

Paraphrase Acquisition For Information Extraction

Yusuke Shinyama

Satoshi Sekine

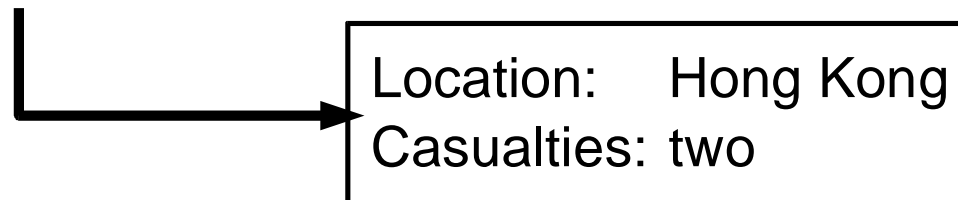
Computer Science Department

New York University



What Is Information Extraction?

- Extract certain kinds of information from articles in a pre-defined domain.
 - “Hong Kong reported two more deaths and 61 fresh cases of SARS Friday as governments across the world took tough steps to stop the killer virus at their borders.”



- Performed by template pattern matching.
 - “COUNTRY reported NUMBER more deaths.”



A Problem

- We need to prepare various patterns to capture the same kind of events.
 - “Hong Kong reported two more deaths and 61 fresh cases of SARS Friday as governments across the world took tough steps to stop the killer virus at their borders.”
 - “The government has announced that two more people have died in Hong Kong after contracting the SARS virus and 61 new cases of the illness have been detected.”



Our Goal

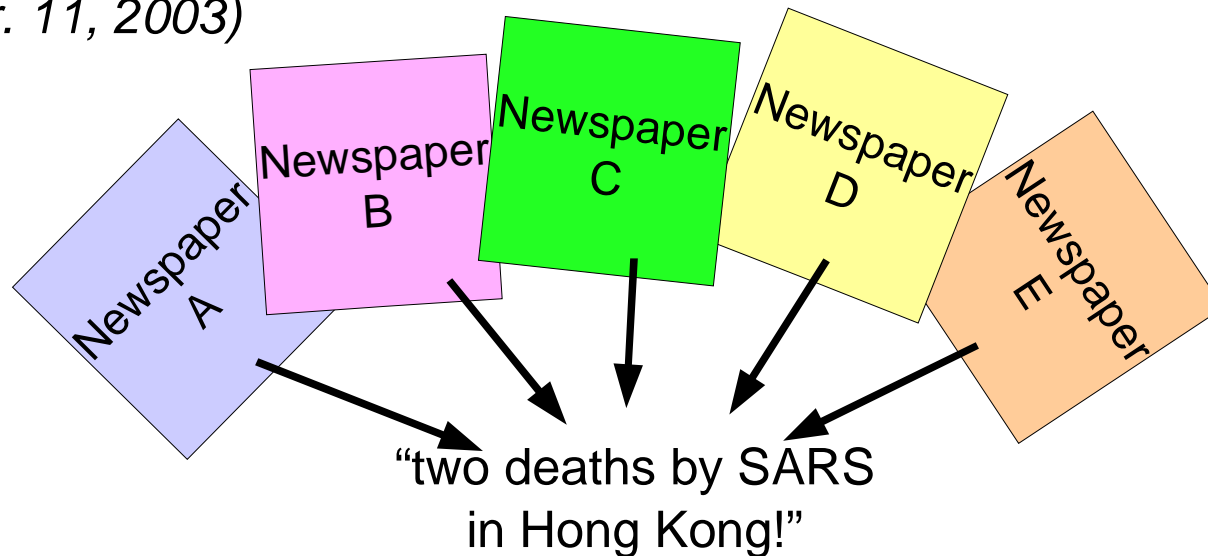
- We want to obtain various expressions in a specific domain.
- Toward this goal, we are trying to connect two expressions which describe the same event; i.e. *domain dependent paraphrases*.
 - “COUNTRY reported NUMBER more deaths.”
 - “NUMBER more people died in COUNTRY.”



Basic Idea

- Some events are reported more than once on the same day, in different forms, in different newspapers.

(Apr. 11, 2003)



Basic Idea

- Actually...

- “Hong Kong reported two more deaths and 61 fresh cases of SARS Friday as governments across the world took tough steps to stop the killer virus at their borders.” *[Channel News Asia, Apr. 11, 2003]*

- “The government has announced that two more people have died in Hong Kong after contracting the SARS virus and 61 new cases of the illness have been detected.” *[Reuters, Apr. 11, 2003]*



Basic Idea

- Even though the form of expressions change, Named Entities don't.
 - “Hong Kong reported two more deaths and 61 fresh cases of SARS Friday as governments across the world took tough steps to stop the killer virus at their borders.”
 - “The government has announced that two more people have died in Hong Kong after contracting the SARS virus and 61 new cases of the illness have been detected.”



Basic Idea

- The portions of sentences which share these Named Entities should be paraphrases.
 - “Hong Kong reported two more deaths and 61 fresh cases of SARS Friday as governments across the world took tough steps to stop the killer virus at their borders.”
 - “The government has announced that two more people have died in Hong Kong after contracting the SARS virus and 61 new cases of the illness have been detected.”

anchors



Basic Idea

- Our tools
 - How to find sentences which report the same event?
 - Firm techniques have been developed for IR/TDT.
 - How to identify Named Entities?
 - There are a bunch of Named Entity taggers.
 - How to extract portions of sentences?
 - Dependency tree representation.



1. Finding Similar Sentences

- Two phases of matching process:
 - Article level matching
 - TF/IDF based technique for TDT (Topic Detection and Tracking).
 - Sentence level matching
 - Similar method to the article level matching.
- We used two Japanese newspapers as sources in this experiment.



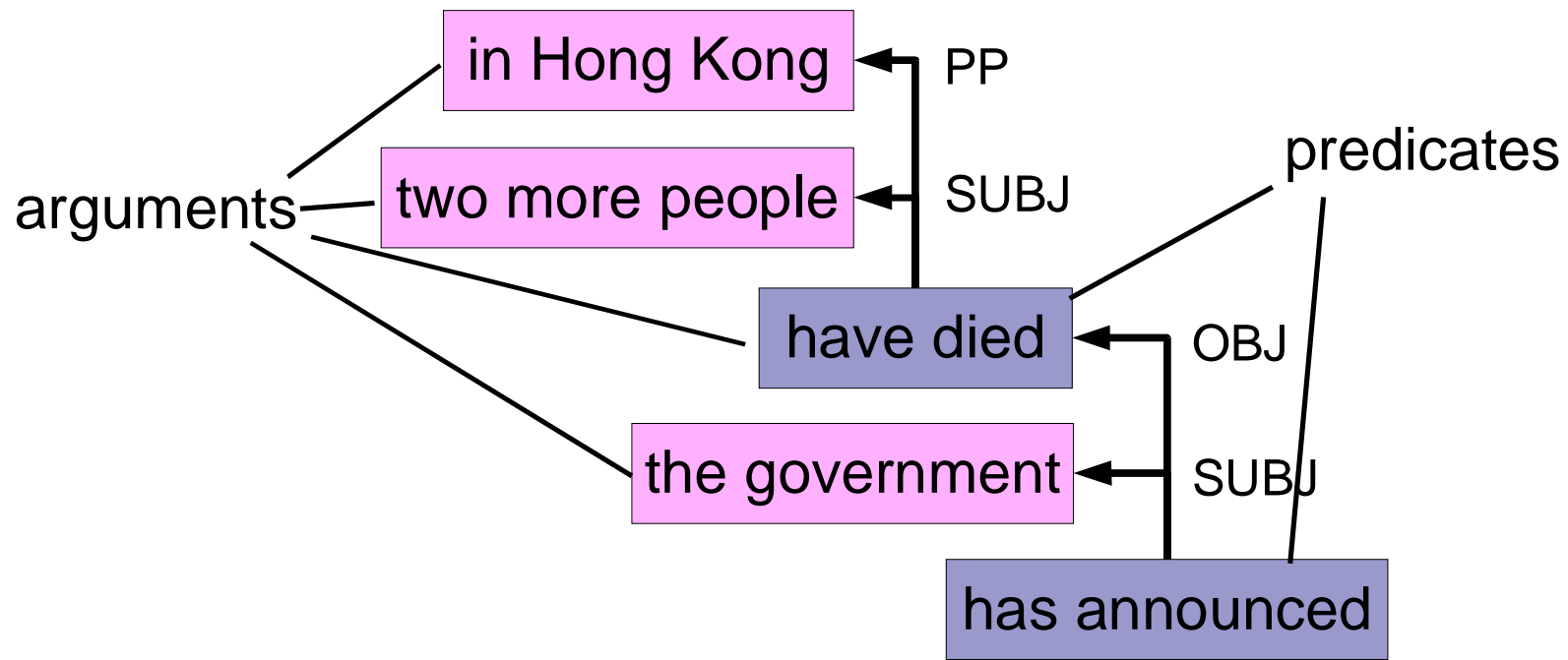
2. Identifying Anchors

- We used 6 categories of Named Entities, and performed simple coreference resolution.
 - “The government has announced that two more people have died in Hong Kong after contracting the SARS virus and 61 new cases of the illness have been detected.”
 - “Hong Kong reported two more deaths and 61 fresh cases of SARS Friday as governments across the world took tough steps to stop the killer virus at their borders.”



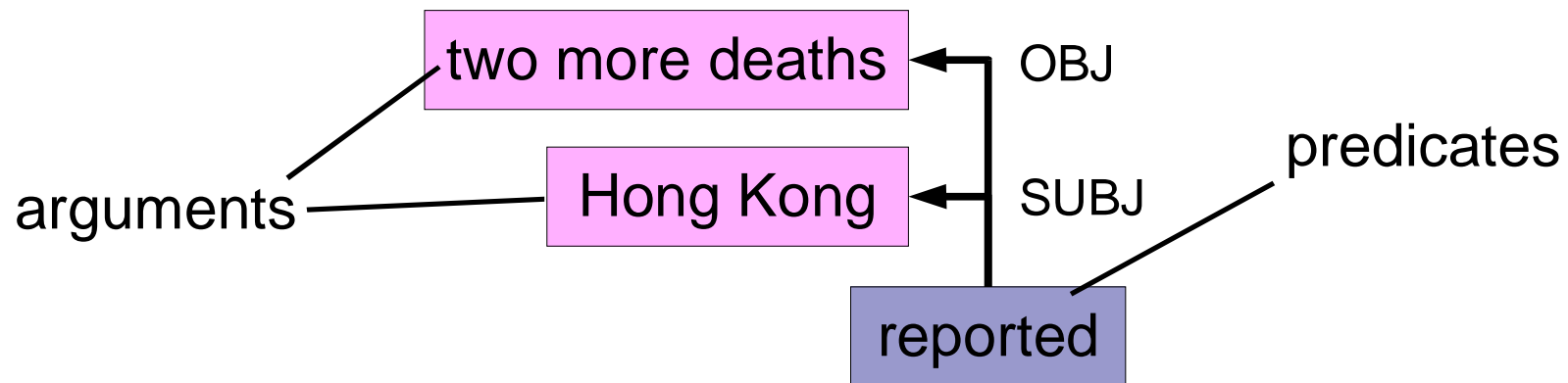
3. Dependency Analysis

- “The government has announced that two more people have died in Hong Kong after
...”
...



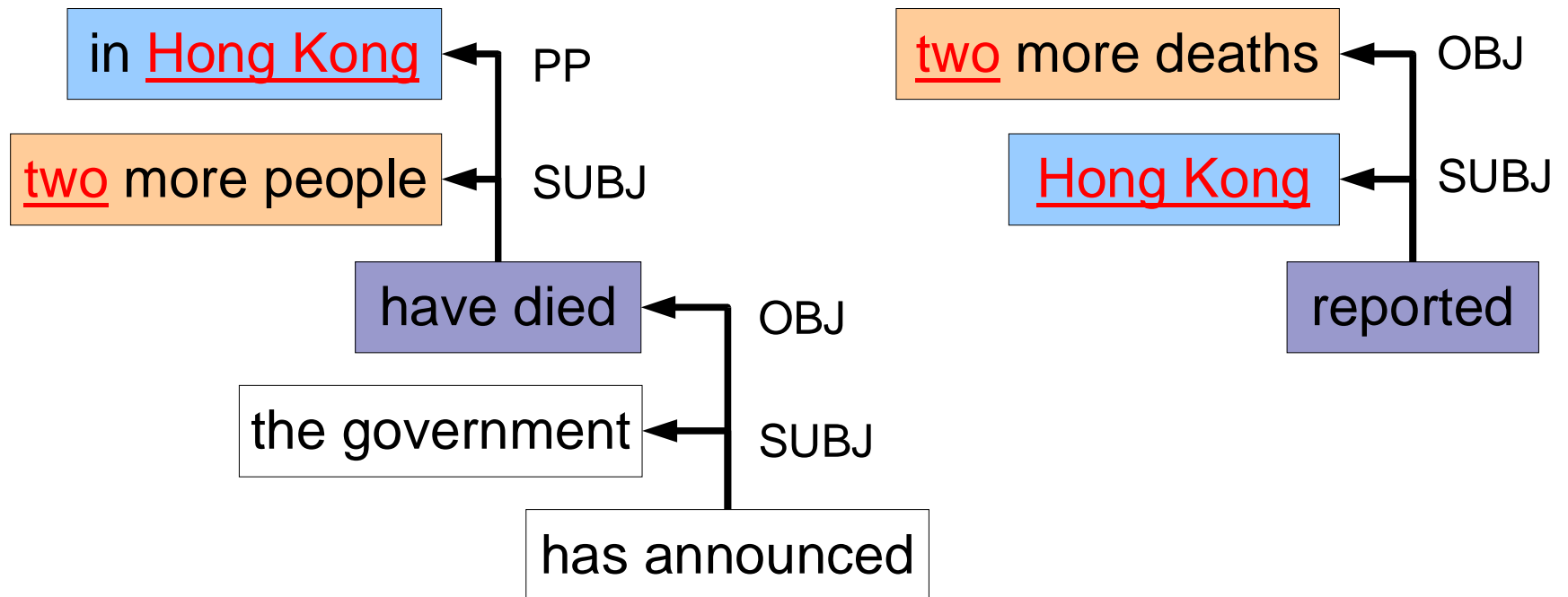
3. Dependency Analysis

- “Hong Kong reported two more deaths and
”
...



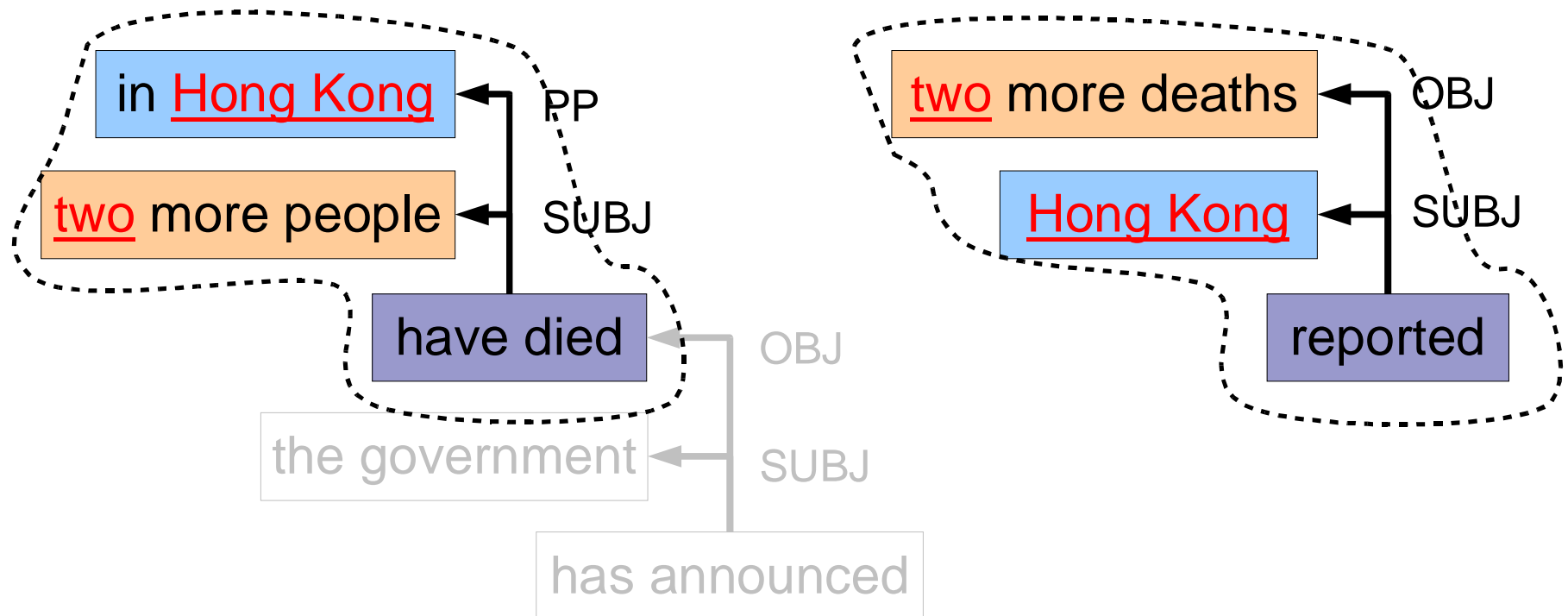
4. Extracting Portions

- Identify the corresponding nodes which share the anchors.



4. Extracting Portions

- Pull the subtrees extending from predicates which share the anchors.



4. Extracting Portions

- Finally we obtain a pair of paraphrases in these two articles:
 - “COUNTRY reported NUMBER more deaths.”
 - “NUMBER more people died in COUNTRY.”



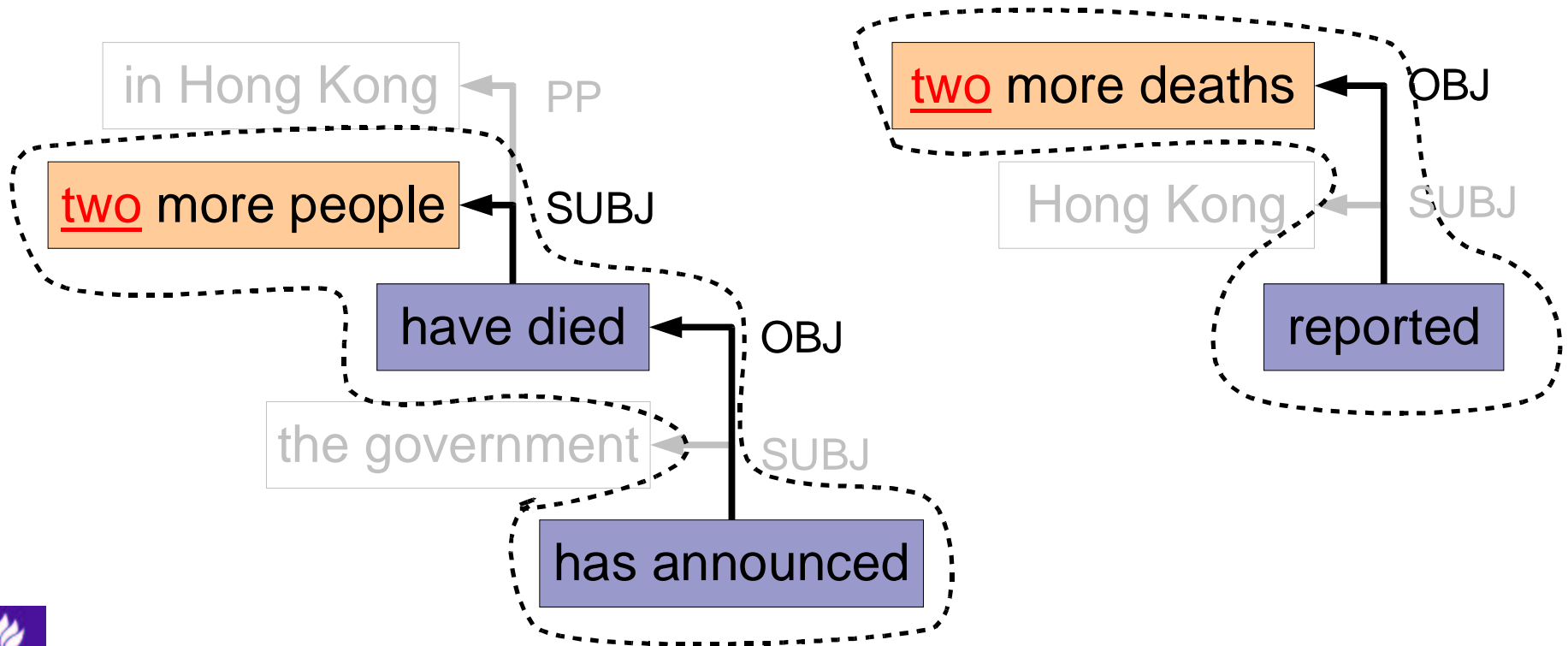
Related Work

- [Barzilay, 03] performed word sequence alignment for clustered sentences.
- [Lin, 01] used dependency trees and mutual information of word distribution as anchors.
- [Shinyama, 02] also used dependency trees and Named Entities, but obtained limited forms of expressions.
 - This task is the improvement to [Shinyama, 02].



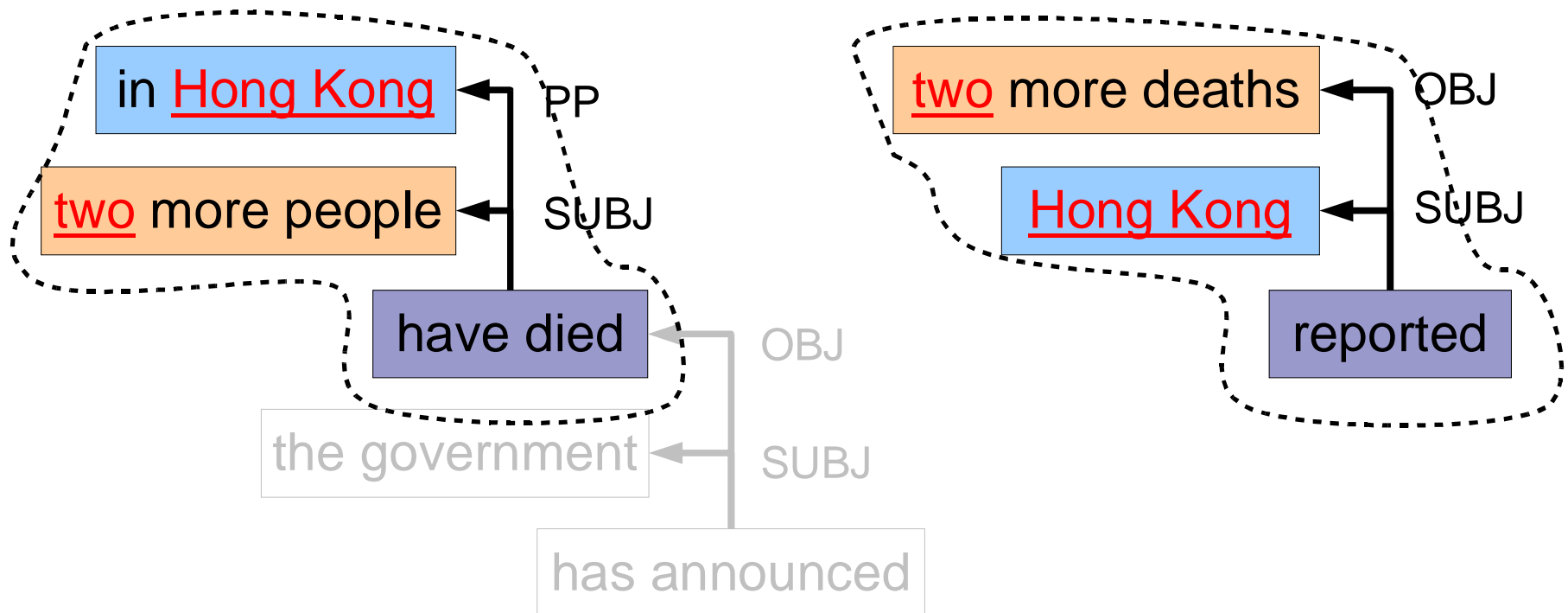
Difference from [Shinyama, 02]

- [Shinyama, 02] obtained only a single thread of dependency trees.



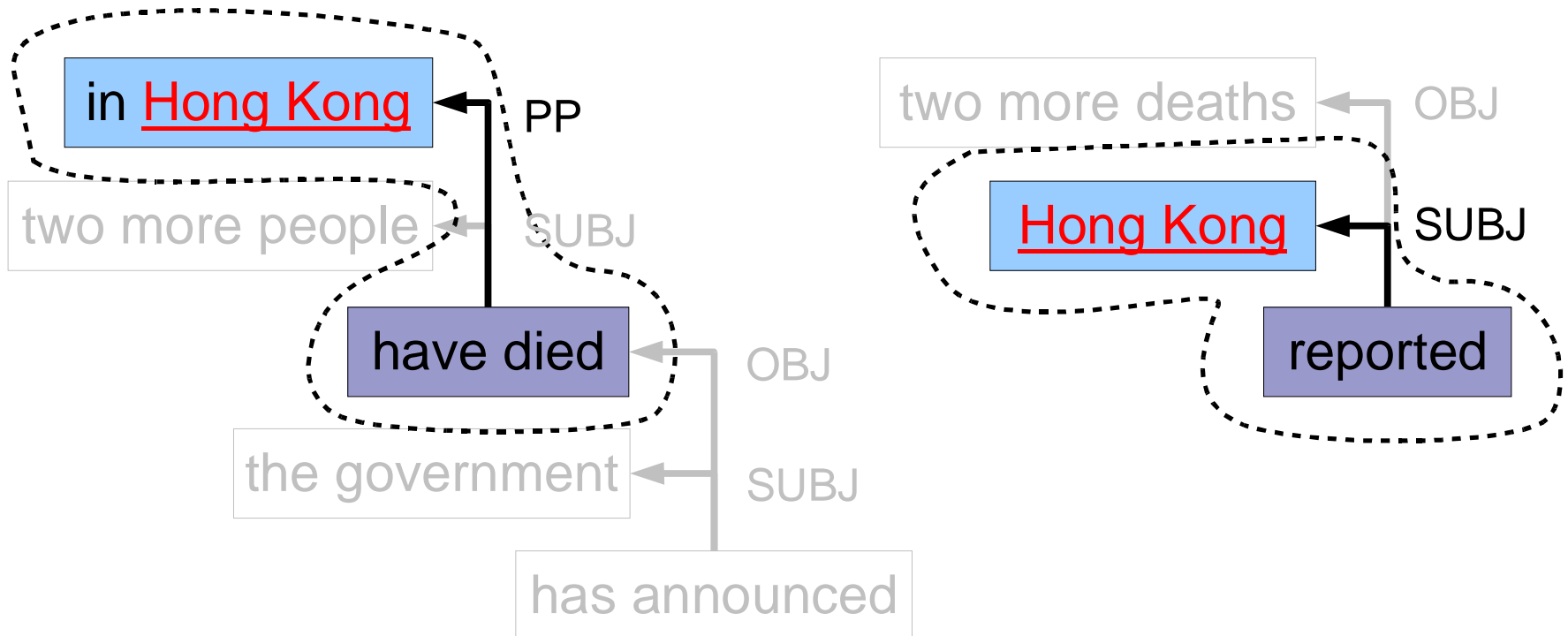
Difference from [Shinyama, 02]

- In the current method, we can obtain more interesting expressions which have branches.



A Problem

- There is a problem that this method may obtain wrong (but grammatical) paraphrases:



A Problem

- [Shinyama, 02] relied on frequency counts of pre-obtained expressions to filter out incomplete expressions.
 - Dependent on Information Extraction patterns.
 - So sparse for all possible subtrees.
- We tried to eliminate wrong paraphrases by maximizing the number of anchors with keeping the variety of expressions.
 - Filter out subtrees whose predicates don't have a “natural combination” of arguments.



Filter Incomplete Expressions

- Argument Structure Database

- We obtained argument structures for every predicate (verbs and adjectives) from one-year news articles in advance:

“report”(12)	SUBJ:5, OBJ:8, PP:4, ...
“announce”(8)	SUBJ:7, OBJ:6, ...
“die”(10)	SUBJ:10

- Calculate the relative frequency for all possible subtrees and pick ones where the score of each predicate exceeds a certain threshold.

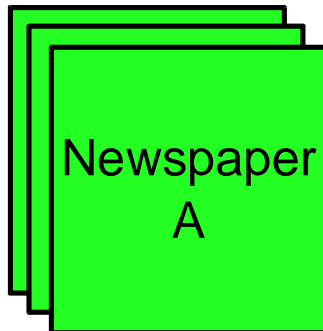


Coreference Resolution

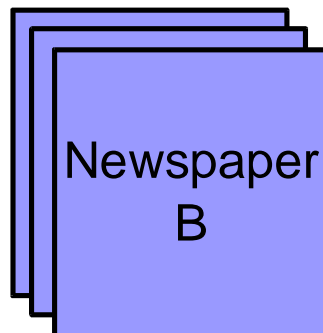
- [Shinyama, 02] had a problem in finding shared Named Entities.
 - “Prime-Minister-Junichiro-Koizumi”
 - “Mr. Koizumi”
- To solve this problem, we introduced a simple coreference resolution mechanism based on substring matching (LCS) on characters.



Experiments



We used two Japanese newspapers which report murder cases (*Mainichi* and *Nikkei*).

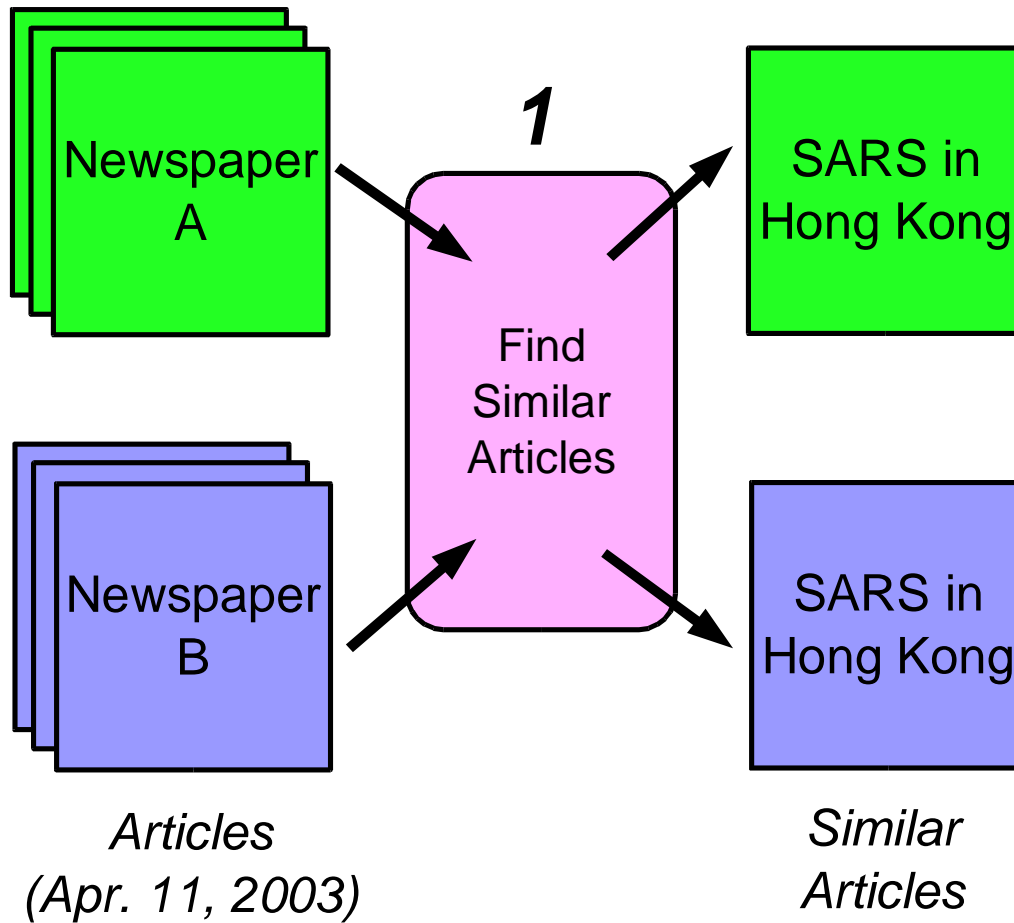


292,755 articles in total (one-year).

Articles
(Apr. 11, 2003)



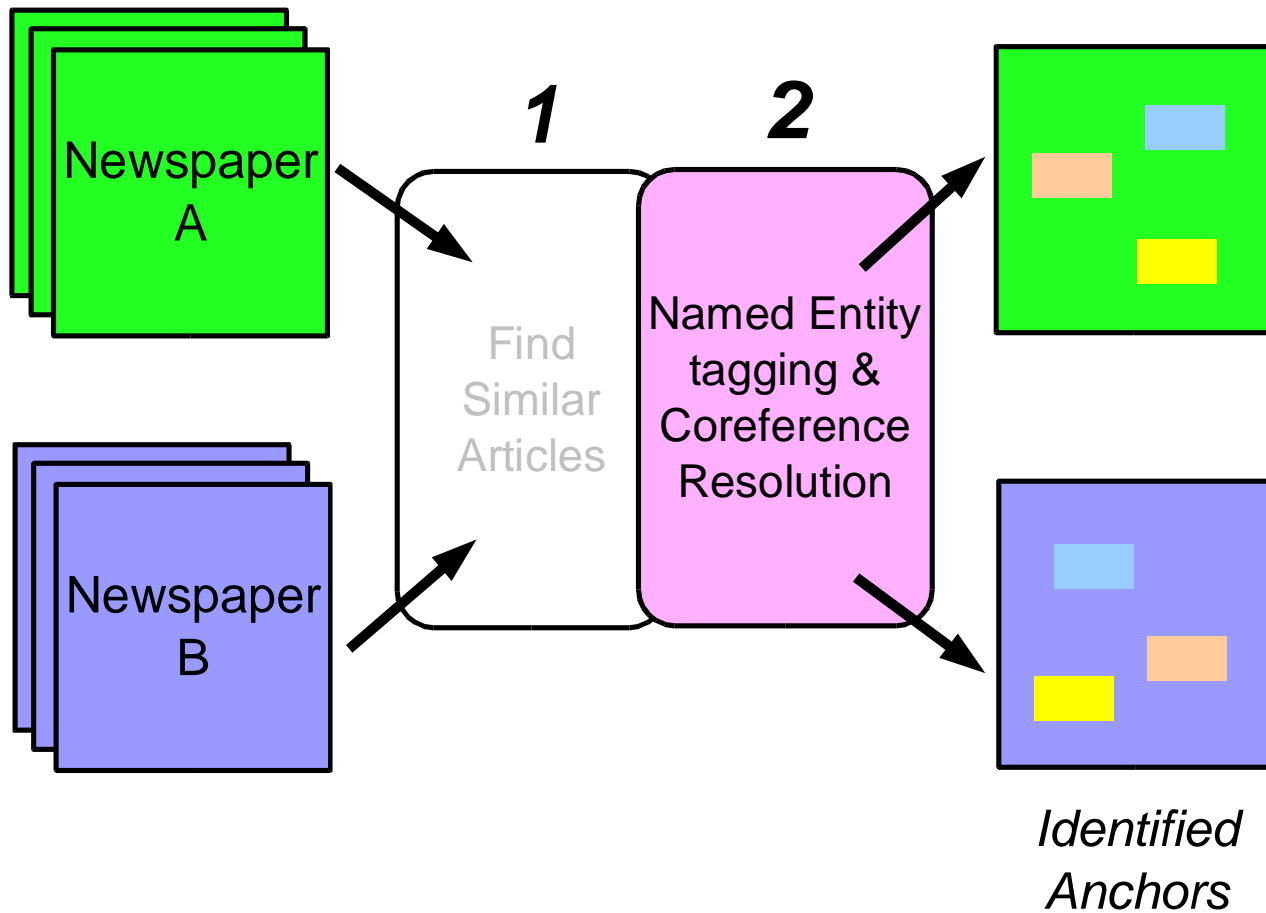
Experiments



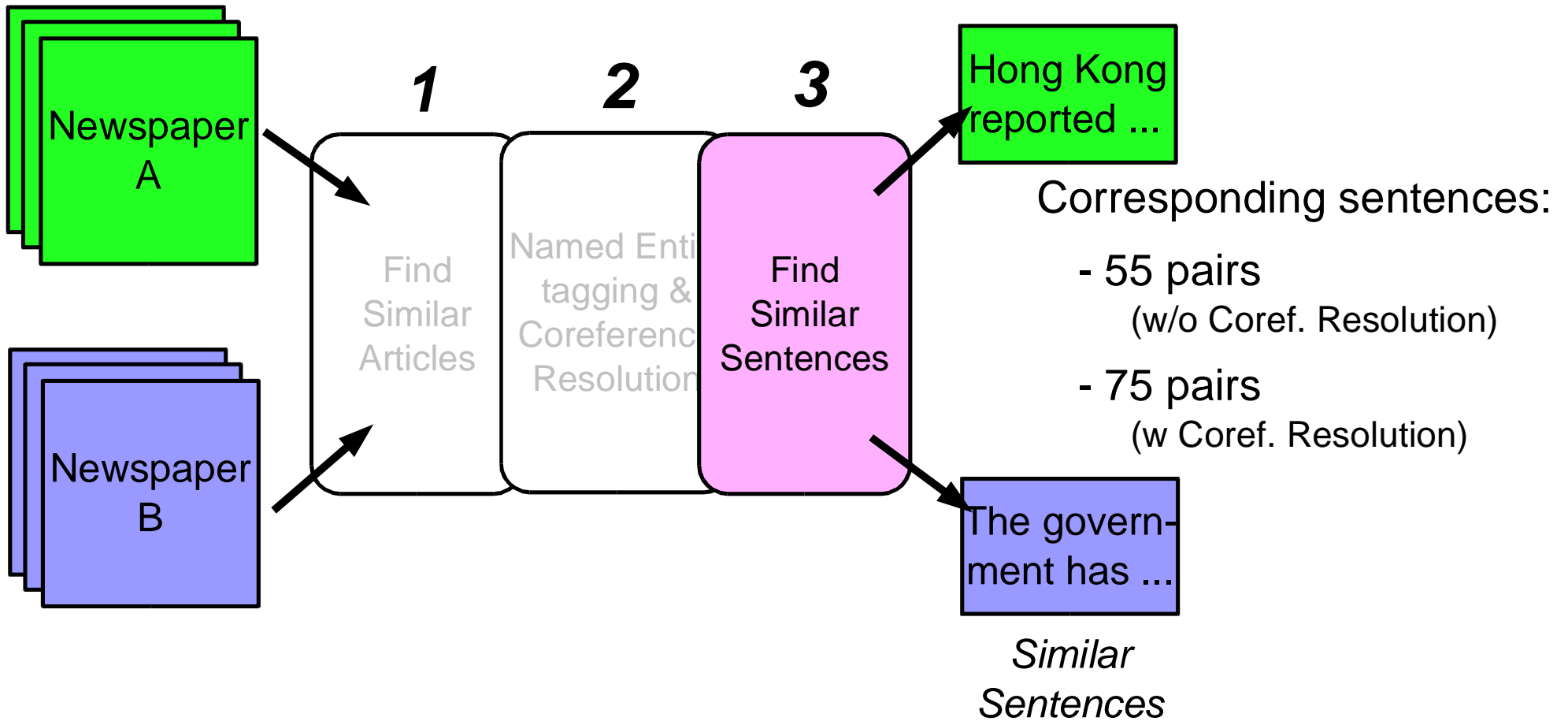
195 corresponding article pairs (390 articles) were automatically obtained.



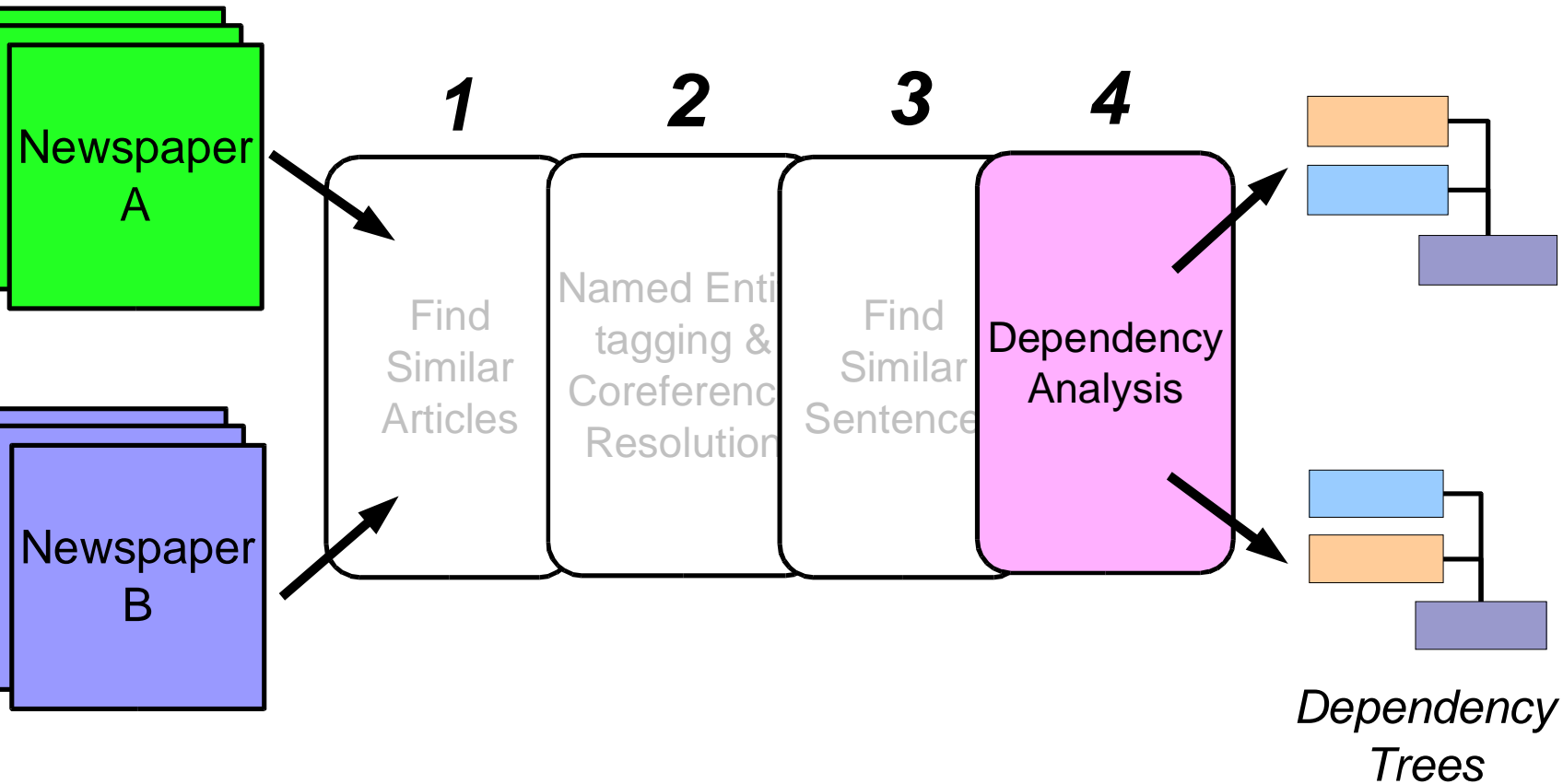
Experiments



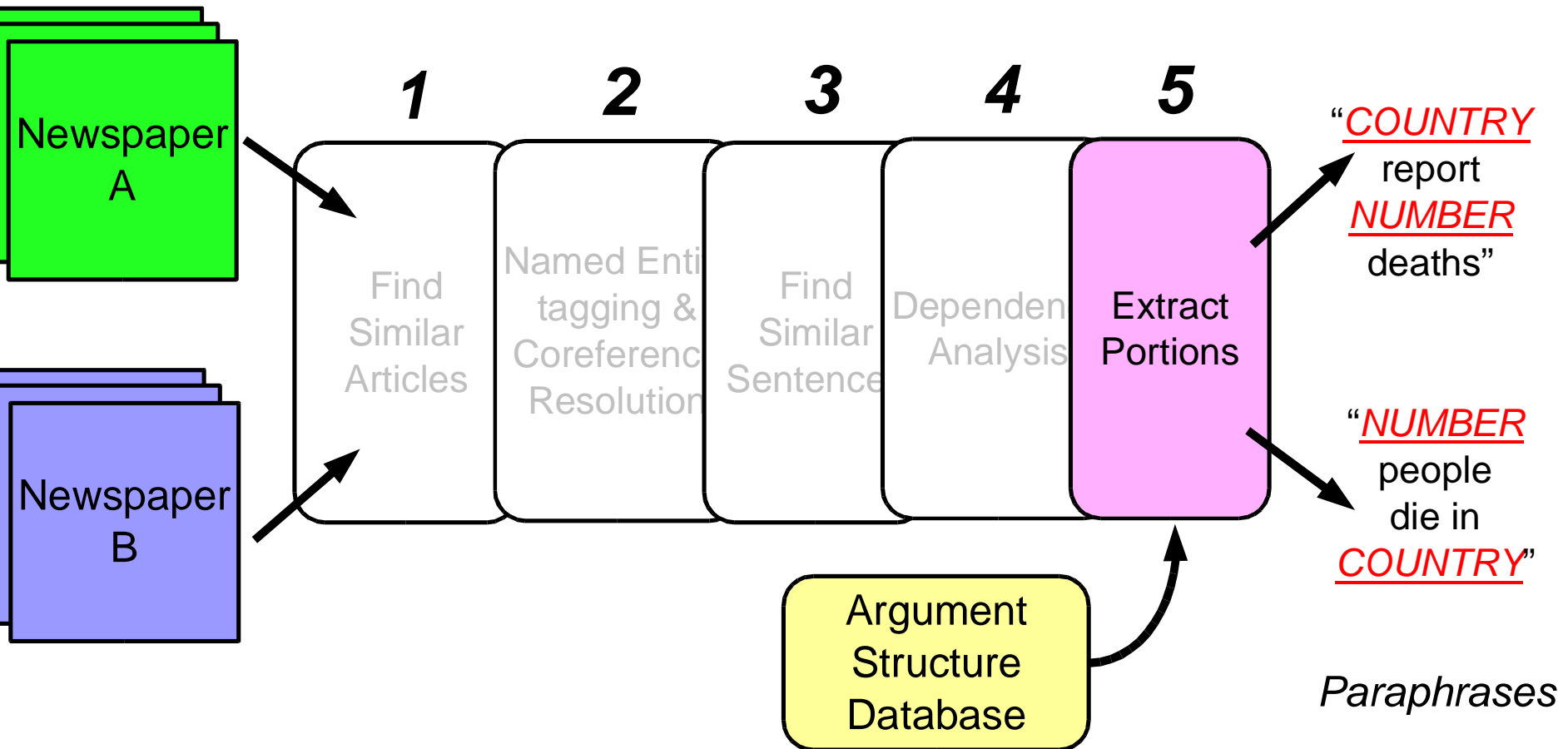
Experiments



Experiments



Experiments



Experiments

- We manually evaluated obtained sentences and paraphrases from the top 20 articles:

	<i>Obtained</i>	<i>Precision</i>	<i>Recall</i>
Sentences (93)	(w/o coref.) 55	75% (41)	44%
	(w coref.) 75	69% (52)	56%
Phrases (>100)	(w/o coref., w/o ASD) 106	24% (25)	
	(w/o coref., w ASD) 32	56% (18)	
	(w coref., w ASD) 37	62% (23)	



Obtained Paraphrases

- Correct: [Translated from Japanese]
 - Sample 1:
 - “PERSON1 killed PERSON2.” (*Satsugai Shita*)
 - “PERSON1 let PERSON2 die from loss of blood.” (*Shikketsushi Sasetta*)
 - Sample 2:
 - “PERSON1 shadowed PERSON2.” (*Bikou Shita*)
 - “PERSON1 kept his eyes on PERSON2.” (*Me Wo Tsuketa*)



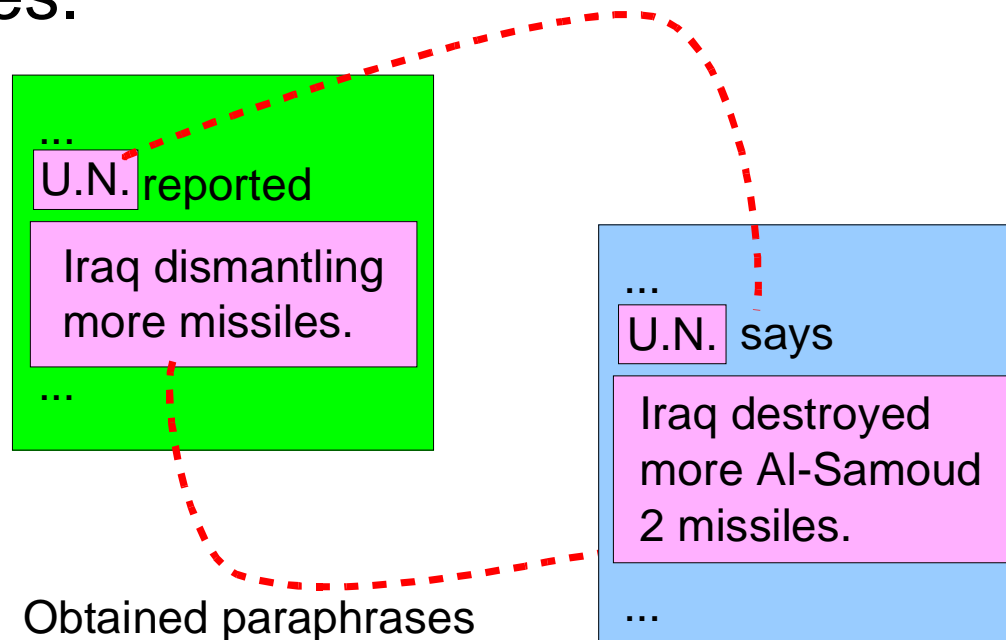
Obtained Paraphrases

- Incorrect: [Translated from Japanese]
 - Sample 1:
 - “PERSON1 fled to LOCATION1.” (*Toubou Shita*)
 - “PERSON1 fled to LOCATION1 and lay in wait.”
(*Toubou Shi, Harikonde Ita*)
 - Sample 2:
 - “PERSON1 cohabited with PERSON2.” (*Dousei*)
 - “PERSON1 murdered in the cohabitation room with PERSON2.”
(*Dousei Site Ita Heya De Satsugai*)



Future Work

- We want more paraphrases (in number and variation).
 - Apply obtained paraphrases to obtain other paraphrases.



Future Work

- More anchors will help.
 - [Sekine, 02] is expanding the categories of Named Entities:
 - Product names (e.g. “Frappuccino”)
 - Disease names (e.g. “SARS”)
 - Regulation names
(e.g. “Marine Mammal Protection Act”)
 - Language names (e.g. “Cebuano”, “Hindi”)
 - ...



Conclusion

- We obtained paraphrases from news articles.
 - From different newspapers on the same day.
 - Named Entity as anchors.
 - Extracted a portion of dependency trees as paraphrases.
- Two techniques were introduced:
 - Argument structure database.
 - Coreference resolution.

