

Automatic Paraphrase Acquisition for Information Extraction

Committee:

Prof. Ralph Grishman

Prof. Dan Melamed

Prof. Satoshi Sekine

**Thesis Proposal
Yusuke Shinyama**



Backgrounds

- What is Information Extraction?
 - e.g. you're seeking M&A events:

Adobe Systems, a leading seller of software for editing and managing documents, announced on Monday that it had acquired Macromedia for \$3.4 billion in stock.

(New York Times, Apr 19, 2005)

Yahoo has purchased online photo-sharing service Flickr, less than a week after the Internet giant launched a beta test of a new blogging tool.

(ZDNet UK, Mar 21, 2005)



Backgrounds

- What is Information Extraction?
 - e.g. you're seeking M&A events:

When	Buyer	Buyee	Price
Apr. 19	Adobe Systems	Macromedia	\$3.4 billion <i>(New York Times, Apr 19, 2005)</i>
Mar. 21	Yahoo	Flickr	- <i>(ZDNet UK, Mar 21, 2005)</i>



Backgrounds

- How to capture event? - by patterns:
 - **C1** acquired **C2** [for **M3**]
 - **C1** purchased **C2**

Adobe Systems, a leading seller of software for editing and managing documents, announced on Monday that it had acquired **Macromedia** for **\$3.4 billion** in stock.

(New York Times, Apr 19, 2005)

Yahoo has purchased online photo-sharing service **Flickr**, less than a week after the Internet giant launched a beta test of a new blogging tool.

(ZDNet UK, Mar 21, 2005)

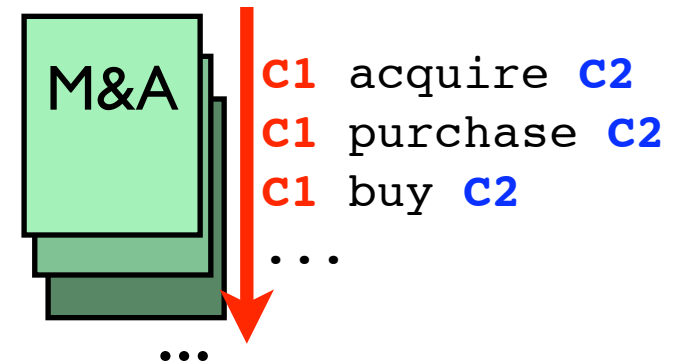


Backgrounds

- What's important in IE application?

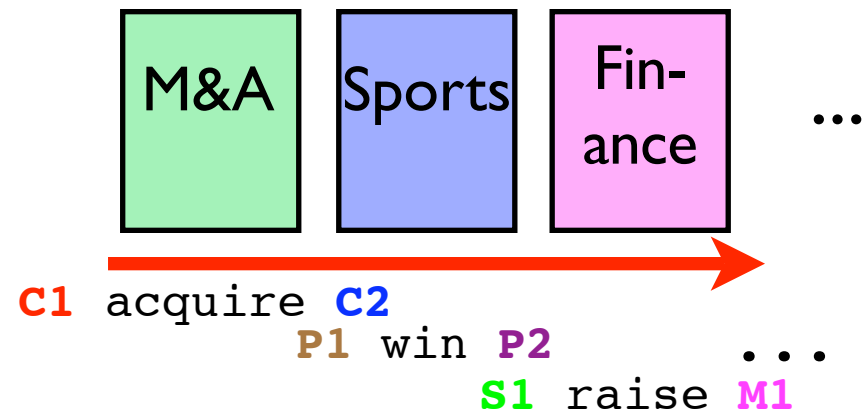
- Depth

- The accuracy and coverage for a particular scenario (e.g. M&A).



- Breadth

- The number of such scenarios.



Backgrounds

- Major problems in IE:
 - Depth
 - How we can get many different patterns which capture all different wordings?
 - Creating by hand - too many!
 - Breadth
 - How we can get patterns for many different scenarios?
 - Creating by hand - too many!
 - We cannot cover all patterns by hand.
 - Learn the patterns automatically.

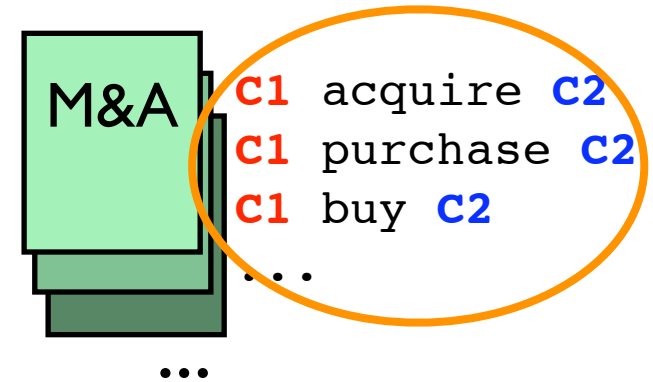


Paraphrases for IE

- Why is paraphrase acquisition useful?

- Depth

- Indeed, different patterns for the same scenario = paraphrases!



- What is a paraphrase?

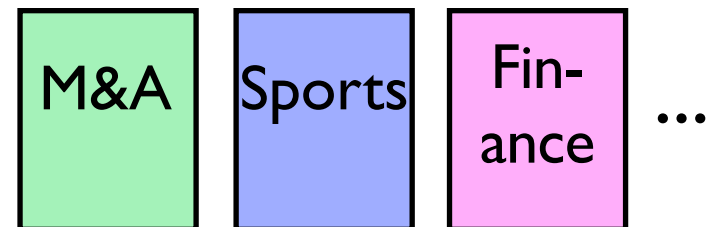
- Different expressions which state the same facts.

- Finding paraphrases \approx finding patterns



Paraphrases for IE

- Why is paraphrase acquisition useful?
 - Breadth
 - M&A
 - Sports
 - Finance
 - ...



C1 acquire **C2**
P1 win **P2**
S1 raise **M1**



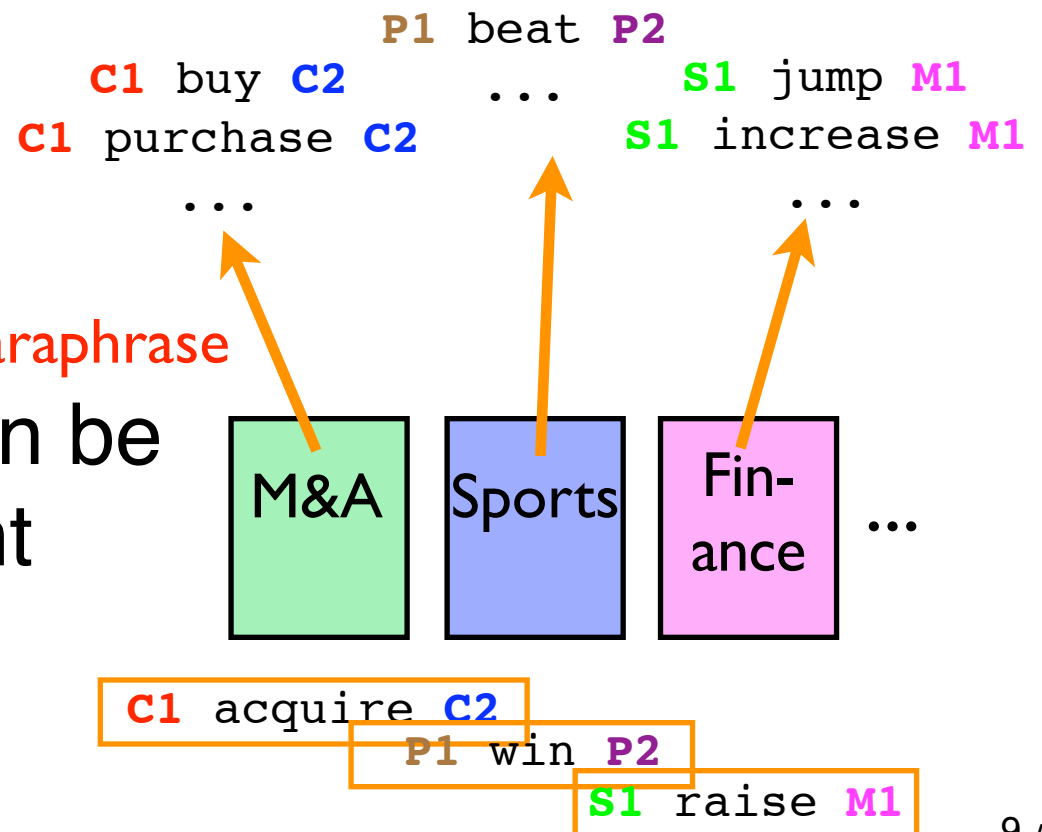
Paraphrases for IE

- Why is paraphrase acquisition useful?

- Breadth

- M&A
- Sports
- Finance
- ...

- Paraphrases can be found in different scenarios.



Paraphrases for IE

- My research objective:
 - “Obtain relevant paraphrases which can improve a repertory of IE patterns automatically from news articles.”
- Issues to solve:
 - Formal definition of paraphrases.
 - How to obtain them?
 - How to use them for IE?



What Is a Paraphrase?

- Definition of paraphrase
 - Must be application specific.
 - General definition is impossible!
 - I am doomed to face these problems:
 - “What time is it?” /
“Do you have a watch?” (Pragmatically same)
 - “U.S. invasion of Iraq” /
“U.S. liberation of Iraq” (Different value judgment)
 - “He helped people commit suicide” /
“He killed people” (Incompatible observation)



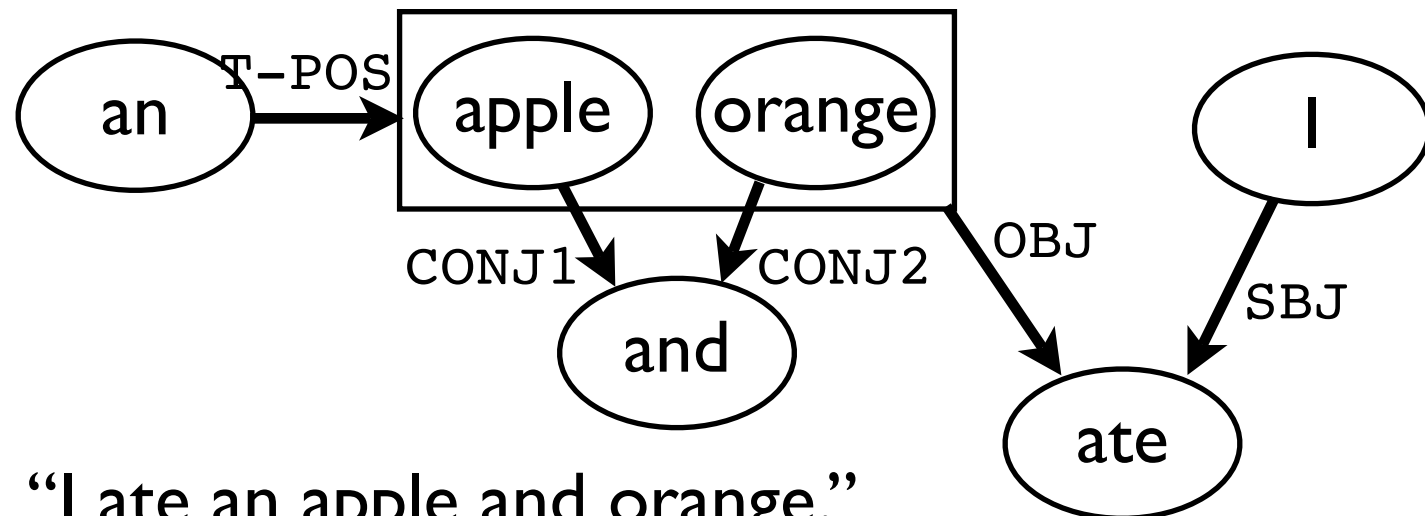
What Is a Paraphrase?

- **Paraphrase** (revisited) :
 - = Different expressions which state the same facts.
 - What is “expression”?
 - Word / Phrase / Sentence
 - Other
 - What is “fact”?
 - Truth condition
 - Viewpoint
 - Implication (entailment)
 - Other



What Is a Paraphrase?

- Expression
 - Predicate-argument structure
 - GLARF [Meyers, 02]
 - Allows to handle various phenomena uniformly.



"I ate an apple and orange."



What Is a Paraphrase?

- Fact

- Extracted Table by IE

- Event type (e.g. “M&A”)
- Involved entities (e.g. “Adobe”, “Macromedia”)

M&A :

When	Buyer	Buyee	Price
Apr. 19	Adobe Systems	Macromedia	\$3.4 billion

- 90% agreement between 2 researchers.
 - 35/39 article pairs.



What Is a Paraphrase?

- How to get rid of contexts?
 - The meaning of an expression might be affected by its context.

Here is the most surprising news ...

Adobe acquired Macromedia ...

Adobe's chief executive Bruce Chizen said ...



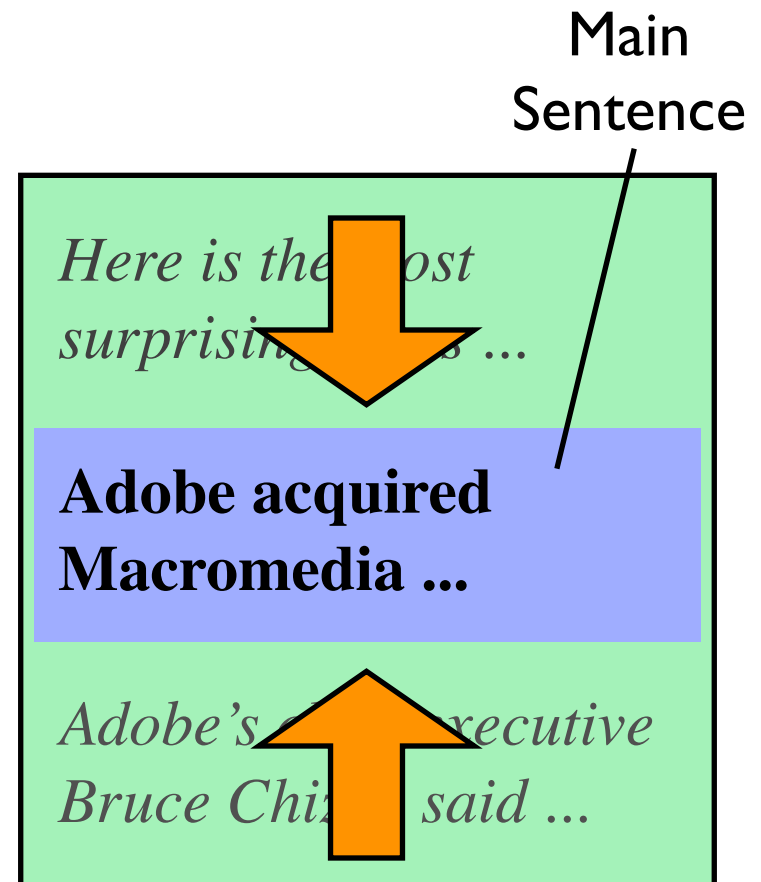
What Is a Paraphrase?

- How to get rid of contexts?

- Assumption:

- There is a main sentence which has the most important fact.
- A main sentence is independent in meaning and represents the article.

- We can “compress” an article into one sentence.

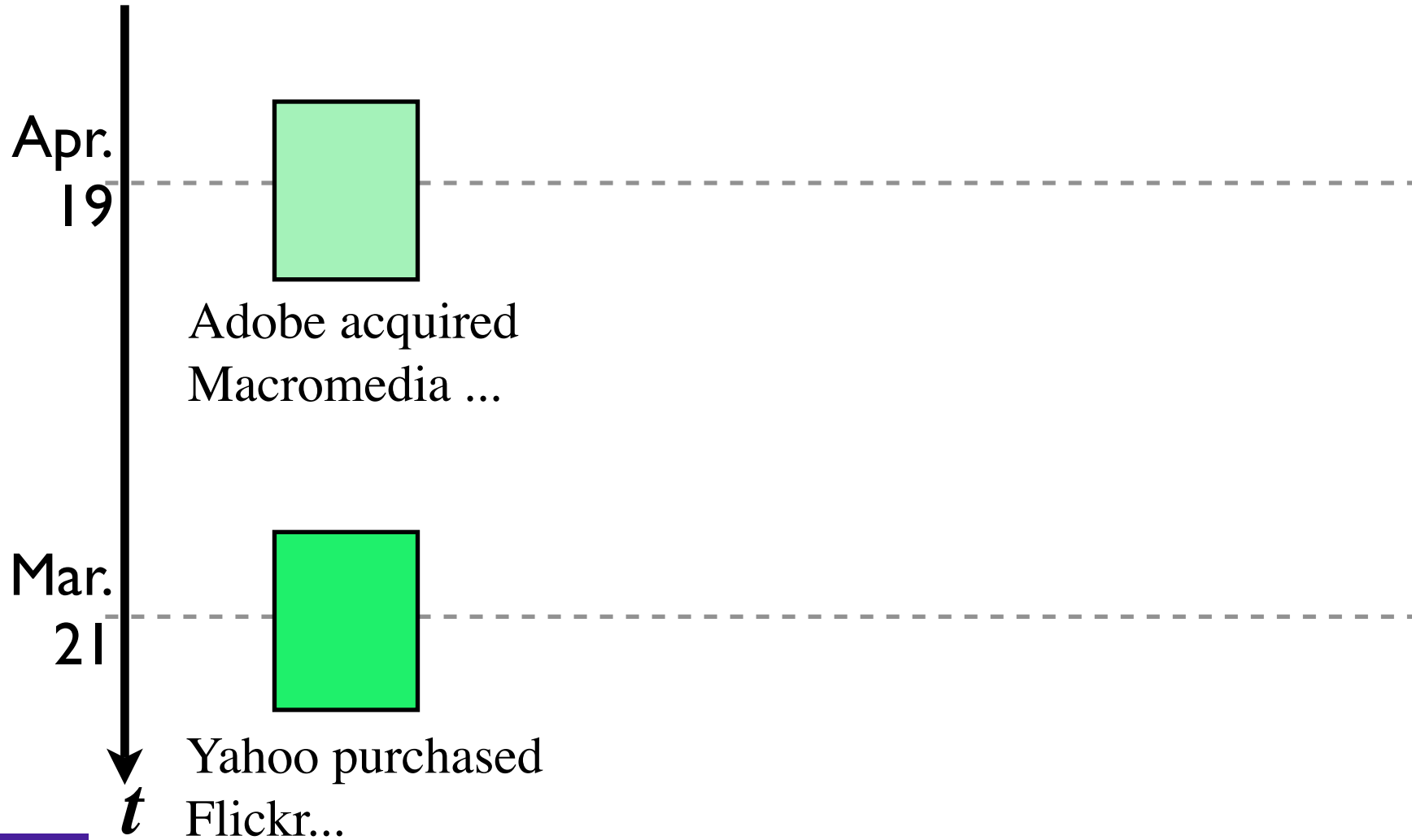


Obtain Paraphrases

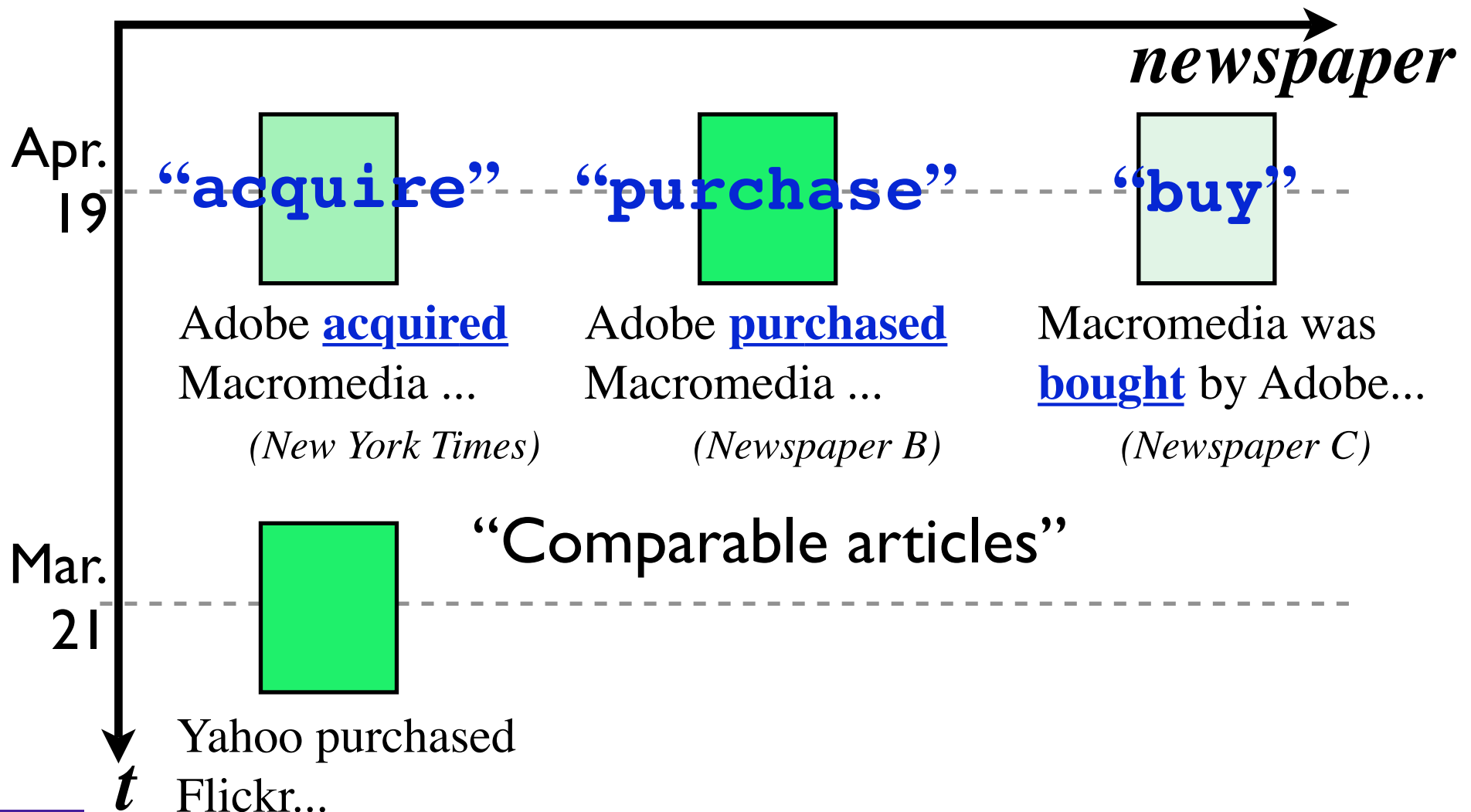
- How?
 - Dictionary-based methods
 - WordNet [Green, 04]
 - NOMLEX [Meyers, 98]
 - Corpus-based methods:
 - Parallel Corpus [Barzilay, 01] [Pang, 03]
 - Non-parallel Corpus [Lin, 01] [Ravichandran, 02]
 - Comparable Corpus [Shinyama, 03]



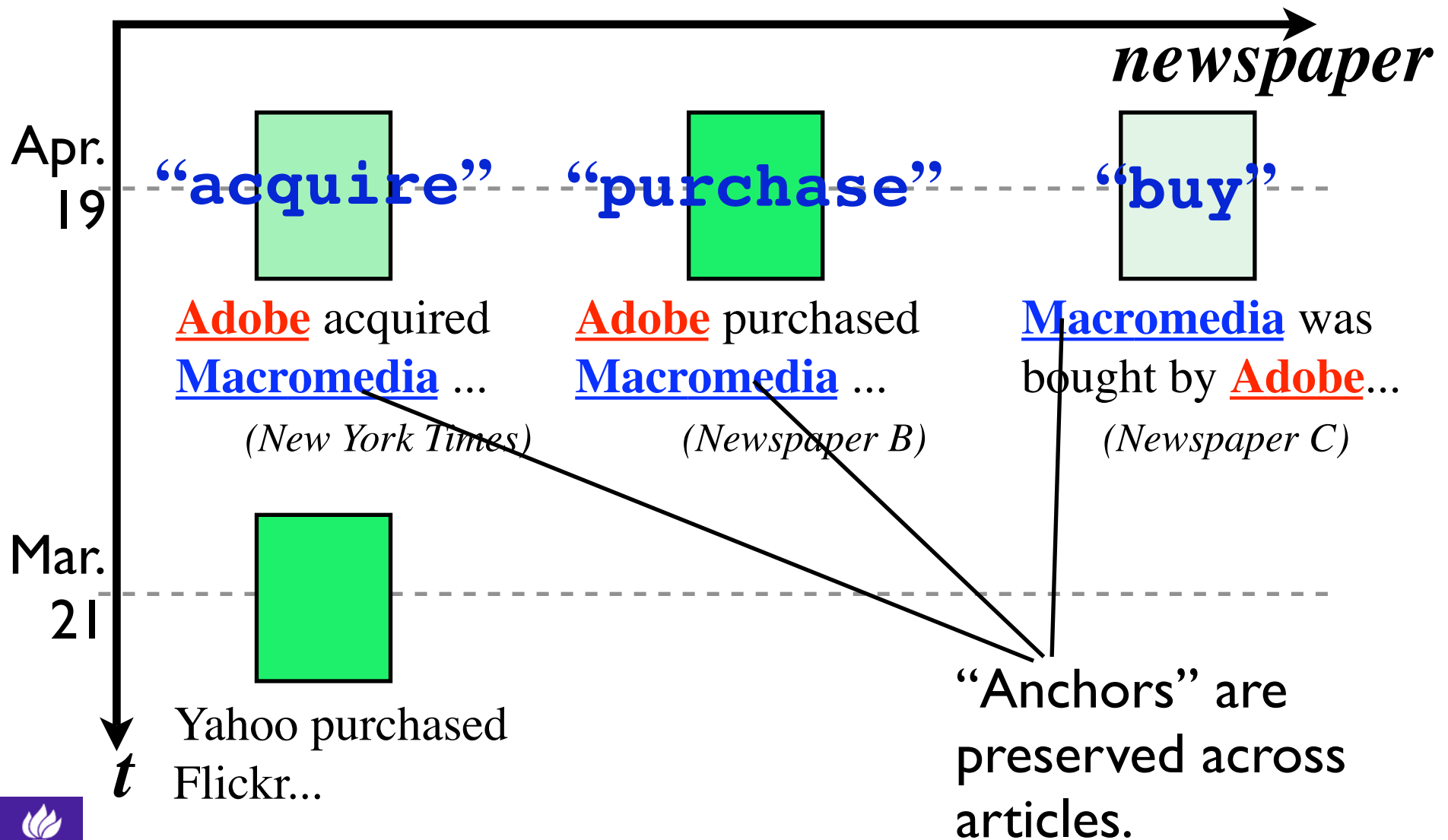
Obtain Paraphrases



Obtain Paraphrases



Obtain Paraphrases

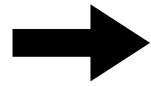
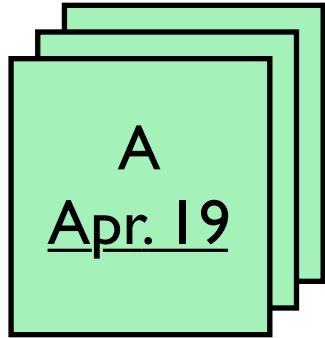


Obtain Paraphrases

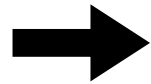
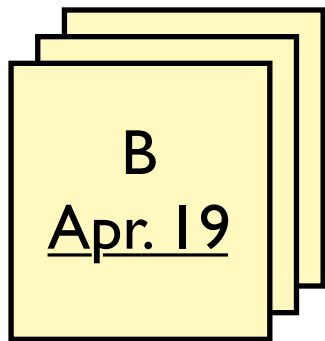
- Overall method:
 1. Obtain pairs of comparable articles from different newspapers **on the same day.**
 - Use Simple TF/IDF-based method.
 2. Extract the main sentence pair.
 3. Extract the paraphrase pair using anchors.
 - Parse / GLARF-convert the main sentence.
 - Extract the portion of the GLARF structure which shares the anchors.



Experiments



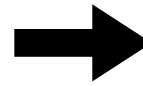
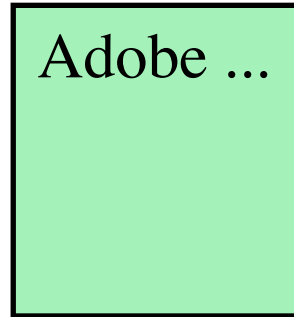
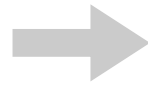
Find comparable articles
(Simple TF/IDF-based method)



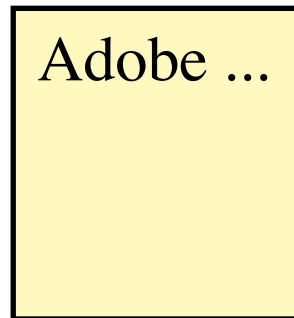
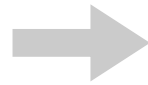
Different
Newspapers



Experiments



Pick main sentences
(First sentence)

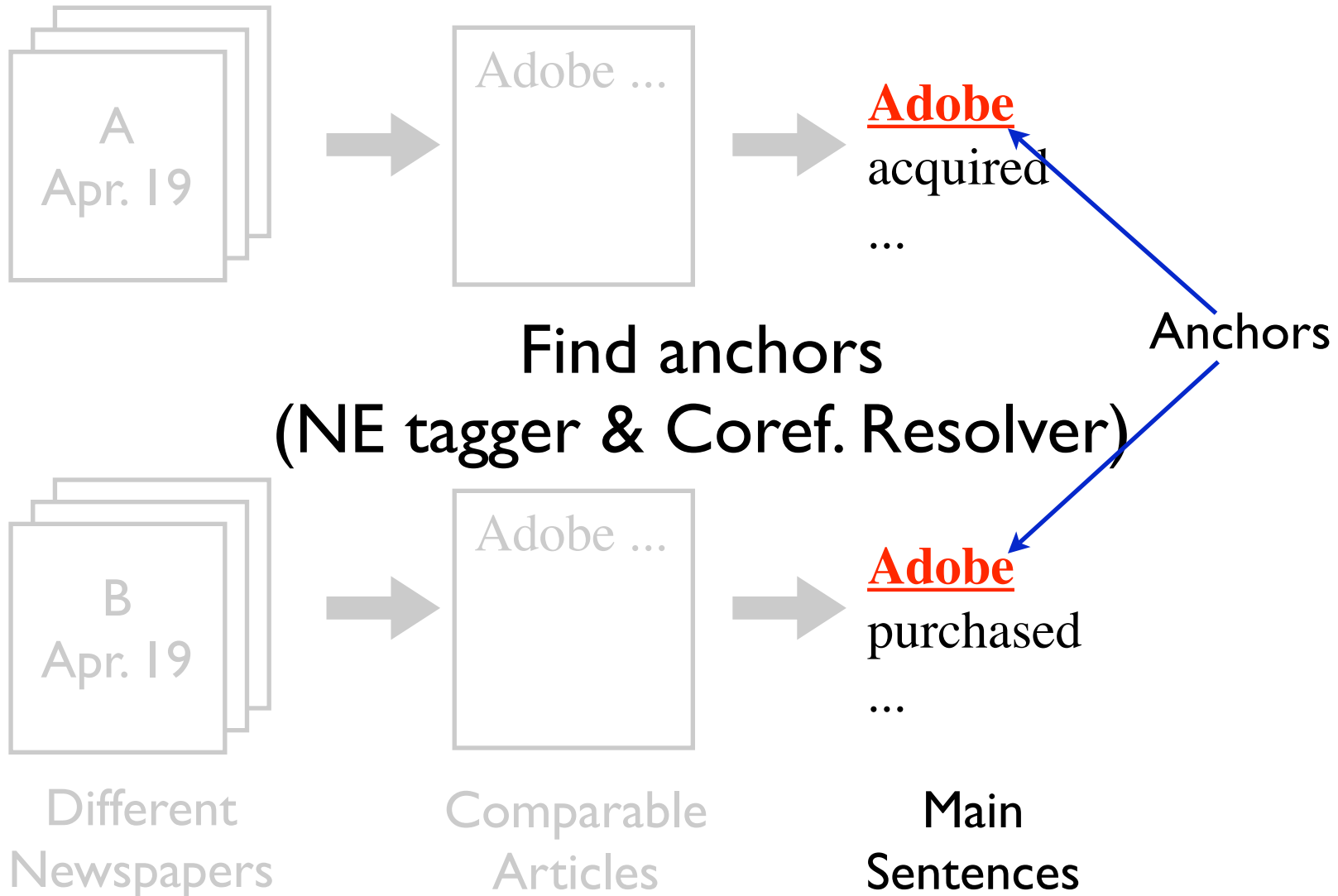


Different
Newspapers

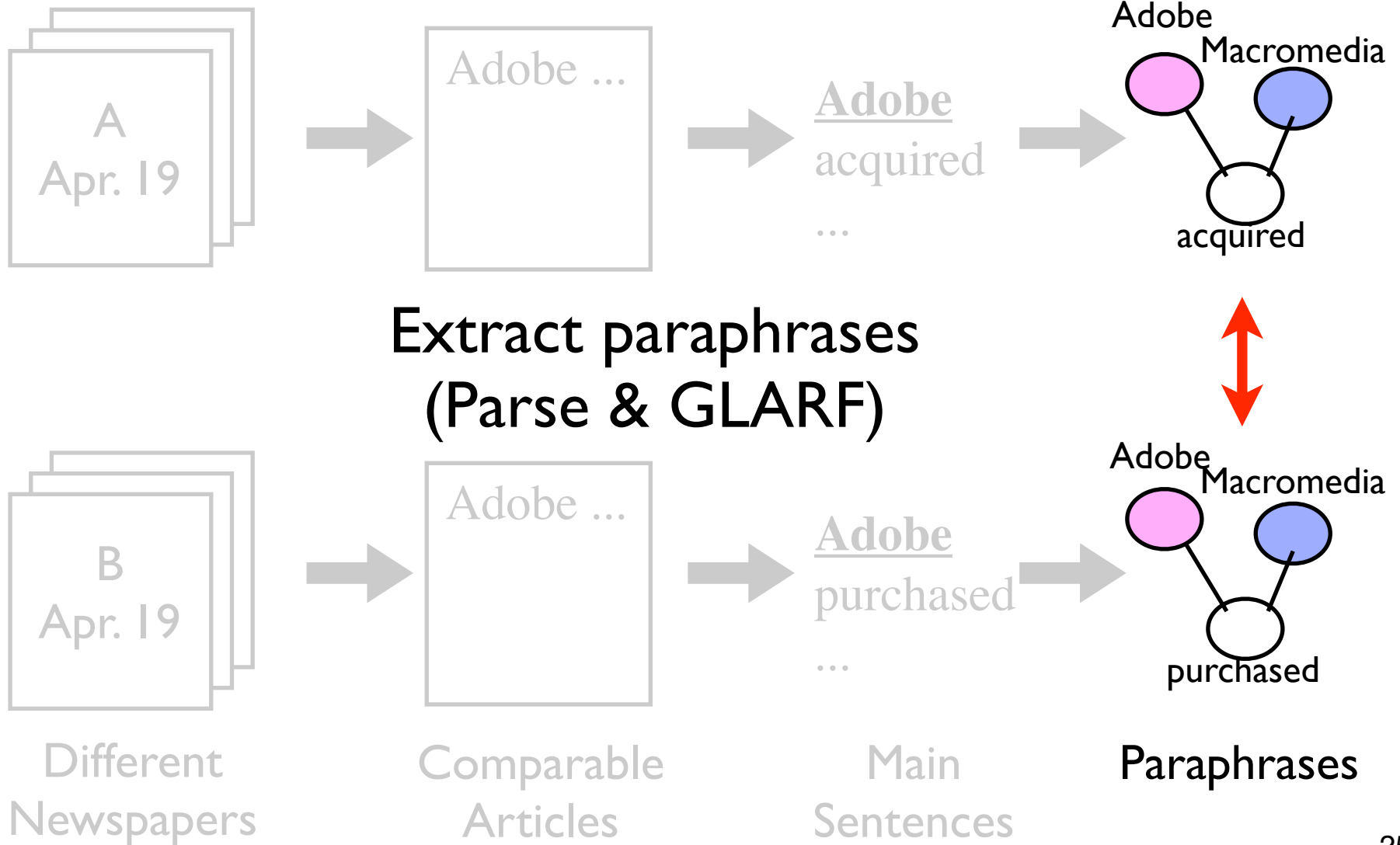
Comparable
Articles



Experiments

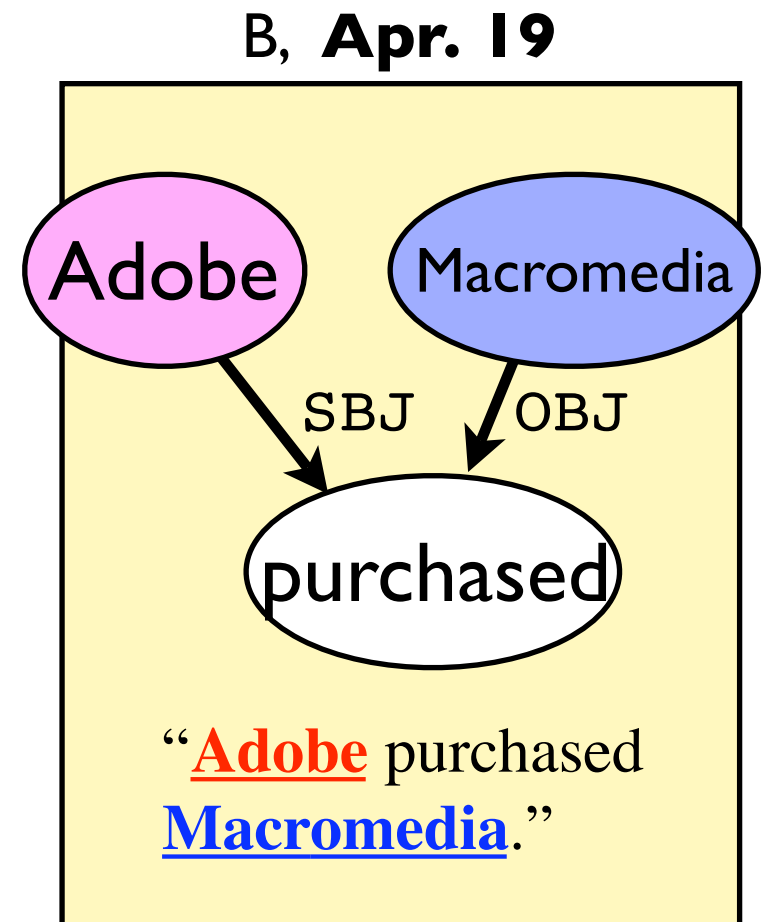
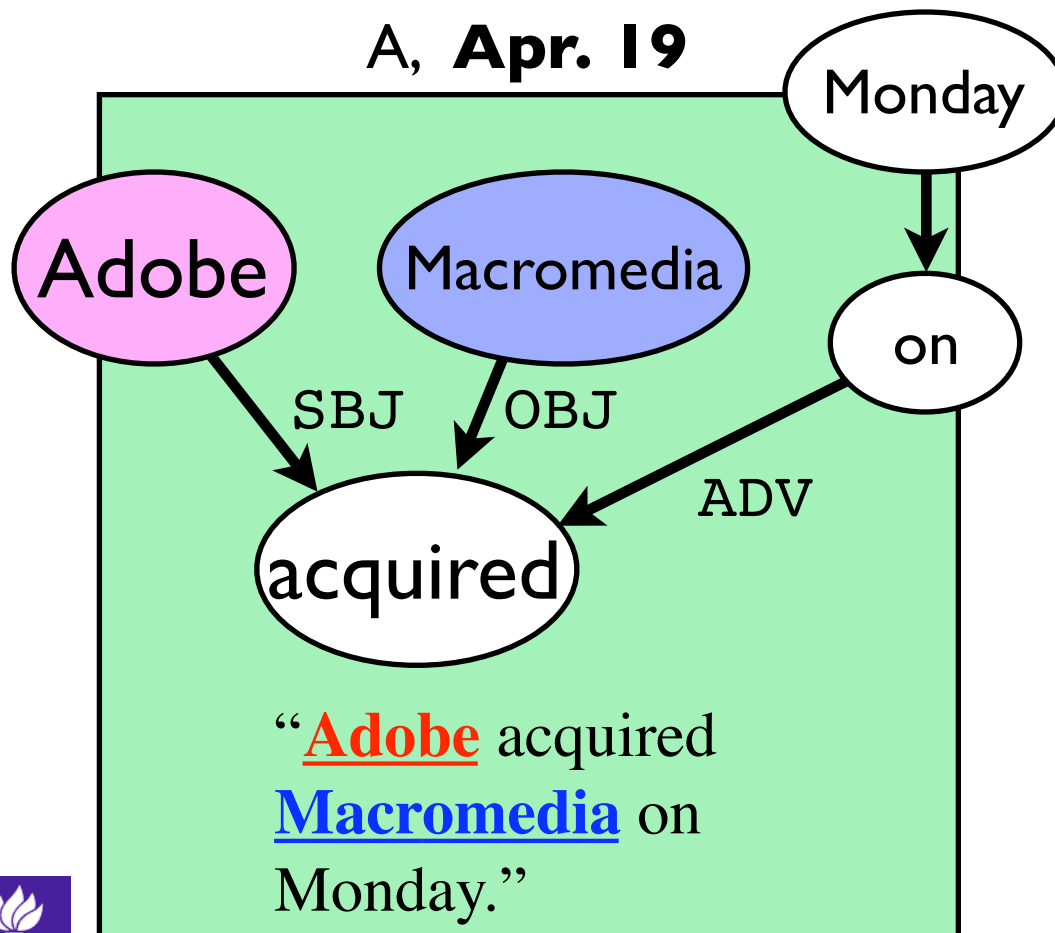


Experiments



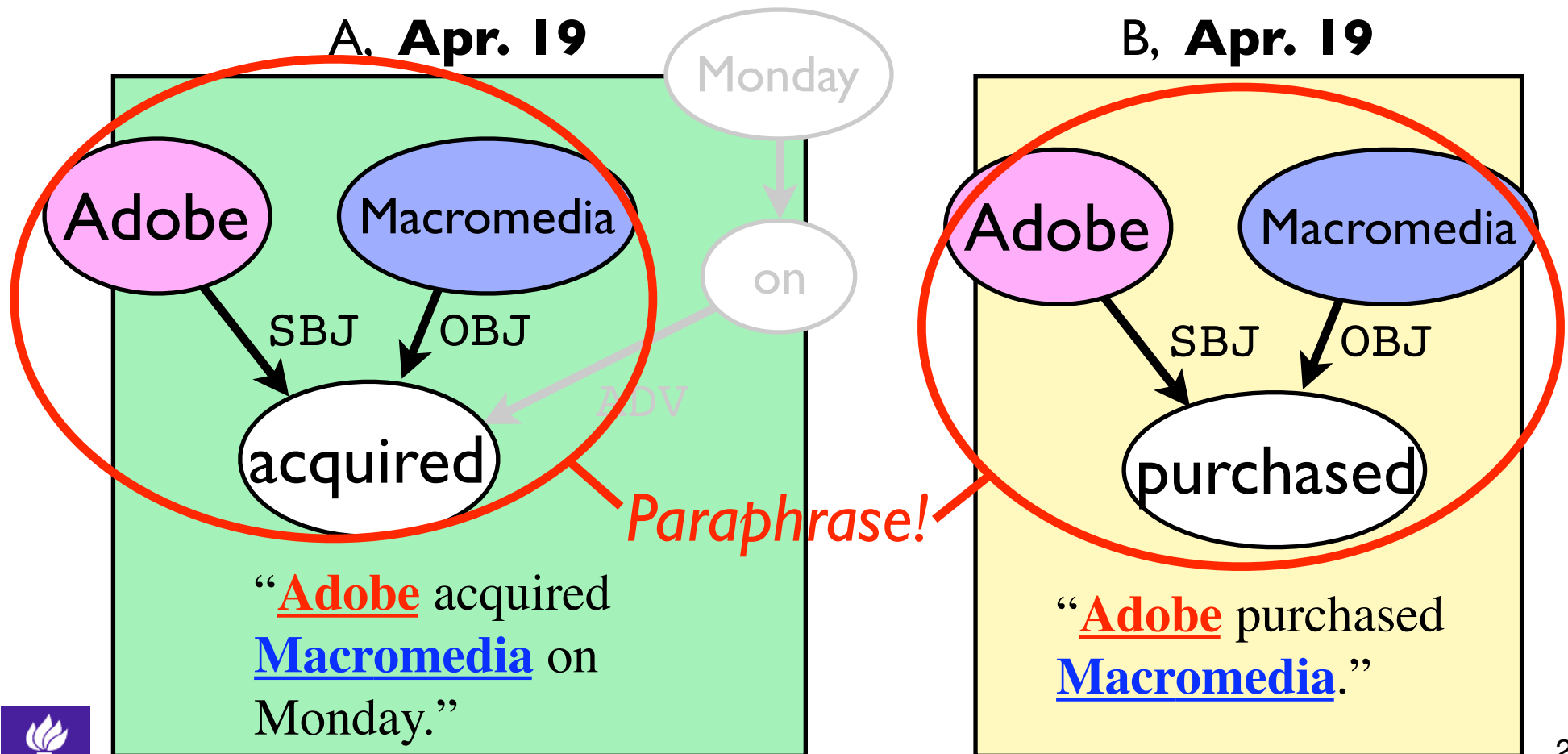
Extract Paraphrase

- Take the minimum spanning tree:



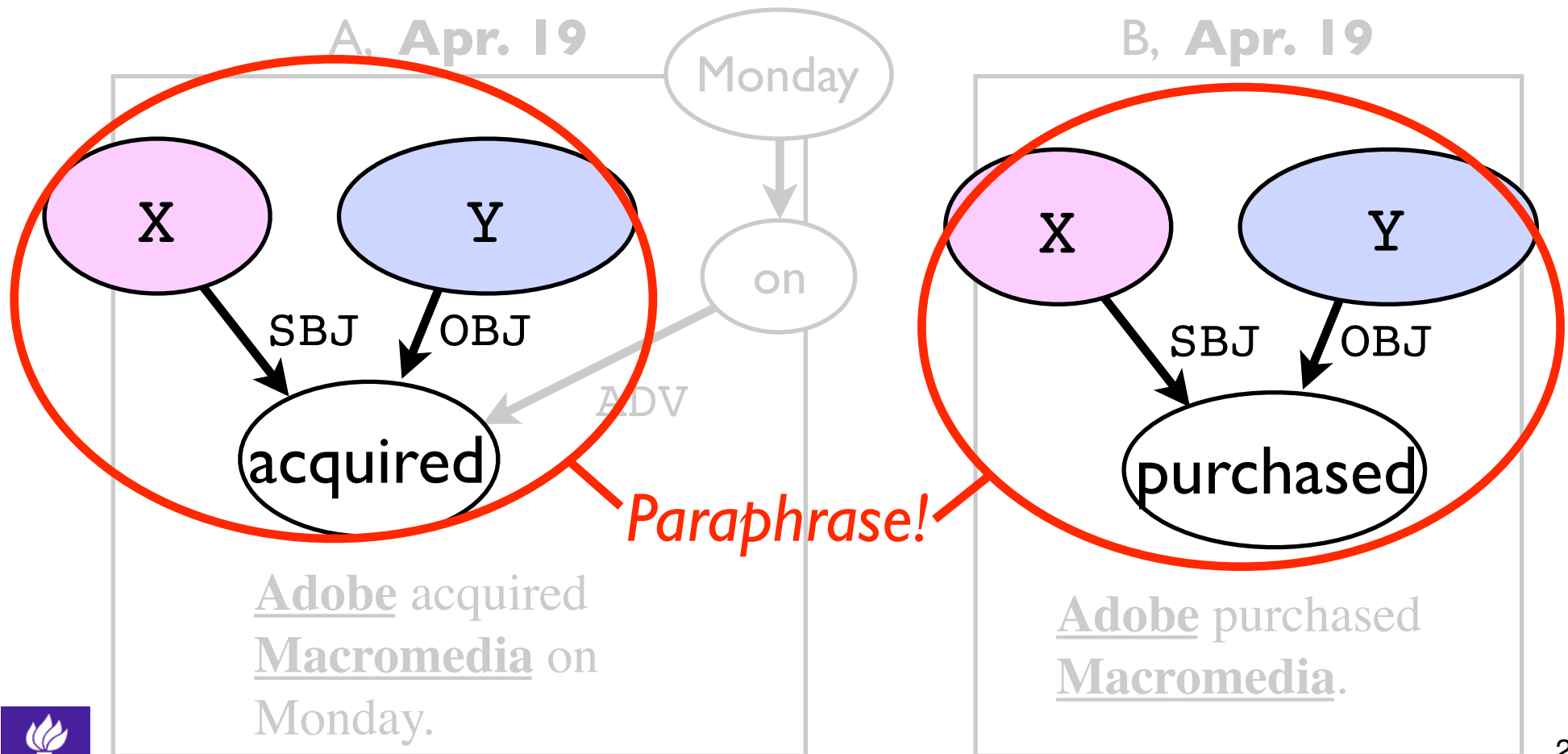
Extract Paraphrase

- Take the minimum spanning tree:



Extract Paraphrase

- Generalize anchors to variables.



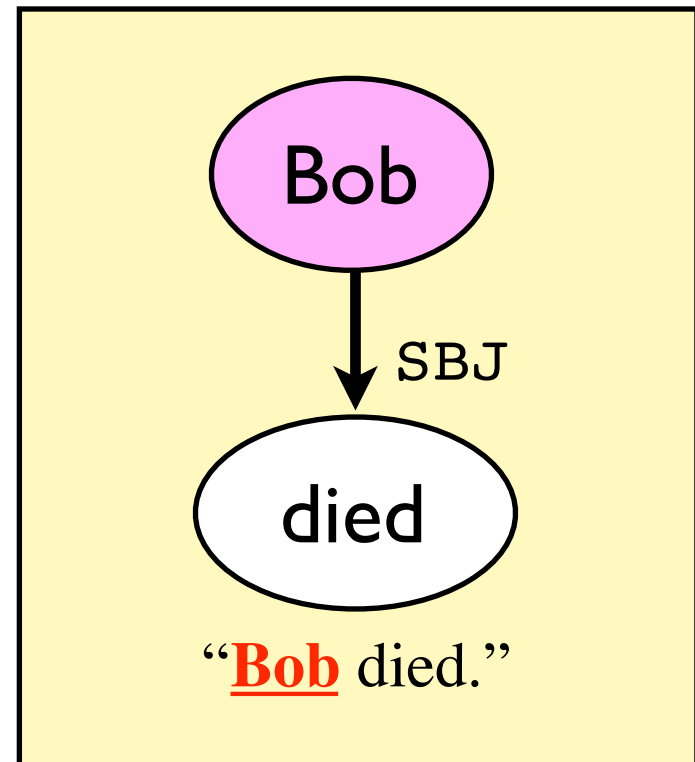
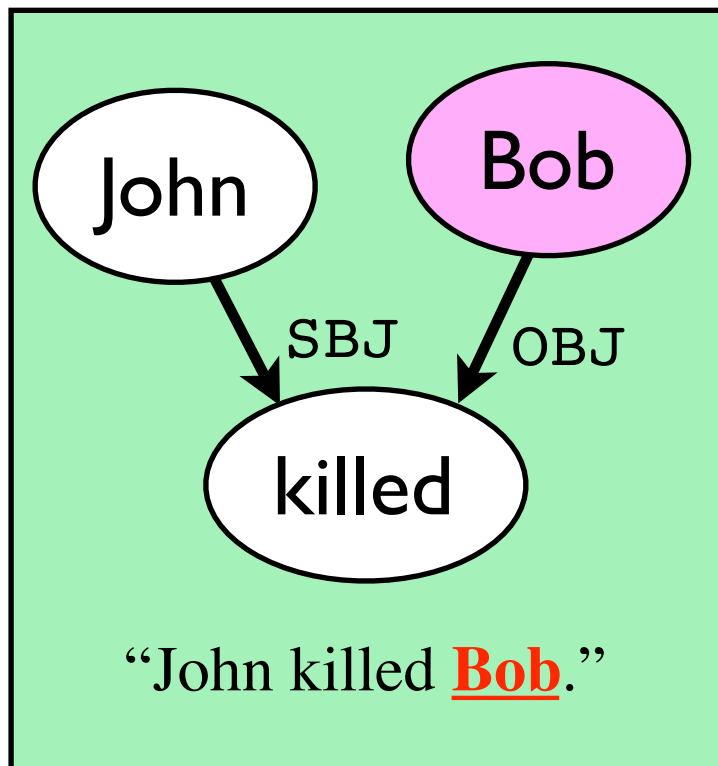
Extract Paraphrase

- Generalize anchors to variables.
 - A. **X** acquired **Y**
 - B. **X** purchased **Y**
 - (in the form of GLARF tree)
- Focused on two types of structures:
 - One-anchor / Two-anchor
 - Three or more anchors are too specific.
 - A. “**X** acquired **Y** from **Z** on **W**”
 - B. “**Y** who filled in **Z** was traded **W** to **X**”



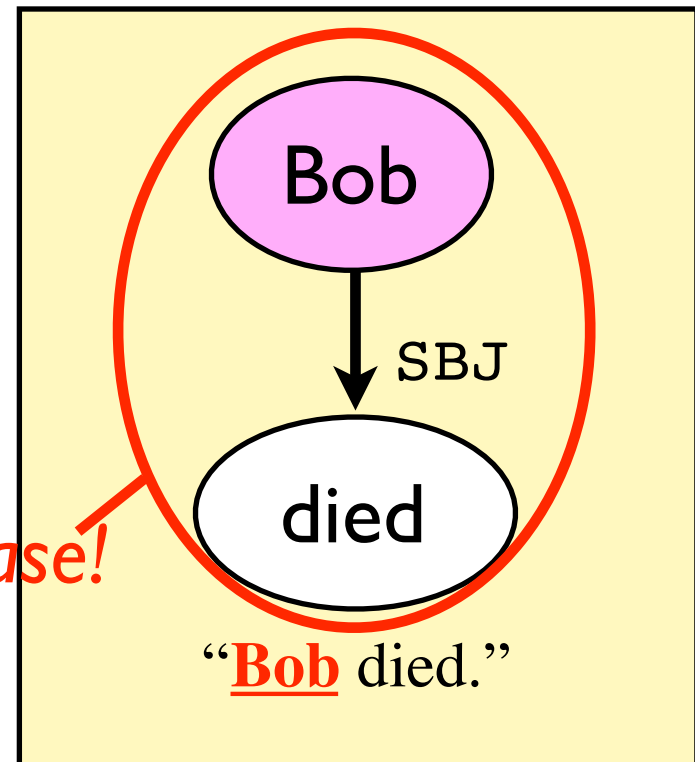
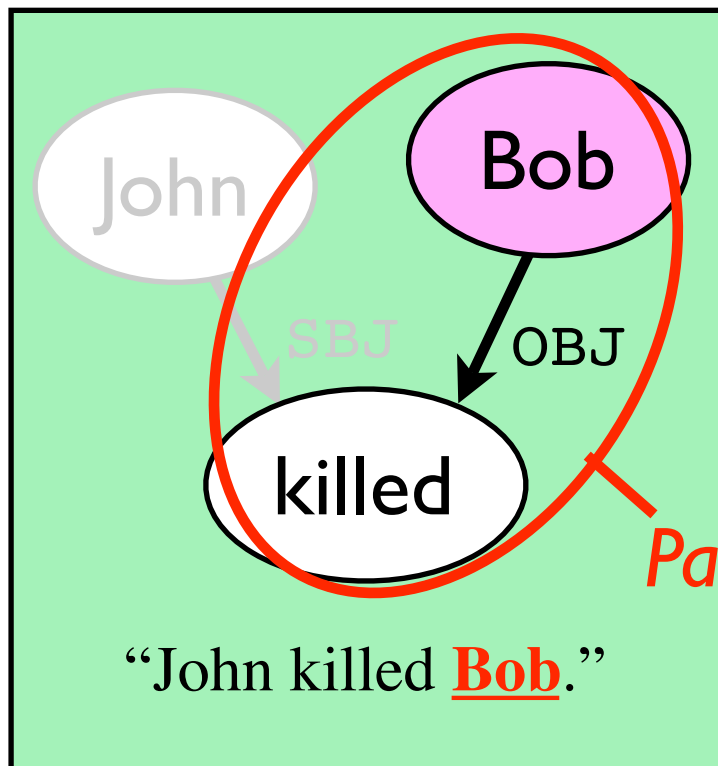
Extract Paraphrase

- Paraphrase which has one anchor.
 - “John killed **Bob**”
 - “**Bob** died.”



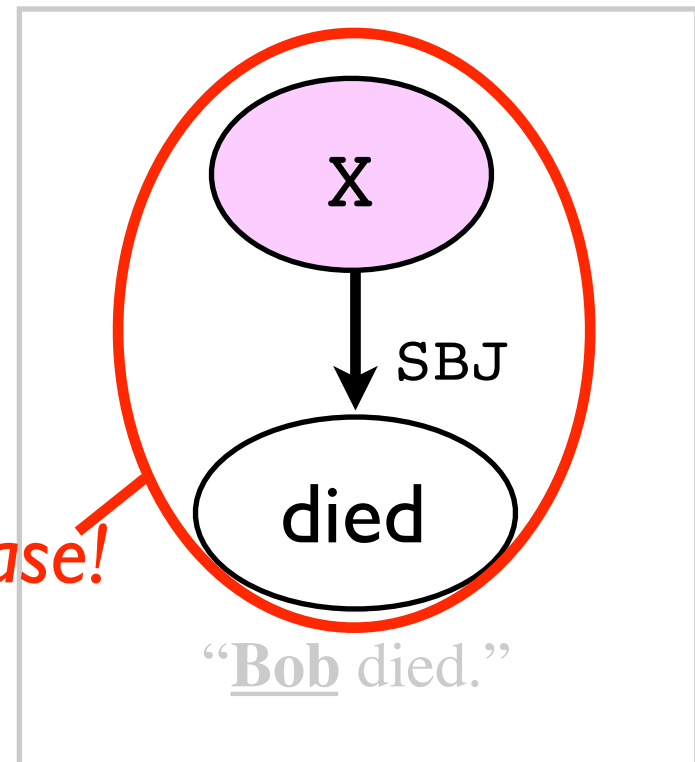
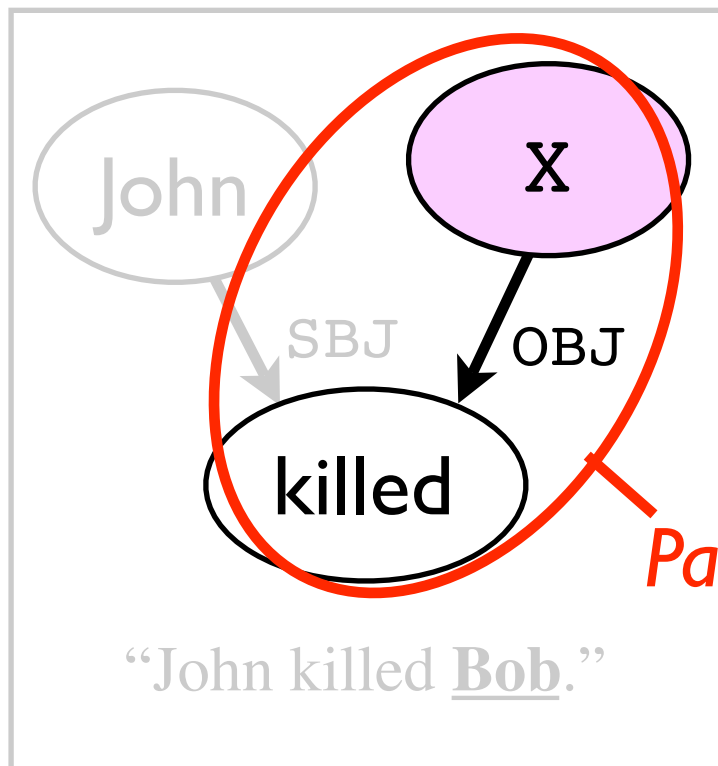
Extract Paraphrase

- Paraphrase which has one anchor.
 - “John killed **Bob**”
 - “**Bob** died.”



Extract Paraphrase

- Paraphrase which has one anchor.
 - ... killed **X**
 - **X** died



Paraphrase!



Prelim. Experiment

- Source: 5-year NYT, Reuters, Xinhua
 - About 500,000 articles in total.
 - 13,036 comparable article pairs found.
 - 54,257 paraphrase pairs obtained.

Paraphrase	Obtained Pairs	Unique Pairs	Freq. Unique Pairs
One-anchor	53,163	45,306	53 (freq. ≥ 4)
Two-anchor	1,094	987	41 (freq. ≥ 2)



Prelim. Experiment

- Evaluation
 - Correct: two facts are same (for IE).
 - Marginal: one is implied by the other.
 - “X voted on Y” / “X rejected Y”.
 - Incorrect: two facts are different (for IE).

Paraphrase	Obtained	Correct	Marginal	Incorrect
One-anchor	53 (freq. ≥ 4)	30 (57%)	11	12
Two-anchor	41 (freq. ≥ 2)	35 (85%)	1	5



Prelim. Experiment

- Obtained paraphrases (one-anchor):

Freq.	Newspaper A	Newspaper B
38	X announced	X said
13	X reported	X said
9	X said	X warned
8	X today	X said incorrect
8	X said	X urged
7	X declared	X said
7	X arrived	X said incorrect

6	X took	X won

4	killed X	X died
4	accused X	charged X



Prelim. Experiment

- Obtained paraphrases (two-anchor):

Freq.	Newspaper A	Newspaper B
21	X announced Y	X said Y
5	X said Y	X told Y
4	X said Y	X suggested Y
3	X ruled Y	X upheld on Y
3	X declared Y	X said Y
3	X accused Y	X said Y marginal
2	X accused Y	X leashed into Y
2	X broke Y	X set Y incorrect
2	X ruled Y	X said Y
2	X jumped in Y	X rose in Y



Error Analysis

- Event mismatch (8 out of 17)
 - “**x** arrived” / “**x** said”
 - A. “Turkish President Suleyman Demirel **arrived** Thursday for a two-day visit aimed at boosting trade with this former communist country.”
 - B. “Turkish President Suleyman Demirel **said** Thursday that Turkey and Romania are bent on realizing a close cooperation in regional and international platforms , especially in the economic domain.”



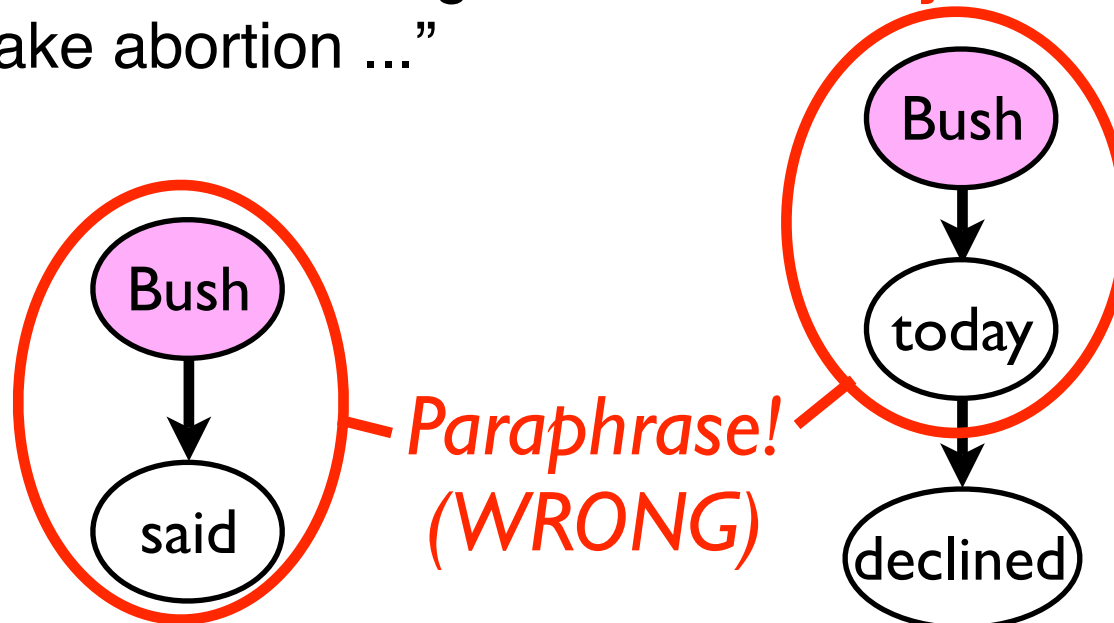
Error Analysis

- Event mismatch (8 out of 17)
 - Can be reduced by improving sentence matching.
 - **For Accuracy:** Filter if two sentences are significantly different.
 - **For Coverage:** Choose one from the first three sentence instead of only the first one.



Error Analysis

- Persistent parse error (5 out of 17)
 - “**x** today” / “**x** said”
 - A. “Texas Gov. George W. Bush **said** Wednesday that a pro-abortion rights ...”
 - B. “Texas Gov. George W. Bush **today** declined to make abortion ...”



Error Analysis

- Bad generalization (4 out of 17)
 - “**X** broke **Y**” / “**X** set **Y**”
 - A. “Teenage sensation Ian Thorpe **broke** the world record for the men’s short-course 400-meter freestyle.”
 - B. “Ian Thorpe **set** a world record in the men’s 400 meters freestyle.”
 - “**X** broke **the record**” / “**X** set **a record**” would be fine.
 - Replacing a whole NP with a variable is bad.



Error Analysis

- Partially generalized paraphrases?
 - Head word of **Y** is same.

Freq.	Newspaper A	Newspaper B
2	jumped Y in X	rose Y in X
2	X left for Y	X left on Y
1	X show at Y	X stand at Y
1	X held onto Y	X regained Y
1	X scored Y	X won Y
1	X confirmed after Y	X ending years of Y



Error Analysis

- Partially generalized paraphrases?
 - Preserve the head word as restriction.

Freq.	Newspaper A	Newspaper B
2	jumped <i>Y_percent</i> in X	rose <i>Y_percent</i> in X
2	X left for <i>Y_visit</i>	X left on <i>Y_visit</i>
1	X show at <i>Y_trial</i>	X stand at <i>Y_trial</i>
1	X held onto <i>Y_control</i>	X regained <i>Y_control</i>
1	X scored <i>Y_victory</i>	X won <i>Y_victory</i>
1	X confirmed after <i>Y_debate</i>	X ending years of <i>Y_debate</i>



Error Analysis

- Partially generalized paraphrases?
 - Introduce various kinds of restrictions:
 - No restriction:
X got **Y** / **X** made **Y**.
 - Head word of **Y** is specified:
X got **Y_start** / **X** made **Y_start**.
 - The literal string of **Y** is specified:
X got ***a_strong_start*** /
X made ***a_strong_start***.
 - NE category is specified:
PERSON got ***a_strong_start*** /
PERSON made ***a_strong_start***.



Error Analysis

- Low recall
 - 500k articles → 100 paraphrases.
 - Only that's it?
 - Find more anchors:
 - Improve cross-document coreference.
 - Pronouns are currently not regarded as anchors.
 - Use single-document coreference resolution.
 - Use more source articles.
 - Obtain news articles from the web.



Research Plan

- My Contribution
 - i. Paraphrases for interpretation: IE.
 - Text generation: [Barzilay, 01]
 - Machine translation: [Pang, 03]
 - Q&A system: [Lin, 01] [Ravichandran, 02]
 - ii. Wide domain: news articles.
 - iii. Deeper analysis: GLARF and coreference.
 - Word sequence: [Barzilay, 01] [Ravichandran, 02]
 - MiniPAR: [Lin, 01]



Research Plan

1. Obtain better paraphrases
 - a. Improve main sentence extraction.
 - b. Find more anchors / articles.
 - c. Partially generalized paraphrases.
- **What I did not address:**
 - NP-NP paraphrases:
 - ex. “Microsoft” /
“the Redmond-based software giant”?



Research Plan

2. Perform experiments with various kinds of news articles.
 - Use data from the web.
 - Obtain domain-specific paraphrases.
3. Evaluate for actual IE application.
 - Try to cluster expressions by its similarity.
 - Possibility of fully-automatic IE?

