A Large-Scale System for Annotating and Querying Quotations in News Feeds

Jisheng Liang

Evri Inc.

jisheng@evri.com
Challenges

• How can semantic technologies be applied to the IR problems?
• How to address scalability and effectiveness of search?
  – by applying IR technologies?
• How to allow web users to exploit the expressiveness of the semantic data?
Overview

• Quotation extraction from news feeds
  – NLP and semantic analysis

• Indexing of the semantically annotated quotes
  – Scale to large data size

• Efficient and flexible querying capability
  – Entity-based and faceted search
What did <Speaker> say about <Subject>?

Speaker or Subject can be specified as:

- Specific Entity
- Facet – category of entities
- Keywords
- Boolean combination of the above

What did <Obama> say about any [Basketball_Player] or [Basketball_Team]
- Decompose sentences into SVO triples
- Entities annotated with IDs and facets from the entity repository
Entity Repository

- ~2M entities
  - people, places, organizations, products, events, concepts, etc.
  - Covered in news/blogs regularly
- Sources
  - Wikipedia
  - Structured/semi-structured sources
- Expanding continuously
  - Automated detection and import of emerging new entities
    - e.g. Susan Boyle, Chile earthquake
  - Live update of entity properties
    - e.g. teams trade players
Entity Properties

- Unique identifier
- Type and facets in a taxonomy
  - Person/Sports/Athlete/Basketball_Player
- Description
- Synonyms and aliases
- Type- and facet-specific attributes
  - e.g. birth date, birth place for person
- Relation properties
  - e.g. basketball player → teams, leagues
Quote Extraction

Sentences

Linguistic Analysis

Syntactic structures

Entity Disambiguation

Annotations

Quotation Detection

Candidates

Attribution & Collapsing

(Speaker, Quote)
Quote Detection

• Detect quotation verbs
  – e.g. acknowledge, add, argue, caution, say, suggest, urge, etc.

• Verify verb subject is person

• Check quotation marks
  – Quote could span multiple sentences
  – Rule out non-quotes, e.g. movie titles
Quote Annotation

• Coreference Resolution
  – Pronouns, e.g. “He said”
  – Aliases, e.g. “said Gates”
  – Definite noun, e.g. “the president said”

• Entity Disambiguation
  – Assign entity ID
  – Assign facets based on context

• Identify speaker entity

• Mark-up quote boundary
Quote Indexing

• Collapse into a triple of (Speaker, Verb, Quote)
• Store the triples in an inverted index, i.e. Lucene
  – Each triple as a document
  – Each element in a field
• Highly-scalable, fast search engine
+ Semantic Annotations
Example

Nash said, ``I would love to meet him, obviously, and to play hoops with the President would be kind of fun.''

**Speaker Field:**
Name = Steve Nash  
Entity ID = 0x49c26  
Facet = [Basketball Player]

**Quote Field:**
**Keywords:** love, meet, Barack Obama, obviously, play, hoops, president, fun

**Entity 1:**
Name = Barack Obama  
ID = 0x49c26  
Facets = [Politician], [Country Leader]
Query Examples

• Find recent quotes by <Steve Nash>
• What did any [Basketball Player] say about <President Obama>?
• What did any [Athlete] say about <the president>?
• What are people saying about <President Obama> and “hoops”?
Public APIs

• **REST APIs**
  - http://api.evri.com/v1/quotes/about?entityURI=/person/barack-obama-0x16f69&speaker=facet/basketball_player

• **XML or JSON output**

• **Documentation**
  http://www.evri.com/developer/rest#API-GetQuotations
Performance

- >10 million quotes
  - 60K new quotes added each day
  - In same index with over 0.5 billion SVO triples
- Query execution time (Uncached)
  - Average: 109 ms
  - Median: 54 ms
Links

- Evri Portal
  - www.evri.com

- Developer APIs
  - http://www.evri.com/developer
Future Work

- Improve relevance ranking
  - Authoritativeness, timeliness, novelty, diversity
- Identify “Pull Quote” from each article
- Sentiment analysis