

Collecting Paraphrases Automatically

Yusuke Shinyama

Proteus Project
Department of Computer Science
New York University



Overview

- Our goal
 - Why paraphrases are useful?
- How to collect paraphrases automatically?
 - Basic Idea
 - Underlying techniques
 - Overall procedure
- Experiment results
- Conclusion and future work



Our Goal

- Make a better Information Extraction system.



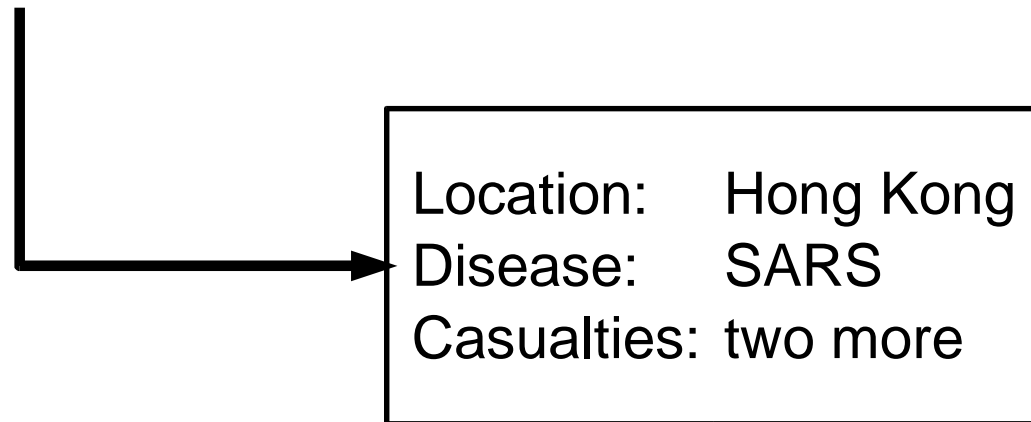
What is Information Extraction?

- Extract specific information from articles in a pre-defined domain:
 - Personnel affairs (person, company, post)
 - Infectious diseases (date, location, name)
 - Natural disasters (date, location, casualties)
 - Sports results (date, team, score)
 - Weather reports, cooking recipes, etc.



What is Information Extraction?

- Infectious diseases:
 - “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.”



What is Information Extraction?

- Infectious diseases:
 - “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.”
 - This is done by pattern matching.
 - “NUMBER people die in COUNTRY”
 - ↓
 - Casualties
 - ↓
 - Location



What is Information Extraction?

- Infectious diseases:

- “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.”

- This is done by pattern matching.

- “NUMBER people die in COUNTRY”

two more

Hong Kong



The Problem

- How about this article?
 - “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 Problem

- How about this article?
 - “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.”
 - We would need a different pattern!
 - “COUNTRY reports NUMBER deaths”
 - ↓
Location
 - ↓
Casualties



The Problem

- Currently, we prepare many patterns manufactured by hand.
- But this is a huge task.
 - It is very hard (or almost impossible) to cover all possible expressions.



Our Goal (again)

- Make a better Information Extraction system.
(without a great deal of labor)



Our Goal (again)

- We are trying to obtain (learn) these patterns automatically.
- Two approaches
 - Sudo's approach:
 - Obtain patterns directly from corpora.
 - My approach:
 - Grow a set of good patterns from “seeds”.



Why Paraphrases?

- Paraphrasing: expressing one thing in various ways.
 - “die” = “kick the bucket”
 - “close to” = “around the corner”



Why Paraphrases?

- Information Extraction patterns which capture the same information can be also considered as paraphrase:
 - “X people die in Y”



Why Paraphrases?

- Information Extraction patterns which capture the same information can be also considered as paraphrase:

– “X people die in Y”

discover a
paraphrase

“Y reports X deaths”

an alternative pattern



Why Paraphrases?

- Collecting paraphrases is useful for other applications:
 - Machine Translation
 - Finding appropriate expressions depending on the contexts from a large expression library.
 - Text Summarization
 - Replacing an expression to a more concise.
 - Information Retrieval, Question & Answering, etc.



How to Find Paraphrases?

- Search news articles on the same day:
 - Various newspapers describe a single event in a different way.
 - There are events which are reported in more than one newspaper.



How to Find Paraphrases?

- Actually...
 - “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*)
 - “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*)



How to Find Paraphrases?

- Anchors: Names, Locations and Numbers
 - “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*)
 - “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*)



How to Find Paraphrases?

- Extract portions which share the anchors.
 - “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*)
 - “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*)



How to Find Paraphrases?

- Generalize the obtained expressions for future use.
 - “NUMBER people die in COUNTRY”
 - “COUNTRY report NUMBER deaths”



How to Find Paraphrases?

- Our method breaks into three parts:
 1. Find similar articles from different newspapers.
 2. Extract appropriate portions from corresponding articles.
 3. Generalize obtained expressions to paraphrases.



Our Tools

1. To find similar articles:
 - Topic Detection and Tracking
2. To extract portions from sentences:
 - Coreference resolution
 - Dependency analysis (parsing)
3. To generalize expressions:
 - Named Entity Tagging



1. Find Similar Articles

- Topic Detection and Tracking (TDT)
 - Identifies two articles which report the same event.
 - Simple, well-developed technique.



1. Find Similar Articles

- Topic Detection and Tracking (TDT)
 - Count the words in both articles.

	government
	two
	people
	Hong Kong
	SARS
Article 1	virus
(Reuters,	61
Apr. 11, 2003)	cases
	illness

Hong Kong	
two	
deaths	
61	
cases	
SARS	
Friday	
governments	Article 2
world	(Channel News Asia,
steps	Apr. 11, 2003)
virus	



1. Find Similar Articles

- Topic Detection and Tracking (TDT)
 - Count the words in both articles.

Article 1
(Reuters,
Apr. 11, 2003)

government
two
people
Hong Kong
SARS
virus
61
cases
illness

7 words
match

Hong Kong
two
deaths
61
cases
SARS
Friday
governments
world
steps
virus

Article 2
(Channel News Asia,
Apr. 11, 2003)



2. Extract Portions

- Coreference resolution: Identifies anchors, but...
 - “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.”



2. Extract Portions

- How do we know the appropriate portions from sentences?
 - “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.”



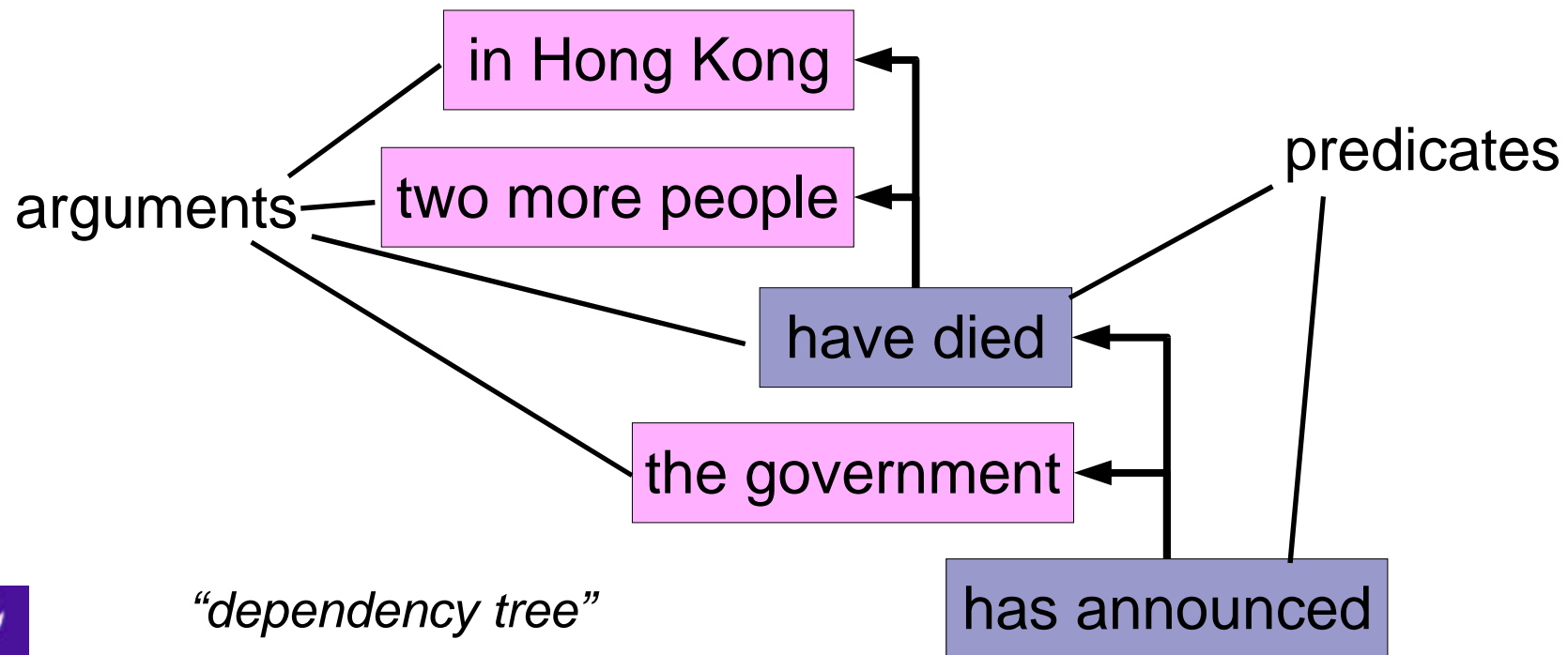
2. Extract Portions

- Simply picking up the words between two corresponding anchors doesn't work...
 - “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.”



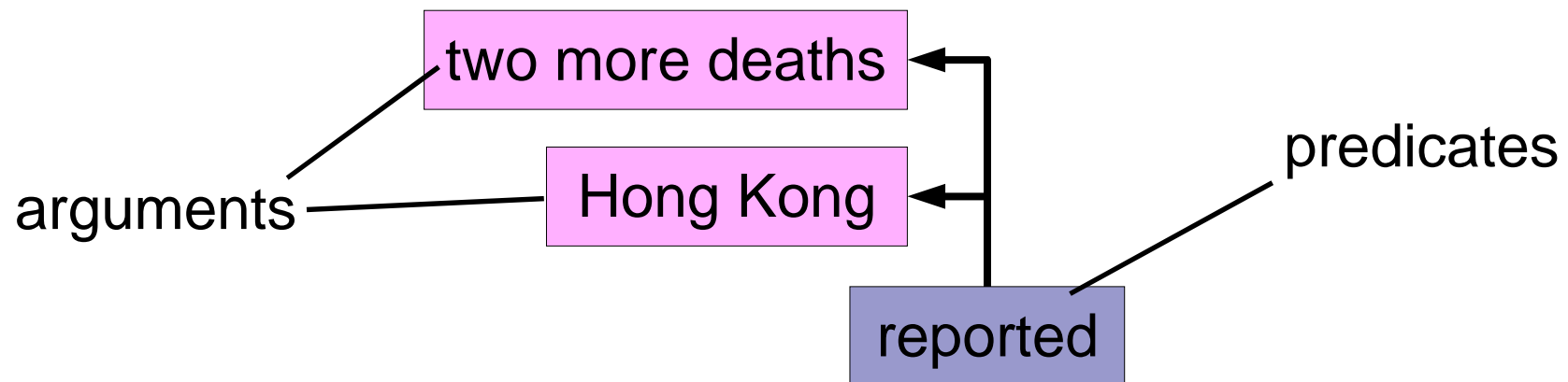
2. Extract Portions

- Dependency analysis (parsing)
 - “The government has announced that two more people have died in Hong Kong after ...”



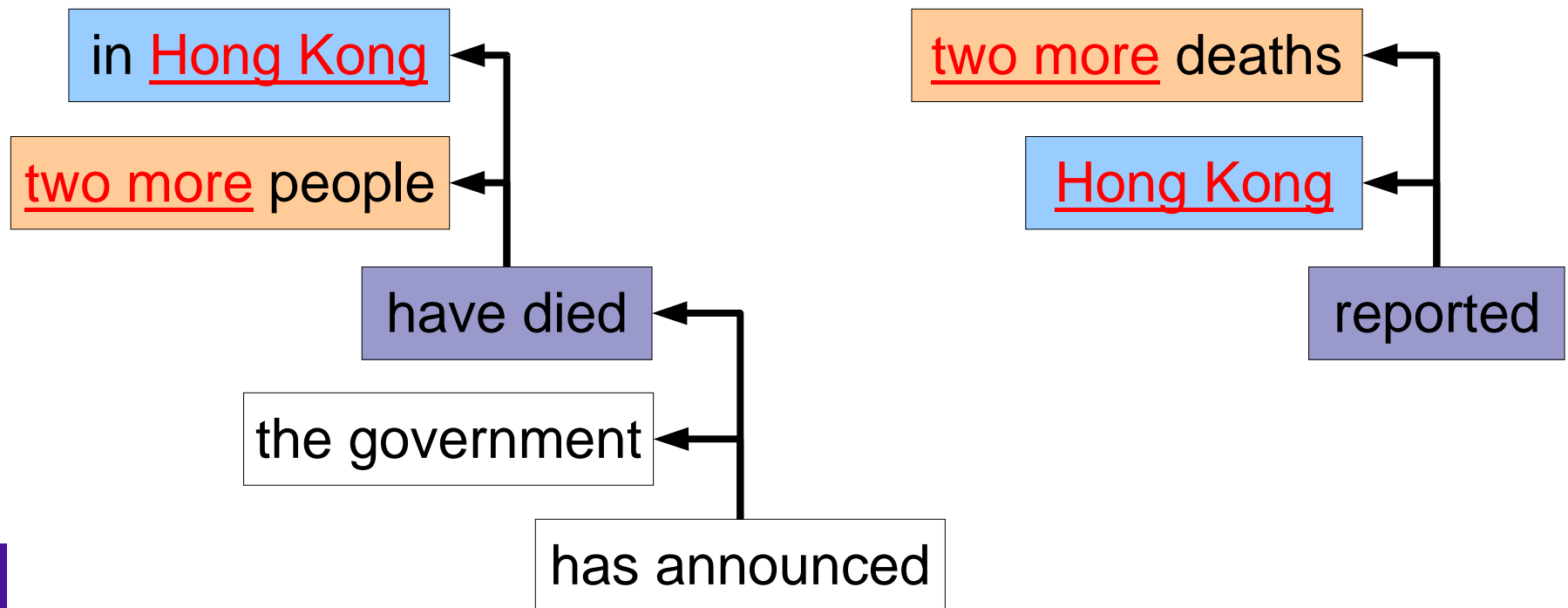
2. Extract Portions

- Dependency analysis (parsing)
 - “Hong Kong reported two more deaths and ...”



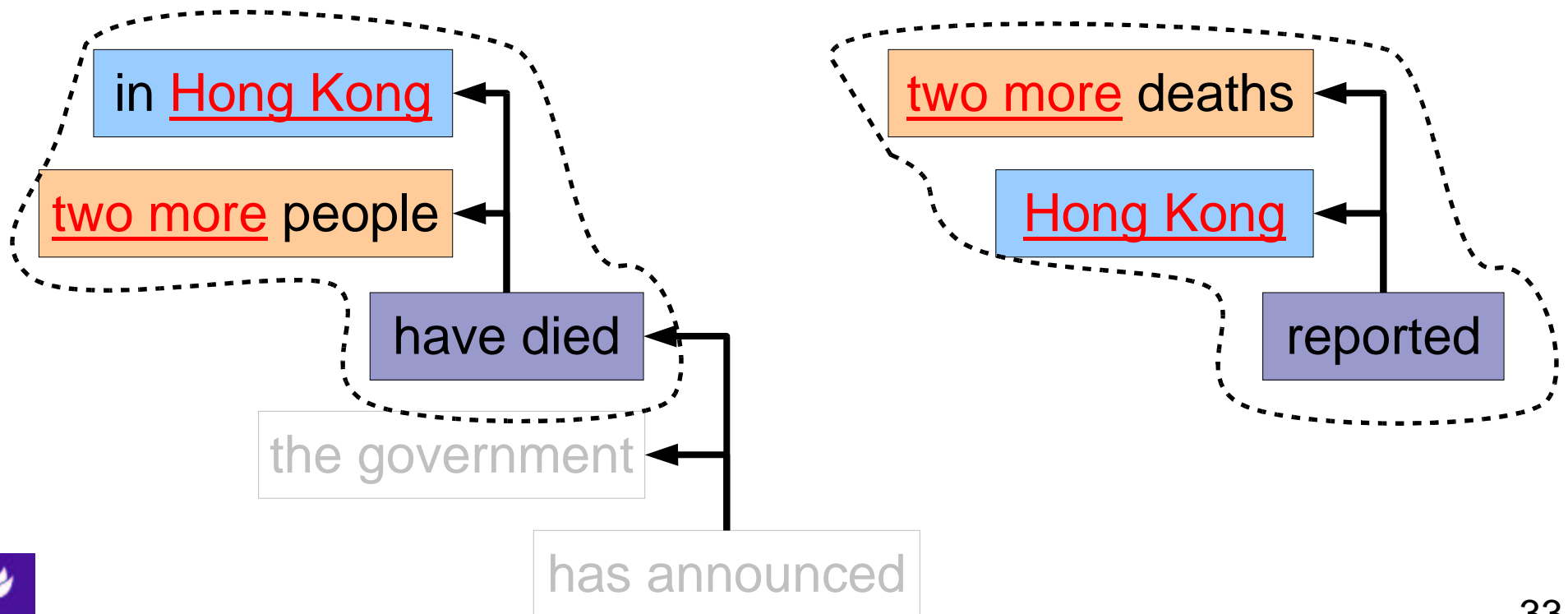
2. Extract Portions

- Dependency analysis (parsing)
 - Identify the corresponding nodes which include anchors.



2. Extract Portions

- Dependency analysis (parsing)
 - Pull the subtrees extending from predicates which share the anchors.



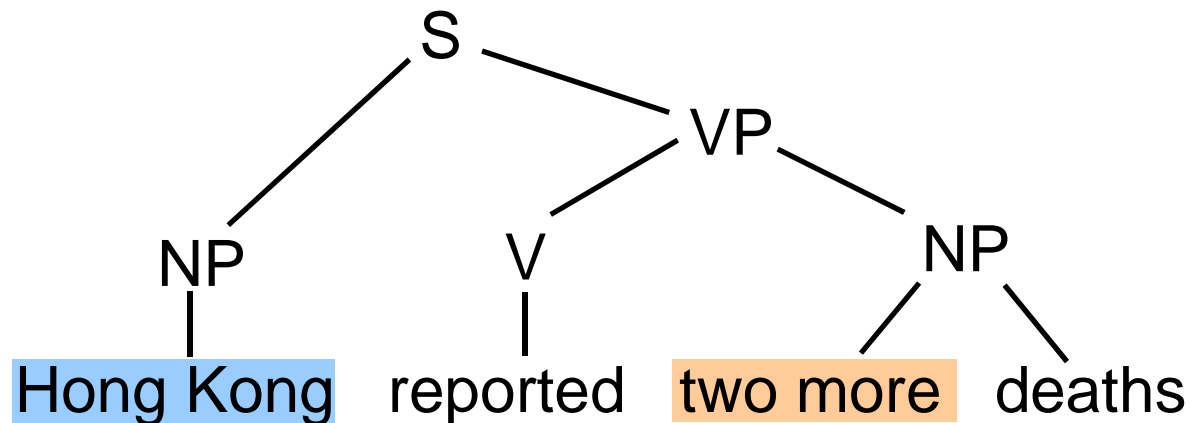
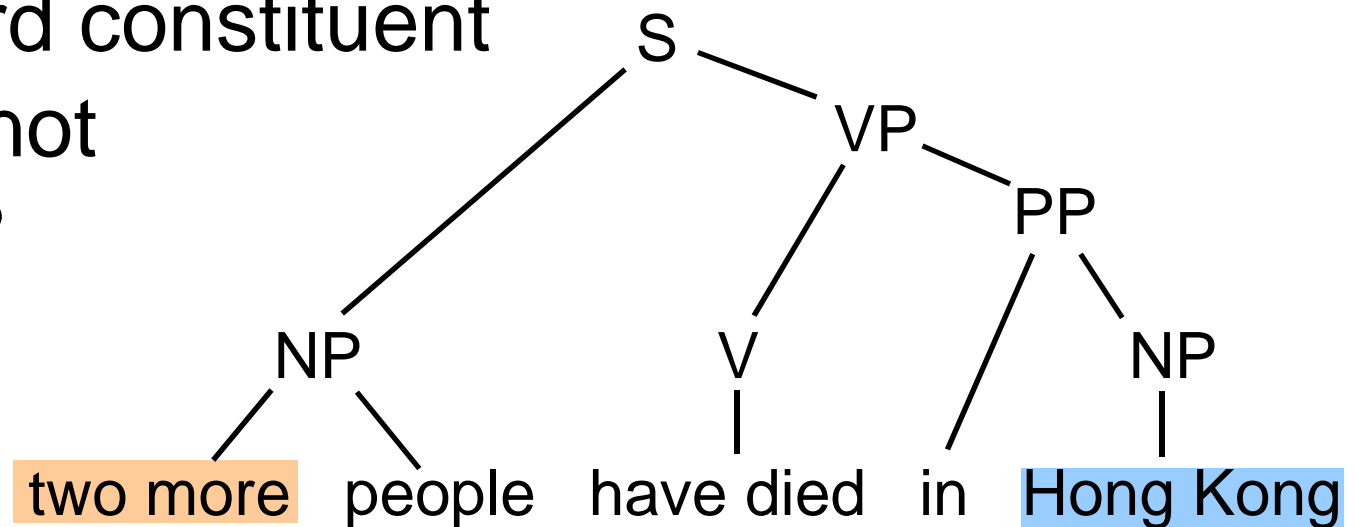
2. Extract Portions

- We obtain two corresponding portions.
 - “two more people die in Hong Kong”
 - “Hong Kong report two more deaths”



2. Extract Portions

- Why standard constituent grammar is not appropriate?



2. Extract Portions

- Advantages in dependency analysis:
 - More “normalized” form.
 - (somewhat) language independent.
 - We are using Japanese articles.



3. Generalize Expressions

- Named Entity Tagging
 - Similar to Part-of-Speech Tagging
 - “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.”
 - “two more” : **NUMBER**
 - “Hong Kong” : **COUNTRY**
 - “SARS” : **DISEASE**
 - “61” : **NUMBER**

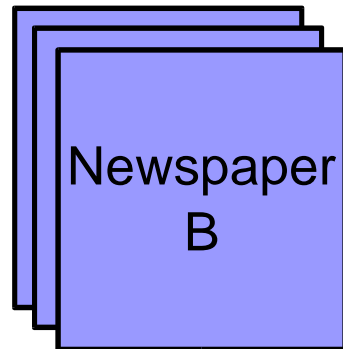
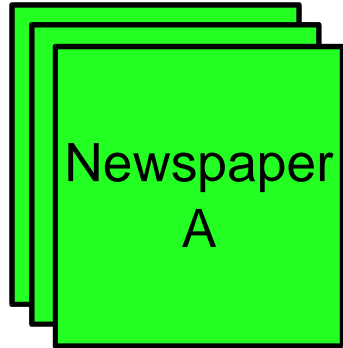


3. Generalize Expressions

- Apply Named Entity Tagging to the obtained portions.
 - “NUMBER people die in COUNTRY”
 - “COUNTRY report NUMBER deaths”
- Finally we obtain paraphrases!



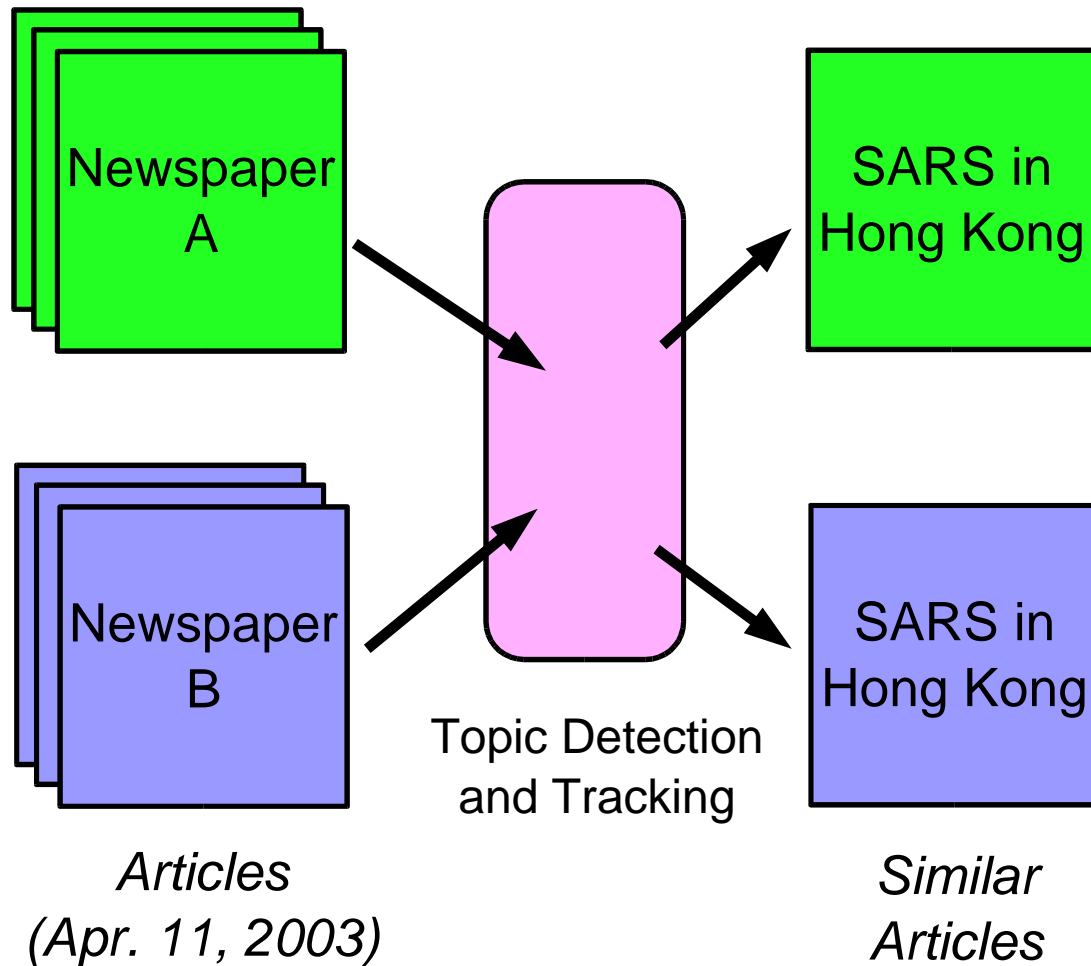
Overall Procedure



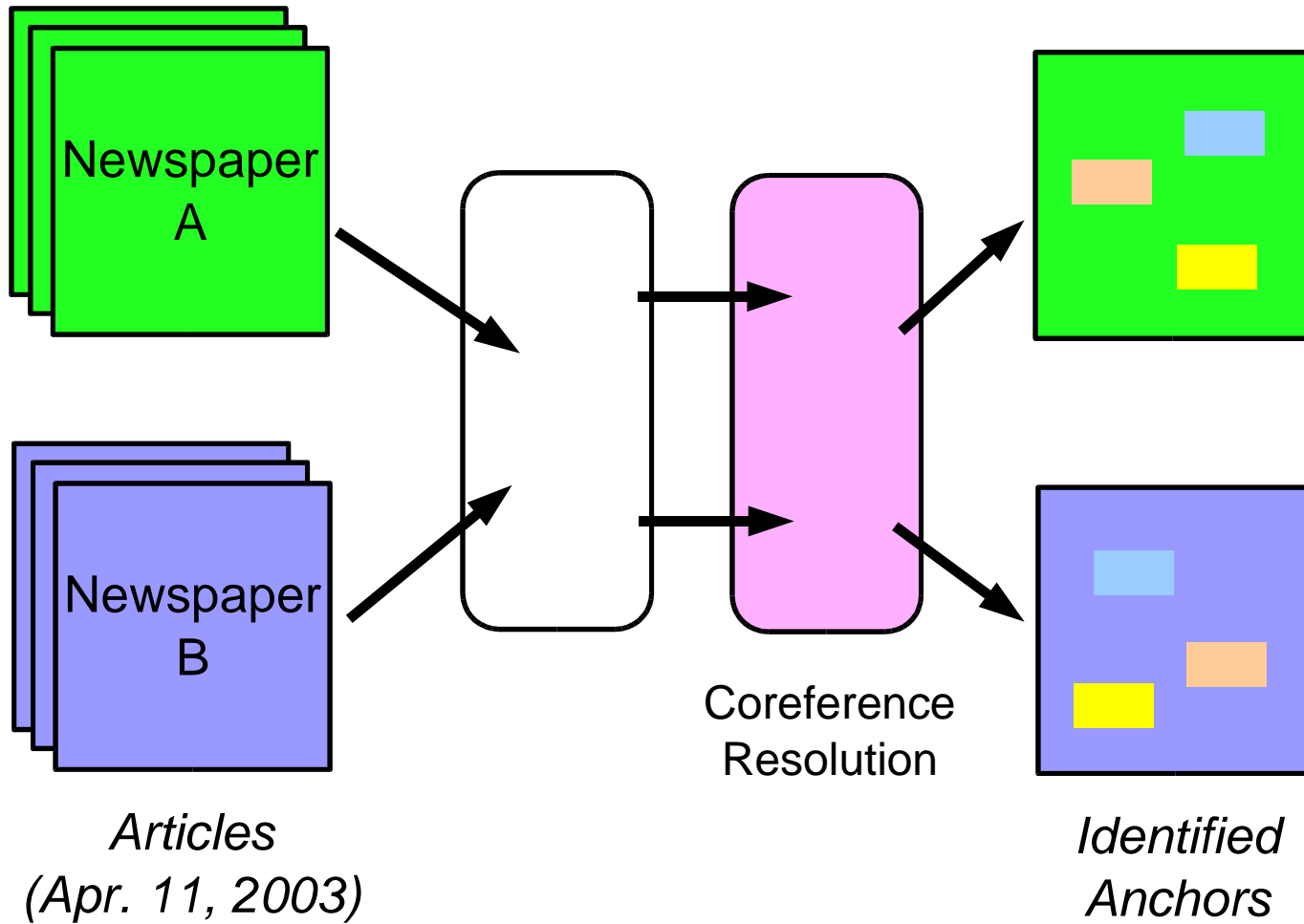
Articles
(Apr. 11, 2003)



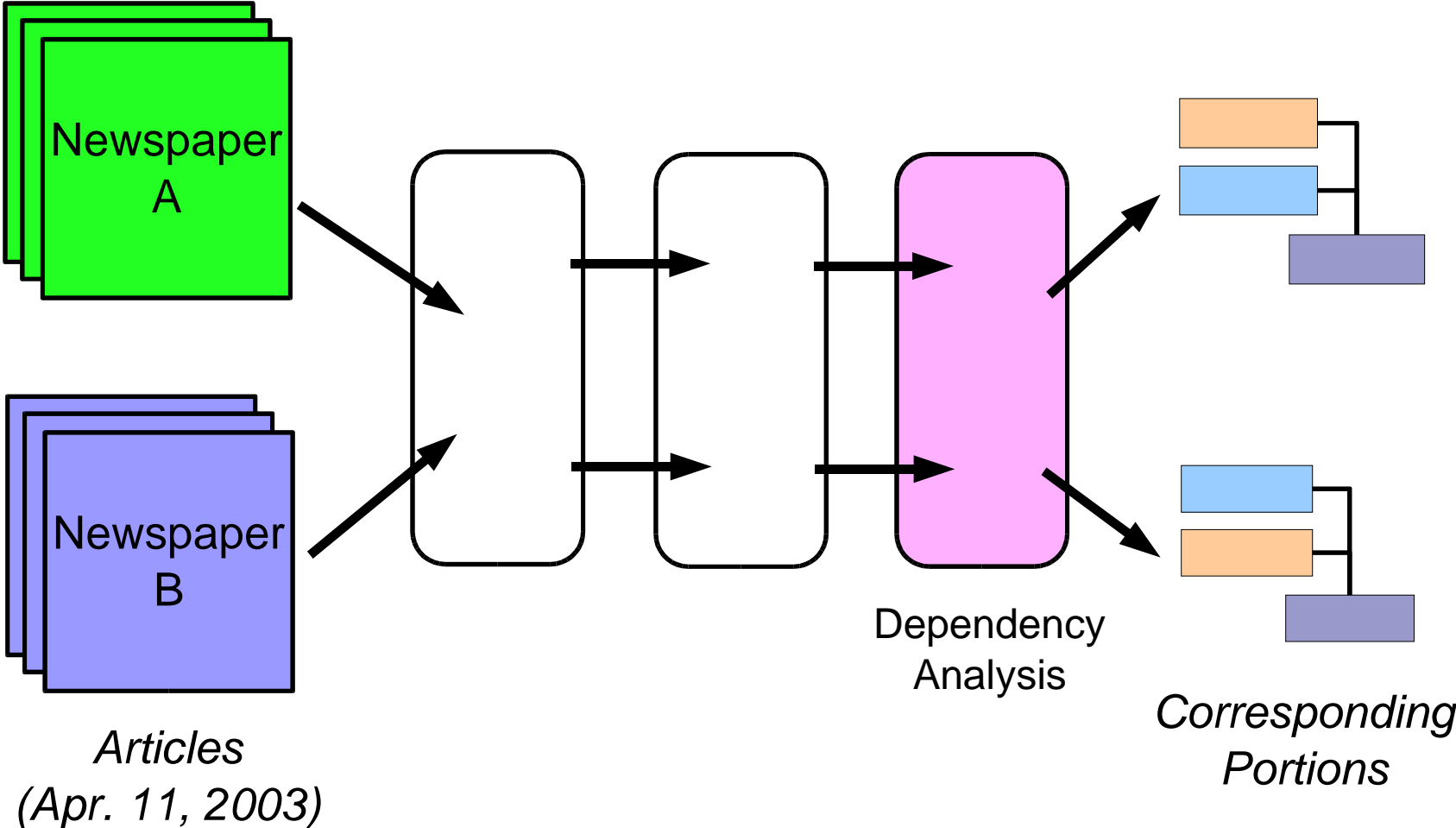
Overall Procedure



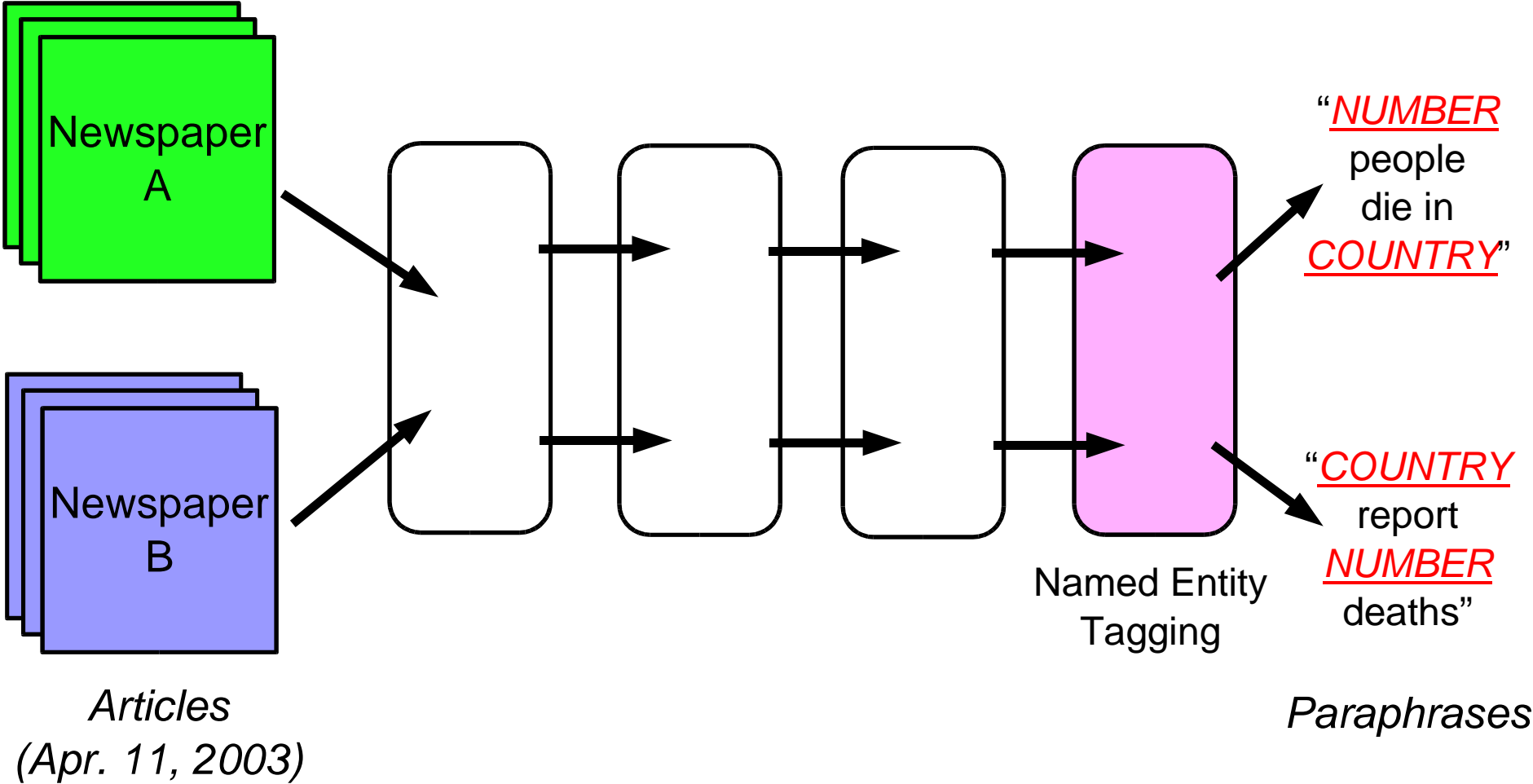
Overall Procedure



Overall Procedure



Overall Procedure



Experiments

- We used Japanese news articles which report murder cases.
- From one-year articles:
 - 195 corresponding pairs (390 articles) obtained.
- From 20 corresponding pairs (40 articles) :
 - 37 paraphrases were obtained.
 - 23 paraphrases were correct. (precision= 62%)
 - (reviewed by human)



Experiments

- Obtained paraphrases
 - Sample 1:
 - “PERSON1 killed PERSON2.”
 - “PERSON1 let PERSON2 die from loss of blood.”
 - Sample 2:
 - “PERSON1 shadowed PERSON2.”
 - “PERSON1 kept his eyes on PERSON2.”



Real World Problems

- Easy
 - “North Korea said Wednesday that it had reactivated its nuclear facilities and is going ahead with their operation “on a normal footing.” (*New York Times*)
 - “North Korea said Wednesday it had restarted and put on a “normal footing” the atomic facilities at the center of its suspected nuclear weapons program.” (*Reuters*)



Real World Problems

- Easy
 - “it had reactivated its nuclear facilities.”
 - “it had restarted the atomic facilities.”
 - “it” = “it” (North Korea)
 - “its nuclear facilities” = “the atomic facilities”



Real World Problems

- Hard
 - “World chess champion Garry Kasparov played his supercomputer opponent to a draw Tuesday in the second game of their Man vs. Machine showdown.” (*ABC News*)
 - “Chess grandmaster Garry Kasparov broke a spell in his contests against computers on Tuesday when he drew the second game of his 6-game match in New York against world champion program Deep Junior.” (*CNN Europe*)



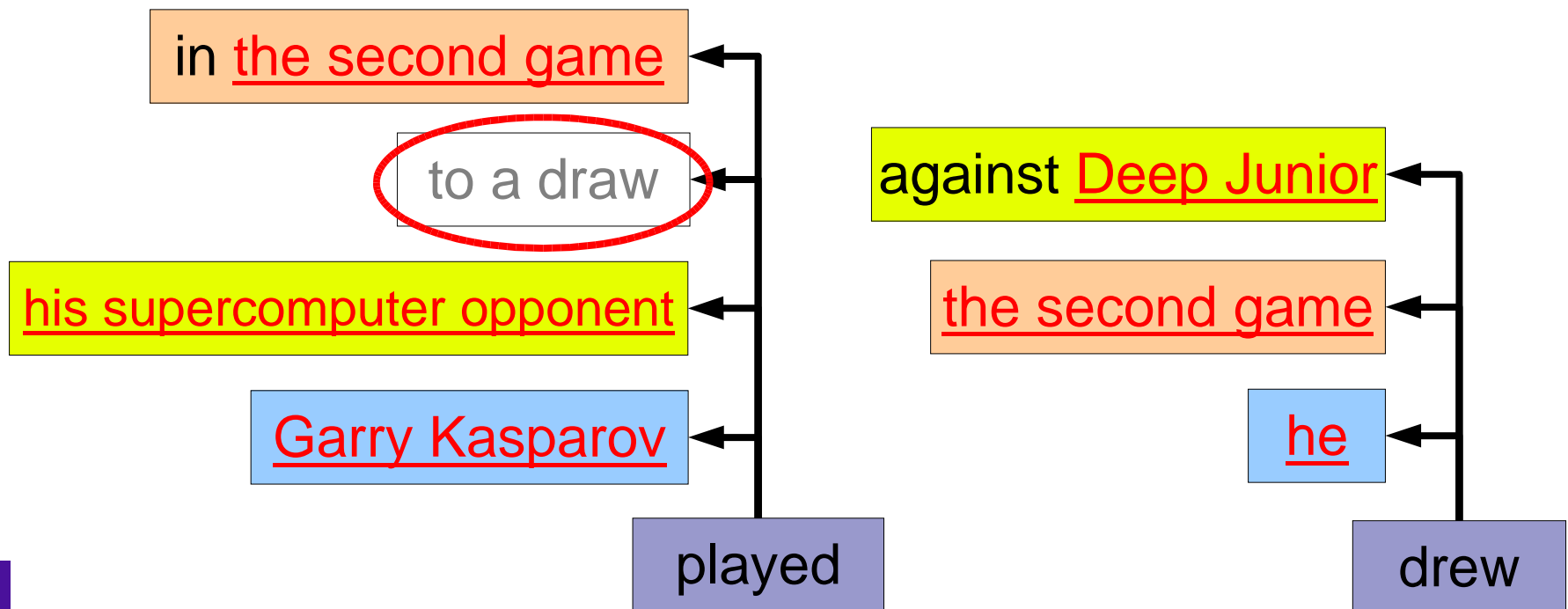
Real World Problems

- Hard
 - “Garry Kasparov played his supercomputer opponent to a draw in the second game.”
 - “he drew the second game against Deep Junior.”
 - “Garry Kasparov” = “he”
 - “the second game” = “the second game”
 - “his supercomputer opponent” = “Deep Junior”



Real World Problems

- Hard
 - How to include an extra node “to a draw” which doesn't include an anchor?



Real World Problems

- To solve these problems:
 - Better coreference resolution
 - “he” = “Garry Kasparov”, “the virus” = “SARS”, etc.
 - Identify typical forms of expressions.
 - “play to a draw”, “draw against”, etc.
 - Collect typical expressions from corpora and calculate the “typicalness” of obtained portions.
 - Reduce computational costs in matching portions.



Real World Problems

- Extremely hard
(headlines about the SQL Slammer worm)
 - “Slammer worm fastest ever seen”
(*San Jose Mercury News*)
 - “Slammer attack almost over after 10 minutes”
(*ZDNet News*)
 - “Slammer – the first “Warhol” worm?”
(*CNET News*)



Real World Problems

- Extremely hard
 - Irregular syntax
 - No subject, no verb, etc.
 - Vagueness
 - “fastest” = “10 minutes” ?
 - Metaphor
 - “In the future everybody will be famous for 15 minites.”
 - *Andy Warhol*



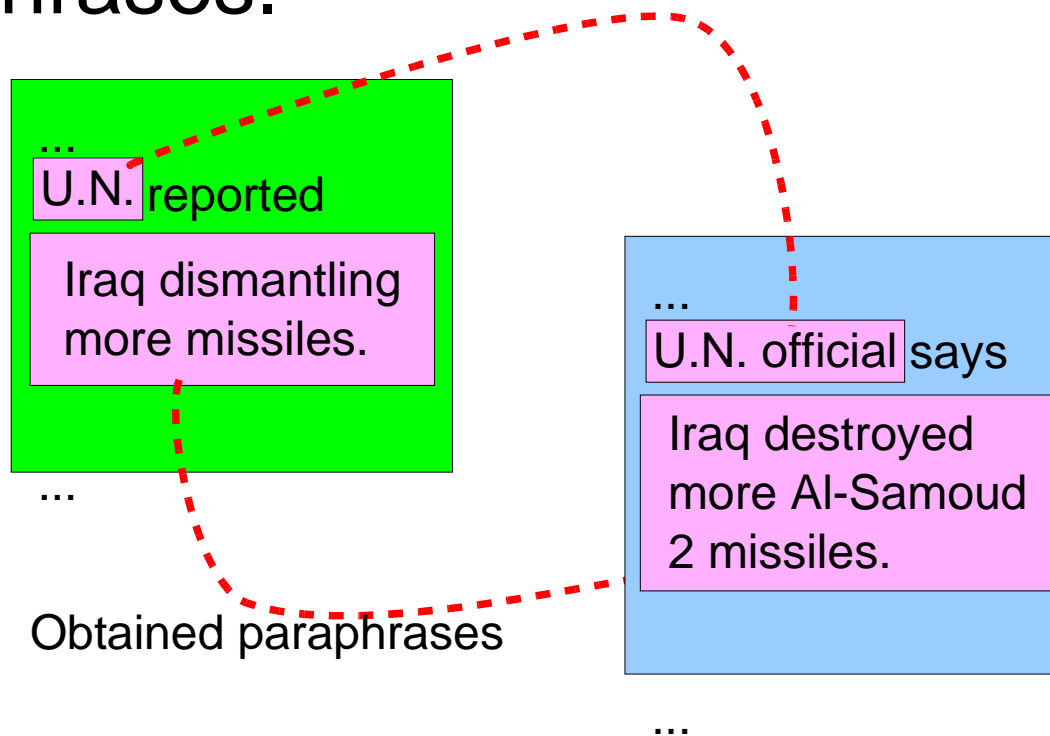
Conclusions

- Why paraphrases are useful?
 - For various kinds of applications including Information Extraction.
- How to collect paraphrases?
 - Obtain from similar articles.
 - Find anchors and extract the corresponding portions.
 - Use dependency analysis.
 - Generalize obtained expressions.



Future Work (1) – Macro Level

- Apply obtained paraphrases to obtain larger paraphrases.



Future Work (2) – Micro Level

- Apply obtained paraphrases to obtain smaller paraphrases.
 - Names which have different forms.
 - “IBM”, “International Business Machines”
 - Names which refer to the same entity depending on its context.
 - “Pyongyang said Wednesday it had reactivated its nuclear facilities.”
 - “North Korea said Wednesday it had restarted the atomic facilities.”



Proteus Project

- Our web page:
 - <http://nlp.cs.nyu.edu/>

