

What Is Paraphrase?

- Expressing one thing in various ways:
 - *Bush says he'll deliver \$20 billion to NY (CNN)*
 - *Bush, in New York, affirms \$20 billion aid pledge (New York Times)*
 - *Bush reassures New York of \$20 billion (Washington Post)*
- Variation in lexicon and syntax.
- Makes understanding more difficult.



Our Goal

- Why do we want to collect paraphrases?
 - Mainly for Information Extraction (IE).
 - Obtain the relationship of IE patterns automatically.
 - Can apply to other applications (IR, QA, MT etc.)
- We hope to develop a system for collecting paraphrase automatically from corpora.

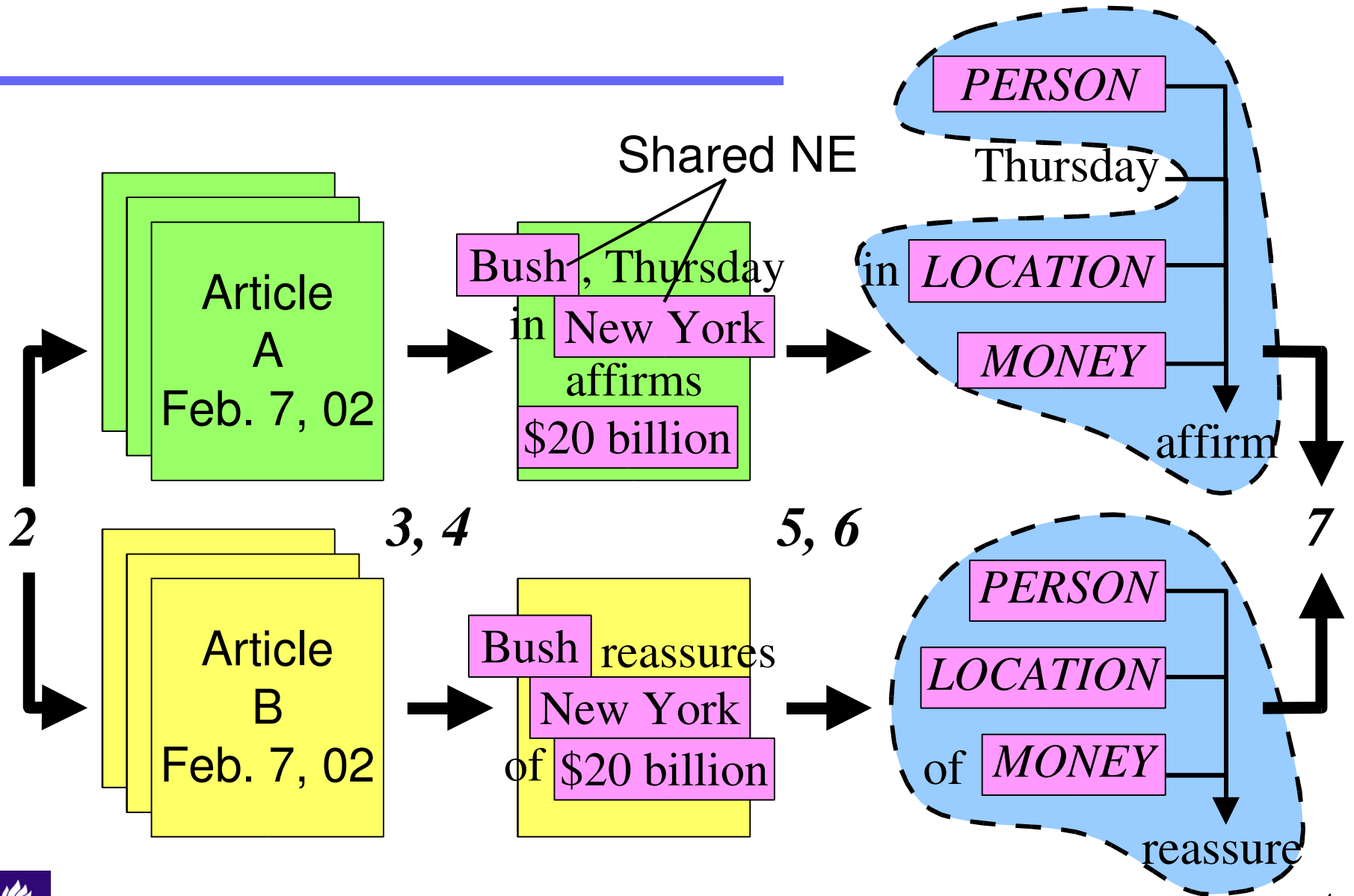


Concepts

- Obtain from news articles.
 - Two sentence reporting the same event can contain paraphrases.
- Use Named Entity (NE) as a clue for finding paraphrases.
 - NE is preserved across comparable expressions.
 - “Bush”, “New York”, “\$20 billion” ...



Method



Method

1. Obtain articles of a certain domain using IR. [Murata, 94]
2. Find article pairs on same event. [Papka, 99]
3. Perform NE tagging. [Uchimoto, 00]
4. Find sentence pairs using NEs.
5. Mark the comparable NEs.
6. Generate dependency trees for each sentence.
7. Extract a pair of paraphrases.



Limiting the Expressions

- Want to avoid recognizing these expressions as paraphrases:
 - “Bush has expressed his confidence in Koizumi’s reforms.”
 - “Bush and Koizumi watched a demonstration of horseback archery.”
- Obtain IE patterns for the domain automatically using statistical analysis in advance. [Sudo, 01]
- Find paraphrases only among IE patterns.



Experiments

- We used two Japanese newspapers (Mainichi, 111373 articles and Nikkei, 181086 articles)

Domain	Arrest	Personnel
Article pairs	294	289
Sentences	4445	5962
Obtained IE patterns	725	157
Paraphrase pairs	53	83



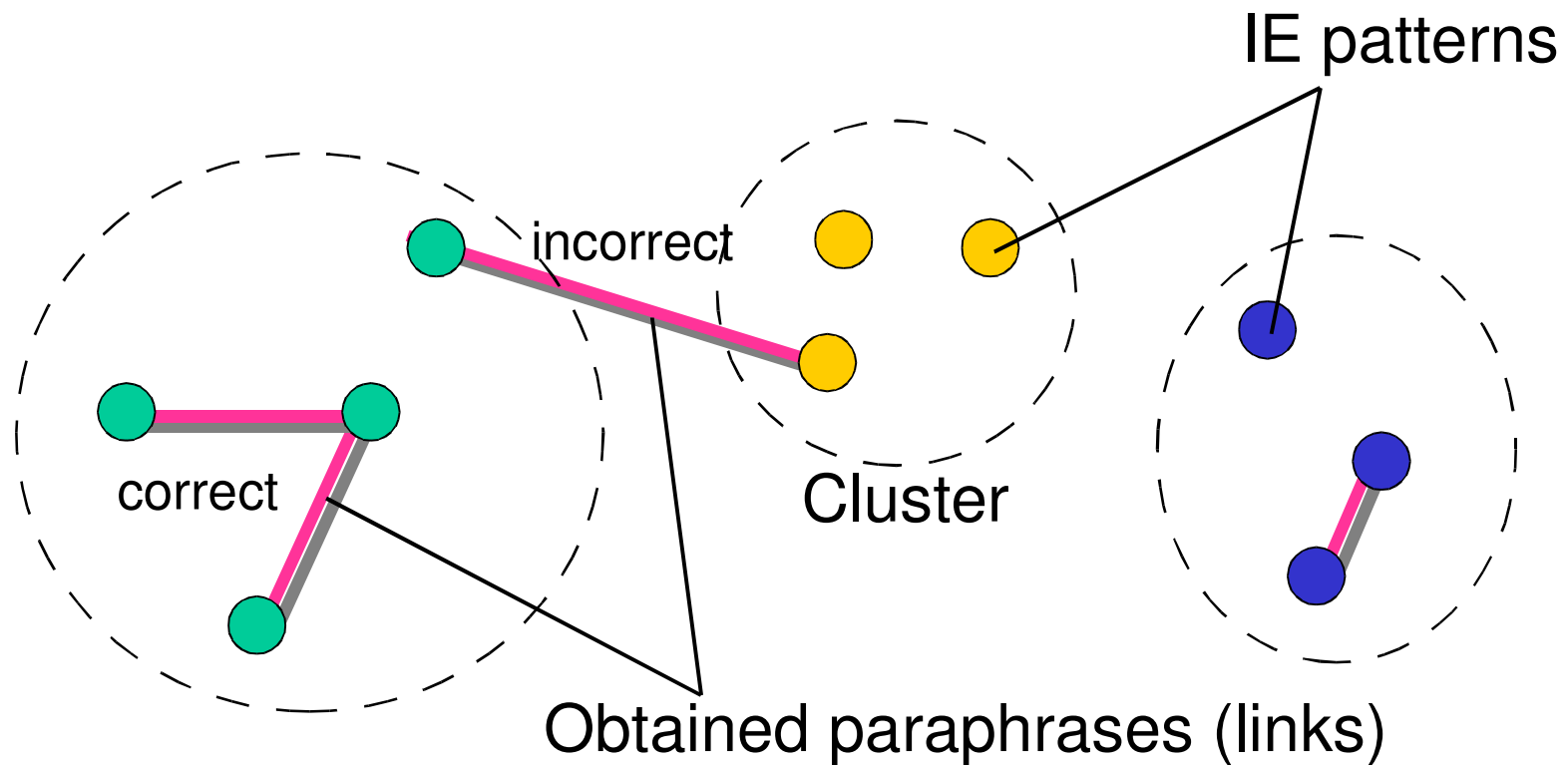
How to Evaluate?

- What is a correct paraphrase pair?
 - Do they describe the same event?
 - If we use them in an actual IE application, do they capture the same information?
 - Correct
 - *ORGANIZATION*₁ decides [*something*].
 - *ORGANIZATION*₁ confirms [*something*].
 - Wrong
 - [*someone*] is promoted to *POST*₁.
 - *POST*₁ is promoted.



How to Evaluate?

- Cluster IE patterns by hand.
 - (Patterns in the same cluster can be regarded as paraphrases)



Results - Precision

- How many obtained links connect patterns within a cluster?

Domain	Arrest	Personnel
Clusters	111	20
Paraphrase pairs (links)	53	83
Correct	26	78
Precision	49%	94%



Results - Coverage

- How many additional links are necessary to connect all the patterns in each cluster?
(How well do the links form clusters?)

$$\text{Coverage} = 1 - \frac{\textit{Additional links necessary}}{\textit{Total links necessary}}$$

Domain	Arrest	Personnel
Additional links necessary	230	57
Total links necessary	252	109
Coverage	9%	47%



Results - Sample paraphrases

- Correct
 - *PERSON*₁ admits [*something*].
 - *PERSON*₁
 - *PERSON*₁ testifies [*something*].
 - *PERSON*₁
 - [*someone*] is promoted to *POST*₁.
 - *POST*₁
 - The promotion to *POST*₁ is decided.
 - *POST*₁



Results - Sample paraphrases

- Incorrect
 - *PERSON*₁ is arrested.
 - *PERSON*₁
 - *PERSON*₁ conspires.
 - *PERSON*₁
 - *PERSON*₁ is promoted.
 - *PERSON*₁
 - *PERSON*₁ hold successively [*something*].
 - *PERSON*₁



Discussion

- Precision in the arrest domain was low.
 - Short and varied IE patterns obtained.
 - Many patterns include only one NE.
 - NE matching (coreference) problem.
 - “New York City”, “NYC”, “the city”
- Limited forms of IE patterns. [Sudo, 01]
- Low coverage is not a big problem.
 - We can get lots of newspapers (compared to [Barzilay, 01]).



Future Work

- Extend the forms of IE patterns.
- Use common nouns (not only NEs) for sentence matching.
- Use contextual information to find paraphrases.
- Try to obtain without IE patterns.

