Approximate subgraph matching for detection of topic variations

Mitja Trampuš
Dunja Mladenić
AI Lab, Jožef Stefan Institute, Slovenia
Mining Diversity

• Web content varies in many aspects, e.g.
  o Topical
  o Social (author, target audience, people written about)
  o Geographical (publisher, places written about)
  o Sentiment (positive/negative)
  o Writing style (structure, vocabulary)
  o Coverage bias

• This work: (micro-) **topical diversity**
  o Macroscopic = largely solved
  o Microscopic = challenge
Task:

Given a collection of texts on a topic,

- identify a common template
- align texts to the template

Suicide attack kills 1 police in Afghanistan

Published December 23, 2010 | Associated Press

KABUL, AFGHANISTAN -- A suicide bomber struck on Thursday, killing one officer, officials said.

The attacker detonated his explosives as he approached Kunduz, said Abdul Rahman Sayedkhili, the provincial police chief. Two other police officers were also wounded in the blast.

The bombing is the second suicide attack in the agricultural and marketing center that controls Tajikistan. On Dec. 19, insurgents stormed a compound and killed two Afghan army soldiers and five police in a daylight assault.

The Taliban have increased their attacks in their drive to take control of major cities and towns. They have also stepped up attacks in Pakistan, where 103 officers were killed in a suicide blast last month.

Suicide bomber kills 2 Shiite pilgrims in Iraq

The Associated Press
Monday, December 13, 2010; 8:53 AM

BAGHDAD -- An Iraqi official says a suicide bomber has killed two Shiite pilgrims north of Baghdad during an important religious ritual for the Muslim sect.

Mohammed Maarouf, the mayor of Balad Ruz town in the northern Diyala province, says the bomber detonated an explosives belt during a body check and killed the policeman searching him and a woman standing nearby. Thirteen people were injured, Maarouf said.

Diyala police spokesman, Maj. Ghalib al-Karkhi, confirmed the death toll.

The ritual, known as Ashoura, marks the anniversary of the 7th century death of Imam Hussein. His death in a battle sealed Islam's historic Sunni-Shiite split.

The ancient divide has provided the backdrop for Iraq's sectarian bloodshed after the 2003 U.S.-led war.
Template representation

• Syntactic
  o **info1**: X people were killed / killed X people / resulted in X casualties
  o **info2**: blew up Y / destroyed Y / attacked in a Y

• Semantic
  o `kill(bomber, people); count(people, X)`
  o `destroy(bomber, Y)`
Prerequisite: Semantic Graph

1. **Believers**
   - Count
   - Suicide bomber
   - Kill
   - Destroy
   - Church
   - Suicide bomber
   - Wear
   - Vest

2. **Patients**
   - Count
   - Receive treatment
   - Kill
   - Demolish
   - Hospital

3. **Police officer**
   - Count
   - Slaughter
   - Bomber
   - Blow up
   - Police station
Mining Templates

• Template := subgraph with frequent specializations
  o Specializations implied by background taxonomy (WordNet)
  o Threshold frequency manually defined
Semantic Graph Construction

1. Data: Google News crawl
2. HTML cleanup
3. Named entity tagging
4. Pronoun resolution (he/she/him/her)
5. Named entity consolidation (Barack Obama vs President Obama)
6. Parsing, triple/fact/assertion extraction (for now: subj-verb-obj only)
7. Ontology/taxonomy alignment
8. Merging triples into a graph
Approximate subgraph matching

FREQUENT SUBTREE MINING

GENERALIZE
Approximate subgraph matching
Preliminary results

• 5 test domains; for each:
  o ~10 graphs, ~10000 nodes
  o 10-60 seconds

• At min. support 30%
  o 20 maximal patterns, 9 manually judged as interesting
Bombing attacks (8 patterns in total)
weekday \(\leftarrow\) kill \(-\) person \(-\) kill \(\rightarrow\) attack \(-\) take \(\rightarrow\) place
himself \(\leftarrow\) blow \(-\) suicide bomber \(-\) explode \(\rightarrow\) device
himself \(\leftarrow\) blow \(-\) suicide bomber \(-\) blow \(\rightarrow\) building

Court sentencings (7 patterns in total)
correctional institution \(\leftarrow\) be \(-\) person \(-\) face \(\rightarrow\) year \(\leftarrow\) be \(-\) sentence
innocent \(\leftarrow\) be \(-\) person \(-\) face \(\rightarrow\) year \(\leftarrow\) be \(-\) sentence

Award ceremonies (2 patterns in total)
period of time \(\leftarrow\) have \(-\) person \(-\) be \(\rightarrow\) feeling

Political visits (3 patterns in total)
summit \(\leftarrow\) attend \(-\) he \(-\) hold \(\rightarrow\) talk
\[\mid\mid\`\leftarrow\) be \(-\) leader
\[\mid\`\leftarrow\) tell \(-\) communicator
\[\`\leftarrow\) express \(\rightarrow\) feeling
need \(\leftarrow\) stress \(-\) he \(-\) hold \(\rightarrow\) talk
\[\`\leftarrow\) attend \(\rightarrow\) summit
\[\`\leftarrow\) be \(-\) leader
leader \(\leftarrow\) meet \(-\) he \(-\) travel \(\rightarrow\) France

Worker layoffs (0 patterns in total)
Conclusion

• Future work:
  o Mapping text -> semantics
    • Other ontologies?
  o Interestingness measure for assertions and patterns
  o Evaluation (precision, recall; multiple domains)
  o Alternative approaches to generalizing subgraphs

• Template extraction is achievable, but not easy
• Human filtering of results hard to avoid
• Current approach reasonably fast
Can we extract *all* relations?

Kind of ...

Thousands of small quakes resumed 18 months ago and continue to rattle Mammoth Lakes, June Lake and other Mono County resort towns. The temblors, most measuring 1 to 3 on the Richter scale, started beneath Mammoth Mountain.

**Subject – Verb – Object**

**Triplets**

**Semantic Graph**
Templates - why

• Interpret content
  o news archives: structure/annotate old texts, enable semantic search
  o wikipedia: suggestions for infobox entries

• Generate content
  o wikipedia: a starting point for new articles / a checklist of information to be included

• No normative definition of “good template”
Evaluation

• Qualitative
  o Usage-specific
  o Not useful for tuning algorithms

• Quantitative
  o Precision
  o Recall