



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.
 - Topical
 - Social (author, target audience, people written about)
 - Geographical (publisher, places written about)
 - Sentiment (positive/negative)
 - Writing style (structure, vocabulary)
 - Coverage bias
- This work: (micro-)**topical diversity**
 - Macroscopic = largely solved
 - 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

 Print  Email  Share  Comments (0)

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

The attacker detonated his explosives as he entered a market in Kunduz, said Abdul Rahman Sayedkhaili, the provincial governor. Two others were also wounded in the blast.

The bombing is the second suicide attack in Afghanistan in an agricultural and marketing center that controls a major trade route through Tajikistan. On Dec. 19, insurgents stormed a market in Afghanistan, killing Afghan army soldiers and five police in a day.

The [Taliban](#) have increased their attacks in Afghanistan, targeting a route through [Pakistan](#) to ones running through Afghanistan.

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.

TOOLBOX

 Resize

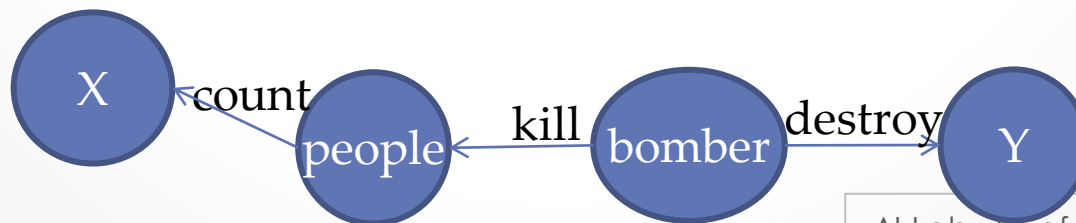
 Print

 E-mail

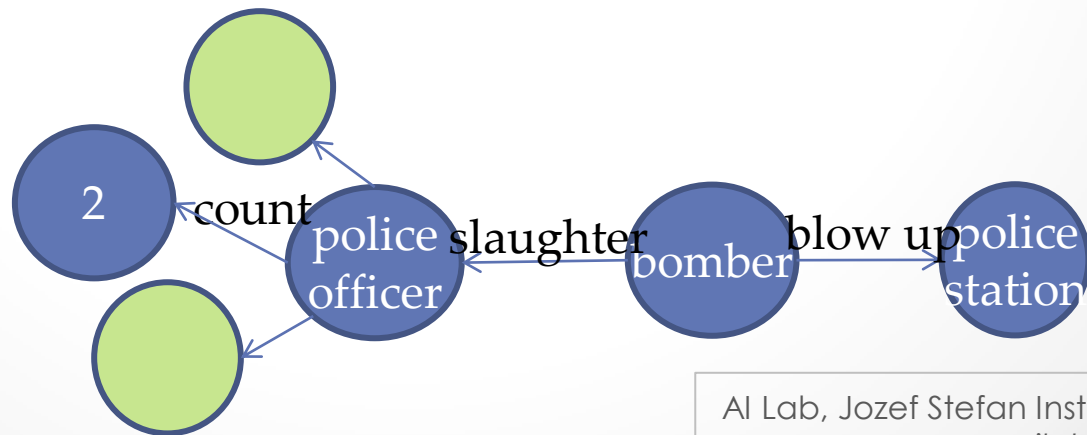
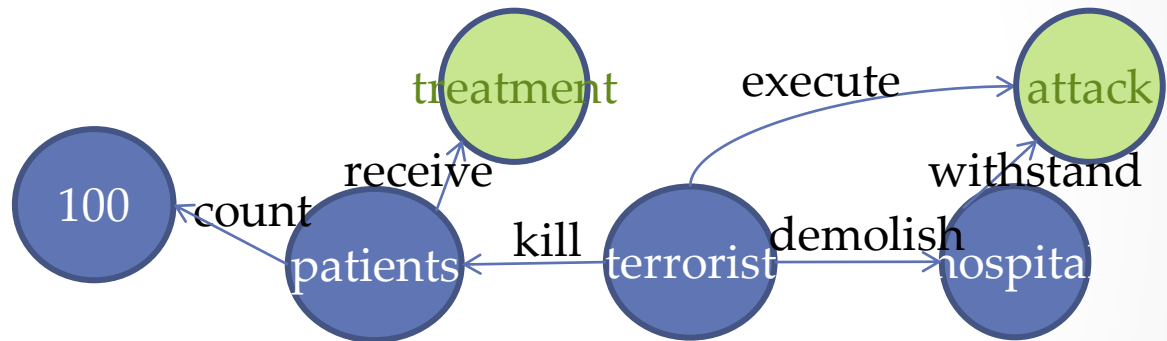
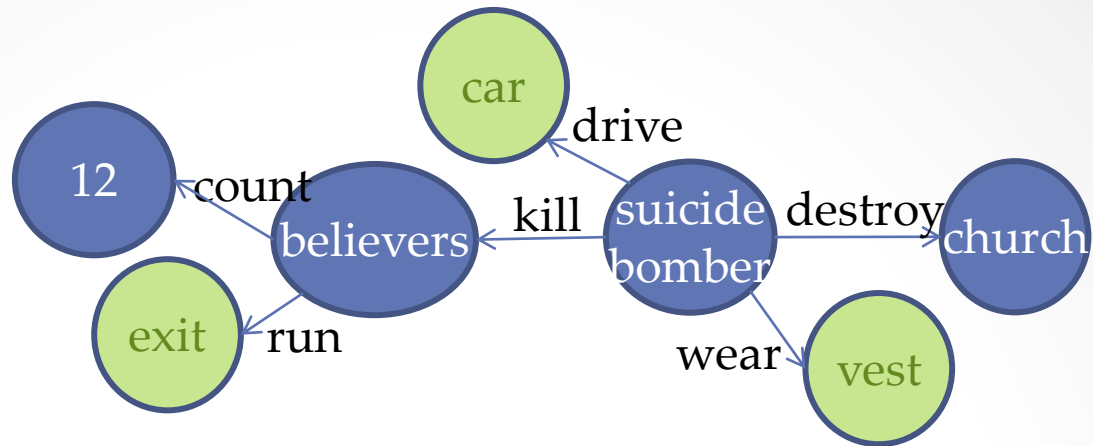
 Reprints

Template representation

- Syntactic
 - **info1:** X people were killed / killed X people / resulted in X casualties
 - **info2:** blew up Y / destroyed Y / attacked in a Y
- Semantic
 - kill(bomber, people); count(people, X)
 - destroy(bomber, Y)

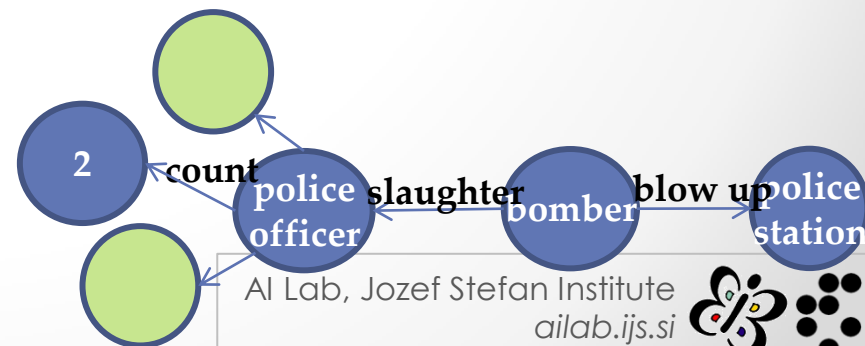
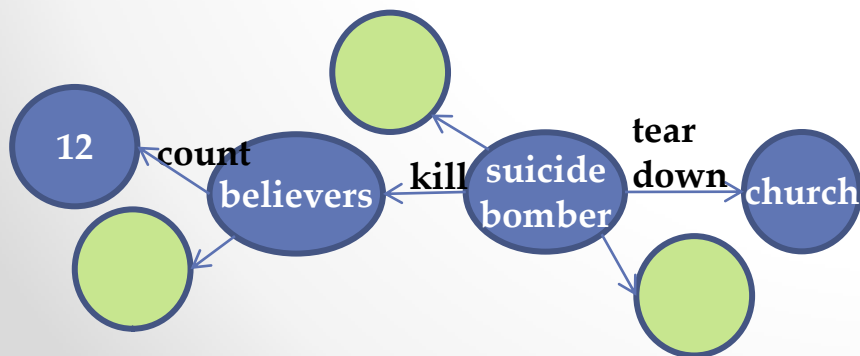
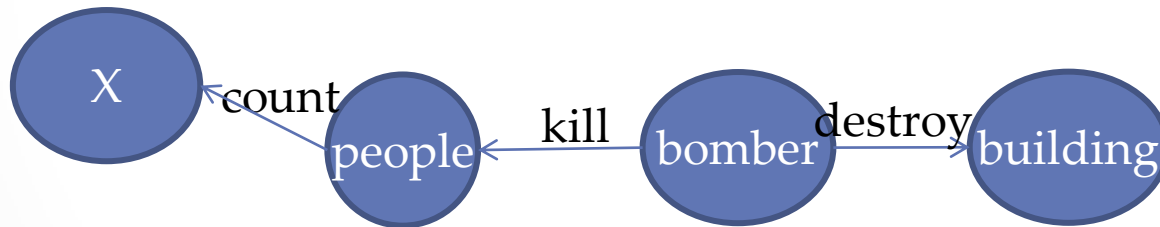


Prerequisite: Semantic Graph



Mining Templates

- Template := subgraph with frequent specializations
 - Specializations implied by background taxonomy (WordNet)
 - Threshold frequency manually defined

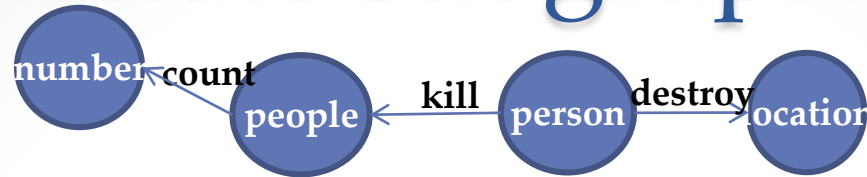


Semantic Graph Construction

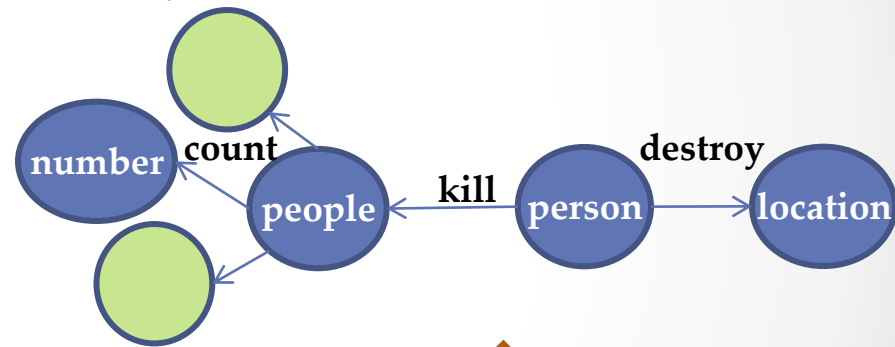
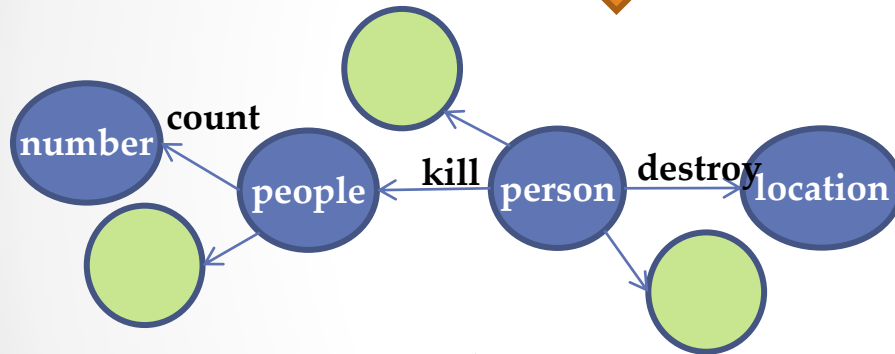
1. Data: Google News crawl
2. HTML cleanup
3. Named entity tagging
4. Pronoun resolution (he/she/him/her)
5. Named entity consolidation (Barack Obama vs President Obama)
6. Parsing, triple/fact/assertion extraction (for now: *subj-verb-obj* only)
7. Ontology/taxonomy alignment
8. Merging triples into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



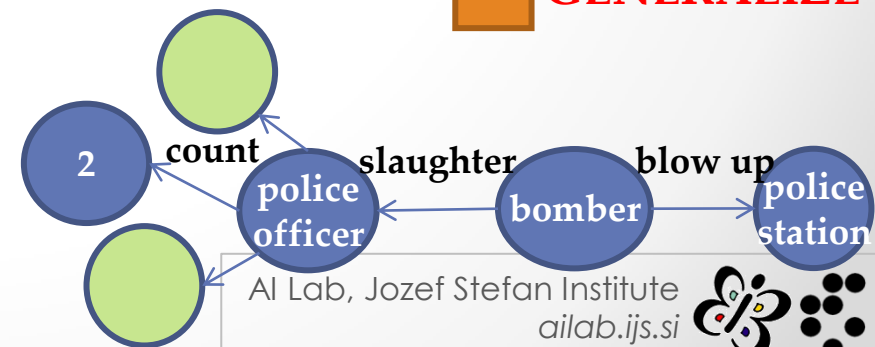
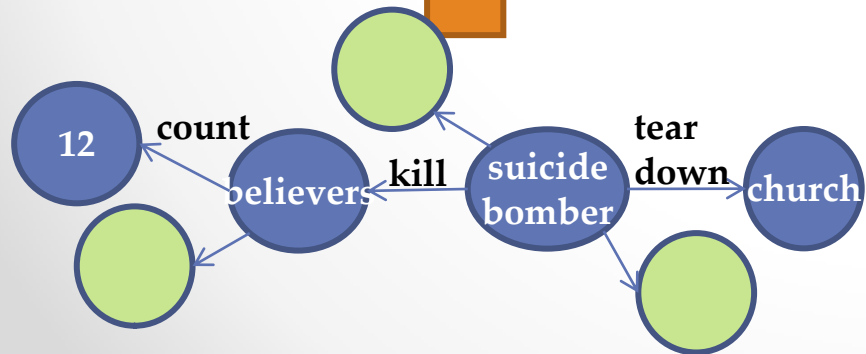
Approximate subgraph matching



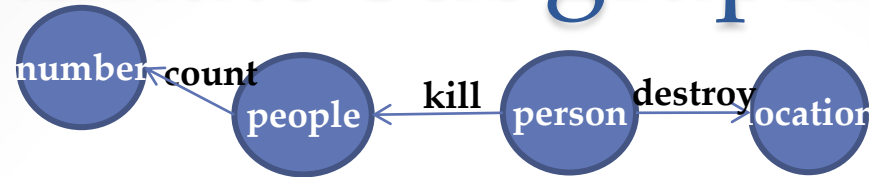
**FREQUENT
SUBTREE MINING**



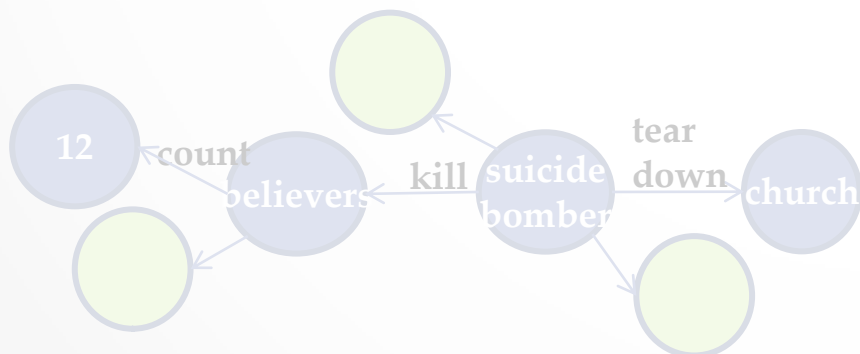
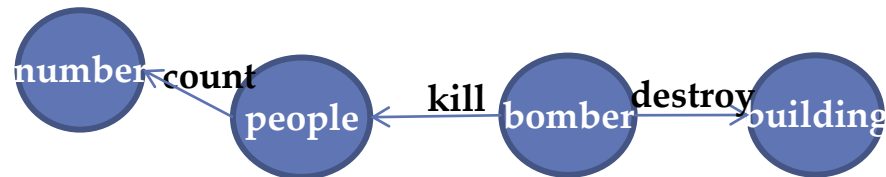
GENERALIZE



Approximate subgraph matching



SPECIALIZE



Preliminary results

- 5 test domains; for each:
 - ~10 graphs, ~10000 nodes
 - 10-60 seconds
- At min. support 30%
 - 20 maximal patterns, 9 manually judged as interesting



Bombing attacks (8 patterns in total)

weekday ←kill- person -kill→ attack -take→ place

himself ←blow- suicide bomber -explode→ device

himself ←blow- suicide bomber -blow→ building

Court sentencings (7 patterns in total)

correctional institution ←be- person -face→ year ←be- sentence

innocent ←be- person -face→ year ←be- sentence

Award ceremonies (2 patterns in total)

period of time ←have- person -be→ feeling

Political visits (3 patterns in total)

summit ←attend- he -- hold→ talk

||`-be→ leader

|`--tell→ communicator

`---express → feeling

need ←stress - he - hold→ talk

|`-attend → summit

`--be→ leader

leader ←meet- he -travel→ France

Worker layoffs (0 patterns in total)

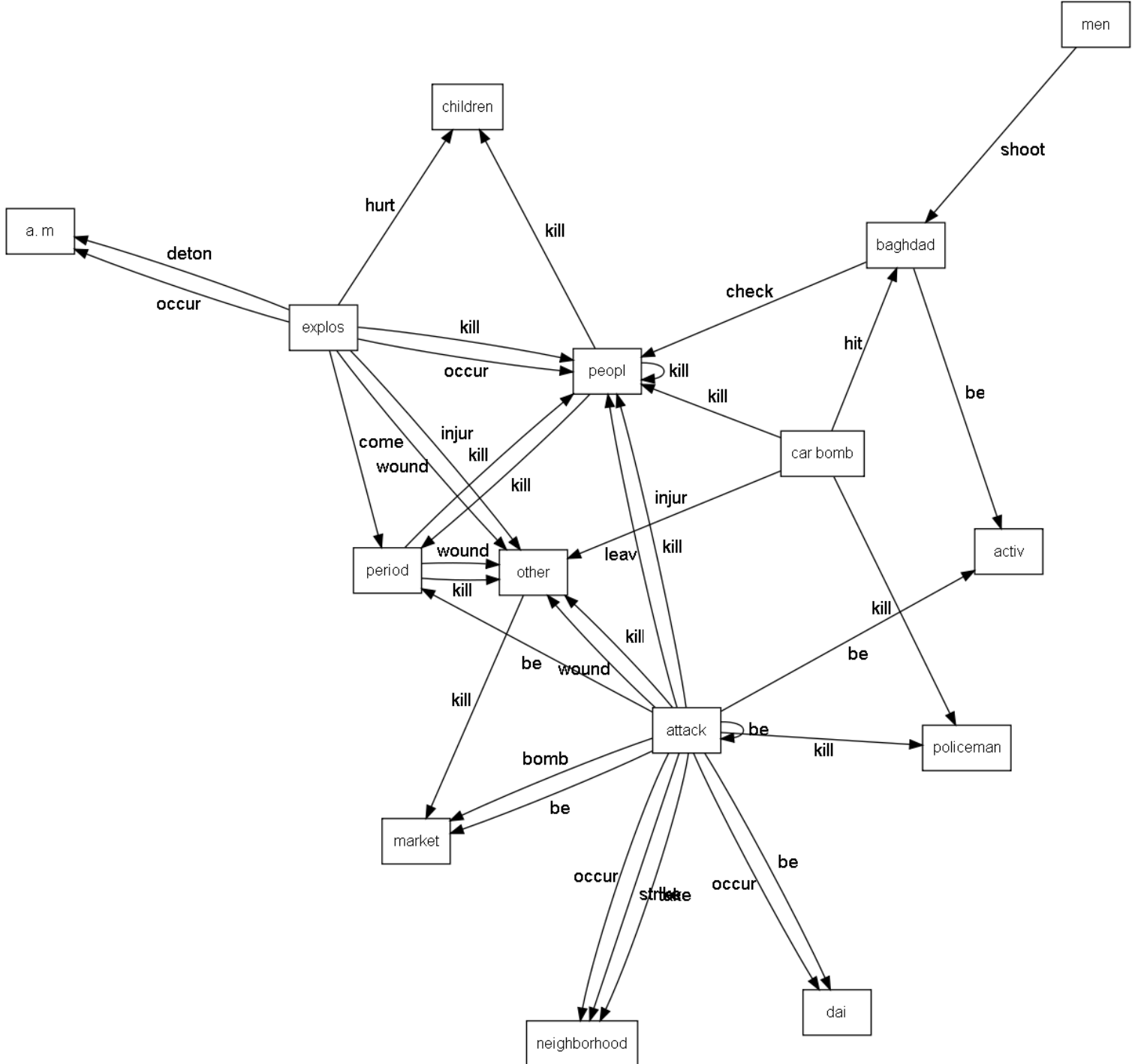
Conclusion

- Future work:
 - Mapping text -> semantics
 - Other ontologies?
 - Interestingness measure for assertions and patterns
 - Evaluation (precision, recall; multiple domains)
 - Alternative approaches to generalizing subgraphs
- Template extraction is achievable, but not easy
- Human filtering of results hard to avoid
- Current approach reasonably fast



Q?





Templates - why

- Interpret content
 - news archives: structure/annotate old texts, enable semantic search
 - wikipedia: suggestions for infobox entries
- Generate content
 - wikipedia: a starting point for new articles / a checklist of information to be included
- No normative definition of “good template”



Evaluation

- Qualitative
 - Usage-specific
 - Not useful for tuning algorithms
- Quantitative
 - Precision
 - Recall

