Subject prediction using semantic embedding

Rob Koopman and Shenghui Wang
OCLC EMEA

Agenda

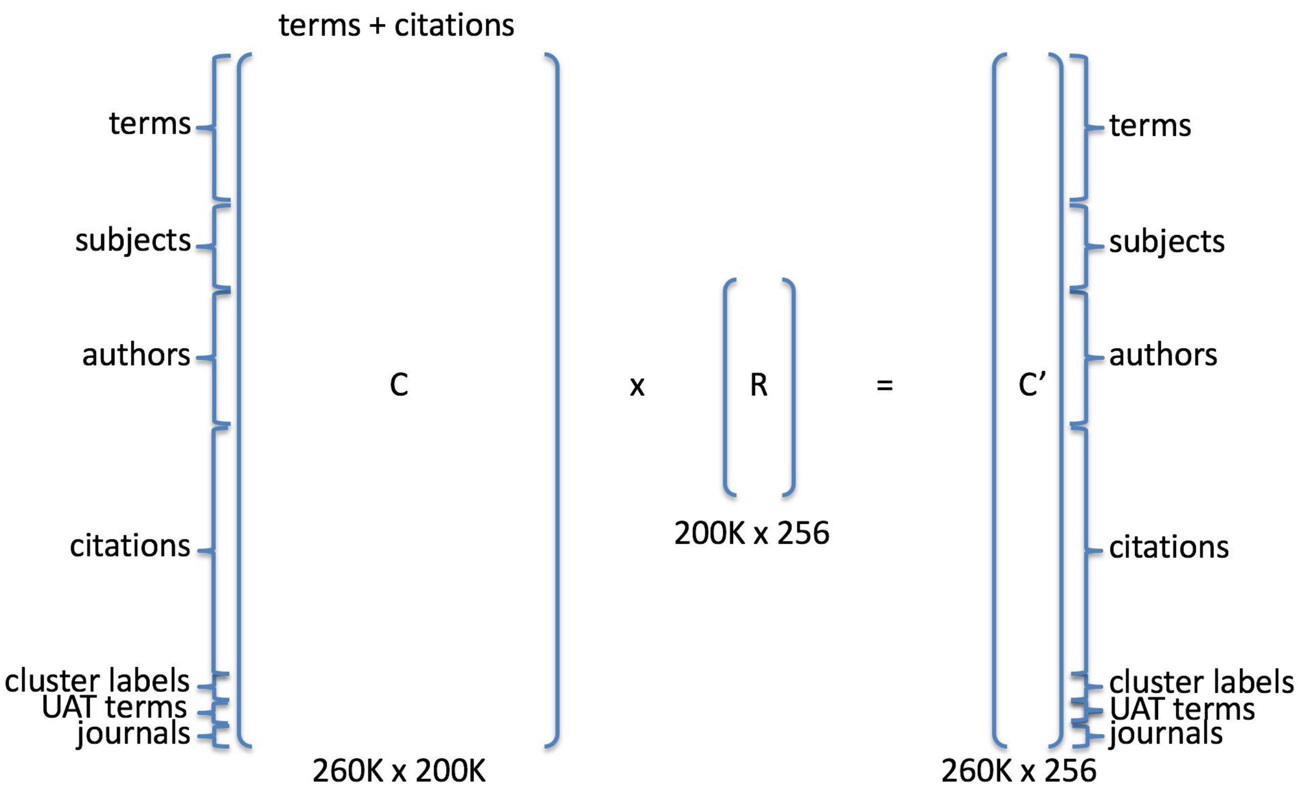
1. Introduction: semantic embedding
2. Ariadne random projection
3. Automatic subject assignment
4. Dataset
5. Evaluation

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, sahlgren2008]: 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



C: a co-occurrence matrix

R: a random matrix of +/-1

C': approximation of C
after 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

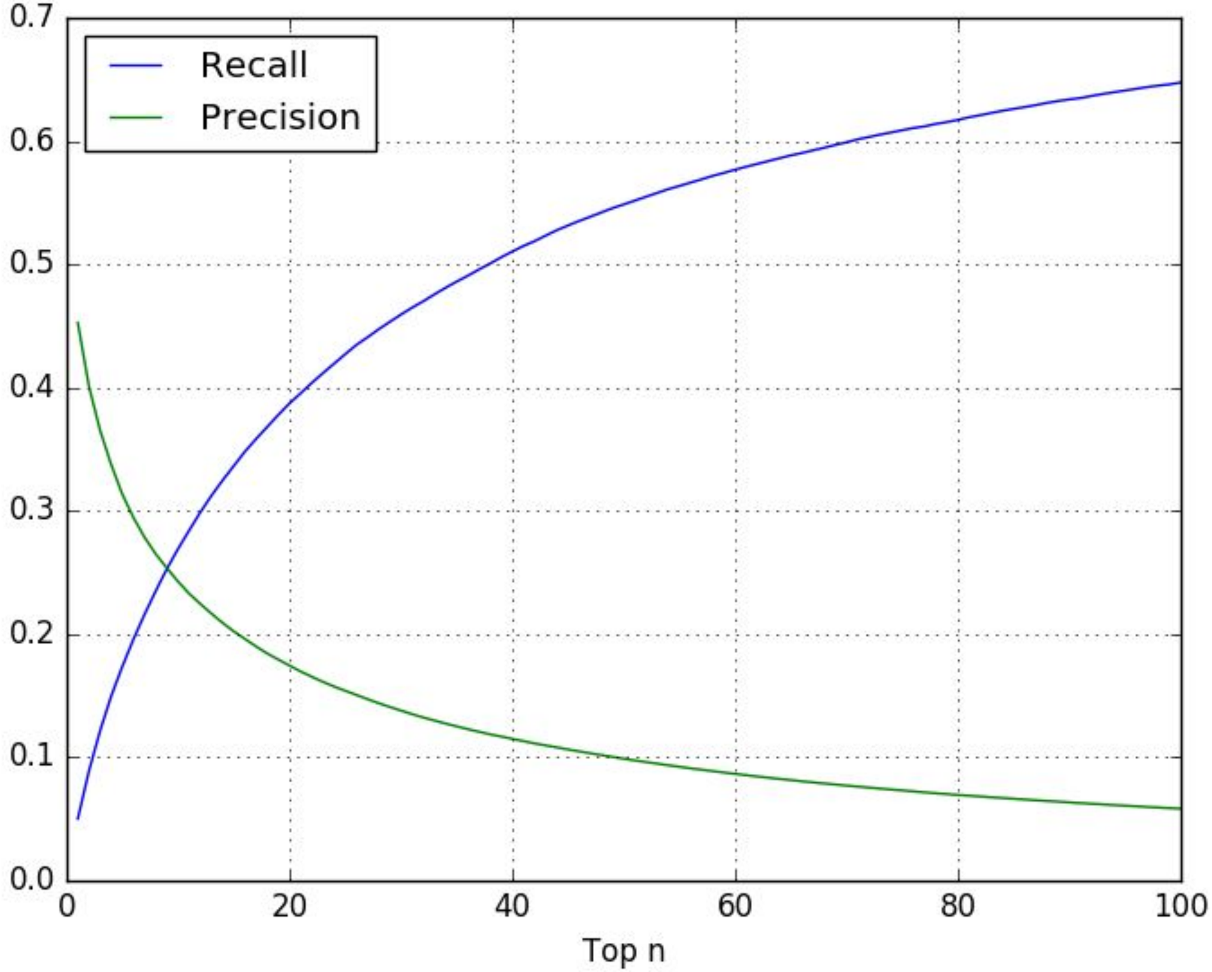
Automatic subject prediction

- 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/)
- 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* $^4\Sigma^-$) by Millimeter/Submillimeter Velocity Modulation Spectroscopy

The millimeter/submillimeter spectrum of the molecular ion $FeCO^+$ ($X\ ^4\Sigma^-$) 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 $^4\Pi$ excited state, numerous centrifugal distortion terms were needed for the spectral analysis. The value of γ_s, 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

Actual	Cosine	Predicted (top 10)	Cosine
astrochemistry	0.5091	**ism:molecules**	0.6223
interstellar	0.1801	**astrochemistry**	0.5091
ism:molecules	0.6223	**chemistry**	0.5010
line:identification	0.3815	**molecular data**	0.4847
chemistry	0.5010	irc+10216	0.4644
envelope	0.3356	ism:abundances	0.4639
hydrocarbons	0.0662	radio lines:ism	0.4613
methods:laboratory	0.3394	clouds	0.4115
molecular data	0.4847	rotational excitation	0.4062
stars:agb and post agb	0.2518	molecular processes	0.3946

Information retrieval using subjects

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

	First result	<=10	<=20	<=30
Human	11%	41%	53%	60%
Machine	7%	33%	47%	56%

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 were 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 ...