Annotating documents with relevant Wikipedia concepts

Janez Brank, Gregor Leban, Marko Grobelnik
Motivation

• Semantic annotation / enrichment
• Wikification
  • Wikipedia pages ≈ Concepts
  • Text and links provide contextual information about each concept and relations between concepts
  • Multilingual, cross-language links between concepts
  • General-purpose, freely available
• Disambiguation
• Paralellization
• Main application: NewsFeed / EventRegistry
  • To augment documents with concepts before further processing (e.g. clustering)
Mentions and candidate annotations

• Given an input document (that we want to annotate), which words/phrases refer to some concept from the Wikipedia?

• Use the internal links from the Wikipedia to identify such phrases
  • If some Wikipedia page contains a link with the anchor text \( a \) and target page \( t \)...
  • Whenever \( a \) occurs in our input document, we consider that as a (possible) mention of the concept \( t \), and \( t \) is a candidate annotation for this input document
Disambiguation

• Problem: links with the same anchor text $a$ can point to different targets $t$

• If $a$ appears as a mention in our input document, which of those target concepts should we annotate the document with (if any)?

• Two approaches to disambiguation
  • Local disambiguation: disambiguate each mention separately
  • Global disambiguation: disambiguate all the mentions in the input document together
    • Intuition: the document as a whole is about some topic, therefore the annotations should (mostly) be about that topic as well
    • Document is about cars → “Tesla Inc.” is more likely than “Tesla (band)”

• Our wikifier uses a pagerank-based global disambiguation approach described by Zhang and Rettinger (2014)
Pagerank-based disambiguation

• Construct a mention-concept graph
  • Bipartite graph: left vertices = mentions, right vertices = concepts
  • Transition probabilities: \( P(a \rightarrow t) = \frac{\text{[number of links with anchor text } a \text{ and target } t]}{\text{[number of links with anchor text } a]} \)

Elon Musk
Tesla
**chief executive**
**electric-car**

Elon Musk
Nikola Tesla
Tesla Inc.
Tesla (band)
**Electric vehicle**

**In Pivotal Moment, Tesla Unveils Its First Mass-Market Sedan**
Elon Musk, Tesla’s chief executive, delivered 50 cars to employees chosen to be the first owners. The electric-car maker faces a challenge in meeting the sizable demand.
Pagerank-based disambiguation

• Add concept-concept links
  • Transition probabilities: $P(c \rightarrow c') \propto SR(c, c')$
  • $SR = \text{"semantic relatedness" } = 1 - \frac{\ln \max(|L_c|,|L_{c'}|) - \ln |L_c \cap L_{c'}|}{\ln N - \ln \min(|L_c|,|L_{c'}|)}$

where $L_c$ is the set of Wikipedia pages that contain a link to $c$
Pagerank-based disambiguation

• Compute pagerank
  • Iteration: $PR_{new}(u) = \tau PR_0(u) + (1 - \tau) \sum_v PR_{old}(v) P(v \rightarrow u)$
  • Baseline pagerank: $PR_0(u) = 0$ if $u$ is a concept vertex
  • For a mention vertex: $PR_0(u) \propto$ [number of Wikipedia pages containing $u$ as the anchor-text of a link] / [number of Wikipedia pages containing $u$].

Elon Musk
Tesla
chief executive
electric-car

Elon Musk
Tesla Inc.
Tesla (band)
Electric vehicle

In Pivotal Moment, Tesla Unveils Its First Mass-Market Sedan
Elon Musk, Tesla's chief executive, delivered 30 cars to employees chosen to be the first owners. The electric-car maker faces a challenge in meeting the sizable demand.
Pagerank-based disambiguation

• If a mention has several candidate annotations, use the one with the highest pagerank
  • We say that this mention supports this annotation

• Intuition: pagerank flows into a concept vertex $c$
  • From mentions $a$ for which links with the anchor-text $a$ often point to the target page $c$
  • And from other concepts $c'$ that are semantically closely related to $c$

• Thus a set of semantically related concepts (that are adequately supported by some mentions) will boost each other and come out on top
  • ...which is just what global disambiguation is about
Highly ambiguous mentions

- Some mentions appear as the anchor text of links to a very large number of different pages
  - Including all these pages as concept vertices in the mention-concept graph would introduce noise
  - The graph would become huge and pagerank computation would be too slow

- Heuristics to deal with this:
  - If the entropy $H(\text{link target} \mid \text{anchor text} = a)$ is above a certain threshold (e.g. 3 bits), ignore the mention as being too ambiguous
  - Use only the 20 most frequently occurring concepts
  - If the mention consists entirely of stopwords (e.g. top 200 most frequent words in the language), ignore it
  - Ignore concepts that belong to certain WikiData categories (e.g. lists)
Miscellaneous heuristics

• Alternative definitions of semantic relatedness:
  • Instead of comparing sets of immediate predecessors in the Wikipedia link graph, we can use immediate successors or all neighbours (predecessors + successors)

• Two-stage disambiguation process:
  • Use a second scoring function to re-rank the top e.g. 20 candidates (by pagerank) before choosing which one to use as the annotation
  • $score(c | a) = w_1 f(P(c | a)) \cdot PR(c) + w_2 S(c, d) + s_3 LS(c, a)$
    • $f(x) = 1 \text{ or } x \text{ or } \log(x)$
    • $P(c | a) =$ probability that the target of a link is $c$ given that its anchor text is $a$
    • $S(c, d) =$ cosine similarity between the input document $d$ and the Wikipedia page for $c$
    • $LS(c, a) =$ cosine similarity between the context of $a$ (in $d$) and the context of links to $c$ (in the Wikipedia)
Implementation

• Suitable for parallel processing
  • Multiple input documents can be processed in parallel, independently of each other
• Can work with any language for which a (sufficiently large) Wikipedia is available
• Our implementation is available on http://wikifier.org/
  • Currently handling about 500,000 requests per day (total length of input documents: 1.2 GB per day) with plenty of CPU time to spare
  • Supports 134 languages (all languages for which a Wikipedia of at least 1000 pages is available)
    • This is too small for good coverage, but ~60 languages have a Wikipedia of at least 100,000 pages, which can already be useful
• Can optionally return WikiData/DbPedia class memberships
Comparison of different wikifiers

• Manually annotated set of 1393 news articles, originally prepared by the authors of AIDA

• For a given wikifier $w$, consider its set of annotations $A_w = \{ (d, c) :$ the wikifier $w$ has annotated the document $d$ with the concept $c \}$.
  
  • We can use the $F_1$-measure (or precision, recall, etc.) to compare two such sets to measure the agreement between wikifiers and/or the gold standard.

• Overall there is little agreement between different wikifiers, which suggests that wikification as a task is too poorly/vaguely defined

  • What does it mean for a concept $c$ to be relevant to /mentioned in the document $d$?
  
  • What sort of concepts do we want? Just entities? Everything?

<table>
<thead>
<tr>
<th></th>
<th>Gold</th>
<th>JSI</th>
<th>AIDA</th>
<th>Waikato</th>
<th>Babelfy</th>
<th>Illinois</th>
<th>Spotlight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gold</td>
<td>1.000</td>
<td>0.593</td>
<td>0.723</td>
<td>0.372</td>
<td>0.323</td>
<td>0.476</td>
<td>0.279</td>
</tr>
<tr>
<td>JSI</td>
<td>1.000</td>
<td>0.625</td>
<td>0.527</td>
<td>0.431</td>
<td>0.489</td>
<td>0.363</td>
<td></td>
</tr>
<tr>
<td>AIDA</td>
<td>1.000</td>
<td>0.372</td>
<td>0.352</td>
<td>0.434</td>
<td>0.356</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Waikato</td>
<td>1.000</td>
<td>0.481</td>
<td>0.564</td>
<td>0.474</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Babelfy</td>
<td>1.000</td>
<td>0.434</td>
<td>0.356</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Illinois</td>
<td></td>
<td>1.000</td>
<td>0.376</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spotlight</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.000</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: $F_1$ measure of agreement between the various wikifiers and the gold standard.
Conclusions and future work

• Efficient, highly parallel implementation of wikification based on global disambiguation

• Planned and/or possible future extensions:
  • Ignore a user-specified set of pages and/or categories when processing the Wikipedia
  • Allow the user to define additional sets of annotations (unrelated to the Wikipedia), along with phrases (mentions) that trigger them
  • Combine local and global disambiguation approaches
    • Local = e.g. based on the similarity between the context of a mention and of links that point to the candidate concept
    • Perhaps using word2vec instead of a plain bag-of-words representation
  • Improved handling of languages whose Wikipedia is small and has poor coverage
    • Use links from other-language Wikipedias to generate candidate annotations
    • Use cross-language information from WikiData to combine link-graphs of different-language Wikipedias into a common large link-graph