A brief Introduction to Entity Linking

Tianxiang Sun (孙天祥)
What is entity linking?

- Entity Linking (EL) aims to link entity mentions in texts to knowledge bases
- Also called Named Entity Disambiguation (NED)
- Non-trivial: entity mentions are usually ambiguous
- A demo

Napoleon [Napoleon] was the emperor of the First French Empire. He was defeated at Waterloo [Battle of Waterloo] by Wellington [Arthur Wellesley, 1st Duke of Wellington] and Blücher [Gebhard Leberecht von Blücher]. He was banned to Saint Helena [Saint Helena], died of stomach cancer, and was buried at Invalides [Les Invalides].
• Formulation
  • Input: document $D = \{w_1, \ldots, w_n\} (+\{m_i\} \text{ if end-to-end})$
  • Output: list of mention-entity pairs $\{(m_i, e_i)\}$

• A EL system typically performs two tasks:
  • NER / Mention Detection (MD)
    • Ent-to-End
    • Disambiguation-only
  • Entity Disambiguation (ED)
    • Candidate selection / generation (usually heuristics)
    • Scoring (Ranking) candidates
      • local & global
• Outline
  • Models
    • Modules
    • Neural models
    • Symbol-neural hybrid model
  • Related topics
    • Distant learning
    • Entity typing
  • Datasets, metrics, and platform
• Outline
  • Models
    • Modules
      • Neural models
      • Symbol-neural hybrid model
  • Related topics
    • Distant learning
    • Entity typing
  • Datasets, metrics, and platform
• Modules in pipeline (Disambiguation-Only)
  • Candidate selection
    • Dictionary
    • Anchors statistic
    • Surface matching heuristic
  • Scoring candidates
    • Entity embedding
    • Local compatibility (modeling the selected mention and its context)
    • Global coherence (modeling other mentions and their candidates)
• Candidate selection
  • Dictionary (Hoffart et al., 2011; Yamada et al., 2016; Cao et al., 2017; Cao et al., 2018)
    • Constructed from knowledge bases, e.g., DBpedia, YAGO, etc.
    • Examples:
      “Apple” for Apple Inc.
      “Big Apple” for New York City
  • Anchors statistic (Ganea et al., 2017; Kolitsas et al., 2018)
    • Mention-entity prior: \( P(e|m) = \frac{|A_{e,m}|}{|A_{*,m}|} \)
    • Computed from mention entity hyperlink count statistic from Wikipedia etc.
    • Also as a feature for disambiguation
  • Surface matching heuristic (Le and Titov, 2019)
• Scoring candidates
  • (Kolitsas et al., 2018)
  • Entity-mention compatibility
    • Entity embedding
    • Context-Independent features
    • Context-Dependent features
    • Mention-entity prior
    • Global features
• Scoring candidates
  • (Kolitsas et al., 2018)
  • Entity-mention compatibility
  • Entity embedding
  • Context-Independent features
  • Context-Dependent features
  • Mention-entity prior
  • Global features
• Scoring candidates
  • (Kolitsas et al., 2018)
  • Entity-mention compatibility
  • Entity embedding
  • Context-Independent features
  • Context-Dependent features
  • Mention-entity prior
  • Global features
Scoring candidates

(Kolitsas et al., 2018)

Entity-mention compatibility

Entity embedding

Context-Independent features

Context-Dependent features

Mention-entity prior

Global features
• Scoring candidates

  • (Kolitsas et al., 2018)
  • Entity-mention compatibility
  • Entity embedding
  • Context-Independent features
  • Context-Dependent features
  • Mention-entity prior
  • Global features
• Scoring candidates
  • \((\text{Kolitsas et al., 2018})\)
  • Entity-mention compatibility
  • Entity embedding
  • Context-Independent features
  • Context-Dependent features
  • Mention-entity prior
  • Global features
• Scoring candidates
  • (Kolitsas et al., 2018)
  • Entity-mention compatibility
  • Entity embedding
    • Context-Independent features
    • Context-Dependent features
    • Mention-entity prior
  • Global features
• Scoring candidates – Entity embedding

  • Jointly map words / mentions and entities into the same continuous vector space.

  • (Yamada et al., 2016; Ganea et al., 2017)

1. Skip-gram model (for words)

\[
P(w_{t+j}|w_t) = \frac{\exp(V_{w_t}^T U_{w_{t+j}})}{\sum_{w \in W} \exp(V_{w_t}^T U_w)}
\]

2. KB graph model (extend word embedding matrix V and U for entities)

\[
P(e_o|e_i) = \frac{\exp(V_{e_i}^T U_{e_o})}{\sum_{e \in E} \exp(V_{e_i}^T U_e)}
\]

3. Anchor context model (let words and entities interact with each other via anchors)

\[
P(w_o|e_i) = \frac{\exp(V_{e_i}^T U_{w_o})}{\sum_{w \in W} \exp(V_{e_i}^T U_w)}
\]
• Scoring candidates – Entity embedding
  • Jointly map words / mentions and entities into the same continuous vector space.
  • (Yamada et al., 2016; Ganea et al., 2017)
  • Based on word2vec pre-trained vectors
    \[ J(z; e) := \mathbb{E}_{w^+ \mid e} \mathbb{E}_{w^-} [h(z; w^+, w^-)] \]
    \[ h(z; w, v) := [\gamma - \langle z, x_w - x_v \rangle]_+ \]
    \[ x_e := \arg \min_{z: \|z\|=1} J(z; e) \]
  
  • where \( w^+ \sim \hat{p}(w \mid e) \propto \#(w, e) \) and \( w^- \sim q(w) \)
  • Let vectors of positive words are closer to the embedding of entity \( e \).
• Scoring candidates – Entity embedding
  • Map words / mentions and entities into different vector space.
  • [(Cao et al., 2017)]
  • Based on Skip-gram and CBOW
  • Learn representations for words, entities, and mention senses.
Scoring candidates

--

Entity embedding

- Jointly map words / mentions and entities into the same continuous vector space.

- (Yamada et al., 2016; Ganea et al., 2017; Cao et al., 2017)

Knowledge Base

Anchor

$$d_1$$: In the 1996 action film [[Independence Day (US)]] [[Independence Day]], the United States military uses alien technology captured …

$$d_2$$: … holds annual [[Independence Day (US)]] celebrations and other festivals …

$$d_3$$: … early Confederate [[Memorial Day]] celebrations were simple, somber occasions for veterans and their families to honor the dead …

Mention Sense Mapping

$$g(July 4th, e_1)$$

Text Representation Learning

$$P(C(w_i | w_i) \cdot P(C(m_1 | s_j^*)$$

$$w_i / s_j^*$$

Text Space

Representation Learning

$$s_{Independence Day (film)}$$

Knowledge Space
• Scoring candidates – Local feature (modeling mentions, contexts, and entities)

  • Mention-entity prior: \( P(e|m) = \frac{|A_{e,m}|}{|A_{*,m}|} \)

  • Context-Independent feature
    • String similarity (Cao et al., 2018)
    • Char BiLSTM (Kolitsas et al., 2018)

  • Context-Dependent feature
    • Average over context words (Yamada et al., 2016; Cao et al., 2017)
    • BiLSTM (Kolitsas et al., 2018; Le and Titov, 2019)
    • Attention (Ganea et al., 2017; Kolitsas et al, 2018; Cao et al., 2018)
- Scoring candidates – Global feature (modeling other mentions and their candidates)
  - Hand-crafted feature like number of shared incoming links... (Hoffart et al., 2011)
  - Bag-of-Words (Yamada et al., 2016)
  - Voting-based (Kolitsas et al, 2018)
  - Markov chain (Delpeuch et al., 2019)
  - CRF (Ganea et al., 2017)
  - GCN (Cao et al., 2018)

All mentions in a document shall be on the same topic!
• Outline

• Models
  • Modules
  • Neural models
    • Symbol-neural hybrid model
• Related topics
  • Distant learning
  • Entity typing
• Datasets, metrics, and platform
• A local model (Ganea et al., 2017)

Training objective (max-margin loss)

$$\theta^* = \arg \min_{\theta} \sum_{D \in \mathcal{D}} \sum_{m \in \mathcal{D}} \sum_{e \in \Gamma(m)} g(e, m),$$

$$g(e, m) := [\gamma - \Psi(e^*, m, c) + \Psi(e, m, c)]_+$$
• A global model (Cao et al., 2018)

\[ c_{m_i, e_j} = \sum_{w_k \in c(m_i)} \alpha_{k,j} w_k \]

\[ \{ \text{sim}(e_j, m_i) | m_i \in N(m_i) \} \]
An end-to-end Model (Kolitsas and Ganea, 2018)

"At training time, for each input document we collect the set $M$ of all (potentially overlapping) token spans $m$ for which $|C(m)| \geq 1.$"
• Outline
  • Models
    • Modules
    • Neural models
    • Symbol-neural hybrid model
  • Related topics
    • Distant learning
    • Entity typing
  • Datasets, metrics, and platform
• DeepType ([Raiman and Raiman, 2018](#))

• Associate with each entity a series of types (e.g. Person, Place, etc.) that if known, would rule out invalid answers, and therefore ease linking.

<table>
<thead>
<tr>
<th>Entity</th>
<th>jaguar</th>
<th>Jaguar</th>
<th>jungle</th>
<th>jungle</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type</td>
<td>Animal</td>
<td>Road vehicle</td>
<td>Region</td>
<td>Music</td>
</tr>
<tr>
<td>only link Prob.</td>
<td>0.29</td>
<td>0.60</td>
<td>0.35</td>
<td>0.17</td>
</tr>
<tr>
<td>Prob. w/. types</td>
<td>1.0</td>
<td>0.0</td>
<td>1.0</td>
<td>0.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Entity</th>
<th>jaguar</th>
<th>Jaguar</th>
<th>highway</th>
<th>Highway</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type</td>
<td>Animal</td>
<td>Road vehicle</td>
<td>Physical object</td>
<td>Film</td>
</tr>
<tr>
<td>only link Prob.</td>
<td>0.29</td>
<td>0.60</td>
<td>0.85</td>
<td>0.04</td>
</tr>
<tr>
<td>Prob. w/. types</td>
<td>1.0</td>
<td>0.0</td>
<td>1.0</td>
<td>0.0</td>
</tr>
</tbody>
</table>
DeepType (Raiman and Raiman, 2018)

- Terminology
  - Relation (e.g. instance of)
  - Type
    A label defined by a relation, e.g., the type applied to all children of Human connected by instance of is IsHuman.
- Type Axis: a set of mutually exclusive types
- Type System: type axes + type labelling function

George Washington  \(\rightarrow\)  IsA Topic  \(\rightarrow\)  \{Person, Politics\}

Washington, D.C.  \(\rightarrow\)  IsA Topic  \(\rightarrow\)  \{Place, Geography\}
• DeepType (Raiman and Raiman, 2018)

• Type System
  • \( A \): the assignment for the boolean discrete variables that define the type system.
    \( A_i = 1 \) if the \( i \)-th parent-child relation gets included in the type system.
    \[
    A = \{0, 1, 0, 1, 1, \ldots \}
    \]
  • Optimize: heuristic search / stochastic optimization (mixed integer problem)

• Type Classifier
  • \( \theta \): continuous variables that parameterize the classifier to fit to the type system.
  • Optimize: gradient descent

• Objective: solve \( A \) and \( \theta \)
  \[
  \max_A \max_\theta S_{\text{model}}(A, \theta) = \frac{\sum_{(m, e_{GT}, e_m) \in M} \mathbb{1}_{e_{GT}}(e^*)}{|M|}.
  \]
• DeepType (Raiman and Raiman, 2018)
  • Discrete optimization of the type system
    • Define an objective to measure how good a solution is
    • There is a trade-off
      • Disambiguation power
        Measure the improvement of entity linking accuracy of the solution.
      • Learnability
        Measure how learnable the type axes in the selected solution.
    • Regularization

\[
J(\mathcal{A}) = (S_{\text{oracle}} - S_{\text{greedy}}) \cdot \text{Learnability}(\mathcal{A}) + \\
S'_{\text{greedy}} - |\mathcal{A}| \cdot \lambda.
\]
• **DeepType** *(Raiman and Raiman, 2018)*

  • Objective of type system

\[
J(A) = (S_{oracle} - S_{greedy}) \cdot \text{Learnability}(A) + S_{greedy} - |A| \cdot \lambda.
\]

  • Mention-entity prior: \[\mathbb{P}_{\text{Link}}(e|m) = \frac{\text{LinkCount}(m,e)}{\sum_{j \in \mathcal{E}_m} \text{LinkCount}(m,j)}\]

  • Greedy: predicts only according to the mention-entity prior.

  • Oracle: prunes candidate set to only contain entities whose types match those of \(e_i^{GT}\)

\[
\text{Oracle}(m) = \arg\max_{e \in \mathcal{E}_{m,oracle}} \mathbb{P}_{\text{entity}}(e|m, \text{types}(x)).
\]

\[
S_{oracle} = \frac{\sum_{(m,e^{GT},\mathcal{E}_m) \in M} \mathbb{1}_{e^{GT}(\text{Oracle}(m))}}{|M|}.
\]
• DeepType (Raiman and Raiman, 2018)

• Objective of type system

\[ J(\mathcal{A}) = (S_{\text{oracle}} - S_{\text{greedy}}) \cdot \text{Learnability}(\mathcal{A}) + S_{\text{greedy}} - |\mathcal{A}| \cdot \lambda. \]

• Learnability

\[ \text{Learnability}(\mathcal{A}) = \frac{\sum_{t \in \mathcal{A}} \text{AUC}(t)}{|\mathcal{A}|} \]

• \( \lambda \): per type axis penalty term
• DeepType (Raiman and Raiman, 2018)

  • Objective of type system

  \[ J(\mathcal{A}) = (S_{\text{oracle}} - S_{\text{greedy}}) \cdot \text{Learnability}(\mathcal{A}) + S_{\text{greedy}} - |\mathcal{A}| \cdot \lambda. \]

  • Search methodologies

    • Beam search and greedy selection
    • Cross-entropy method
    • Genetic algorithm
    • ...

• DeepType ([Raiman and Raiman, 2018](#))
  • Discrete optimization of the type system
• Type classifier
  • Classify per-token type
- DeepType (Raiman and Raiman, 2018)
  - Discrete optimization of the type system
  - Type classifier
  - Inference
    - Given Input words $w_0, \ldots, w_L$ and mention $m$ covering words $w_x, \ldots, w_y$
    - Through type classifier, we obtain the type conditional probability for all type axes $i$: $\{P_i(\cdot|w_x, D), \ldots, P_i(\cdot|w_y, D)\}$
    - Aggregate using max-over-time and obtain $P_{i,*}(\cdot|m, D)$
    - Take the prior into consideration, we get the final entity score

$$s_{e,m,D,A,\theta} = P_{\text{Link}}(e|m) \cdot \left(1 - \beta + \beta \cdot \left\{ \prod_{i=1}^{k} (1 - \alpha_i + \alpha_i \cdot P_{i,*}(t_i|m, D)) \right\} \right).$$
• Outline
  • Models
    • Modules
    • Neural models
    • Symbol-neural hybrid model
  • Related topics
    • Distant learning
    • Entity typing
  • Datasets, metrics, and platform
• Distant learning
  • Distant supervision (also referred to weak supervision) assumption:
    
    *If two entities participate in a relation, all sentences that mention these two entities express that relation.*

  • An example:

    *Elevation Partners, the $1.9 billion private equity group that was founded by Roger McNamee*

  • However, the assumption can be violated:

    *Roger McNamee, a managing director at Elevation Partners, ...*
• Distant learning

• When aligning Freebase to Wikipedia and New York Times...

**Table 1.** Percentage of times a related pair of entities is mentioned in the same sentence, but where the sentence does not express the corresponding relation

<table>
<thead>
<tr>
<th>Relation Type</th>
<th>New York Times</th>
<th>Wikipedia</th>
</tr>
</thead>
<tbody>
<tr>
<td>nationality</td>
<td>38%</td>
<td>20%</td>
</tr>
<tr>
<td>place_of_birth</td>
<td>35%</td>
<td>20%</td>
</tr>
<tr>
<td>contains</td>
<td>20%</td>
<td>10%</td>
</tr>
</tbody>
</table>

• ([Riedel et al., 2010](#)) proposed a relaxed assumption:

*If two entities participate in a relation, at least one sentence that mentions these two entities might express that relation.*
• Distant learning in entity linking (**Le and Titov, 2019**)
  
  • Construct distant supervision: surface matching heuristics (measure overlap)
  
  • Positive lists: top candidates from the matching heuristics
  
  • Negative lists: randomly sampled sets of entities
  
  • Multi-Instance Learning (MIL): find the entity should be linked

Can **Bill Clinton** really emerge as a beloved father figure to a frazzled **America**?

<table>
<thead>
<tr>
<th>Bill_Clinton (TV episode)</th>
<th>America (song)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bill_Clinton (president)</td>
<td>Volunteers_of_America</td>
</tr>
<tr>
<td>Bill_Clinton's_victory</td>
<td>United_States_of_America (nation)</td>
</tr>
<tr>
<td>Presidency_of_Bill_Clinton</td>
<td>United_States_of_America (music track)</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

- name-matched candidates
- knowledge-base triples
- name-matched candidates
• Distant learning in entity linking (Le and Titov, 2019)

• During training, we have $\langle m, c, E^+, E^- \rangle$, in testing, $E^- = \emptyset$.

• MIL: we want to train the model to score at least one candidate in $E^+$ higher than any candidate in $E^-$. To achieve this, we employ a max-margin loss

$$l(m, c) = \left[ \max_{e \in E^-} g(e, m, c) + \delta - \max_{e \in E^+} g(e, m, c) \right]_+$$

$$L_1 = \sum_{(m, c) \in D} l(m, c)$$

• Recall that many data points are noisy. $E^+$ may not contain the correct entity.
• Distant learning in entity linking (Le and Titov, 2019)

• Representation for $E^+$
  
  \[ e_{E^+} = \sum_{e \in E^+} \alpha_e e \]

• Noise detection
  
  \[ p_N(1|m, c, E^+) = \sigma\left(\frac{\text{FFN}_f([e_{E^+}, f_{h-1}, b_{h-1}, f_k, b_k])}{T}\right) \]

• Use a binary classifier

• Training
  
  • Down-weight potentially noisy data points. New loss:

\[ L_2 = \sum_{(m,c) \in D} p_N(0|m, c, E^+)l(m, c) + \]

\[ \eta \times \text{KL}\left(\frac{\sum_{(m,c) \in D} p_N(\cdot|m, c, E^+)}{|D|} | p_N^*\right) \]

• Testing: with / without noise detector
• Outline
  • Models
    • Modules
    • Neural models
    • Symbol-neural hybrid model
  • Related topics
    • Distant learning
    • Entity typing
  • Datasets, metrics, and platform
- Entity Typing
- FIGER ([Ling and Weld, 2012](#))
- Fine-grained NER task
- Hierarchical labels
  - `person/`
  - `person/actor`
  - `location/`
  - `location/city`
  
| **person**  
|---|
| actor  
| architect  
| artist  
| athlete  
| author  
| coach  
| director  
| doctor  
| engineer  
| monarch  
| musician  
| politician  
| religious_leader  
| soldier  
| terrorist  
| **organization**  
|---|
| airline  
| company  
| educational_institution  
| fraternity_sorority  
| sports_league  
| sports_team  
| terrorist_organization  
| government_agency  
| government  
| political_party  
| educational_department  
| military  
| news_agency  
| **location**  
|---|
| body_of_water  
| city  
| country  
| county  
| province  
| railway  
| road  
| bridge  
| mountain  
| glacier  
| astral_body  
| cemetery  
| park  
| **product**  
|---|
| camera  
| engine  
| airplane  
| car  
| ship  
| spacecraft  
| train  
| mobile_phone  
| computer  
| software  
| game  
| instrument  
| weapon  
| **art**  
|---|
| written_work  
| film  
| newspaper  
| play  
| music  
| **event**  
|---|
| military_conflict  
| attack  
| natural_disaster  
| election  
| sports_event  
| protest  
| terrorist_attack  
| **building**  
|---|
| airport  
| dam  
| hospital  
| hotel  
| library  
| power_station  
| restaurant  
| sports_facility  
| theater  
| time  
| color  
| award  
| educational_degree  
| title  
| law  
| ethnicity  
| language  
| religion  
| god  
| **chemicalThing**  
|---|
| chemical_thing  
| biological_thing  
| medical_treatment  
| disease  
| symptom  
| drug  
| body_part  
| living_thing  
| animal  
| food  
| **website**  
|---|
| website  
| broadcast_network  
| broadcast_program  
| tv_channel  
| currency  
| stock_exchange  
| algorithm  
| programming_language  
| transit_system  
| transit_line |
• Entity Typing for Entity Linking (ET4EL) (Onoe and Durrett, 2019)
  • Alleviate overfitting
  • Construct entity typing dataset using hyperlinks and Wiki categories
  • Two parts:
    • Entity typing: \( \Phi : (m, s) \rightarrow T. \)
    • Entity linking: \( e = \Omega(\Phi(m, s), C'). \)
• Entity Typing for Entity Linking (ET4EL) (Onoe and Durrett, 2019)

• Entity linking prediction (heuristic, untrained)
  
  • $\Omega$ is defined as the sum of probabilities for each type

\[
e'_c = \sum_i t_i \cdot 1_{T_c} (V^t_i)
\]

\[
e = \arg \max_e (e'_1, \ldots, e'_{|C|})
\]

• No need to access the labeled entity linking data.
• Outline
  • Models
    • Modules
    • Neural models
    • Symbol-neural hybrid model
  • Related topics
    • Distant learning
    • Entity typing
  • Datasets, metrics, and platform
• Datasets
  • AIDA-CoNLL (Hoffart et al., 2011)
    • Text data: CoNLL 2003 NER task
    • Knowledge base: YAGO
  • TAC 2010 (Ji et al., 2010)
    • Text data: news articles from various agencies and Web log data
  • WikiDisamb30 (Raiman and Raiman, 2018)
• Platform
  • GERBIL
• Metrics
  • Disambiguation-only
    • Micro accuracy
    • Macro accuracy
  • End-to-End
    • Micro F1
    • Macro F1
  • InKB v.s. NIL ("unlinkable")
Q & A