

- After generating the variations of the good case frames, we check these variations against the original set.
- Remove any variations that already exist in the original set, but with poor statistics (precision & frequency).

[One More Filter]



[Results]

F>2, Synonyms:

Cutoff	Prec'n	Pr-	Recall	R+	CFs
P>0.1	0.520	0.016	0.507	0.019	1452
P>0.3	0.663	0.030	0.467	0.016	1019
P>0.5	0.784	0.013	0.407	0.010	563
P>0.7	0.831	0.009	0.333	0.006	357
P>0.9	0.817	0.024	0.062	0.005	174

[Results]

F>2, Hyponyms:

Cutoff	Prec'n	P-	Recall	R+	CFs
P>0.1	0.505	0.031	0.513	0.025	3693
P>0.3	0.637	0.056	0.475	0.024	2630
P>0.5	0.769	0.028	0.412	0.015	1263
P>0.7	0.818	0.022	0.336	0.009	748
P>0.9	0.800	0.041	0.062	0.005	382

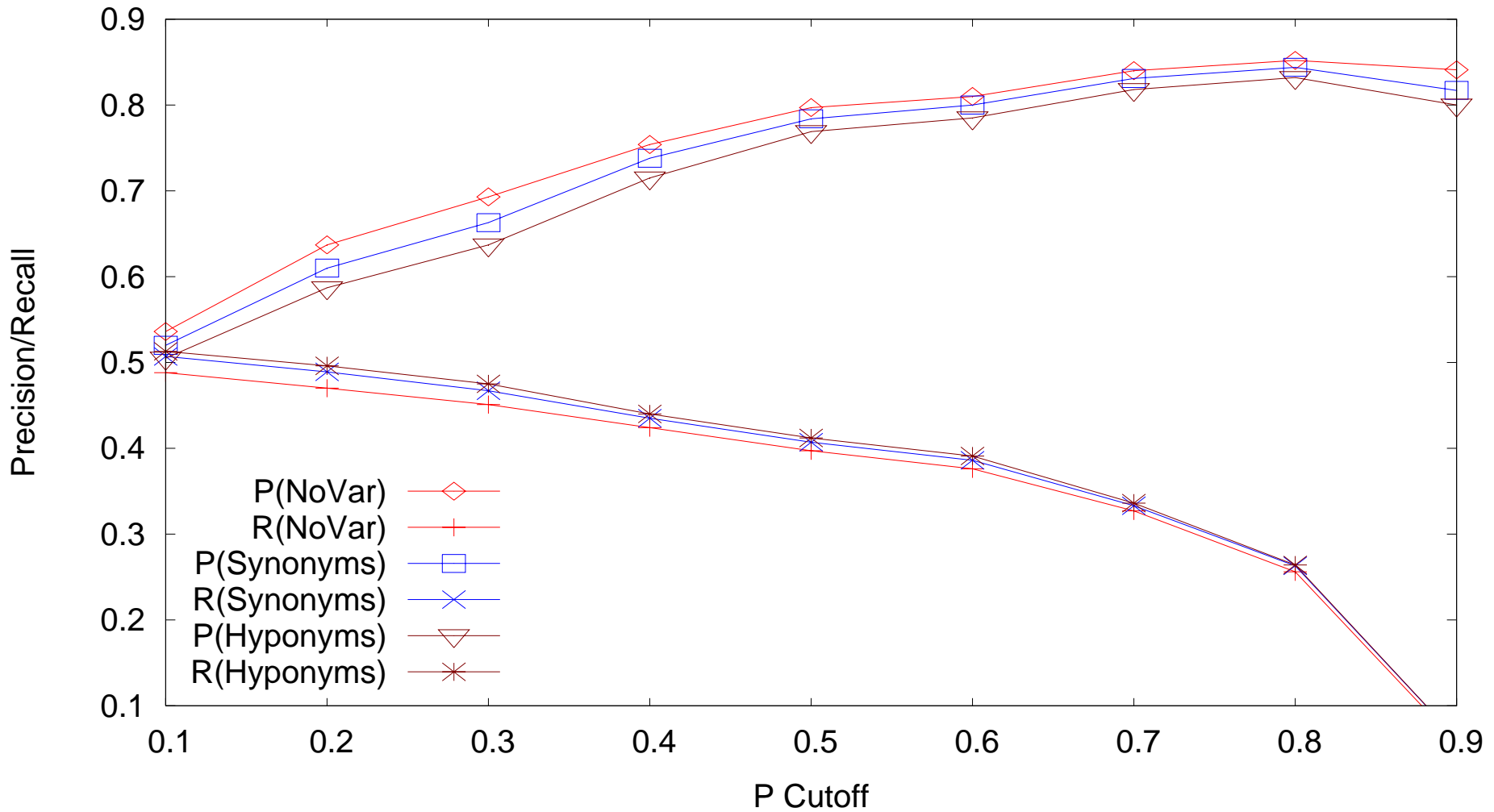
Results

F>2, Synonyms + Hyponyms:

Cutoff	Prec'n	P-	Recall	R+	CFs
P>0.1	0.483	0.053	0.527	0.039	5555
P>0.3	0.586	0.107	0.494	0.043	4223
P>0.5	0.729	0.068	0.430	0.033	1979
P>0.7	0.781	0.059	0.354	0.027	1215
P>0.9	0.754	0.087	0.099	0.042	664

[Results]

Synonyms and Hyponyms

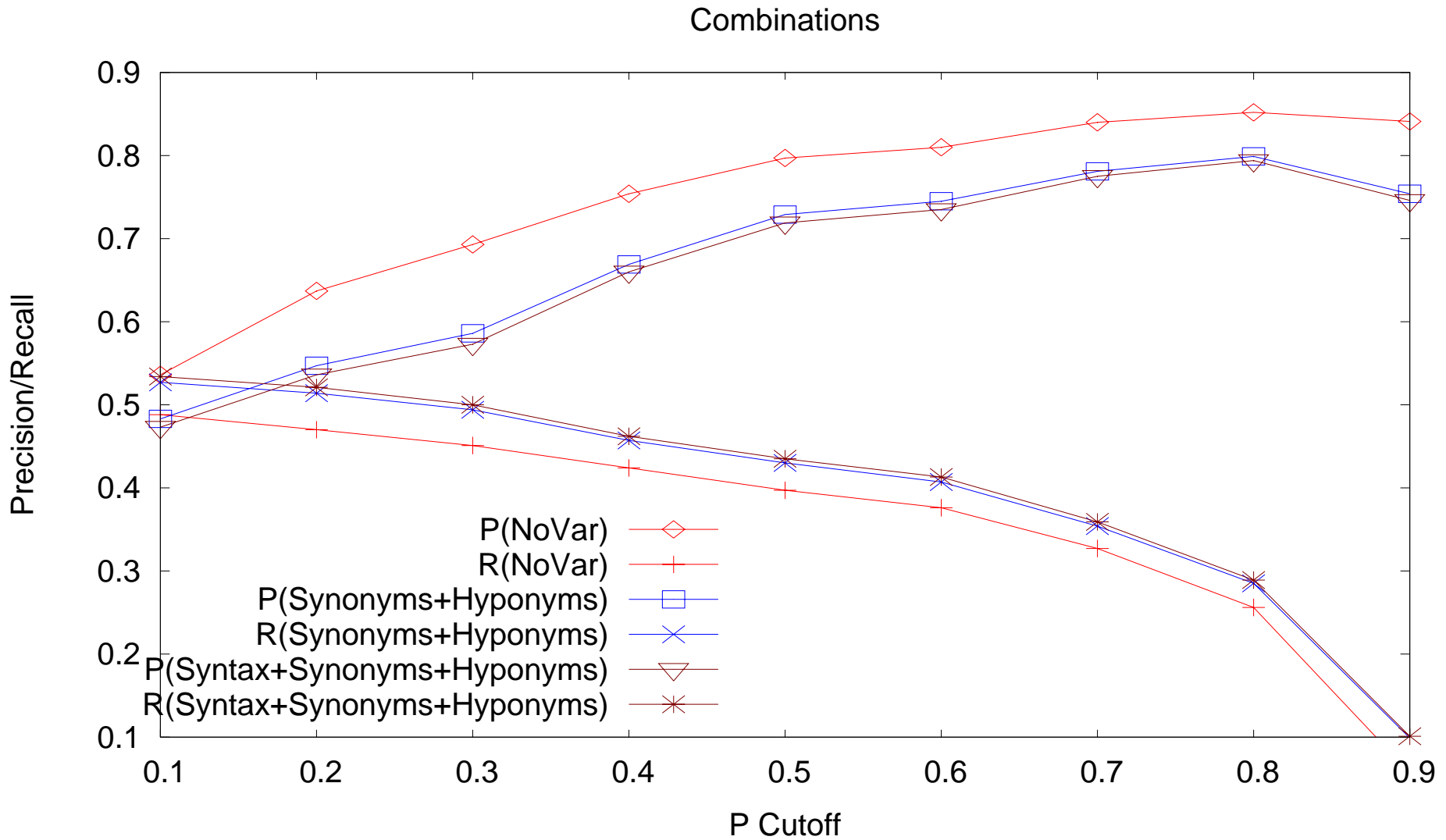


[Results]

F>2, Syntactic + Synonyms + Hyponyms:

Cutoff	Prec'n	P-	Recall	R+	CFs
P>0.1	0.473	0.063	0.534	0.046	7601
P>0.3	0.573	0.120	0.500	0.049	5922
P>0.5	0.719	0.078	0.435	0.038	2904
P>0.7	0.775	0.065	0.359	0.032	1817
P>0.9	0.746	0.095	0.101	0.044	968

Results



[Conclusions]

- We have a technique that gives us high precision in extracting sources.
- Despite the small training set, the syntactic and semantic expansions of case frames allow us to increase coverage a little (with a precision tradeoff), and seems like a promising path to explore.

[Future Work]

- Evaluate performance only on sentences containing opinions.
- Evaluate new patterns on their ability to extract *Authorities* from unannotated text.
- Integrating our patterns and extractions as features to a ML algorithm for opinion detection (Cornell).