SENTENCE SIMPLIFICATION USING SIMPLE WIKIPEDIA

Ameeta Agrawal
Nikolay Yakovets

01 Dec 2011

THE PROBLEM

In complex sentences, facts can be presented with varied and complex linguistic constructions.

…Prime Minister Vladimir V. Putin, the country's paramount leader, cut short a trip to Siberia, returning to Moscow to oversee the federal response. Mr. Putin built his reputation in part on his success at suppressing terrorism, so the attacks could be considered a challenge to his stature….

THE PROBLEM

In complex sentences, facts can be presented with varied and complex linguistic constructions.

main clause

…Prime Minister Vladimir V. Putin, the country's paramount leader, cut short a trip to Siberia, returning to Moscow to oversee the federal response. Mr. Putin built his reputation in part on his success at suppressing terrorism, so the attacks could be considered a challenge to his stature….

THE PROBLEM

In complex sentences, facts can be presented with varied and complex linguistic constructions.

main clause

appositive

…Prime Minister Vladimir V. Putin, the country's paramount leader, cut short a trip to Siberia, returning to Moscow to oversee the federal response. Mr. Putin built his reputation in part on his success at suppressing terrorism, so the attacks could be considered a challenge to his stature….

THE PROBLEM

In complex sentences, facts can be presented with varied and complex linguistic constructions.

main clause

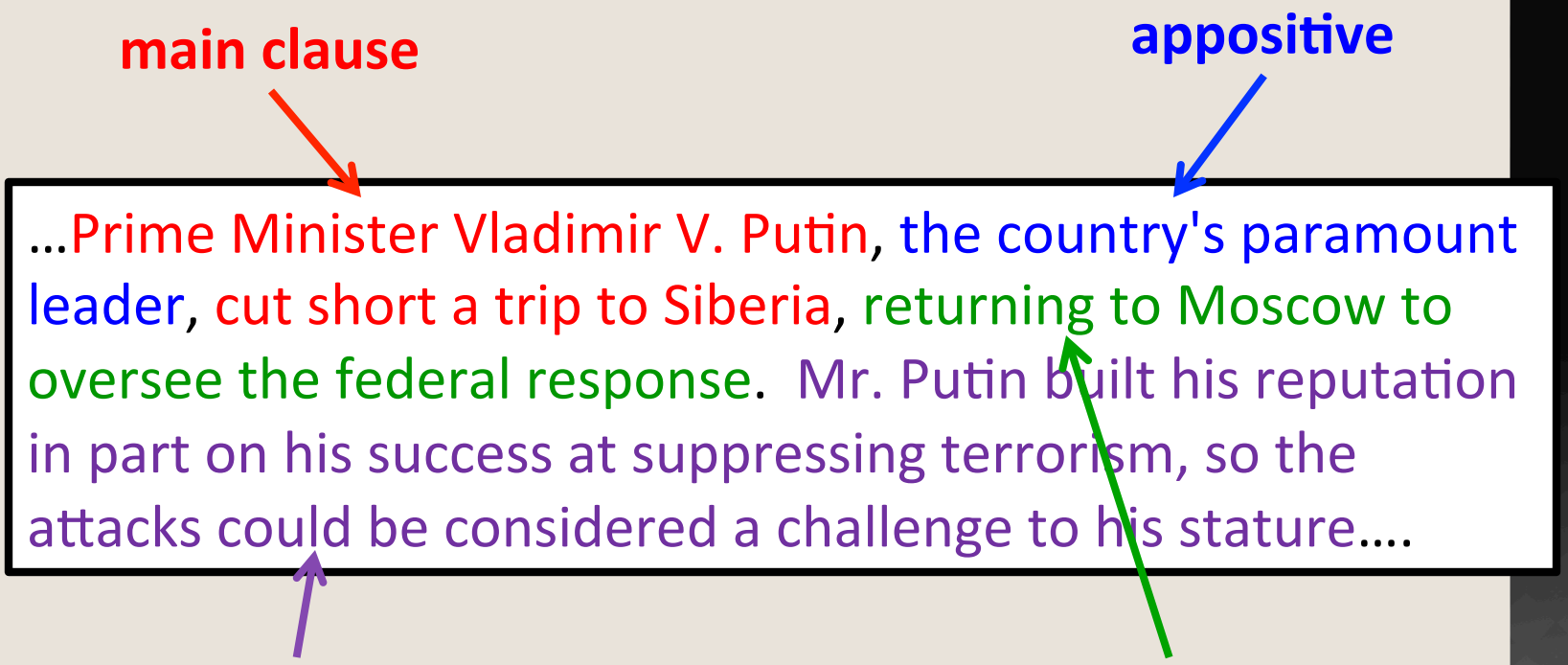
appositive

…Prime Minister Vladimir V. Putin, the country's paramount leader, cut short a trip to Siberia, returning to Moscow to oversee the federal response. Mr. Putin built his reputation in part on his success at suppressing terrorism, so the attacks could be considered a challenge to his stature….

participial phrase

THE PROBLEM

In complex sentences, facts can be presented with varied and complex linguistic constructions.



THE PROBLEM

In complex sentences, facts can be presented with varied and complex linguistic constructions.

Output:

- Prime Minister Vladimir V. Putin cut short a trip to Siberia.
- He is the country's top leader.
- He returned to Moscow to oversee the federal response.
- Mr. Putin built his reputation in part on his success at suppressing terrorism.

SENTENCE SIMPLIFICATION

- Source:
  - **Complex** sentence
- Target:
  - Set of **simple** declarative sentences
  - Easier to read
  - Simpler vocabulary and syntactic structure

WHY?

- Development of reading aids for:
  - People with aphasia (Carroll et al., 1999)
  - Non-native speakers (Siddharthan, 2003)
  - Individuals with low literacy (Watanabe et al., 2009)
- Improve performance of:
  - Parsers (Chandrasekar et al., 1996)
  - Summarizers (Klebanov et al., 2004)
  - Semantic role labelers (Vickrey and Koller, 2008)

INTUITION

- Baseline: substitution of difficult words with more common words
  - Input: Prime Minister Putin, the country's paramount leader…
  - Output: Prime Minister Putin, the country's top leader…
- Sentence structural rules
  - e.g. Rewrite operations: deletion, substitution, insertion, reordering, etc.
  - Input: Prime Minister Putin, the country's paramount leader, cut short a trip to Siberia.
  - Output: Prime Minister Putin is the country's top leader. He cut short a trip to Siberia.
- Challenge: learning the rules
- How do you do this automatically?

RELATED WORK

- Woodsend, Lapata, 2011
  - Wikipedia to induce quasi-synch grammar
  - align phrases and learn syntactic & lexical simplification rules
  - Integer Linear Programming to put the phrases together
- Coster, Kauchak, 2011
  - generate a parallel corpus using Wikipedia and Simple Wikipedia
  - align phrases and dynamic programming to find the best global sentence alignment
- Biran, Brody, Elhadad, 2011
  - use Wikipedia revision histories to learn lexical simplification rules

RELATED WORK

- Woodsend, Lapata, 2011
  - **Wikipedia and Simple Wikipedia** to induce quasi-synch grammar
  - **align phrases** and learn syntactic & lexical simplification rules
  - Integer Linear Programming to put the phrases together
- Coster, Kauchak, 2011
  - generate a parallel corpus using **Wikipedia and Simple Wikipedia**
  - **align phrases** to find the best global sentence alignment
- Biran, Brody, Elhadad, 2011
  - use **Wikipedia and Simple Wikipedia** revision histories to learn lexical simplification rules

ROOM FOR IMPROVEMENT

- Alignment
  - Most alignment previously done at **global** level
  - Useful when input and output sequences are similar and of roughly equal size
  - But consider the example sentences:

…Prime Minister Vladimir V. Putin, the country's paramount leader, cut short a trip to Siberia, returning to Moscow to oversee the federal response. Mr. Putin built his reputation in part on his success at suppressing terrorism, so the attacks could be considered a challenge to his stature…	Prime Minister Vladimir V. Putin cut short a trip to Siberia. He is the country's top leader. He returned to Moscow to oversee the federal response. Mr. Putin built his reputation in part on his success at suppressing terrorism.

  - Sequences not similar, not equal size!

OUR PROPOSED METHODOLOGY

- Alignment
  - Propose to test **local** alignment
  - More useful for dissimilar sequences that are suspected to contain regions of similarity
- Other **sequence alignment** algorithms used in Bioinformatics
- Perhaps **conceptual graphs** as they provide an intuitive and easily understandable means to represent knowledge (??)

EVALUATION CRITERIA

- Two ways:
  - Human judges
  - Automatic evaluation (readability measures)
    - For Wikipedia, Simple Wikipedia and our output:
      - Flesch-Kincaid Grade Level index
      - BLEU: scores the target output by counting *n*-gram matches with the reference
      - TERp: similar to word error rate, allows shifts
    - Closest to Simple Wikipedia = good!

THANKS!
COMMENTS?