Automatic extraction of paraphrasing rules: A survey and plans for future work

Prodromos Malakasiotis, Ph.D. student
Department of Informatics
Athens University of Economics and Business
rulller@aueb.gr

What is Paraphrasing?

- “X is the writer of Y” ≈ “X wrote Y” ≈ “X is the author of Y”.
- “Oswald killed Kennedy” / “Kennedy was killed by Oswald”.
- “Who invented the light bulb?” / “Who was the inventor of the light bulb?”.
- “Edison invented the light bulb” / “Edison’s invention of the light bulb”.
- “Athens is located in Greece” / “Athens is the capital of Greece”.
  - Textual entailment, not really paraphrasing.
- Can be used in:
  - Question Answering, Information Retrieval, Web Search Engines.
  - Natural Language Generation, Automatic Summarization.

Contents of this talk

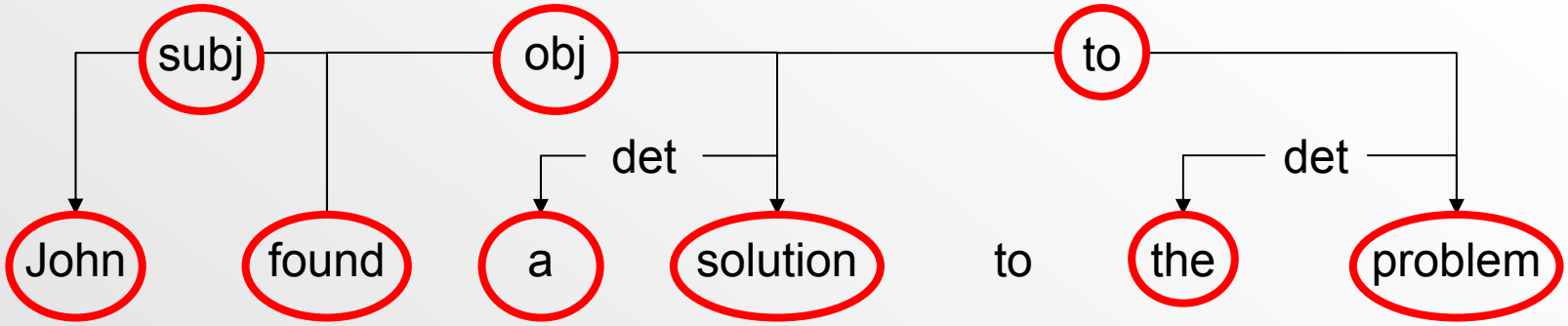
- What is Paraphrasing?

Paraphrasing methods

- Lin & Pantel.
- Barzilay & McKeown.
- Barzilay & Lee.
- Pang et al.
- Ibrahim et al.

- Directions for future work.

Lin & Pantel's method (1 of 3)



"X finds solution to Y"

Lin & Pantel's method (2 of 3)

"X finds a solution to Y"		"X solves Y"	
Slot X	Slot Y	Slot X	Slot Y
commission	strike	committee	problem
committee	civil war	clout	crisis
committee	crisis	government	problem
government	crisis	he	mystery
government	problem	she	problem
he	problem	petition	woe
I	situation	researcher	mystery
legislator	budget deficit	resistance	crime
sheriff	dispute	sheriff	murder

Lin & Pantel's method (3 of 3)

- Good performance despite only approximately correct or occasionally incorrect paraphrases.
  - "X caused Y" ≈ "Y is blamed on X".
  - "X asks Y" ≈ "Y asks X".
  - "X worsens Y" ≈ "X solves Y".
- Requires reliable dependency parser.
  - Computationally expensive.
  - Not always available (e.g. in Greek).

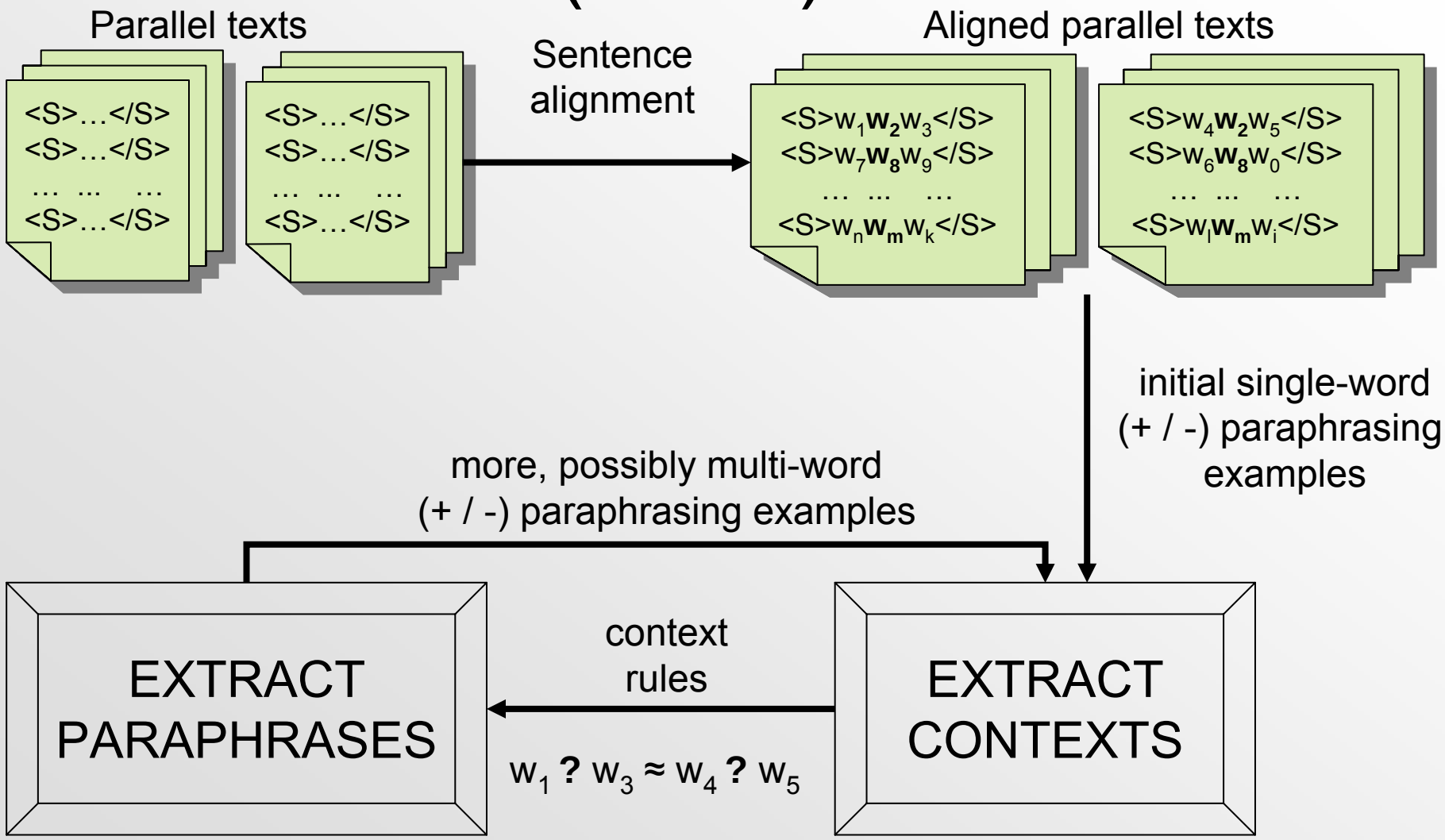
Contents of this talk

- What is Paraphrasing?

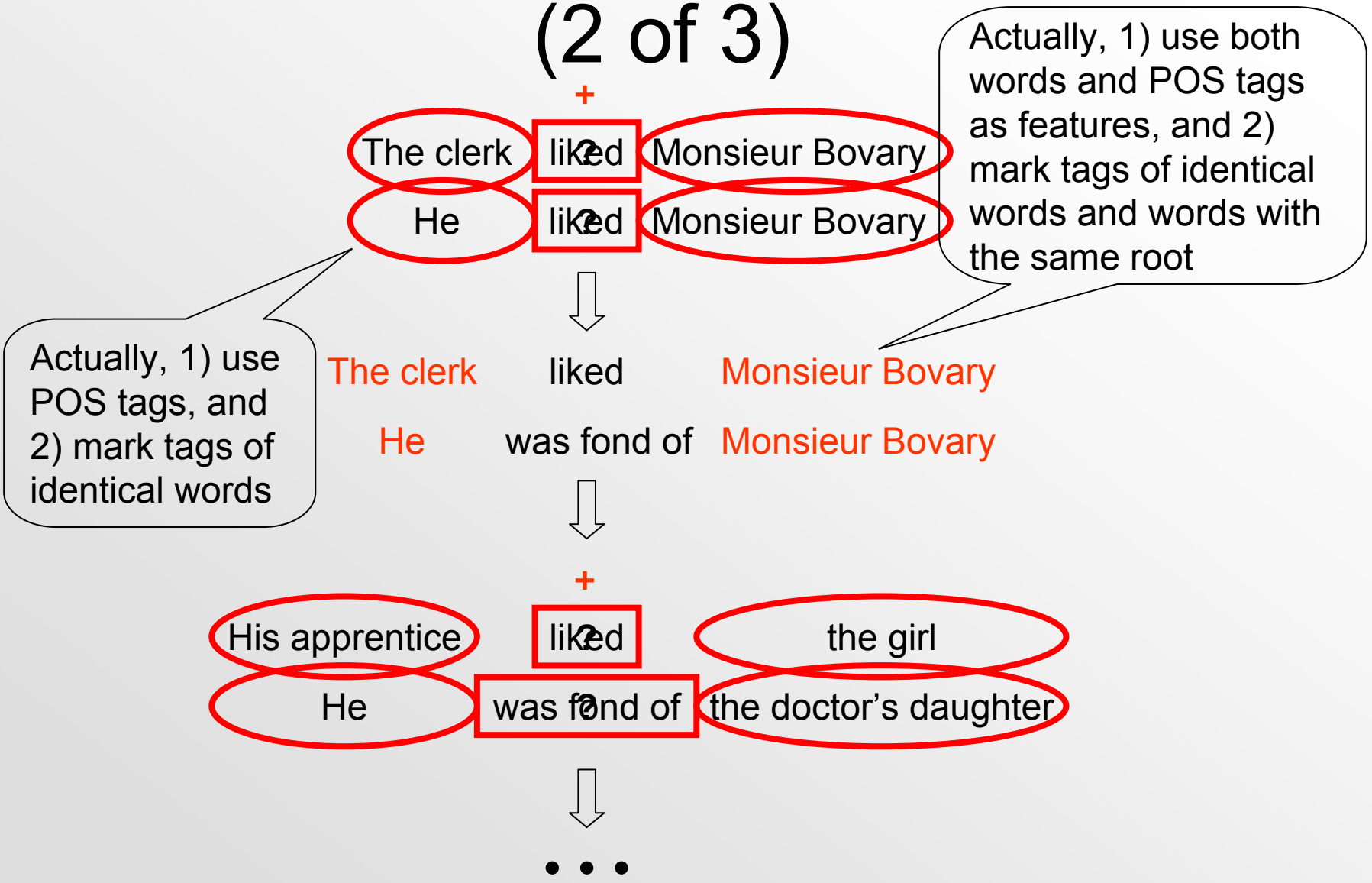
Paraphrasing methods

- **Lin & Pantel.**
  - **Dependency paths with similar slot fillers have similar meanings.**
- Barzilay & McKeown.
- Barzilay & Lee.
- Pang et al.
- Ibrahim et al.
- Directions for future work.

Barzilay & McKeown's method (1 of 3)



Barzilay & McKeown's method (2 of 3)



Barzilay & McKeown's method (3 of 3)

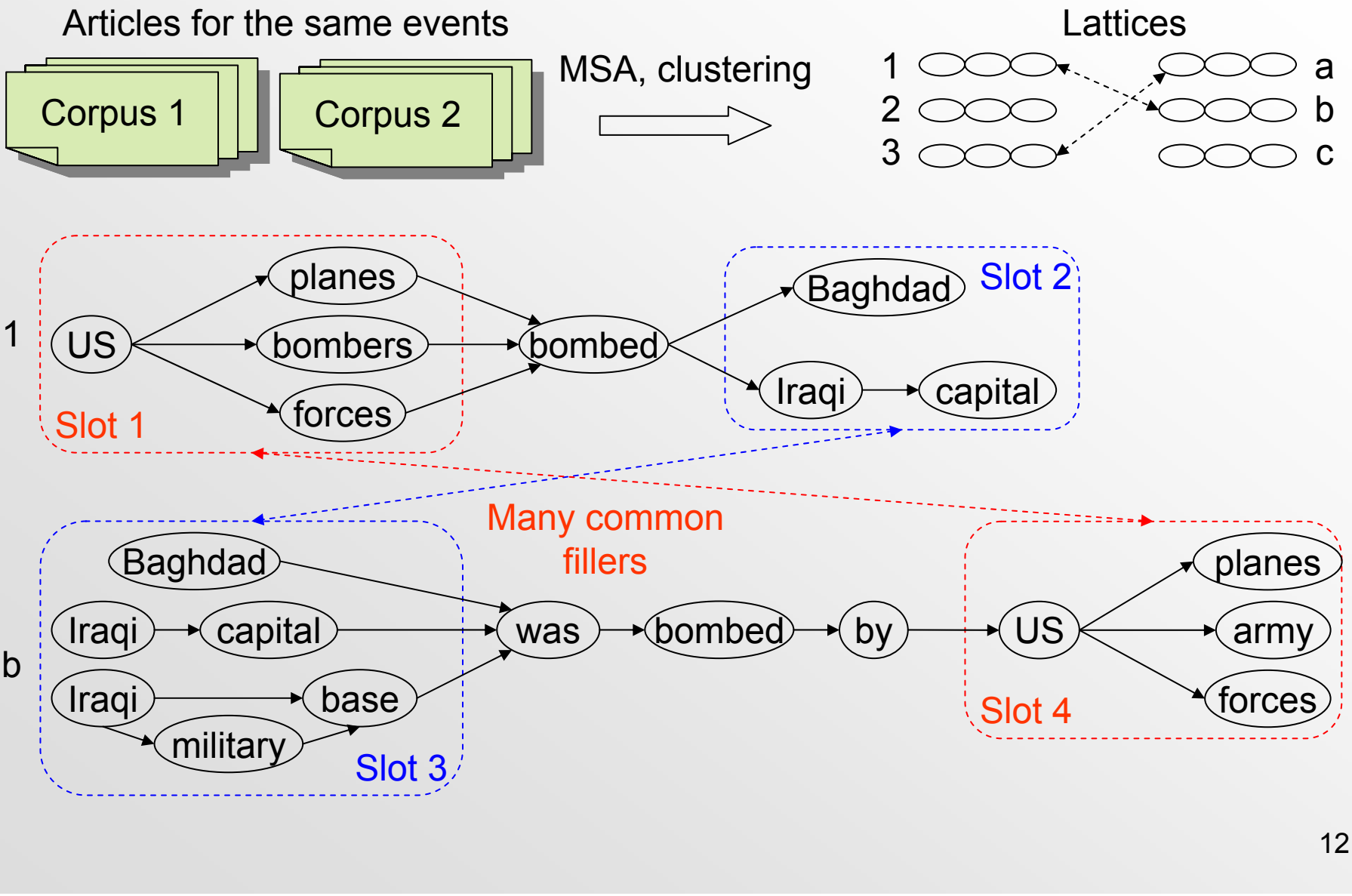
- High precision
  - 86.5% when context not given to human judges.
  - 91.6% when context given to human judges.
- But 70.8% single word paraphrases.
  - In effect low recall
- Requires parallel corpus.
  - Difficult to obtain.
- Requires POS tagger, aligner.
  - Easier to obtain.

Contents of this talk

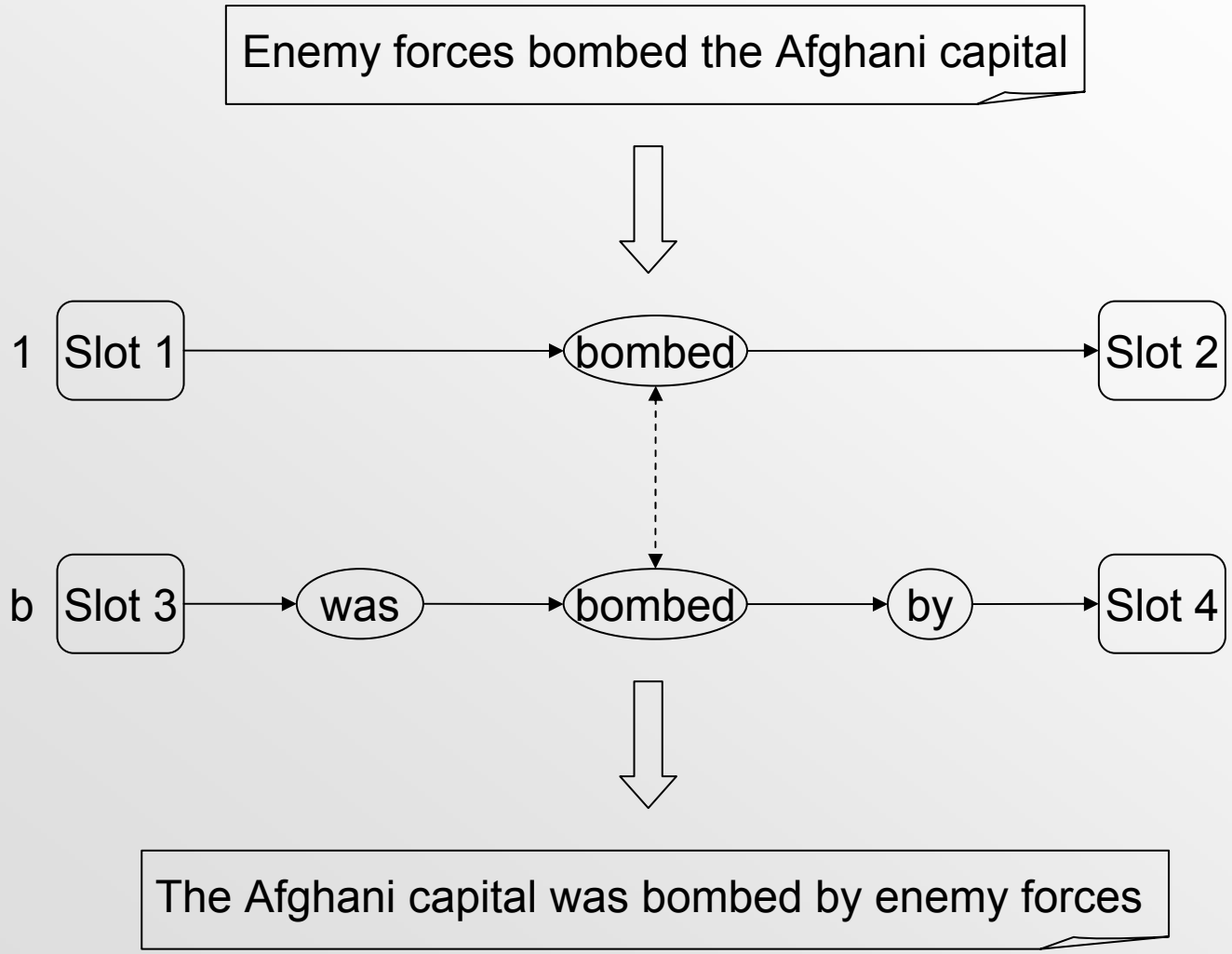
- What is Paraphrasing?
- Lin & Pantel.
  - Dependency paths with similar slot fillers have similar meanings.
- **Barzilay & McKeown.**
  - **Identical words → contexts → more paraphrases → more contexts → …**
- Barzilay & Lee.
- Pang et al.
- Ibrahim et al.
- Directions for future work.

Paraphrasing methods

Barzilay & Lee's method (1 of 3)



Barzilay & Lee's method (2 of 3)



Barzilay & Lee's method (3 of 3)

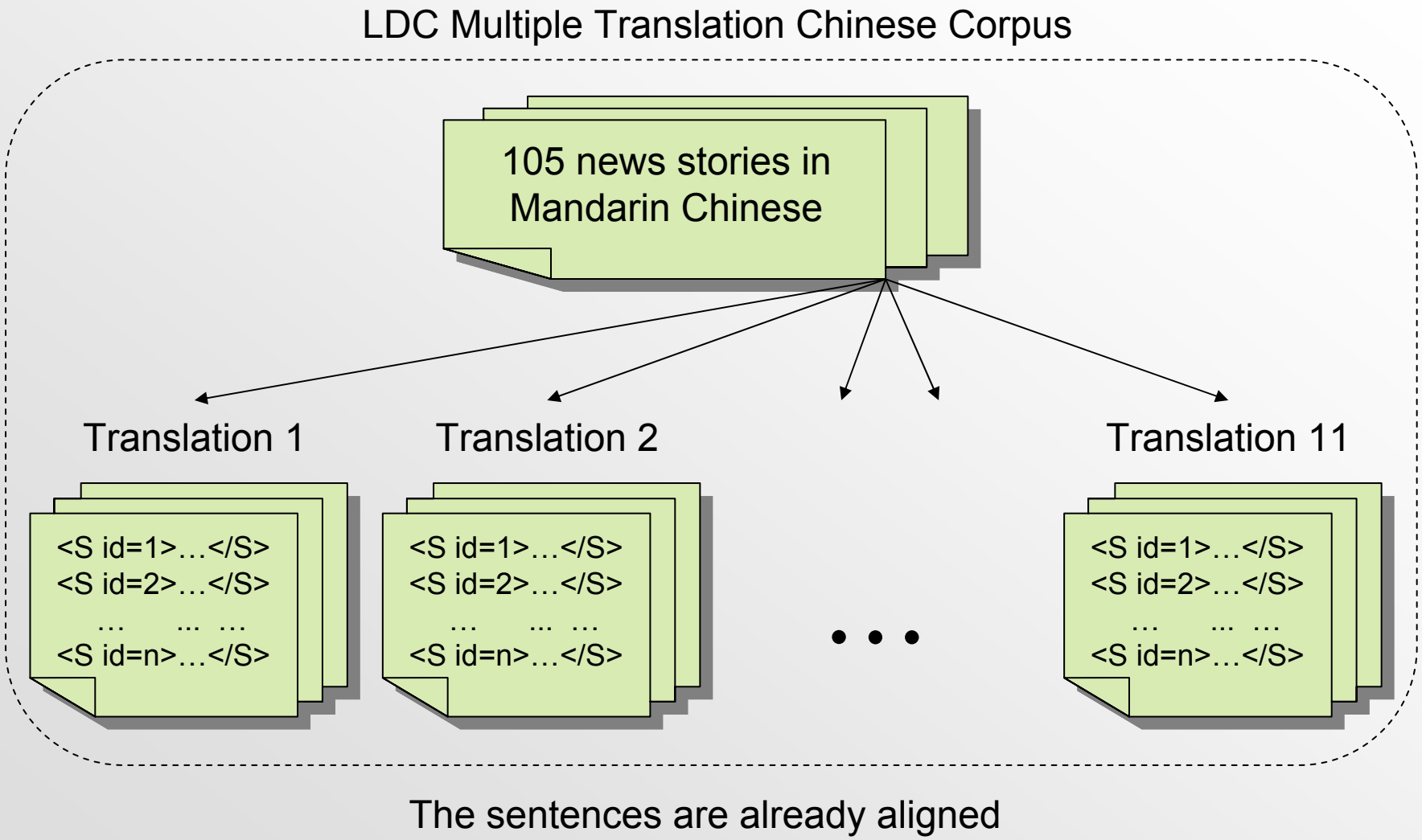
- Relatively high precision (78.5%).
  - Many sentence-level paraphrases.
  - Unknown recall.
  - Seems to outperform Lin & Pantel's method (42.5% precision).
- But able to paraphrase only 12.2% of a set of new sentences.
  - Input does not match any lattice.
  - Precision at the same level (79.7%).
- Does not require dependency parser, POS tagger, aligner, etc.
  - Uses simplistic named-entity (NE) recogniser.
  - NE recognition could help other methods too.

Contents of this talk

Paraphrasing methods

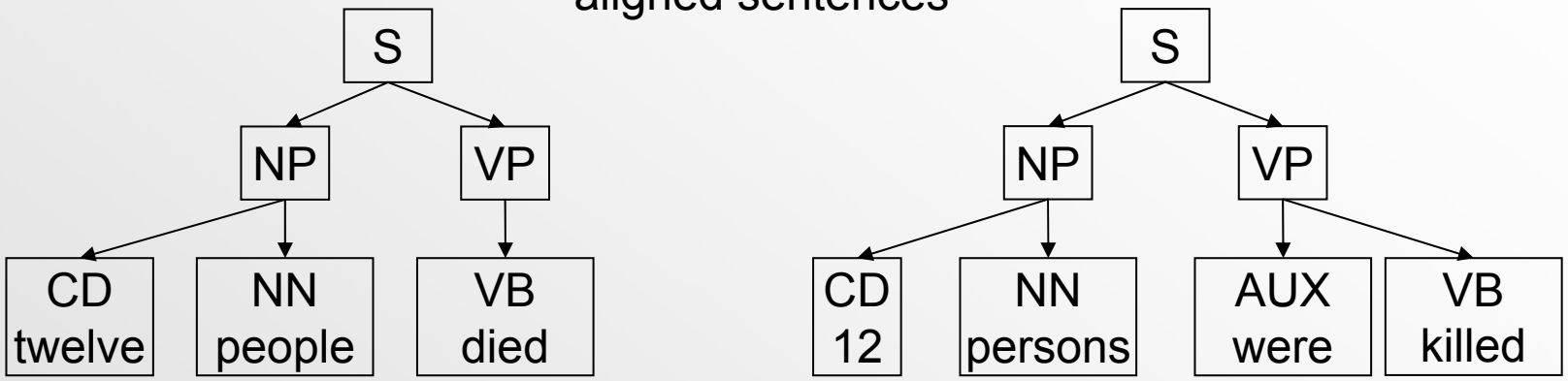
- What is Paraphrasing?
- Lin & Pantel.
  - Dependency paths with similar slot fillers have similar meanings.
- Barzilay & McKeown.
  - Identical words → contexts → more paraphrases → more contexts → …
- **Barzilay & Lee.**
  - **Lattices with common slot fillers tend to correspond to paraphrases.**
- Pang et al.
- Ibrahim et al.
- Directions for future work.

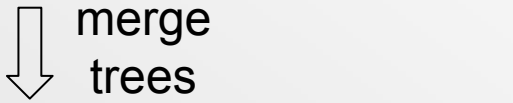
Pang et al.’s method (1 of 4)

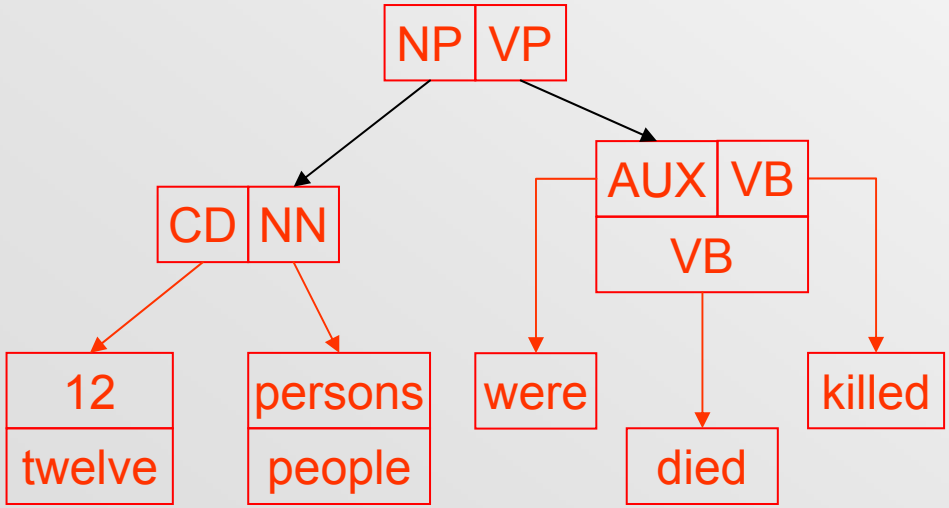


Pang et al.'s method (2 of 4)

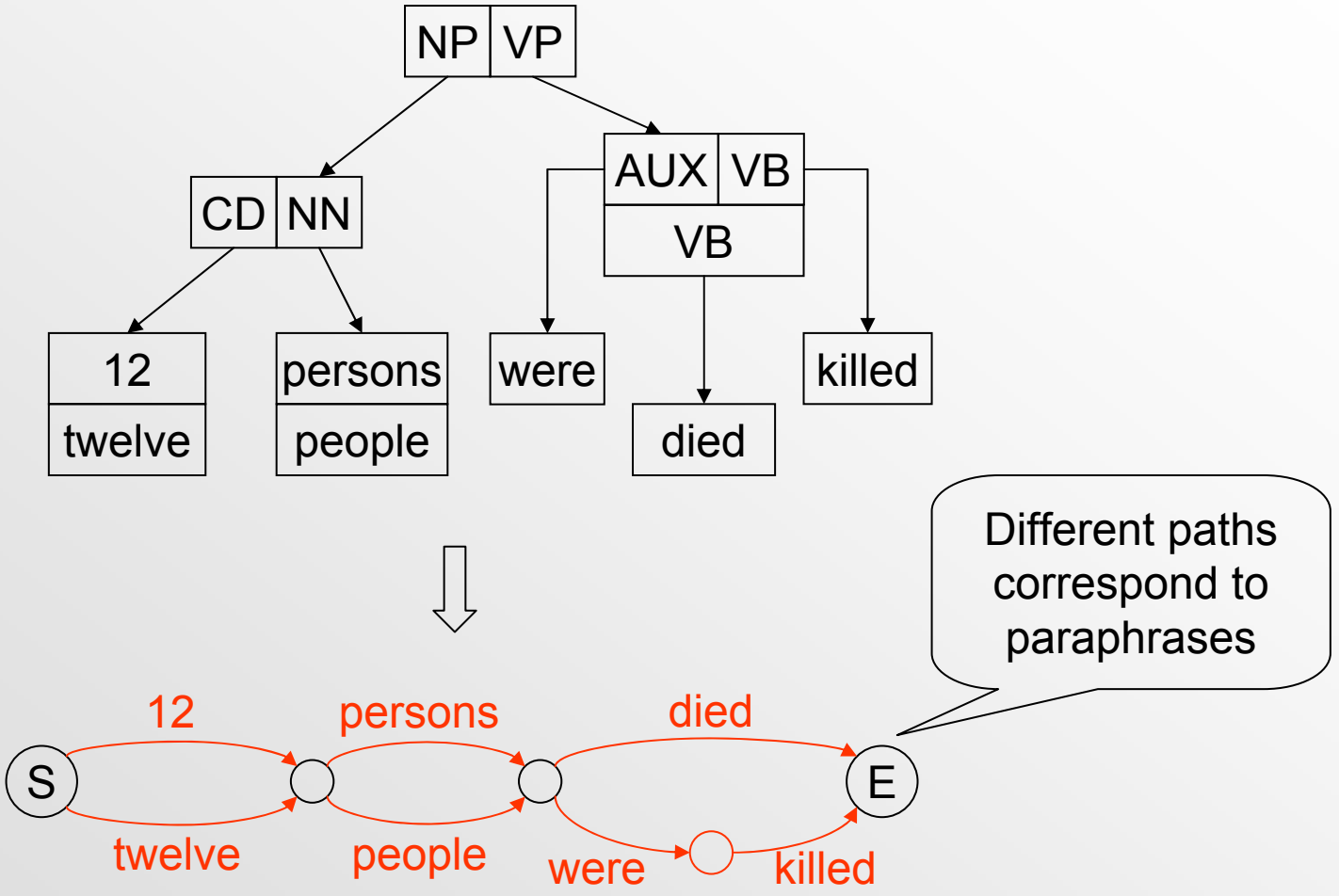
Parse trees of
aligned sentences







Pang et al.'s method (3 of 4)



Pang et al.’s method (4 of 4)

- Better results than Barzilay & McKeown’s method.
  - 81% vs. 66% precision, context not given to human judges.
  - 93% vs. 77% precision, context given to human judges.
- Produces complete sentences not patterns.
- Requires reliable parser, parallel corpus.
  - Difficult to obtain (e.g. in Greek).

Contents of this talk

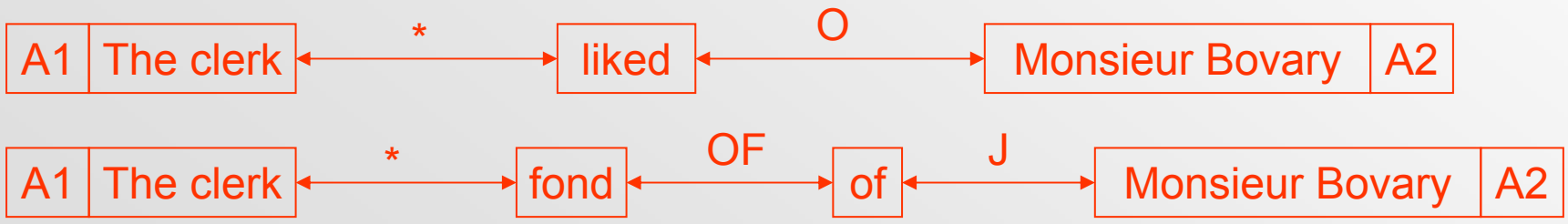
- What is Paraphrasing?
- Lin & Pantel.
  - Dependency paths with similar slot fillers have similar meanings.
- Barzilay & McKeown.
  - Identical words → contexts → more paraphrases → more contexts → …
- Barzilay & Lee.
  - Lattices with common slot fillers tend to correspond to paraphrases.
- **Pang et al.**
  - **Merge parse trees of aligned sentences to extract FSAs.**
  - **Different paths in an FSA correspond to paraphrases.**
- Ibrahim et al.
- Directions for future work.

Paraphrasing methods

Ibrahim et al.'s method (1 of 2)

- Same as Lin & Pantel's method
  - Dependency parse trees.
- Compares only paths from aligned sentences
- Find anchors among nouns and pronouns of the aligned sentences and score them using heuristics.

The clerk liked Monsieur Bovary / The clerk was fond of Monsieur Bovary



"X liked Y" ≈ "X was fond of Y"

Ibrahim et al.'s method (2 of 2)

- Low precision.
  - 40.2% average precision.
  - Up to 47.8% by increasing the threshold.
- Requires dependency parser, parallel corpus.
  - Difficult to obtain (e.g. in Greek).
- Requires aligner.
  - Easier to obtain.
- Reduces search space compared to Lin & Pantel's method.
  - Compares only paths from aligned sentences.
  - Unclear if it overcomes the other problems of Lin & Pantel's method (e.g. "fail" ≈ "succeed").

Contents of this talk

Paraphrasing methods

- What is Paraphrasing?
- Lin & Pantel.
  - Dependency paths with similar slot fillers have similar meanings.
- Barzilay & McKeown.
  - Identical words → contexts → more paraphrases → more contexts → …
- Barzilay & Lee.
  - Lattices with common slot fillers tend to correspond to paraphrases.
- Pang et al.
  - Merge parse trees of aligned sentences to extract FSAs.
  - Different paths in an FSA correspond to paraphrases.
- **Ibrahim et al.**
  - **Dependency paths with similar anchors tend to correspond to paraphrases.**
- Directions for future work.

References

- R. Barzilay and K. McKeown. Extracting paraphrases from a parallel corpus. In *Proceedings of the ACL/EACL,* 2001.
- R. Barzilay and L. Lee. Learning to Paraphrase: An Unsupervised Approach Using Multiple-Sequence Alignment. In *HLT-NAACL 2003, Main Proceedings,* 2003.
- A. Ibrahim, B. Katz, and J. Lin. Extracting structural paraphrases from aligned monolingual corpora. In *Proceedings of the Second International Workshop on Paraphrasing (IWP-2003),* 2003.
- D. Lin and P. Pantel. Discovery of Inference rules for Question Answering. *Natural Language Engineering,* 2001.
- B. Pang, K. Knight, and D. Marcu. Syntax-based Alignment of Multiple Translations: Extracting Paraphrases and Generating New Sentences. In *HLT-NAACL 2003, Main Proceedings,* 2003.