Boosting Entity Linking Performance by Leveraging Unlabeled Documents

Phong Le, Ivan Titov

Xiaofan Yan

Key idea

- Leveraging unlabeled documents:
  - propose a weakly-supervised model which exploits only naturally occurring information: unlabeled documents and Wikipedia
  - First stage
    - Construct a high recall list of candidate entities for each mention in an unlabeled document
  - Second stage
    - Use the candidate lists as weak supervision to constrain our document-level entity linking model

Model

- Candidates generation:
  - use the Wikipedia link graph, restrict vertices to the ones potentially appearing in the document
  - perform message passing with a simple probabilistic model which does not have any trainable parameters
- Document-level disambiguation:
  - train a document-level statistical disambiguation model which treat sentities as latent variables and uses the candidate lists as weak supervision

Candidates generation

- Goal:
  - Not only model fit between an entity and its local context but also model interactions between entities in a document
- Model
  - Use CRF to define score function
  - Score entities independently relying on the candidate lists
  - Local score in Ganea and Hofmann(2017).
  - Similar attention model in Globerson et al. (2016)
  - Feature normalized frequency of mention

Producing weak supervision

- Filter candidates set (ranking)
  - preprocessing technique of Ganea and Hofmann (2017)
  - use Wikipedia to create a link graph which defines the structure of a probabilistic graphical model which we use to rerank the candidate list
  - Select top candidate

Experiments

- Parameters
    - Ganea and Hofmann (2017): Entity embeddings
    - Word2vec word embeddings: local score function and GloVe embeddings8
    - 6 test set
    - AIDA CoNLL ‘testa’ data as development set
    - SpaCy9: extract named entity mentions
    - Baseline: plus unlabeled documents + trained supervisedly

Analysis

- Constraint-driven learning
- Document-level disambiguation model
- Local and global disambiguation
- different NER (named entity recognition) types