Embeddings for KB and text representation, extraction and question answering.

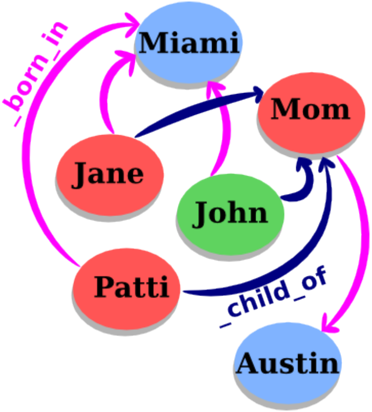
Jason Weston† & Antoine Bordes & Sumit Chopra
Facebook AI Research

External Collaborators:
Alberto Garcia-Duran & Nicolas Usunier & Oksana Yakhnenko

† Some of this work was done while J. Weston worked at Google.

Multi-relational data

- Data is structured as a graph
- Each node = an entity
- Each edge = a relation/fact
- A relation = (*sub*, *rel*, *obj*):
  - *sub* =*subject*,
  - *rel* = *relation type*,
  - *obj* = *object*.
- Nodes w/o features.



We want to also link this to text!!

Embedding Models

KBs are hard to manipulate

- Large dimensions: $10^5/10^8$ entities, $10^4/10^6$ rel. types
- Sparse: few valid links
- Noisy/incomplete: missing/wrong relations/entities

Two main components:

1. Learn low-dimensional vectors for **words** and KB **entities** and **relations**.
2. **Stochastic gradient** based training, *directly trained to define a similarity criterion of interest.*

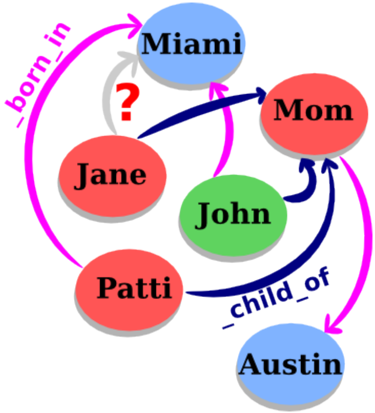
Link Prediction

Add new facts without requiring extra knowledge

From known information, assess the validity of an unknown fact

Goal: We want to model, from data,

$$\mathbb{P}[rel_k(sub_i, obj_j) = 1]$$

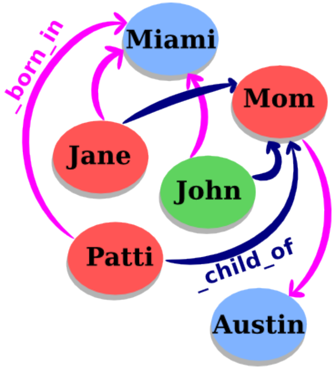


Previous Work

- Tensor factorization (Harshman et al., '94)
- Probabilistic Relational Learning (Friedman et al., '99)
- Relational Markov Networks (Taskar et al., '02)
- Markov-logic Networks (Kok et al., '07)
- Extension of SBMs (Kemp et al., '06) (Sutskever et al., '10)
- Spectral clustering (undirected graphs) (Dong et al., '12)
- Ranking of random walks (Lao et al., '11)
- Collective matrix factorization (Nickel et al., '11)
- Embedding models (Bordes et al., '11, '13) (Jenatton et al., '12) (Socher et al., '13) (Wang et al., '14) (García-Durán et al., '14)

Modeling Relations as Translations (Bordes et al. '13)

Intuition: we want $\mathbf{s} + \mathbf{r} \approx \mathbf{o}$.



Modeling Relations as Translations (Bordes et al. '13)

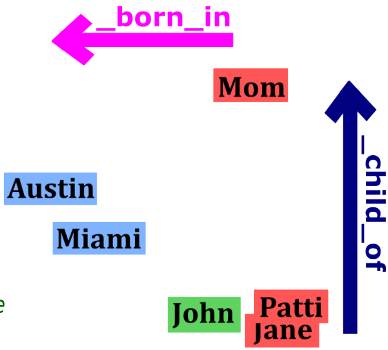
Intuition: we want $\mathbf{s} + \mathbf{r} \approx \mathbf{o}$.

The similarity measure is defined as:

$$d(sub, rel, obj) = ||\mathbf{s} + \mathbf{r} - \mathbf{o}||_2^2$$

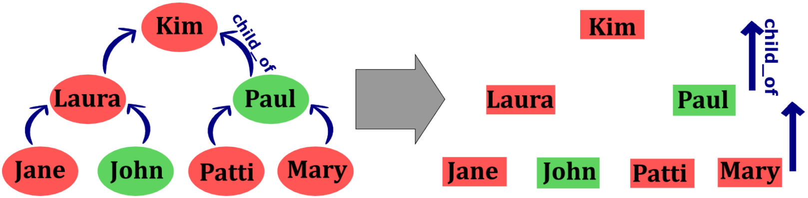
s,**r** and **o** are learned to verify that.

We use a ranking loss whereby true triples are higher ranked.



Motivations of a Translation-based Model

- Natural representation for hierarchical relationships.



- Word2vec word embeddings (Mikolov et al., '13):
  there may exist embedding spaces in which relationships among concepts are represented by translations.

Chunks of Freebase

- **Data statistics**:

	Entities (n_e)	Rel. (n_r)	Train. Ex.	Valid. Ex.	Test Ex.
FB13	75,043	13	316,232	5,908	23,733
FB15k	14,951	1,345	483,142	50,000	59,071
FB1M	1×10^6	23,382	17.5×10^6	50,000	177,404

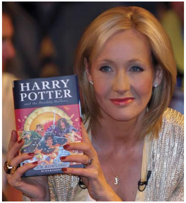
- **Training times for** TransE:
  - Embedding dimension: 50.
  - Training time:
    - on Freebase15k: ≈2h (on 1 core),
    - on Freebase1M: ≈1d (on 16 cores).

Example

"Who influenced J.K. Rowling?"

J. K. Rowling `_influenced_by`

- G. K. Chesterton
- J. R. R. Tolkien
- C. S. Lewis
- Lloyd Alexander
- Terry Pratchett
- Roald Dahl
- Jorge Luis Borges
- Stephen King
- Ian Fleming



Green=Train Blue=Test Black=Unknown

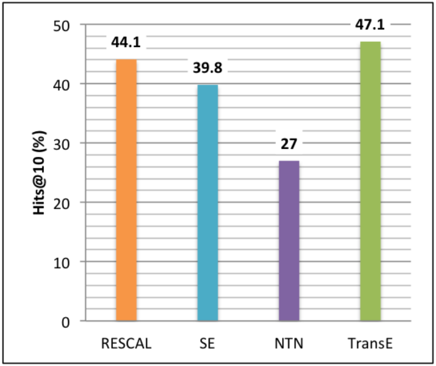
Example

"Which genre is the movie WALL-E?"

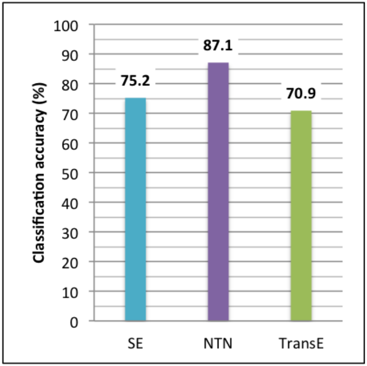
WALL-E	`_has_genre`	Animation
		Computer animation
		Comedy film
		Adventure film
		Science Fiction
		Fantasy
		Stop motion
		Satire
		Drama

Benchmarking

Ranking on FB15k



Classification on FB13



On FB1M,TransE predicts 34% in the Top-10 (SE only 17.5%).

Results extracted from (Bordes et al., '13) and (Wang et al., '14)

Refining TransE

- TATEC (García-Durán et al., '14) supplements TransE with a trigram term for encoding complex relationships:

$$d(sub, rel, obj) = \overbrace{\mathbf{s}_1^\top \mathbf{R}\mathbf{o}_1}^{\text{trigram}} + \overbrace{\mathbf{s}_2^\top \mathbf{r} + \mathbf{o}_2^\top \mathbf{r}' + \mathbf{s}_2^\top \mathbf{D}\mathbf{o}_2}^{\text{bigrams}\approx\text{TransE}},$$

with $\mathbf{s}_1 \neq \mathbf{s}_2$ and $\mathbf{o}_1 \neq \mathbf{o}_2$.

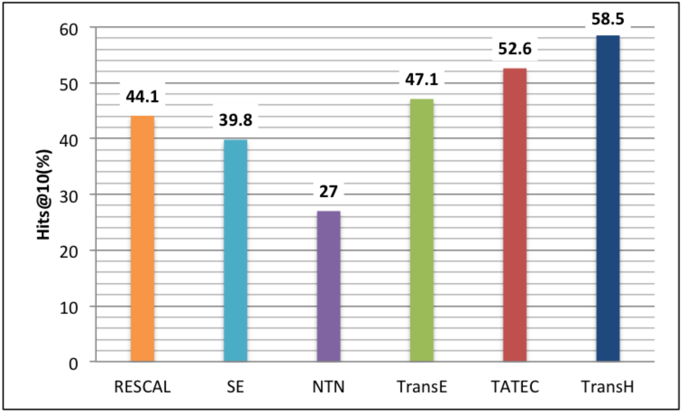
- TransH (Wang et al., '14) adds an orthogonal projection to the translation of TransE:

$$d(sub, rel, obj) = ||(\mathbf{s} - \mathbf{r}_p^\top \mathbf{s}\mathbf{r}_p) + \mathbf{r}_t - (\mathbf{o} - \mathbf{r}_p^\top \mathbf{o}\mathbf{r}_p)||_2^2,$$

with $\mathbf{r}_p \perp \mathbf{r}_t$.

Benchmarking

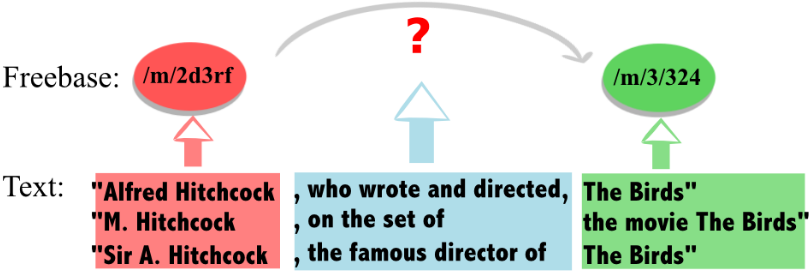
Ranking on FB15k



Results extracted from (García-Durán et al., '14) and (Wang et al., '14)

Relation Extraction

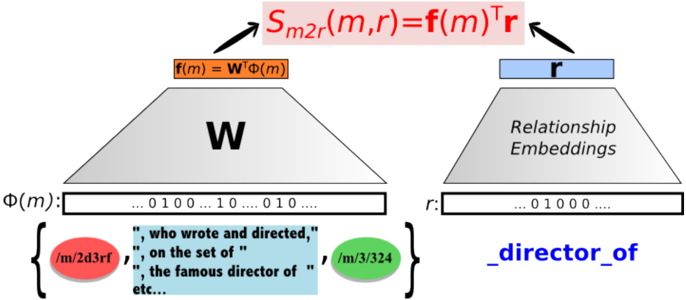
Goal: Given a bunch of sentences concerning the same entity pair, identify relations (if any) between them to add to the KB.



Embeddings of Text and Freebase (Weston et al., '13)

- **Basic Method:** an embedding-based classifier is trained to predict the relation type, given text mentions $\mathcal{M}$ and (sub, obj):

$$r(m, sub, obj) = \arg\max_{rel'} \sum_{m \in \mathcal{M}} S_{m2r}(m, rel')$$



Classifier based on WSABIE (Weston et al., '11).

Embeddings of Text and Freebase (Weston et al., '13)

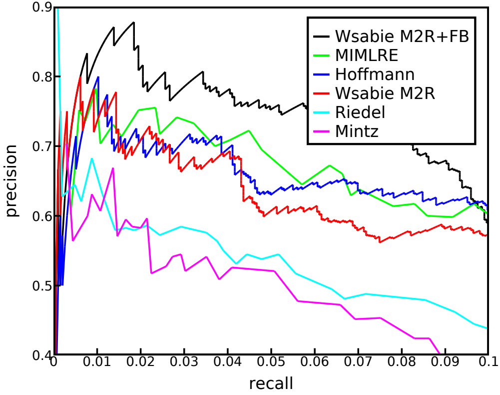
- **Idea:** improve extraction by using both text + available knowledge (= current KB).
- A model of the KB used to help extracted relations agree with it:

$$r'(m, sub, obj) = \arg\max_{rel'} \Big(\sum_{m \in \mathcal{M}} S_{m2r}(m, rel') - d_{KB}(sub, rel', obj) \Big)$$

with $d_{KB}(sub, rel', obj) = ||\mathbf{s} + \mathbf{r}' - \mathbf{o}||_2^2$

Benchmarking on NYT+Freebase

Exp. on NY Times papers linked with Freebase (Riedel et al., '10)



Precision/recall curve for predicting relations

A new embedding method, Wang et al., EMNLP'14, now beats these.

Open-domain Question Answering

- **Open-domain Q&A**: answer question on any topic
  $\longrightarrow$ query a KB with natural language

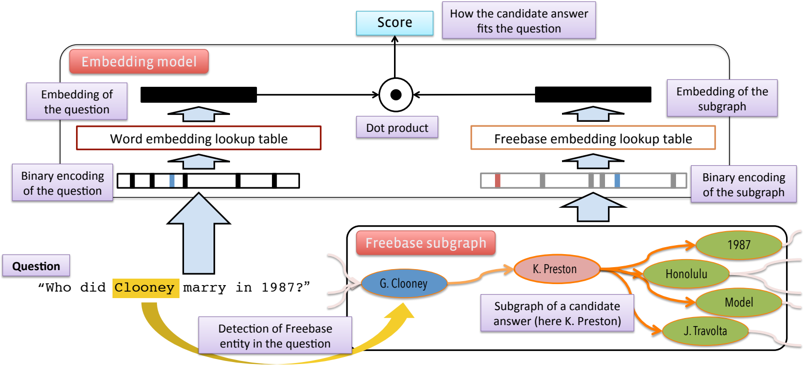
Examples

"What is `cher`'s son's name ?"	`elijah_blue_allman`
"What are `dollars` called in `spain` ?"	`peseta`
"What is `henry_clay` known for ?"	`lawyer`
"Who did `georges_clooney` marry in `1987` ?"	`kelly_preston`

- Recent effort with semantic parsing (Kwiatkowski et al. '13) (Berant et al. '13, '14) (Fader et al., '13, '14) (Reddy et al., '14)
- Models with embeddings as well (Bordes et al., '14)

Subgraph Embeddings (Bordes et al., '14)

- Model learns embeddings of questions and (candidate) answers
- Answers are represented by entity and its neighboring subgraph



Training data

- Freebase is automatically converted into Q&A pairs
- Closer to expected language structure than triples

Examples of Freebase data

(sikkim, location.in_state.judicial_capital, gangtok)
what is the judicial capital of the in state sikkim ? – gangtok

(brighouse, location.location.people_born_here, edward_barber)
who is born in the location brighouse ? – edward_barber

(sepsis, medicine.disease.symptoms, skin_discoloration)
what are the symptoms of the disease sepsis ? – skin_discoloration

Training data

- All Freebase questions have rigid and similar structures
- Supplemented by pairs from clusters of paraphrase questions
- Multitask training: similar questions $\leftrightarrow$ similar embeddings

Examples of paraphrase clusters

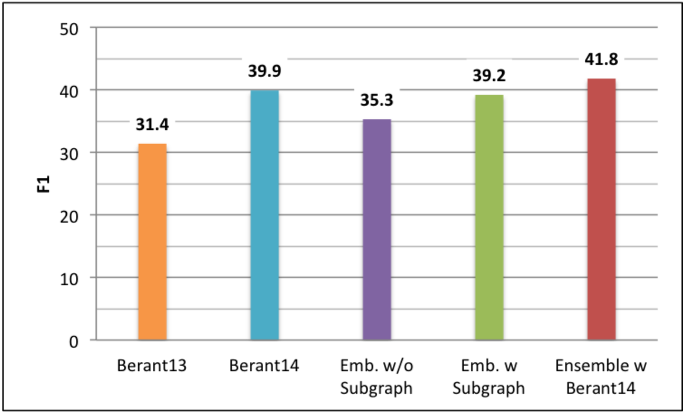
what are two reason to get a 404 ?
what is error 404 ?
how do you correct error 404 ?

what is the term for a teacher of islamic law ?
what is the name of the religious book islam use ?
who is chief of islamic religious authority ?

what country is bueno aire in ?
what countrie is buenos aires in ?
what country is bueno are in ?

Benchmarking on WebQuestions

Experiments on WebQuestions (Berant et al., '13)



F1-score for answering test questions

New result: Wang et al. reports 45.3 on same data.

Conclusion

- Embeddings are efficient features for many tasks in practice
- Training with SGD scales & parallelizable (Niu et al., '11)
- Flexible to various tasks: multi-task learning of embeddings
- Supervised or unsupervised training
- Allow to use extra-knowledge in other applications

Current limitations

- Compression: improve the memory capacity of embeddings and allow for one-shot learning of new symbols
- Beyond linear: most supervised labeling problems are well tackled by simple linear models. Non-linearity should help more.