Linkify

Enhanced Reading Experience by Augmenting Text Using Linked Open Data

Ikuya Yamada\textsuperscript{1,2,3} Tomotaka Ito\textsuperscript{1,2} Shinnosuke Usami\textsuperscript{1,2} Tomoya Toyoda\textsuperscript{1,2} Shinsuke Takagi\textsuperscript{1,2} Hideaki Takeda\textsuperscript{3} Yoshiyasu Takefuji\textsuperscript{2}

\textsuperscript{1}Studio Ousia \textsuperscript{2}Keio University \textsuperscript{3}National Institute of Informatics
Background

Kyary Pamyu Pamyu is a Japanese model and singer. Her public image is associated with Japan's kawaisa culture centered in the Harajuku, Tokyo.

- We frequently encounter unfamiliar entities while reading text.
- Tedious steps are required to retrieve detailed information about an entity.
  - Select, copy, and paste text into the search box...
  - Especially problematic on recent touchscreen devices.
Kyary Pamyu Pamyu is a Japanese model and singer. Her public image is associated with Japan's kawaii culture centered in the Harajuku, Tokyo.

- Linkify automatically converts entity names into links
Kyary Pamyu Pamyu is a Japanese model and singer. Her public image is associated with Japan's kawaisa culture centered in the Harajuku, Tokyo.

- Linkify automatically converts entity names.
- When a user selects a link, a popup window containing the summarized information of the entity is displayed.
Outline

- Converting Entity Names into Links
  - Entity Linking
  - Evaluating Helpfulness of Entities
    - Machine-learning features
    - Dataset
    - Results
- Rendering the Widget
- Demonstration
1. **DOM Iterator** sends the page text to the server when the web browser renders a web page

2. *The server-side system* extracts entity names from the text and sends back to the browser

3. **Link converter** converts the extracted entity names into links
Converting Entity Names into Links

1. *Entity Linker* extracts entity names and disambiguates them to entries in Wikipedia

2. *Entity Helpfulness Evaluator* detects whether the detected entities are likely helpful to users
**Entity Linking**

Kyary Pamyu Pamyu is a Japanese model and singer. Her public image is associated with Japan's kawai culture centered in the Harajuku, Tokyo.

- **Entity Linking**: The task of linking entity names to entries in a knowledge base (KB) (e.g., Wikipedia)
- Recently the task has received considerable attention
  - Many research papers (2006-) [Cucerzan 2007, Milne et al. 2008, etc.]
  - **Open Source Implementations**: Wikipedia Miner, DBpedia Spotlight, TAGME, AIDA, Apache Stanbol, Illinois Wikifier, etc.

Our implementation of Wikipedia Miner is used
Evaluating Helpfulness of Entities

- We use supervised machine-learning to assign binary labels to entities detected by an entity linking system [1].
- A binary classification task to detect whether an entity is helpful to readers

Machine-learning Features for Evaluating Helpfulness of Entities

Features are classified into six categories

- Link probability
- Entity
- Entity class
- Topical coherence
- Textual
- Mention occurrence

List of machine-learning features

<table>
<thead>
<tr>
<th>Feature Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Link probability features</td>
<td></td>
</tr>
<tr>
<td>LINK_PROB(m)</td>
<td>The probability that m appears as an anchor text in KB</td>
</tr>
<tr>
<td>LINK_PROB_MAX(e, M)</td>
<td>The maximum LINK_PROB(m) of mentions that link to e</td>
</tr>
<tr>
<td>LINK_PROB_AVG(e, M)</td>
<td>The average LINK_PROB(m) of mentions that link to e</td>
</tr>
<tr>
<td>Entity features</td>
<td></td>
</tr>
<tr>
<td>GENERALITY(e)</td>
<td>The minimum depth of e from the root in the category structure of KB</td>
</tr>
<tr>
<td>WIKI_STATS(e)</td>
<td>The average number of page views of e</td>
</tr>
<tr>
<td>IN_LINKS(e)</td>
<td>The number of KB articles that have a link to e</td>
</tr>
<tr>
<td>OUT_LINKS(e)</td>
<td>The number of links pointing to other KB articles in e</td>
</tr>
<tr>
<td>CATEGORIES(e)</td>
<td>The number of categories assigned to e</td>
</tr>
<tr>
<td>Entity class features</td>
<td></td>
</tr>
<tr>
<td>FREEBASE_{class}(e)</td>
<td>Is e assigned a class named class in Freebase?</td>
</tr>
<tr>
<td>DBPEDIA_{class}(e)</td>
<td>Is e assigned a class named class in DBpedia?</td>
</tr>
<tr>
<td>SCHEMA_ORG_{class}(e)</td>
<td>Is e assigned a class named class in Schema.org?</td>
</tr>
<tr>
<td>Topical coherence feature</td>
<td></td>
</tr>
<tr>
<td>COHERENCE(e, E)</td>
<td>Representing how e is related to the central topics of D</td>
</tr>
<tr>
<td>Textual features</td>
<td></td>
</tr>
<tr>
<td>CAPITALIZED(m)</td>
<td>Is m capitalized?</td>
</tr>
<tr>
<td>MENTION_LEN(m)</td>
<td>The number of tokens in m</td>
</tr>
<tr>
<td>TITLE_LEN(e)</td>
<td>The number of tokens in e’s title</td>
</tr>
<tr>
<td>MENTION_IN_TITLE(m, e)</td>
<td>Does e’s title contain m?</td>
</tr>
<tr>
<td>TITLE_IN_MENTION(m, e)</td>
<td>Does m contain e’s title?</td>
</tr>
<tr>
<td>MENTION_EQ_TITLE(m, e)</td>
<td>Does e’s title exactly equal m?</td>
</tr>
<tr>
<td>EDIT_DISTANCE(m, e)</td>
<td>The similarity between m and e’s title</td>
</tr>
<tr>
<td>Mention occurrence features</td>
<td></td>
</tr>
<tr>
<td>SPREAD(m, D)</td>
<td>The number of tokens between the first and last m occurrences in D</td>
</tr>
<tr>
<td>FREQUENCY(m, D)</td>
<td>The number of times m appears in D</td>
</tr>
</tbody>
</table>


Link Probability Features

*Link probability* is the probability that a mention is used as an anchor text in Wikipedia

\[ LINK_PROB(m) = \frac{|C_a(m)|}{|C_t(m)|} \]

- \( C_a(m) \): A set of entities that contain mention as an anchor
- \( C_t(m) \): A set of entities that contain mention

- It reflects Wikipedia contributors’ judgment of whether or not the mention is helpful to readers
- Three derivative features:
  - \( LINK_PROB \): The link probability of the corresponding mention
  - \( LINK_PROB_AVG \): The average link probability of all mentions that link to the entity
  - \( LINK_PROB_MAX \): The maximum link probability in all mentions that link to the entity
Takeshita Street is a street lined with fashion boutiques, cafes in Harajuku in Tokyo, Japan.

Her public image is associated with Japan's kawaii culture centered in the Harajuku, Tokyo.

Department Store and Museum is a department store located in the Harajuku.

\[ LINK\_PROB(\text{Harajuku}) = \frac{2}{3} \]
Link Probability Features

*Link probability* is the probability that a mention is used as an anchor text in Wikipedia

\[
LINK\_PROB(m) = \frac{|C_a(m)|}{|C_t(m)|}
\]

- \(C_a(m)\): A set of entities that contain mention as an anchor
- \(C_t(m)\): A set of entities that contain mention

- It reflects Wikipedia contributors’ judgment of whether or not the mention is helpful to readers
- Three derivative features:
  - *LINK\_PROB*: The link probability of the corresponding mention
  - *LINK\_PROB\_AVG*: The average link probability of all mentions that link to the entity
  - *LINK\_PROB\_MAX*: The maximum link probability of all mentions that link to the entity
Entity Class Features

Entity class features are derived from the classes of an entity in DBpedia, Freebase, and Schema.org

Kyary Pamyu Pamyu
wikipedia/Kyary_Pamyu_Pamyu

DBpedia:/ontology/Person
DBpedia:/ontology/Artist
Freebase:/people/person
Freebase:/music/artist

Harajuku
wikipedia/Harajuku

DBpedia:/ontology/Place
Freebase:/location/location
Freebase:/location/neighborhood

Each class is used as a feature
Topical Coherence Feature

*Topical coherence feature* represents how an entity is related to the topics of the document

- Measured by how a target entity is related to other detected entities
- Relatedness between two entities is calculated based on the Wikipedia in-link-based similarity

\[
REL(e_1, e_2) = \frac{\log(\max(|c(e_1)|, |c(e_2)|) - \log(|c(e_1) \cap c(e_2)|))}{\log(|KB|) - \log(\min(|c(e_1)|, |c(e_2)|))}
\]

\(e\): Entity; \(KB\): Entities in KB; \(c(e)\): Entities having a link to \(e\)

- The average relatedness between the entity and other entities in the document [Milne 2008] is added as a feature

\[
COHERENCE(e, E) = \frac{\sum_{e' \in E'} REL(e, e')}{|E'|}
\]

\(e\): Entity; \(E\): Set of entities in the document
Kyary Pamyu Pamyu is a Japanese model and singer. Her public image is associated with Japan's kawaisa culture centered in the Harajuku, Tokyo.

\[
REL(Kyary Pamyu Pamyu, \text{kawaisa culture}) = 0.4 \\
REL(Kyary Pamyu Pamyu, \text{Harajuku}) = 0.2
\]

\[
COHERENCE(Kyary Pamyu Pamyu, E) = \frac{0.4 + 0.2}{2} = 0.3
\]
Dataset for Evaluating Helpfulness of Entities

The dataset for our task was developed using crowd-sourcing service

- Based on the *IITB EL dataset*
  - Contains annotated entities on news articles for entity linking tasks
  - Popular in entity linking studies
- We added an annotation to each entity in the IITB EL dataset that represents whether it is helpful to readers
- Amazon Mechanical Turk was used to hire annotators online
- A total number of 33,567 annotations were obtained
- The dataset is publicly available for further studies
  https://github.com/studio-ousia/el-helpfulness-dataset/
Classify a keyword in a document

We are developing a tool that converts notable keywords in a document into links that enable users to easily retrieve further information on keywords.

Question:
Do you think converting the highlighted keyword into a link is helpful, interesting, or relevant enough for readers of this document and could improve readers’ overall user experience?

This data will be used for improving the accuracy of the keyword detection of our tool.

Document

Scientists have produced a novel type of nanoparticle that they say could make it possible to dramatically increase magnetic-based data storage on future generations of computer hard drives. The researchers at Brown University and Sandia National Laboratories have announced new ways to create iron-platinum nanorods and nanowires for magnetic media. In doing so, the new materials could make possible devices that do not

They hope to make particles shaped less like wires and more like bricks, which align well but need further refinement to self-assemble in a fully parallel pattern.

Keyword

Brown University

Target Wikipedia article

Brown University

Brown University is an American private Ivy League research university located in Providence, Rhode Island. Founded in 1764 prior to American independence from the British Empire as the College in the English Colony of Rhode Island and Providence Plantations early in the reign of King George III (1760–1820), Brown is the third oldest institution of higher education in New England and seventh oldest in the United States. The university consists of The College, Graduate School, Alpert Medical School, and the School of Engineering.

Answer:

- Very helpful: Converting this keyword into a link is highly helpful for readers.
- Helpful: Converting this keyword into a link might be helpful for some readers.
- Rarely helpful: Converting this keyword into a link is rarely helpful for readers.
- Not helpful: Converting this keyword into a link is not helpful for readers.

Annotation screen displayed to the annotators
The Performance of Evaluating Helpfulness of Entities

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>LinkProb</td>
<td>0.6980</td>
<td>0.7761</td>
<td>0.7349</td>
</tr>
<tr>
<td>TAGME</td>
<td>0.7936</td>
<td>0.7782</td>
<td>0.7857</td>
</tr>
<tr>
<td>Milne&amp;Witten</td>
<td>0.8079</td>
<td>0.7904</td>
<td>0.7989</td>
</tr>
<tr>
<td>Our Method</td>
<td>0.8697</td>
<td>0.8419</td>
<td>0.8554</td>
</tr>
</tbody>
</table>

- Our method performs significantly better than existing methods by 5.7-12% F1
Rendering a Widget

- The widget content is retrieved from *Linked Open Data*
- The content is dynamically customized using entity classes
  - *Map* and *Images* for *Hollywood*
  - *YouTube* and *Twitter* for *Lady Gaga* (singer)
- *DBpedia Ontology Classes* are used as the entity classes
Linkify is implemented as add-ons of *desktop* & *mobile* web browsers
DEMONSTRATION
The entity linking system works in production service

TechCrunch: Linkify’s SDK Wants to Make Mobile Searches Less Cumbersome

Semanticweb.com: Linkify is Working to Make Mobile Searches Easier
THANK YOU!