Semantic Tagging and Text Classification for Archives and Content Production Systems

Chris Newell
• Lead R&D Engineer
• Natural Language Processing
• Internet Research and Future Services

15 June 2021
The BBC’s Royal Charter mandates “a centre of excellence” for research and development in the electronic distribution of audio-visual media

- 200 research engineers, scientists and designers
- Research Labs in London and Salford

Work on every aspect of broadcast and online services
- Production
- Distribution
- Audiences
SEMANTIC TAGGING AND TEXT CLASSIFICATION

Natural Language Processing at BBC R&D

- Semantic Tagging
- Text Classification
- Speech to Text
- Text Summarisation
- Sentiment Analysis
- Quote Extraction

Applied to:

- Web content
- News feeds
- TV subtitles
- Radio transcripts
- Social media
Description of a document in terms of unambiguous entities and themes

- Linked to additional sources of information
- Tagging can consist of either
  - all the entities and themes mentioned in the text
  - the primary topic(s) of the text
Motivation for Semantic Tagging

- **World Service Archive Project** (2011-2014)
- 70,000 radio programmes from over 60 years
- Metadata sparse, inaccurate or non-existent
- Efficient access and re-use of content is impossible without metadata
SEMANTIC TAGGING AND TEXT CLASSIFICATION

World Service Archive Approach

- Speech-to-text technology used to create transcripts
- Tags extracted using an automated tagging system called Mango
- Tagging supported search and navigation in prototype
- Crowd sourcing used to improve the tagging
- Mango is a semantic tagging system that identifies entities and themes e.g. for the text

  “The German chancellor, Angela Merkel called on the Russian president to exercise his influence on the separatists to enforce the ceasefire”
Mango is a semantic tagging system that identifies entities and themes e.g. for the text

“The German chancellor, Angela Merkel called on the Russian president to exercise his influence on the separatists to enforce the ceasefire”
Mango Overview

- Mango is a semantic tagging system that identifies entities and themes e.g. we want to identify:
  “The German chancellor, Angela Merkel called on the Russian president to exercise his influence on the separatists to enforce the ceasefire”

- Mango uses DBpedia identifiers:
  - http://dbpedia.org/resource/Angela_Merkel
  - http://dbpedia.org/resource/President_of_Russia
# SEMANTIC TAGGING AND TEXT CLASSIFICATION

**DBpedia entry for Angela_Merkel**

<table>
<thead>
<tr>
<th>Property</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>dbpedia-owl:birthDate</td>
<td>1954-07-17</td>
</tr>
<tr>
<td>dbpedia-owl:birthPlace</td>
<td>dbpedia:West_Germany</td>
</tr>
<tr>
<td></td>
<td>dbpedia:Hamburg</td>
</tr>
<tr>
<td>dbpedia-owl:alias</td>
<td>Merkel, Angela Dorothea (full name)</td>
</tr>
<tr>
<td>dbpedia-owl:almaMater</td>
<td>dbpedia:Leipzig_University</td>
</tr>
<tr>
<td>dbpedia-owl:office</td>
<td>Minister of the Environment</td>
</tr>
<tr>
<td></td>
<td>Chancellor of Germany</td>
</tr>
<tr>
<td></td>
<td>Leader of the Christian Democratic Union</td>
</tr>
<tr>
<td></td>
<td>Member of the Bundestag</td>
</tr>
<tr>
<td></td>
<td>Minister of Women and Youth</td>
</tr>
<tr>
<td>dbpedia-owl:party</td>
<td>dbpedia:Christian_Democratic_Union</td>
</tr>
</tbody>
</table>
Mango Components

- Mango has two components:
  - Spotter
  - Disambiguator

- Both created using a complete dump of Wikipedia
  - https://dumps.wikimedia.org/enwiki
The spotter uses an index which maps surface forms to entities.

For example:
- “Birmingham City F.C.”
- “Birmingham City”
- “Birmingham”
- “The Blues”

All map to:
- http://dbpedia.org/resource/Birmingham_City_F.C.
SEMANTIC TAGGING AND TEXT CLASSIFICATION

Ambiguity Problem

- However many surface forms are ambiguous
- “Birmingham” could refer to either:
  - http://dbpedia.org/resource/Birmingham_City_F.C.
  - http://dbpedia.org/resource/Birmingham
  - http://dbpedia.org/resource/Birmingham,_Alabama
  - http://dbpedia.org/resource/Birmingham_(horse)
  - etc...
SEMANTIC TAGGING AND TEXT CLASSIFICATION

Mango Schematic

Spotter

DBpedia Index

Candidate URIs

Disambiguator

word2vec Model
SEMANTIC TAGGING AND TEXT CLASSIFICATION

Word2Vec:
A vector space representation of words

/Germany
bundesrepublik
/Commission
/Europe
migration

/Al_Gore
election
president
/United_states
federal
america

ham
cheese
/Sandwich
bread

weather
climate
/Global_warming
/Fossil_fuel
SEMANTIC TAGGING AND TEXT CLASSIFICATION
Comparing Candidates against Documents

/Birmingham_City_F.C.

\[ \theta \]

Document vector
UK and US have an 'indestructible relationship', PM says. The alliance between the US and the UK should be known as the "indestructible relationship", Boris Johnson has told the BBC after meeting US President Joe Biden for the first time. He said he had "terrific" talks with Mr Biden, who has travelled to Cornwall for the G7 summit of world leaders. The summit begins later, with vaccines and climate change on the agenda.

<table>
<thead>
<tr>
<th>Surface form</th>
<th>Topic</th>
<th>Label</th>
<th>Score</th>
<th>Count</th>
<th>Types</th>
</tr>
</thead>
<tbody>
<tr>
<td>US President Joe Biden</td>
<td><a href="http://dbpedia.org/resource/Joe_Biden">http://dbpedia.org/resource/Joe_Biden</a></td>
<td>US President Joe Biden</td>
<td>0.6650</td>
<td>2</td>
<td>Agent, Person, Person, OfficeHolder</td>
</tr>
<tr>
<td>BBC</td>
<td><a href="http://dbpedia.org/resource/BBC">http://dbpedia.org/resource/BBC</a></td>
<td>Bbc</td>
<td>0.5612</td>
<td>1</td>
<td>Organization, Agent, Company, Organisation</td>
</tr>
<tr>
<td>vaccines</td>
<td><a href="http://dbpedia.org/resource/Vaccine">http://dbpedia.org/resource/Vaccine</a></td>
<td>Vaccines</td>
<td>0.5555</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>G7 summit</td>
<td><a href="http://dbpedia.org/resource/Group_of_Seven">http://dbpedia.org/resource/Group_of_Seven</a></td>
<td>G7 Summit</td>
<td>0.5262</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Cornwall</td>
<td><a href="http://dbpedia.org/resource/Cornwall">http://dbpedia.org/resource/Cornwall</a></td>
<td>Cornwall</td>
<td>0.5120</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>climate change</td>
<td><a href="http://dbpedia.org/resource/Climate_change">http://dbpedia.org/resource/Climate_change</a></td>
<td>Climate Change</td>
<td>0.5115</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Boris Johnson</td>
<td><a href="http://dbpedia.org/resource/Boris_Johnson">http://dbpedia.org/resource/Boris_Johnson</a></td>
<td>Boris Johnson</td>
<td>0.5091</td>
<td>1</td>
<td>Person, OfficeHolder, Agent, Person</td>
</tr>
</tbody>
</table>
SEMANTIC TAGGING AND TEXT CLASSIFICATION

World Service Archive
- BBC journalists manually tag every published article with their primary topics in our Content Production System.
- The tags affect where a story might appear.
- Automated tag suggestion is used to improve working efficiency and encourage consistent tagging.
Starfruit Tagging System

- Starfruit is an automated tag suggestion system for text
- Tags use a dictionary of terms called BBC Things
  - entities
  - themes
  - storylines
Starfruit Tagging System

- Starfruit uses a multilabel text classifier
  - Trained using 5 years of articles which have been manually tagged by BBC journalists
- Statistical approach using TF-IDF and Support Vector Machines
- 11k unique tags
- Retrained every 48 hours
### Topics Results

<table>
<thead>
<tr>
<th>Preferred Label</th>
<th>URI</th>
<th>Disambiguation Hint</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tokyo 2020</td>
<td><a href="http://www.bbc.co.uk/things/256b153b-5a34-4320-8c4a-a1cc8544d7fc#id">http://www.bbc.co.uk/things/256b153b-5a34-4320-8c4a-a1cc8544d7fc#id</a></td>
<td>Summer Olympics Competition, Olympic games</td>
<td>0.7848</td>
</tr>
<tr>
<td>Olympics</td>
<td><a href="http://www.bbc.co.uk/things/0d205538-c52e-41c7-a4a1-50ef6c1da59#id">http://www.bbc.co.uk/things/0d205538-c52e-41c7-a4a1-50ef6c1da59#id</a></td>
<td>Summer Olympics Recurring Competition</td>
<td>0.7660</td>
</tr>
<tr>
<td>Japan</td>
<td><a href="http://www.bbc.co.uk/things/3f53c272-5b8f-4a4f-99de-a857d6726c5b#id">http://www.bbc.co.uk/things/3f53c272-5b8f-4a4f-99de-a857d6726c5b#id</a></td>
<td>Country</td>
<td>0.7655</td>
</tr>
<tr>
<td>Asia</td>
<td><a href="http://www.bbc.co.uk/things/ba90754a-9033-4e9c-990b-d1139e5070a3#id">http://www.bbc.co.uk/things/ba90754a-9033-4e9c-990b-d1139e5070a3#id</a></td>
<td>world news region</td>
<td>0.5613</td>
</tr>
</tbody>
</table>
**SEMANTIC TAGGING AND TEXT CLASSIFICATION**

**Starfruit Mentions**
All entities and themes mentioned in the text

<table>
<thead>
<tr>
<th>Mentions Results</th>
<th>Preferred Label</th>
<th>URI</th>
<th>Disambiguation Hint</th>
<th>Surface Forms</th>
<th>Score</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tokyo 2020</td>
<td><a href="http://www.bbc.co.uk/things/256b153b-5a34-4320-8c4a-a1cc8544d7fc#id">http://www.bbc.co.uk/things/256b153b-5a34-4320-8c4a-a1cc8544d7fc#id</a></td>
<td>Summer Olympics Competition, Olympic games</td>
<td>Tokyo 2020</td>
<td>0.7848</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Olympics</td>
<td><a href="http://www.bbc.co.uk/things/0d205538-c52e-41c7-a4a1-50ef6bc1da59#id">http://www.bbc.co.uk/things/0d205538-c52e-41c7-a4a1-50ef6bc1da59#id</a></td>
<td>Summer Olympics Recurring Competition</td>
<td>Olympics, Olympic</td>
<td>0.7660</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Japan</td>
<td><a href="http://www.bbc.co.uk/things/3f53c272-5b8f-4a4f-99de-a857e6726c5b#id">http://www.bbc.co.uk/things/3f53c272-5b8f-4a4f-99de-a857e6726c5b#id</a></td>
<td>Country</td>
<td>Japan</td>
<td>0.7655</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Sexism</td>
<td><a href="http://www.bbc.co.uk/things/ea10d9be-b2fs-46e5-9711-183b6480c675#id">http://www.bbc.co.uk/things/ea10d9be-b2fs-46e5-9711-183b6480c675#id</a></td>
<td>Discrimination based on a person’s sex or gender.</td>
<td>sexist</td>
<td>0.4930</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Figure Skating</td>
<td><a href="http://www.bbc.co.uk/things/485f9f4-f87e-4513-b133-46a5dfcc4d28#id">http://www.bbc.co.uk/things/485f9f4-f87e-4513-b133-46a5dfcc4d28#id</a></td>
<td>Sports discipline</td>
<td>figure skating</td>
<td>0.4803</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>The Football Association</td>
<td><a href="http://www.bbc.co.uk/things/d51567ad-4fc4-475b-8642-d8e9be47fb6#id">http://www.bbc.co.uk/things/d51567ad-4fc4-475b-8642-d8e9be47fb6#id</a></td>
<td>Governing body for English football</td>
<td>Football Association</td>
<td>0.4463</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Harassment</td>
<td><a href="http://www.bbc.co.uk/things/abf2a87d-ad5b-4fb-9497-c99a53edc24b#id">http://www.bbc.co.uk/things/abf2a87d-ad5b-4fb-9497-c99a53edc24b#id</a></td>
<td>Behaviour</td>
<td>harassment</td>
<td>0.3515</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

**Tokyo 2020: Japan Olympics minister Seiko Hashimoto appointed head of Games**

Japan’s Seiko Hashimoto has been appointed the next Tokyo 2020 president, after her predecessor quit over sexist comments he made.

The former Olympics Minister is a seven time Olympian herself, having competed as a cyclist and a speed skater.

Former chief Yoshiro Mori resigned following a backlash, after he was quoted as saying women talk too much.
## SEMANTIC TAGGING AND TEXT CLASSIFICATION

### Starfruit Applications

<table>
<thead>
<tr>
<th>User</th>
<th>System</th>
<th>Application</th>
</tr>
</thead>
<tbody>
<tr>
<td>BBC News</td>
<td>Content Production System</td>
<td>News Articles</td>
</tr>
<tr>
<td>Content Distribution</td>
<td>Content Enrichment Chassis</td>
<td>iPlayer Content</td>
</tr>
<tr>
<td>BBC News</td>
<td>Wolftech</td>
<td>News Planning</td>
</tr>
</tbody>
</table>
We did a piece with Sally Cornes from Chorley last autumn. She was “the dancer with cancer” she was ace. Her parents were amazing too. She died last week. She was about to start a clinical trial involving only 50 kids across Europe two days before NYE, but she fell ill on Xmas eve. Her parents wonder if we’d like to interview them & film her funeral today to raise continued importance of research into childhood cancers at a time when funding is plummeting. She had a very active tic-toc account where she posted lots of content. Stuff on Jupiter too.
SEMANTIC TAGGING AND TEXT CLASSIFICATION

Alternative Tagging Systems

- DBpedia Spotlight  (open source)
- Tagmatic  (adaptive)
- Amazon Comprehend
- Google Natural Language
- Microsoft Azure NLP
- etc

(with usual disclaimer)
WOLFTECH STORY TAGGING

Thank you

Chris Newell
chris.newell@bbc.co.uk

Further Information
https://www.bbc.co.uk/rd/projects/natural-language-processing