Freebase Annotation and Deep Learning Embeddings

Yadollah Yaghoobzadeh (yadollah@cis.lmu.de), Hinrich Schuetze
Center for Information and Language Processing, LMU University of Munich

Motivation

- Freebase is a manually created knowledge base
- Deep learning embeddings are automatically learned representations for words and other linguistic units
- Goal: Explore benefits of combining "top-down" knowledge from Freebase with "bottom-up" deep learning

Approaches

- Approach 1: Learn embeddings for mention strings, e.g. , “Reading F.C.”
  - Problem: Ambiguity and Synonymy
- Approach 2: Learn embeddings for entity-annotated text, e.g., m/016gp5 (Reading F.C.).
  - Problem: requires well performing entity tagger
- Approach 3: Hybrid/iterative approach combining approaches 1 & 2
  - Problem: Finding different subsets of entities that are proper for each approach 1 or 2

Applications

- Improve entity tagging by using deep learning embeddings (in Approaches 1 or 3) as additional features
- Improve precision: below we look at a subclass of entities with low precision
- Improve recall: richer context representation due to embeddings may improve recall
- Predict properties of entities that are not represented in Freebase.

Entity/Non-entity Ambiguity

- Analysis of roughly 0.01% of FACC1 annotations
- Identified all mention strings with the following properties
  - Mention string is ambiguous
  - One of the entities occurs at least 10 times
  - At least 1/3 of instances of mention string are lowercase in Wikipedia
- There were 423 such single word mention strings
  - Examples: Deviant (member of fictional race), Target (retailing company), Western (film genre)
  - Examples of phrases: Free Press (newspaper), Top Ten (musical genre), The Museum (TV series)
- Conclusion: Entity/non-entity ambiguity is a fairly common problem that needs to be solved to achieve good entity disambiguation

Extraction of Entity and Word Embeddings

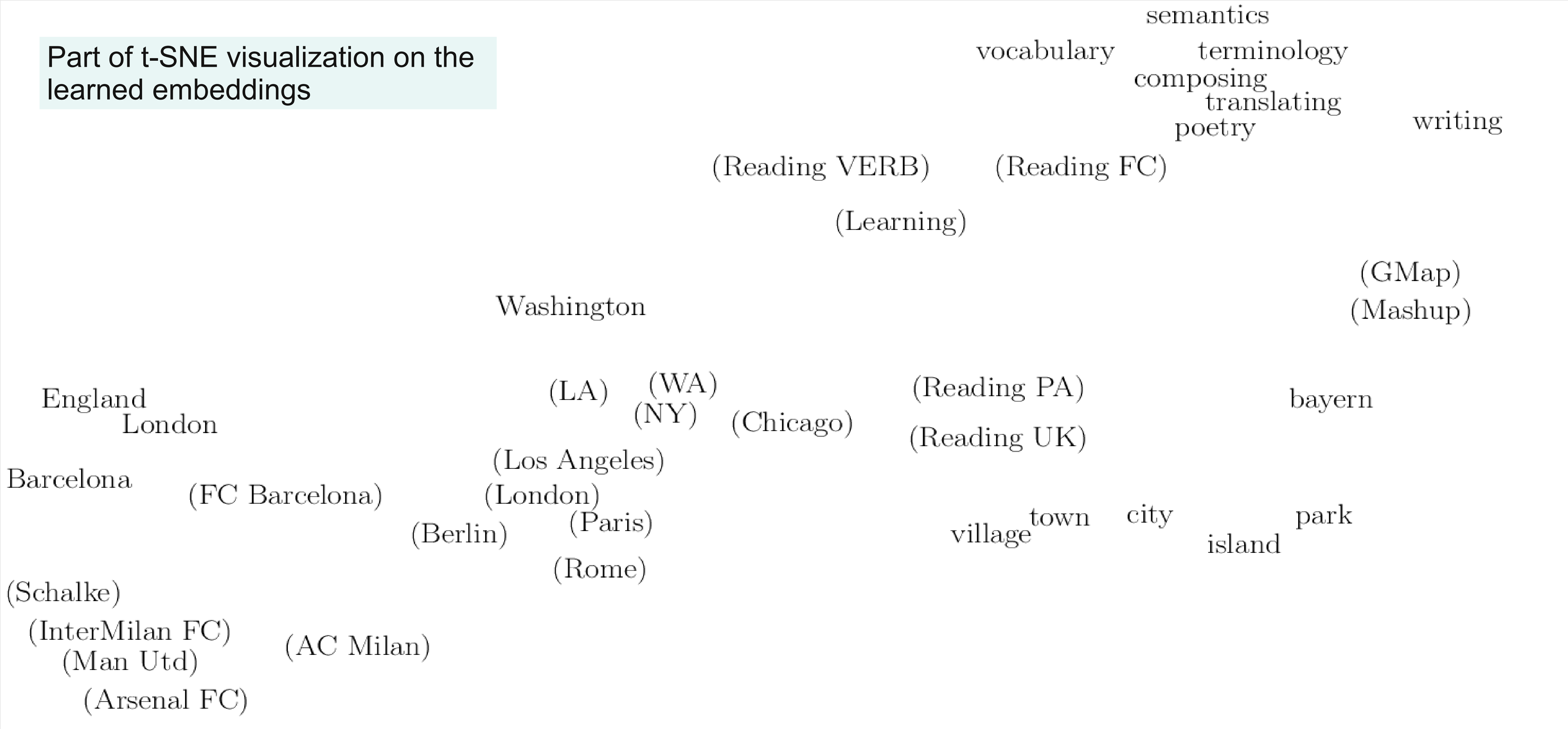
- Creating an entity annotated corpus by merging FACC1 with ClueWeb09
- Computing embeddings for each token in the corpus using word2vec Skip-gram model (Mikolov et al., 2013)
  - 100 dimensional vectors
  - 4.5GB corpus with 453,641 entities and 4,533,988 words
  - Running time: about 8 hours with 20 threads on a 32 cores Intel Xeon CPU 2.27GHz

Table 1: Some “Reading” Entities in Freebase

Freebase MID	Notable Type	Freebase name	What is it?
/m/0b_yz	City/Town/Village	Reading	Large town in England
/m/016gp5	Football team	Reading F.C.	Football Club in England
/m/04yclrg	Quotation Subject	Reading	Verbal Noun of “to read”
/m/0zlgm	City/Town/Village	Reading	a city in Pennsylvania

Table 2: Nearest Neighbors of Some Entities

(Reading, F.C.)	(Reading, VERB)	(Reading, PA)	(Google Map)
reading	(Learning)	(DeKalb, IL)	Mashup
book	(Reading F.C.)	(MorganTown, VA)	iGoogle
(Young Adult Literature)	Syllabi	(WytheCountry, VA)	(Google App Engine)
writing	(Education)	(Edgefield, USA)	(Mediawiki)
(Reading, VERB)	(Writing)	(Northwest Ohio)	(User Interface)


Part of t-SNE visualization on the learned embeddings

Simple Experiment 1 (Manual Labels)

- Classification of the “Reading, VERB” from other “Reading” entities
- Hand annotating 58 sentences
- Computing embedding of each sentence by summing over embeddings of its contained words
- Cross validating a linear SVM
- Accuracy: 79 % vs 69%

Simple Experiment 2 (No Manual Labels)

- Classification of the “Reading, VERB” from “Reading, PA”
- Training a linear SVM on 3500 sentences that contain Freebase entities with notable types “city” or “Quotation Subject”
- Test-set with 57 instances
- Computing the embeddings of sentences by summing over embeddings of its contained words and entities
- Accuracy: 87% vs 76%

Future Work

- Should the Freebase entities considered for entity disambiguation be restricted to certain types?
  - E.g., exclude “Literature / TV / Quotation / Periodical Subject” (the types of m/04yclrg reading). Develop criteria.
- Alternative: joint entity and word sense disambiguation
- Systematic approach to entity/non-entity disambiguation
- Continuous space language modeling for entity disambiguation.
  - E.g., the contexts "He had a few games with ***" vs "He had a few games in ***" are sports club vs location contexts.
- Use deep learning representations for predicting properties of entities that are not listed in Freebase, e.g., "does it snow in X?"