Evaluating the Quality of a Knowledge Base Populated from Text

James Mayfield, JHU Human Language Technology Center of Excellence, and Tim Finin, UMBC

The Problem

- Directly evaluate knowledge bases of entities and facts automatically extracted from text
- Evaluate things, not strings
- Many problems make this hard:
  - How can we assess potentially huge KBs?
  - Can we accommodate different KR systems?
  - Where and how does inference come into play?
  - How can we align entities without unique IDs?
  - How provenance, probabilities and time fit in?
  - Must we test both completeness and soundness?

An Approach

- Make some simplifying assumptions
  - Use a common, fixed KB schema (ontology)
  - KBs are populated from common document collection
  - Record provenance, including entity mentions
  - Either require a standard KB API or map KB to a common KR language (e.g., RDF)
  - Sample KB via evaluation queries, using entity mentions in documents to start the queries
  - Human assessors judge results, which are scored using precision and recall

Sampling for KB Errors

Our sampling approach depends on having a set of queries that test for likely error types, such as:
- Two distinct ground truth entities are conflated
- GT entity is split into several entities
- GT entity is missing from the KB
- Spurious entity is present in the KB
- GT relation is omitted from the KB
- Spurious relation is present in the KB
- Entry point is tied to the wrong KB node

Hard Problems Remain

A number of hard problems arise in evaluating KBs from expressive KR representation languages, e.g.:
- Assigning a semantics for probabilistic KBs
- KBs with disjunctions (Bart attends either Springfield Middle School or Springfield High)
- Skloem individuals
- Indefinite knowledge (Lisa’s age is less than 18)

2012 TAC Cold Start Knowledge Base Population

This approach is being used in 2012 Cold Start TAC task which combines several new features:
- The schema is defined but the KB is initially empty
- Tasks include entity linking and slot filling
- Participants submit a simple export format of their populated KB for evaluation
- Submitted KBs are converted to RDF and evaluated with a set of SPARQL queries
- Assessors score the results using metrics based on precision and recall

Knowledge Base Annotations

- Cold Start’s provenance and certainty features are supported as KB annotations or metadata
- Extracted entities and facts are tied to mentions
- Fact assertions can be qualified with a certainty measure
- Implemented using RDF reification, since these make assertions about RDF triples
- The semantics of the certainty annotations is not defined

2012 Cold Start Submission Format

- CS uses a simple submission format that is then used to generate an RDF KB for evaluation
- Each line is a tab-separated triple with optional annotations for provenance and certainty

Conclusions

- We want to evaluate things, not strings
- Make some reasonable assumptions (e.g., fixed schema, include provenance data, etc.)
- Either require a standard KB API or map KB to a common KR language (e.g., RDF)
- Evaluate results of a set of sampling queries using standard metrics (precision, recall, F)
- To be used in the 2012 TAC Cold Start KBP task

http://tac.nist.gov/p592