Subject prediction using semantic embedding

Rob Koopman and Shenghui Wang
OCLC EMEA
Agenda

1. Introduction: semantic embedding
2. Ariadne random projection
3. Automatic subject assignment
4. Dataset
5. Evaluation
Introduction: Semantic embedding

- Statistical Semantics [furnas1983, weaver1955] based on the assumption of “a word is characterized by the company it keeps” [firth1957]
- Distributional Hypothesis [harris1954, sahlbrown2008]: words that occur in similar contexts tend to have similar meanings

- Word embedding: words are represented in a continuous vector space where semantically similar words are mapped to nearby points (‘are embedding nearby each other’)
- Two main categories of approaches: global co-occurrence count-based methods (e.g. LSA) vs local context predictive methods (e.g. word2vec)
- A desirable property: computable similarity
Ariadne random projection

- Each entity is embedded as a 256-byte vector
- Each document is embedded as the weighted average of word embeddings
- Cosine similarity reflects semantic similarity

C: a co-occurrence matrix
R: a random matrix of +/-1
C’: approximation of C after random projection
Automatic subject prediction

- Our hypothesis: A document is more likely to be indexed with subjects that are most related to it.
- Can embedding-based similarities help us to find suitable subjects?
Experiments:

- Astro dataset: 111k articles published in 59 Astronomy and Astrophysics journals (Downloaded from [http://www.topic-challenge.info/](http://www.topic-challenge.info/))
- 95% for training, 5% for testing
- The training set contains 18791 different subjects on average 9 per article.
- For each testing document, we compute a list of most related subjects
- Measure precision/recall at N
Results:
Actual vs predicted

Laboratory Detection of FeCO+ (X 4Σ−) by Millimeter/Submillimeter Velocity Modulation Spectroscopy

The millimeter/submillimeter spectrum of the molecular ion FeCO+ (X 4Σ−) has been recorded using velocity modulation spectroscopy. The molecular ion was created in an AC discharge of Fe(CO)₅ and argon. Twenty-seven rotational transitions, each consisting of four fine-structure components, were measured in the range 198-418 GHz. The data were fit with a case b Hamiltonian, and rotational, spin-rotation, and spin-spin constants were determined. Because of the presence of higher order spin-orbit interactions, probably caused in part by a nearby 4Π excited state, numerous centrifugal distortion terms were needed for the spectral analysis. The value of γₚ, the third-order spin-rotation constant, was also remarkably large at -72.4 MHz. Rest frequencies for FeCO+ are now available for interstellar and circumstellar searches. This species may be present in molecular clouds, where CO is abundant and gas-phase iron should be in the form of Fe+. Molecular ions such as FeCO+ could be the hidden carriers of metallic elements in such clouds.
## Actual vs predicted

<table>
<thead>
<tr>
<th>Actual</th>
<th>Cosine</th>
<th>Predicted (top 10)</th>
<th>Cosine</th>
</tr>
</thead>
<tbody>
<tr>
<td>astrochemistry</td>
<td>0.5091</td>
<td>ism:molecules</td>
<td>0.6223</td>
</tr>
<tr>
<td>interstellar</td>
<td>0.1801</td>
<td>astrochemistry</td>
<td>0.5091</td>
</tr>
<tr>
<td>ism:molecules</td>
<td>0.6223</td>
<td>chemistry</td>
<td>0.5010</td>
</tr>
<tr>
<td>line:identification</td>
<td>0.3815</td>
<td>molecular data</td>
<td>0.4847</td>
</tr>
<tr>
<td>chemistry</td>
<td>0.5010</td>
<td>irc+10216</td>
<td>0.4644</td>
</tr>
<tr>
<td>envelope</td>
<td>0.3356</td>
<td>ism:abundances</td>
<td>0.4639</td>
</tr>
<tr>
<td>hydrocarbons</td>
<td>0.0662</td>
<td>radio lines:ism</td>
<td>0.4613</td>
</tr>
<tr>
<td>methods:laboratory</td>
<td>0.3394</td>
<td>clouds</td>
<td>0.4115</td>
</tr>
<tr>
<td>molecular data</td>
<td>0.4847</td>
<td>rotational excitation</td>
<td>0.4062</td>
</tr>
<tr>
<td>stars:agb and post agb</td>
<td>0.2518</td>
<td>molecular processes</td>
<td>0.3946</td>
</tr>
</tbody>
</table>
Information retrieval using subjects

- Make an embedding of the human assigned subjects and the top 9 machine assigned subjects of a record in the test set.
- Try to find the records in the data set.

<table>
<thead>
<tr>
<th></th>
<th>First result</th>
<th>&lt;=10</th>
<th>&lt;=20</th>
<th>&lt;=30</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>11%</td>
<td>41%</td>
<td>53%</td>
<td>60%</td>
</tr>
<tr>
<td>Machine</td>
<td>7%</td>
<td>33%</td>
<td>47%</td>
<td>56%</td>
</tr>
</tbody>
</table>
Conclusions

- Humans are on average a bit better in finding the right mix of subject headings than our algorithm.
- In the case of an astronomy paper it is not so easy to judge whether subject headings are correct.
  - In cases where the machine is better than the human our first test is a harsh judge.
  - When the algorithm wrong in the second test the embedding can still be somewhat reasonable.
- Automatic subject assignments can clearly help the user.
- Our algorithm is in general not capable to find all subject headings.
What’s next

Deep Learning for Extreme Multi-label Text Classification
Jingzhou Liu, Wei-Cheng Chang, Yuexin Wu, Yiming Yang

But our method is orders of magnitudes faster ...