A Study in Domain-Independent Information Extraction for Disaster Management

DIMPLE Workshop

In conjunction with LREC 2014, Reykjavik, Iceland

Lars Döhling, Jirka Lewandowski, Ulf Leser

Humboldt-Universität zu Berlin, Germany

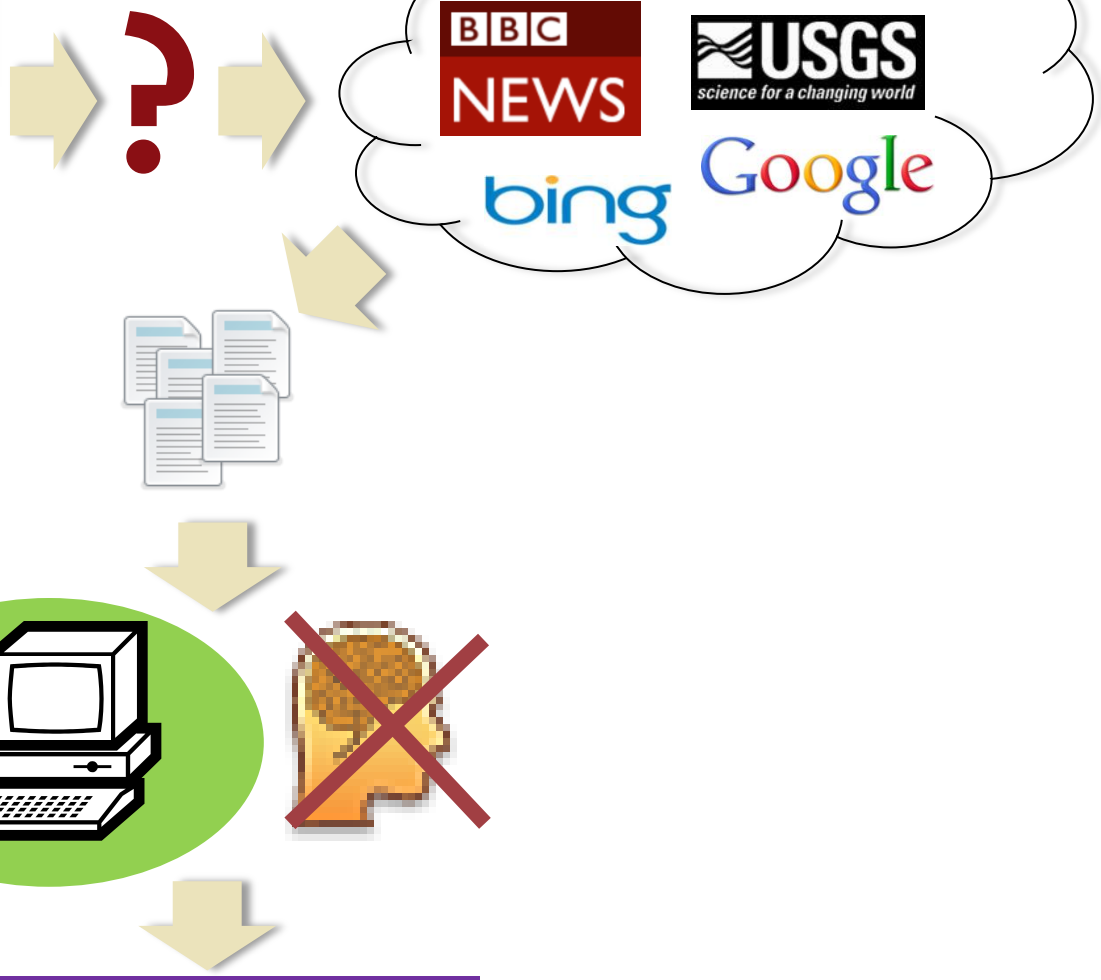
2014-05-31

Motivation


Earthquake


Disease


Drought


?
twitter
BBC NEWS
USGS
science for a changing world
bing
Google
Facts

Information Extraction

Annotated earthquake reports

Training

Extraction model for earthquake reports

Annotated disease reports

Training

Extraction model for disease reports

No annotated data avalable?

Use an **existing** (and **robust**) extraction model

Model robustness?

Extracted Facts

Casualties reported after natural disasters

- Injured, killed, missing, buried, homeless, affected

> The death toll in an earthquake in south-west China is now at least 32, with 467 injuries, state media says.

≥32 people killed

467 people injured

Approach

- N-ary relationship extraction

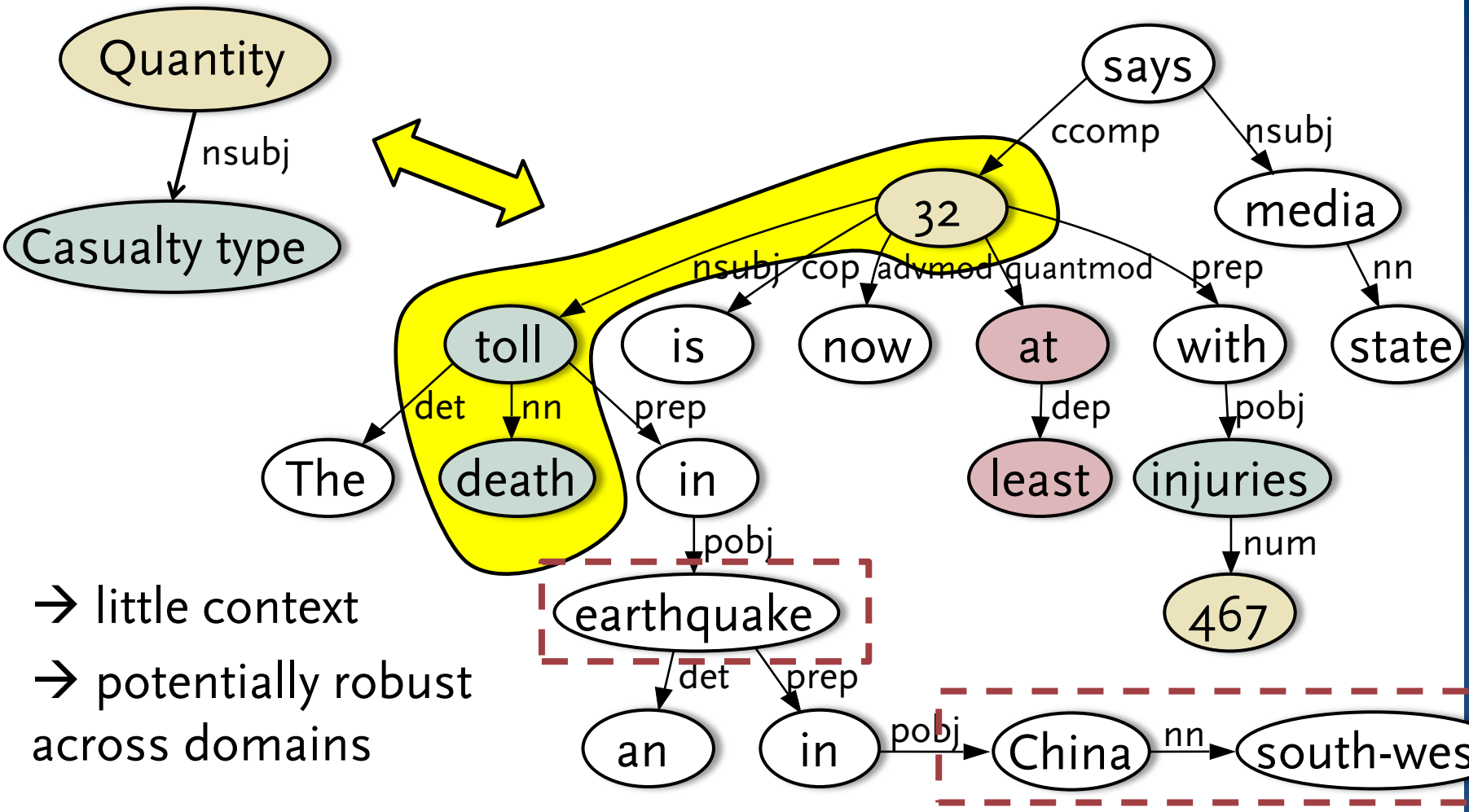
Extraction Process

1. Recognize Entities, e.g. “at least”, “24”, “injured”
   - Regular expression for quantities
   - Dictionary for all others

   → little context → potentially robust across domains

2. Find related entities
   - Pattern matching in dependency graphs
   - Patterns = shortest paths between entities

Shortest Path Patterns

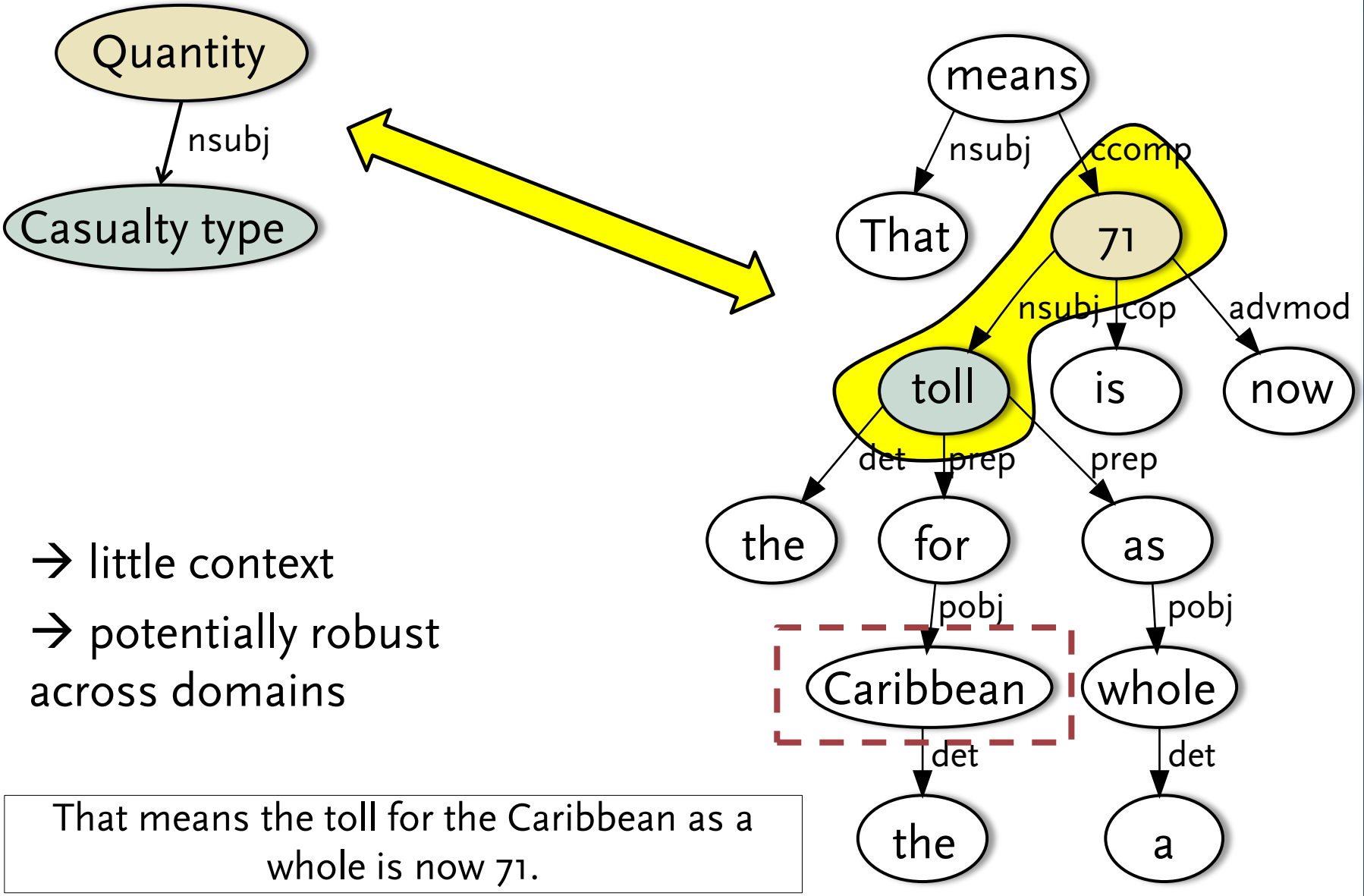


→ little context

→ potentially robust across domains

The death toll in an earthquake in south-west China is now at least 32, with 467 injuries, state media says.

Shortest Path Patterns (cont.)



→ little context

→ potentially robust across domains

That means the toll for the Caribbean as a whole is now 71.

Evaluation

Compare extraction robustness across two domains

- Earthquake
- Flood

Corpora

- English news articles
  - Earthquake: 245 documents, 1277 relationship instances
  - Flood: 412 documents, 1860 relationship instances
- Manually annotated
  - IAA ~82%
- 2/3 training set, 1/3 evaluation set
- Available on request

Intra-/Inter-Domain Results

	Target					
	Earthquake			Flood		
Source	P	R	F1	P	R	F1
Earthquake	.803	.735	**.768**	**–.024**	**–.127**	**–.076**
Flood	**–.051**	**–.074**	**–.064**	.765	.811	**.787**

P/R/F1 measures for the extracted relationship

- 9% F1 drop (avg) when applied across domains
  - Similar sentence structures and wordings
  - Large overlap in dictionaries (~44%) and pattern catalogues (~32%)
  - Only ~4% domain-specific entries
- Recall (–13%) declines more than precision (–5%)
- Flood slightly better due to larger data set

Mixed-Domain Results

Can extraction models benefit from extra domain data?

- Enhance training set by extra domain data

	Target					
Enhanced Source	Earthquake			Flood		
	P	R	F1	P	R	F1
Earthquake	**–.063**	**+.038**	**–.012**			
Flood				**–.061**	**+.012**	**–.028**

P/R/F1 measures for the extracted relationship

- Recall +3% (more patterns, greater dictionary)
- but precision –8% (more noise)
- F1 score is slightly lower (–2%) for mixed domain models

Conclusion

- **Dictionaries and dependency patterns are a good choice for domain-independent extraction models**
- Small F1 decrease of 9% measured
- Stay with single-domain models in the intra-domain scenario

Thank you!