Learning to Paraphrase: An Unsupervised Approach Using Multiple-Sequence Alignment

– Regina Barzilay and Lillian Lee

Presented By: Mohammad Saif

Department of Computer Science, University of Toronto

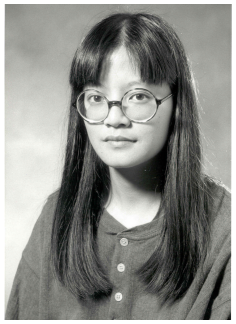
10 King's College Rd., Toronto, M5S 3G4, Canada

E-mail: smm@cs.toronto.edu

Authors

Regina Barzilay

- **Assistant Professor**, Dept. of Electrical Engineering and Computer Science, MIT
- Paraphrasing, Text Summarization, Sentence Alignment, Lexical Choice and lexical Chains
- `http://www.sls.csail.mit.edu/~regina`



Lillian Lee

- **Associate Professor**, Dept. of Computer Science, Cornell University
- Multiple Sequence Alignment, Segmentation, Information Retrieval, Distributional Clustering, and Distributional Similarity.
- `http://www.cs.cornell.edu/home/llee/default.html`

Learning to Paraphrase: An Unsupervised Approach Using Multiple-Sequence Alignment

His name was given as 20-year-old Mohsen Fouad Jaber,
from Khan Yunis in the southern Gaza Strip

He was identified as Mohsen Fouad Jaber, 20,
from Khan Yunis in the southern Gaza Strip

Different lexical realizations conveying (nearly) same information

- Mechanism to automatically generate paraphrases of a sentence
- HLT-NAACL 2003: Main Proceedings, pages 16–23, 2003

Press Articles

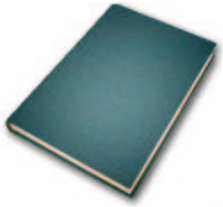
- "Software paraphrases sentences", Kimberly Patch, Technology Research News, December 3/10, 2003
- "Get Me Rewrite! Hold On, I'll Pass You to the Computer.", Anne Eisenberg, The New York Times, December 25, 2003
- ACM TECHNews article 5(588), December 29, 2003

Setting the Stage

- **Approach**: Unsupervised and corpus based
- **Source of Information**: Collection of articles from different news wire agencies about the same events
  - Meaning preserved
  - Use different words to convey meaning
  - Domain dependent paraphrases
- **Relaxing the requirement**
  - Simple sentence alignment not possible
  - Finding alignment an important issue

Comparable Corpora vs. Parallel Translations

Barzilay and McKeown



Non-English Source Text

Not Used

Different English Translations

Used

Barzilay and Lee

Event

Can not Use!

Comparable Corpora

Used

Multiple-Sequence Alignment

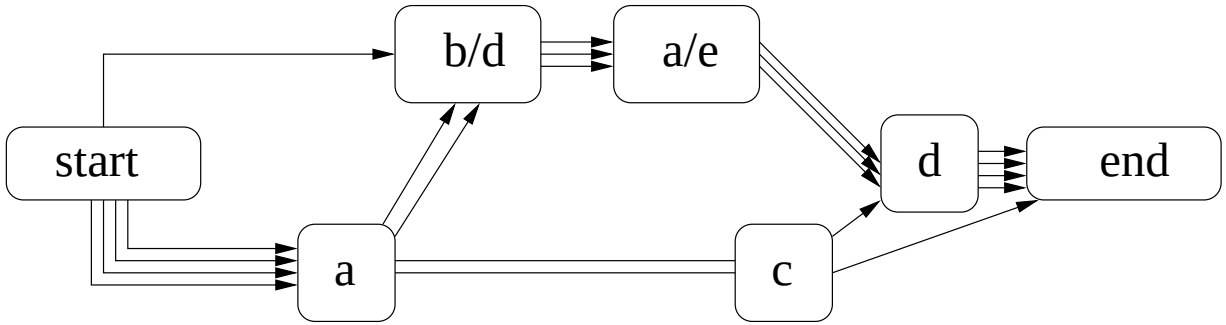
- Input: n strings/sequences, Output: n-row correspondence table
  - rows correspond to sequences
  - columns indicate the elements corresponding to that point

- MSA generated using iterative pairwise alignment
  - polynomial time approximation procedure

- A lattice may be generated from the MSA

a b a _ d
a _ _ c d
a _ _ c _
_ b a _ d
a d e _ d

MSA



LATTICE

Algorithm

- Start with two comparable corpora

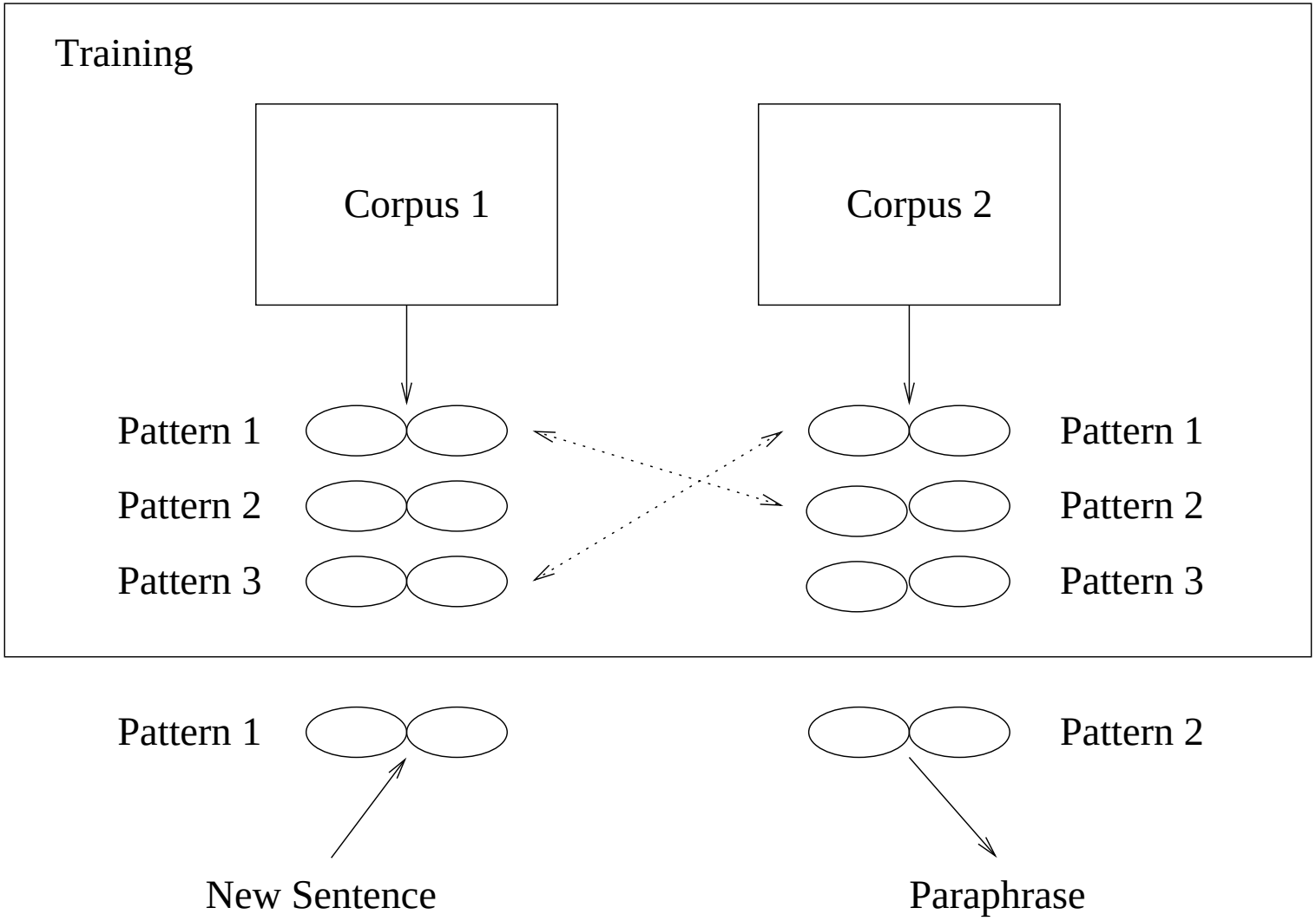
- Identify patterns in each dataset independently
  - Sample pattern:

    X (injured/wounded) Y people, Z seriously ...[1]

- Identify pairs of patterns across the two data sets that represent paraphrases
  - A pattern which may be paired with [1]:

    Y were (wounded/hurt) by X, among them Z were in serious condition ...[2]

System Architecture



Sentence Clustering

- First step in identifying patterns
- Hierarchical complete-link clustering of sentences
  - Similarity metric: word n-gram overlap (n=1,2,3,4)
  - Mismatches on details undesirable
    * Proper nouns, dates and numbers replaced by generic tokens

Sample Sentences from a Cluster

- **A Palestinian suicide bomber blew himself up in** a southern city Wednesday, **killing** two other **people and wounding** 27

- **A suicide bomber blew himself up in** the settlement of Efrat, on Sunday, **killing** himself **and injuring** seven people

- **A suicide bomber blew himself up in** the coastal resort of Netanya on Monday, **killing** three other **people and wounding** dozens more

- **A Palestinian suicide bomber blew himself up in** a garden cafe on Saturday, **killing** ten **people and wounding** 54

Lattices and Patterns

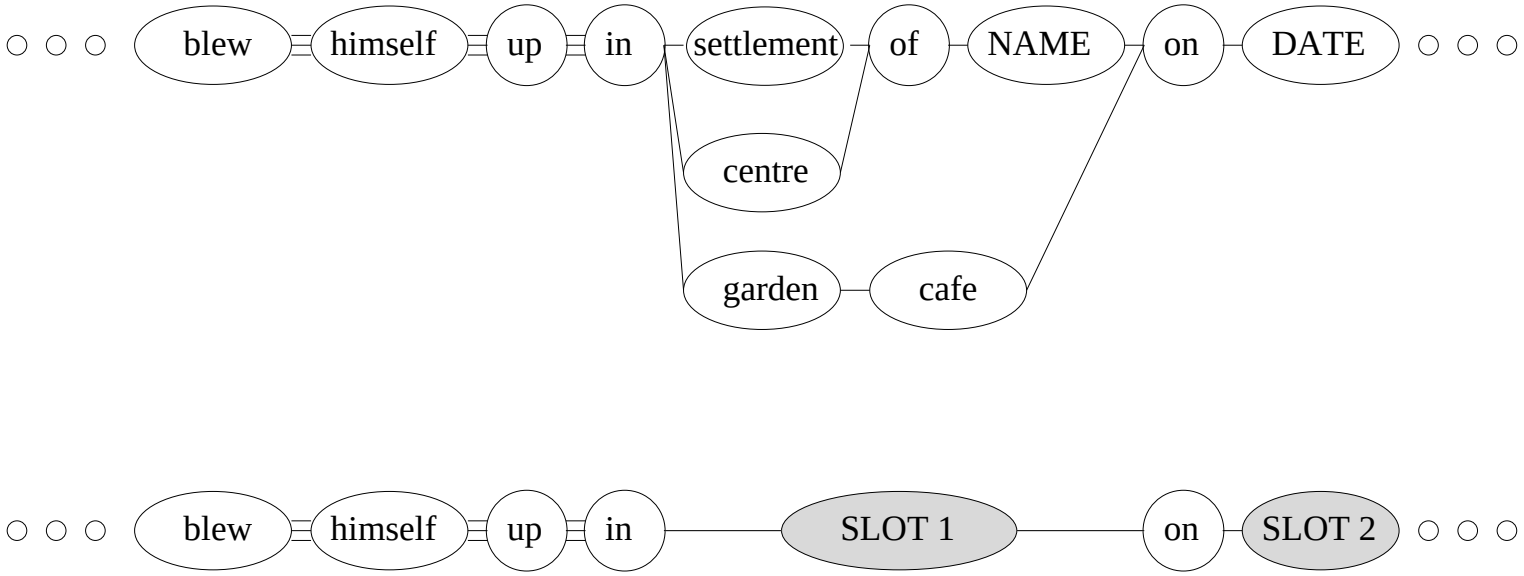
- Lattices learned using Multiple Sequence Alignment
  - Number of edges between nodes corresponds to number of sentences following that path

- Identify Backbone Nodes
  - Nodes shared by more than 50% of the cluster's sentences
  - Replace generic token backbone nodes by slot nodes

- Identify regions of variability
  - Distinguish between
    - **Argument variability**: replace by slots
    - **Synonym variability**: to be preserved

- Condense adjacent slot nodes into one

Lattice and Slotted Lattice



Synonym and Argument Variability

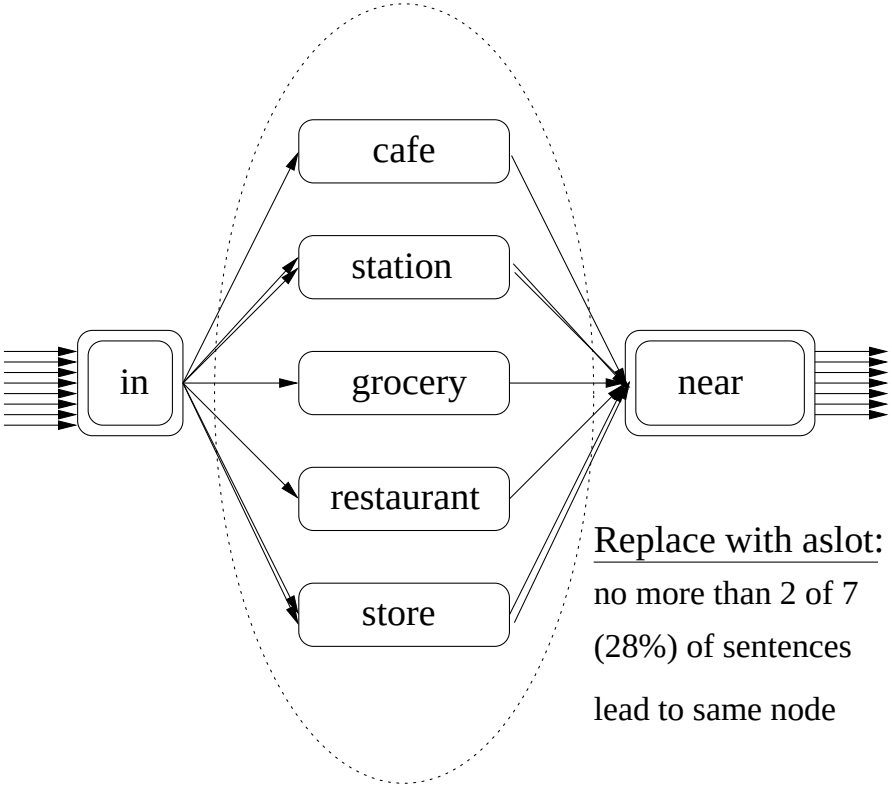
- Arguments cause of more variability than synonyms
- Analyze **split level** of backbone nodes
- Compare with **synonym threshold** s (30)

If s% or less edges go from the backbone node
to all of its follower nodes, insert slot

Else, keep all nodes that are reached by at least s%
of edges going between the two neighboring backbone nodes

Example Argument and Synonym Variability

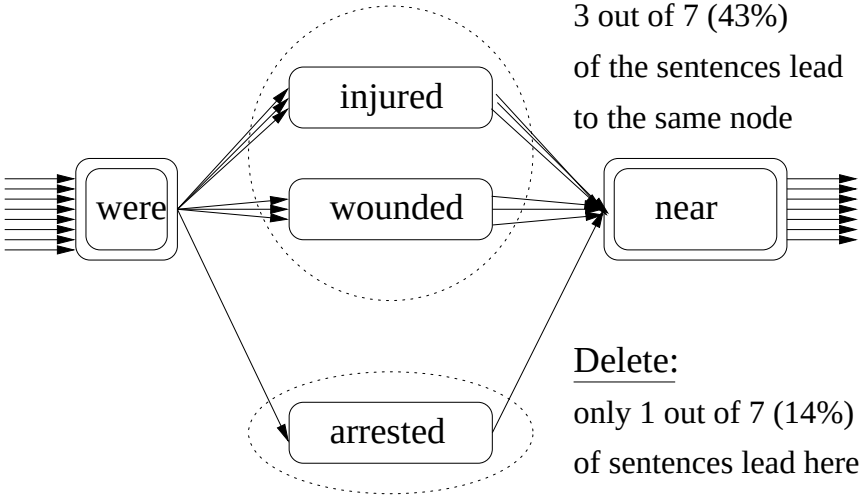
ARGUMENT VARIABILTY



Replace with aslot:

no more than 2 of 7 (28%) of sentences lead to same node

SYNONYM VARIABILTY



Preserve both nodes:

3 out of 7 (43%) of the sentences lead to the same node

Delete:

only 1 out of 7 (14%) of sentences lead here

Lattice Matches

- Parallel corpora
  - Sentence alignment

- Comparable corpora
  - Paraphrases will take same argument values

slot1 bombed slot2

the Israeli fighters bombed Gaza strip

slot3 was bombed by slot4

Gaza strip was bombed by the Israeli fighters

Candidate lattices X and Y

- Retrieve sentences **XX** and **YY** corresponding to **X** and **Y** from the two corpora
- **XX** and **YY** must be from articles written on same day and on same topic
- Lattices paired if degree of "**match**" above threshold
  - count word overlap
  - double the weight for proper names and numbers
  - auxiliaries discarded
  - word order ignored

Generating Paraphrases

- Input: sentence to be paraphrased, say **X**
- Check if exists lattice **XX** that may represent **X**
  (with some error margin)
  - Employ multiple sequence alignment
  - Allow insertion of nodes in lattice with a penalty (-0.1)
  - All other node alignments receive a score of 1
- If **XX** exists, retrieve lattice **YY**, its pair in the other corpus
- Substitute appropriate arguments from **X** into the slots of **YY**

Statistics

- Articles produced between September, 2000 and August 2002 by the Agence France-Presse (AFP) and Reuters news agencies
  - 9MB of articles pertaining to Individual acts of violence in Israel and raids on Palestinian territories
  - 120 articles held out for parameter-training set
- 43 slotted lattices from AFP and 32 from Reuters data
- 25 pairs of matching cross-corpus lattices
- 6,534 template pairs (thanks to multiple paths per lattice)

Template Evaluation

- Judged by native speakers unfamiliar with system
  - Templates are paraphrases if in general one may be substituted for the other (not necessarily vice-versa)

- Lin and Pantel, 2001 and Shinyama et al., 2002 closest work on paraphrasing at sentence level
  - DIRT's templates are much shorter and was implemented on larger corpus
  - 6,534 highest scoring templates selected

- 500 of the two sets of templates selected randomly

- Barzilay and Lee system outperformed DIRT by around **38%** points, as rated by 4 judges

Paraphrase Evaluation

- Baseline System: replace words with synonyms from WordNet
  - Randomly selected from synset obtained by choosing most frequent sense of source word
  - Number of substitutions proportional to that done by Barzilay and Lee system

- 20 articles on violence in Middle East from AFP
  - 59 (**12.2%**) sentences paraphrased out of 484
  - After proper name substitution only 7 of the 59 were found in training set

- Two judges found close to **80%** of the paraphrases accurate

Conclusion

- Barzilay and Lee, 2003 give a mechanism for generating sentence level paraphrases

- Unlike some of the previous work which used parallel translations, comparable corpora is used
  - More abundantly available and in many domains

- **80%** of the paraphrases have been shown to be accurate
  - Given a piece of text, around **12.2%** of the sentences may be expected to be paraphrased

Still some way for automatic rewriting of text but Barzilay and Lee provide a promising start!