Real-time population of Knowledge Bases: Opportunities and Challenges

Ndapa Nakashole
Gerhard Weikum

AKBC Workshop at NAACL 2012
Real-time Data Sources

• In news and social media, the implicit query is:
  – What’s happening right now?

• Batch-oriented KBP methods rely on Web snapshots (e.g., ClueWeb09, ca. 3 years old)

• News aggregators present a timely big picture
  – But they display text snippets and headlines, not relational facts
  – [Google News, …]
Goal: Real-time KBP

• **Goal:** Timely transformation of text into relational facts
  – Enabling fine-grained exploration of the big picture as it emerges

• The big picture is a series of stories and events

• Stories and events are made of facts
Francois Hollande elected as president of France
Koffi Annan warns about Syria
George Zimmerman arrested in Martin murder case
Challenges (1): Relation Discovery

• **Open Set of Relations**
  – Need to discover and maintain a large, dynamically evolving set of relations

• Go **beyond common relations** such as “bornIn”
  – Example interesting relations: firedFrom, hadAffairWith, …

• **Capture only semantically meaningful relations**
  – Discard noisy relations
Challenge (2): Dynamic Entity Discovery

• **For semantic consistency** in the facts we extract
  – Need to map *noun phrases to entities* in a KB
  – E.g., “Jeff Dean” can mean Google engineer or rock musician

• **But, KBs are incomplete** in the entities they contain
  – Jeff Dean the Google engineer doesn’t have a Wikipedia page
  – He is missing in Wikipedia-derived KBs

• **Open set of entities**
  – Need to recognise and handle out-of-KB entities
  – But go above the level of noun phrases
Challenge (3): Extraction under Time Constraints

- Due to need for timely fact extraction
  - Need to produce results under time constraints

- We would like to report the facts soon after they become available as
  - not a few weeks down the line
Our Approach
Approach – Relation Discovery: Semantically-typed patterns

• To identify meaningful relations
  – We introduced Syntactic-Lexical-Ontological (SOL) patterns

• Syntactic-Lexical – surface words and part-of-speech stags
• Ontological – semantic classes as entity placeholders, e.g., <singer, scientist, …>

• Example SOL patterns:
  – <comedian> parodied <person>
  – <musician> wrote hits for <musician>
  – <person> headliner at <event>
Approach – Relation Discovery: Semantically-typed patterns (2)

• **SOL patterns** are arranged them into *synonyms* and a hierarchy of subsumptions

• **Example subsumptions:**
  - wife of => spouse of
  - spouse of => knows

• **We produced ca. 350,000 SOL patterns**
  - Available for download
  - For details see: Nakashole, Weikum and Suchanek at EMNLP 2012
Approach – Dynamic Entity Discovery: Infer types for new entities

• SOL patterns require that entities have types
  – Need to align new entities along ontological dimension
  – Proposal: infer entity types from SOL patterns

• SOL pattern: <singer> released <album>
  – Given: X released Y, Is X of type singer? Not always!
    – Due to: polysemy in syntax
    – Due to: incorrect dependency paths between entity pairs

• But we can approximate likely types
Approach – Time Constraints: Continuous processing model

- Continuously process stream of incoming documents
- Define a time slice for extraction
  - Time window
  - Within time slice, define target recall
  - Redundancy means need not process all documents in a time slice
Thanks!

Poster 14