Language and Domain Independent Entity Linking with Quantified Collective Validation

Han Wang¹, Jin Guan Zheng², Xiaogang Ma¹, Peter Fox¹, and Hong Ji²
(Tetherless World Constellation¹, Computer Science Department², Rensselaer Polytechnic Institute, Troy, NY, USA
(wangh17, zhengj6, max7, pfox, jih)@rpi.edu

• These authors contributed equally to this work.

**WALKING-THROUGH EXAMPLE**

One day after released by the Patriots, Florida born Caldwell visited the Jets. .... The New York Jets have six receivers on the roster: Cotchery, Coles, ...

**MENTION CONTEXT GRAPH (G_m)**

- **G_m** is a light-weight source context representation which simply involves mention co-occurrence.
- **G_m** will have an edge between two mention vertices if both of them fall into a context window in the source document.
- Two mention vertices will be connected via a dashed edge if they are coreferential but are not located in the same context window.

**KNOWLEDGE BASE GRAPH (G_K)**

- **G_K** is a weighted graph that consists of a set of vertices representing the entities and a set of directed edges labeled with relations between entities.
- Information entropy based relation weight computation:
  \[ H(r) = - \sum_{c \in C(r)} P(e_i) \log(P(e_i)) \]
- A "wiki link" relation is added between two entities if one of them appears in the Wikipedia article of the other.

**CANDIDATE RANKING**

- **Salience Ranking (SR)**: measures candidates’ importance without the context using information entropy:
  \[ S_i(r) = \sum_{c \in C(r)} H(r) \frac{S_{\text{src}}(e_i)}{L(c)} \]
- **Context Similarity Ranking (CS)**: measures the structural similarity between candidate graphs using Jaccard Similarity:
  \[ S_m(m_c, e) = \frac{\# \text{src}(m_c) \cap \# \text{src}(e)}{\# \text{src}(m_c) \cup \# \text{src}(e)} \]
- **Candidate Graph Collective Validation (CV)**:
  **Assumption**: A “tighter” relation between two candidates is more likely to be an appropriate representation of the relation between their co-occurring mentions in the source context.
  **Quantitatively** differentiates different types of relations using the calculated relation weights in **G_m**.
  \[ W(G_c(r)) = \sum_{c \in C(V(G_c))} S_i(r) S_m(m_c, e) + \sum_{c \in C(V(G_c))} H(r) \]

**EXPERIMENTS**

**Effectiveness**:
- **SR** injects the impact of different relations into the entity salience through relation weights.
- **SR + CS** further incorporates the source context. It promotes a great number of good candidates to the top of the ranking list, if not dozen.
- **SR + CS + CV** collectively validates the candidates of the target mention and its collaborating mentions by integrating their SR and CS scores as well as the relation weights.

**Error Analysis**:
- Lack of context for the target mention especially in web blogs and discussion forums resulted in poor performance in CS and CV.
- Failed coreference rules introduced false candidates.
- Incomplete KB with missing relations caused difficulties in the final CV step.

**Languages and Domains**

- **Chinese EL**
- **Biomedical EL**
- **Earth Science EL**

- **Biomedical Science Error Analysis**
  - In the experiment dataset, the relations between collaborating mentions were too similar for QCV to function at its full power.

- **Earth Science Error Analysis**
  - DBpedia introduced certain noise as a general KB.
  - Not all the domain-specific relations were clearly defined in DBpedia.