Intelligent Agents

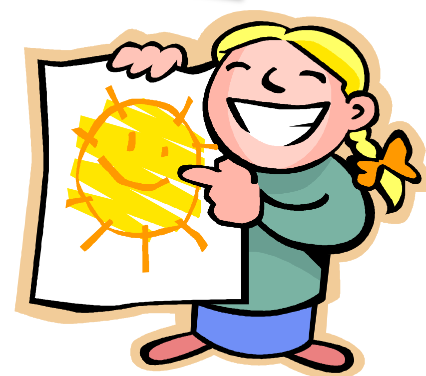
Multi-Relational Latent Semantic Analysis

Ralf Möller
Universität zu Lübeck
Institut für Informationssysteme

Semantics Needs More Than Similarity

Tomorrow will be rainy.

Tomorrow will be sunny.



$\mathit{similar}$(r, sunny)?

$\mathit{antonym}$(rainy, sunny)?

Leverage Linguistic Knowledge

- Can’t we just use the existing thesauri for information about synonyms and antonyms?
  - Knowledge in these resources is never complete
  - Often lack of “membership degree” for relations
    - Various ways to measure “membership degree”

- Goal: Create a continuous semantic representation that
  - leverages existing rich linguistic resources,
  - discovers new relations, and
  - enables us to measure the “degree” of multiple relations (not just similarity)

Roadmap

- Two opposite relations:
  - Polarity Inducing Latent Semantic Analysis
- Multiple relations:
  - Multi-Relational Latent Semantic Analysis
- Relational domain knowledge

- Yih, Zweig & Platt. *Polarity Inducing Latent Semantic Analysis*. In EMNLP-CoNLL-12.
- Chang, Yih & Meek. *Multi-Relational Latent Semantic Analysis*. In EMNLP-13.
- Chang, Yih, Yang & Meek. *Typed Tensor Decomposition of Knowledge Bases for Relation Extraction*. In EMNLP-14.

EMNLP: Empirical Methods in Natural Language Processing
CoNLL: Computational Natural Language Learning
ACL; Annual Meeting of the Association for Computational Linguistics

Problem: Handling Two Opposite Relations

- Can cope to some extent with homonyms and synonyms due to word context
- Embedding techniques cannot clearly distinguish antonyms
  - "Distinguishing synonyms and antonyms is still perceived as a difficult open problem." [Poon & Domingos 09]

- Idea #1: Change the data representation

Synonyms: Different words, same meaning
Homonyms: Same words, different meanings
Antonyms: Opposite words

Hoifung Poon and Pedro Domingos. Unsupervised semantic parsing. In Proceedings EMNLP '09. **2009**.

Polarity Inducing LSA

- Data representation
  - Encode two opposite relations in a matrix using "polarity"
    - Synonyms & antonyms (e.g., from a thesaurus)
- Factorization
  - Apply SVD to the matrix to find latent components
- Measuring degree of relation
  - Cosine of latent vectors

Wen-tau Yih, Geoffrey Zweig, John Platt. Polarity Inducing Latent Semantic Analysis. In Proceedings EMNLP '12. **2012**.

Encode Synonyms & Antonyms in Matrix

- Joyfulness: joy, gladden; sorrow, sadden
- Sad: sorrow, sadden; joy, gladden

Target word: row-vector

	joy	gladden	sorrow	sadden	goodwill
Group 1: "joyfulness"	1	1	1	1	0
Group 2: "sad"	1	1	1	1	0
Group 3: "affection"	0	0	0	0	1

Encode Synonyms & Antonyms in Matrix

- Joyfulness: joy, gladden; sorrow, sadden
- Sad: sorrow, sadden; joy, gladden

Inducing polarity

Target word: row-vector

	joy	gladden	sorrow	sadden	goodwill
Group 1: “joyfulness”	1	1	-1	-1	0
Group 2: “sad”	-1	-1	1	1	0
Group 3: “affection”	0	0	0	0	1

Cosine Score: $+\,Synonyms$

Encode Synonyms & Antonyms in Matrix

- Joyfulness: joy, gladden; sorrow, sadden
- Sad: sorrow, sadden; joy, gladden

Inducing polarity

Target word: row-vector

	joy	gladden	sorrow	sadden	goodwill
Group 1: "joyfulness"	1	1	-1	-1	0
Group 2: "sad"	-1	-1	1	1	0
Group 3: "affection"	0	0	0	0	1

Cosine Score: $-$ *Antonyms*

Problem: How to Handle More Relations?

- Limitation of the matrix representation
  - Each entry captures a particular type of relation between two entities, or
  - Two opposite relations with the polarity trick
- Encoding other binary relations
  - Is-A (hyponym) – ostrich *is a* bird
  - Part-whole – engine is a *part of* car
- Idea #2
  - Encode multiple relations in a 3-way tensor (3-dim array)!

M. Nickel, V. Tresp, and H.-P. Kriegel. A three-way model for collective learning on multi-relational data. In Proceedings of the 28th International Conference on International Conference on Machine Learning, ICML'11, pages 809–816, **2011**.

Encode Multiple Relations in Tensor

- Represent word relations using a tensor
  - Each slice encodes a relation between terms and target words.

	joy	gladden	sadden	feeling
joyfulness	1	1	0	0
gladden	1	1	0	0
sad	0	0	1	0
anger	0	0	0	0

Synonym layer

	joy	gladden	sadden	feeling
joyfulness	0	0	0	0
gladden	0	0	1	0
sad	1	0	0	0
anger	0	0	0	0

Antonym layer

Construct a tensor with two slices

Encode Multiple Relations in Tensor

- Can encode multiple relations in the tensor

1	1	0	0
1	1	0	0
0	0	1	0
0	0	0	0

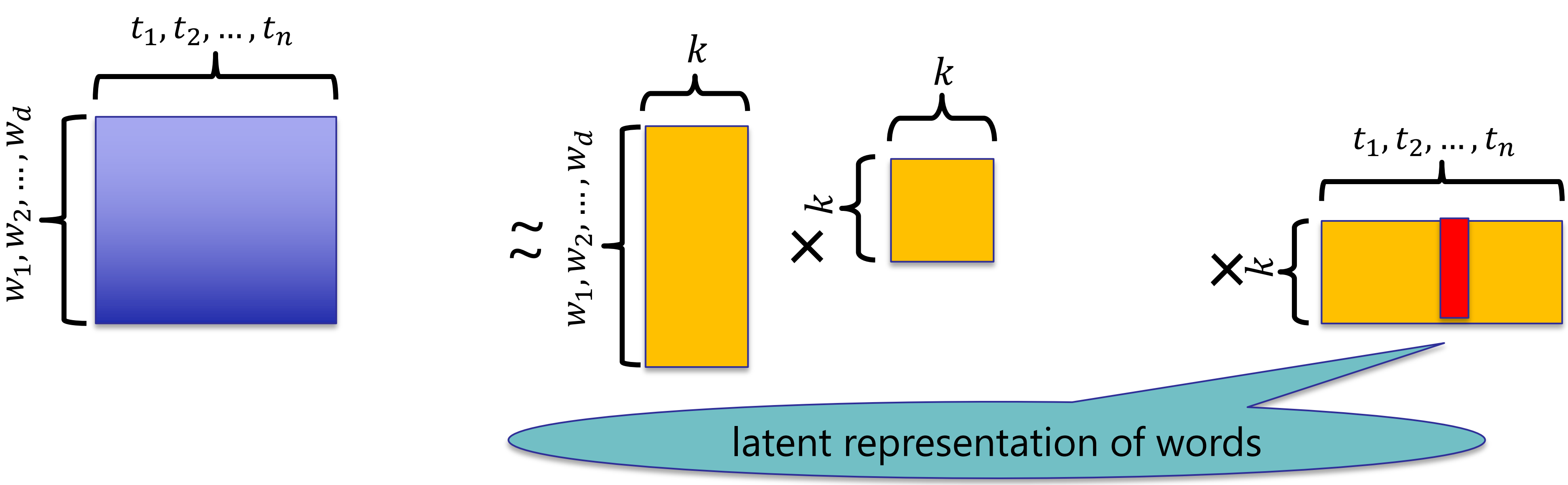
	joy	gladden	sadden	feeling
joyfulness	0	0	0	1
gladden	0	0	0	0
sad	0	0	0	1
anger	0	0	0	1

Hyponym layer

Hyponym IS-A/TYPE-OF hypernym
Metonym: Substitute for another term
(substitute usually used for sth else)

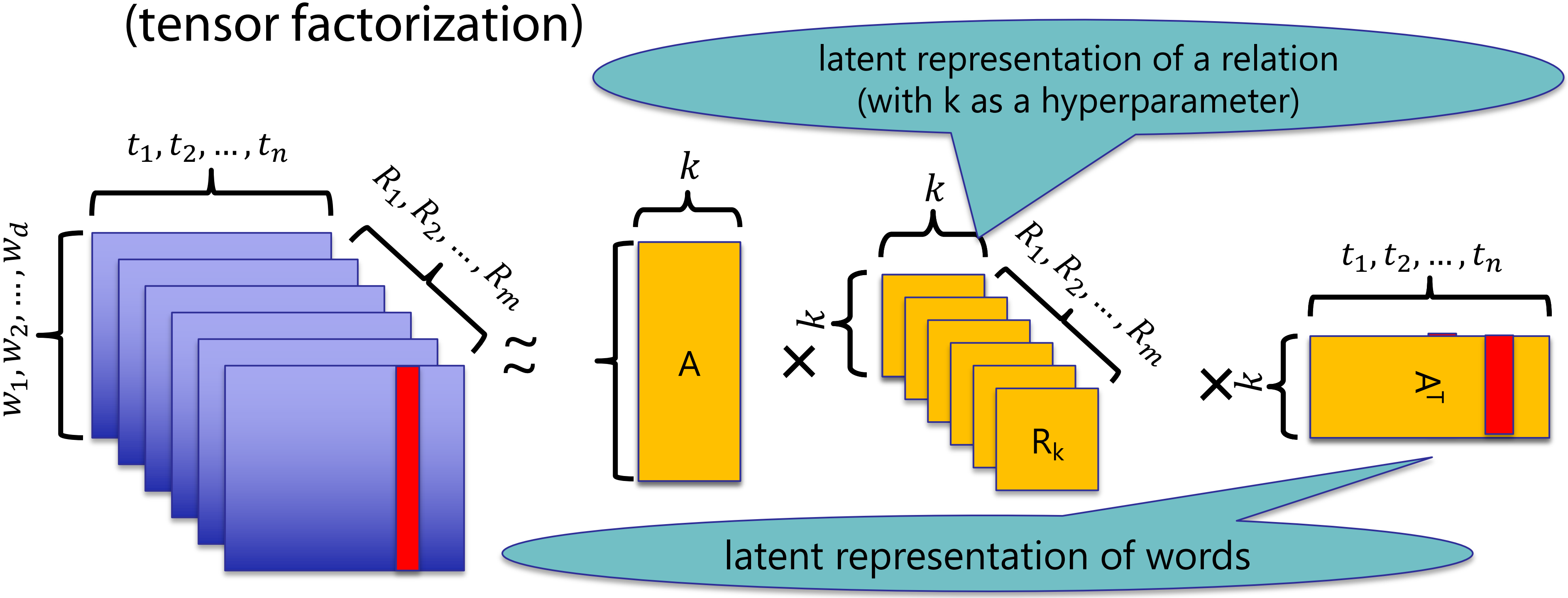
Tensor Decomposition – Analogy to SVD

- Derive a low-rank approximation to generalize the data and to discover unseen relations
- SVD



Tensor Decomposition – Analogy to SVD

- Derive a low-rank approximation to generalize the data and to discover unseen relations
- Apply Tucker decomposition and reformulate the results (tensor factorization)



Ledyard R. Tucker. "Some mathematical notes on three-mode factor analysis". Psychometrika. 31 (3): 279–311, **1966**.

Measure Degree of Relation

- Similarity
  - Cosine of the latent vectors
- Other relations (both symmetric and asymmetric)
  - Take the latent matrix of the *pivot* relation (synonym)
  - Take the latent matrix of the relation
  - Cosine of the latent vectors after projection

Measure Degree of Relation: Raw Representation

- $ant(\text{joy}, \text{sadden}) = \cos(\boldsymbol{w}_{:\text{joy},syn}, \boldsymbol{w}_{:,\text{sadden},ant})$

	joy	gladden	sadden	felling
joyfulness	**1**	**1**	**0**	**0**
gladden	**1**	**1**	**0**	**0**
sad	**0**	**0**	**1**	**0**
anger	**0**	**0**	**0**	**0**

Synonym layer

	joy	gladden	sadden	felling
joyfulness	**0**	**0**	**0**	**0**
gladden	**0**	**0**	**1**	**0**
sad	**1**	**0**	**0**	**0**
anger	**0**	**0**	**0**	**0**

Antonym layer

Measure Degree of Relation: Raw Representation

- $ant(\text{joy}, \text{sadden}) = \cos(\boldsymbol{w}_{:,\text{joy},syn}, \boldsymbol{w}_{:,\text{sadden},ant})$

	joy	gladden	sadden	felling
joyfulness	1	1	0	0
gladden	1	1	0	0
sad	0	0	1	0
anger	0	0	0	0

Synonym layer

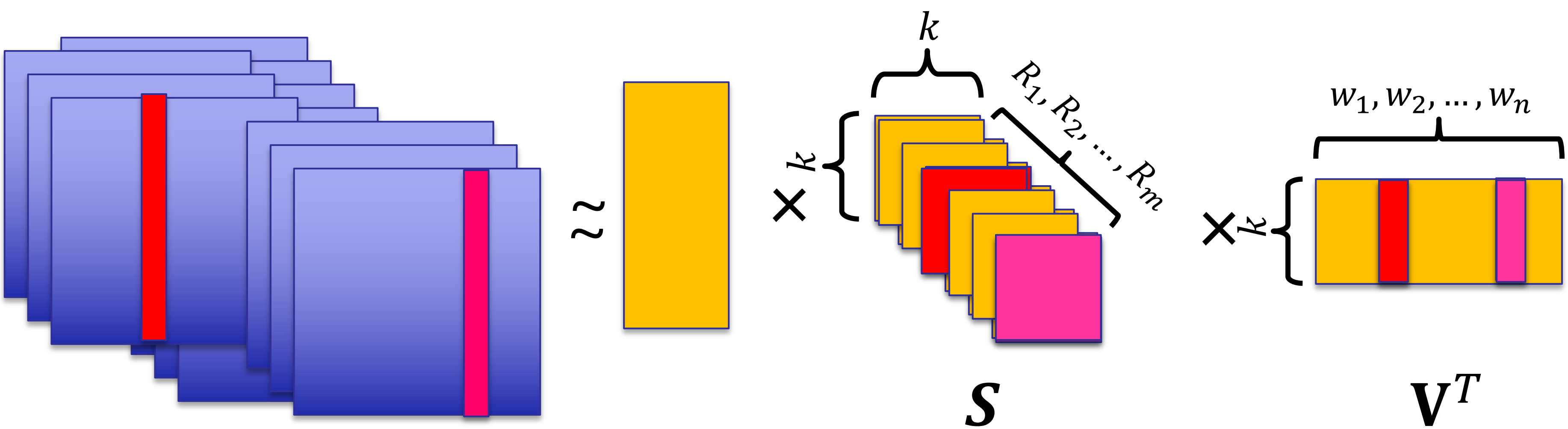
	joy	gladden	sadden	felling
joyfulness	0	0	0	0
gladden	0	0	1	0
sad	1	0	0	0
anger	0	0	0	0

Antonym layer

Measure Degree of Relation: Latent Representation

- $rel(\mathrm{w}_i, \mathrm{w}_j) = \cos\left(\boldsymbol{S}_{:,:,syn}\mathbf{V}_{i,:}^T, \boldsymbol{S}_{:,:,rel}\mathbf{V}_{j,:}^T\right)$

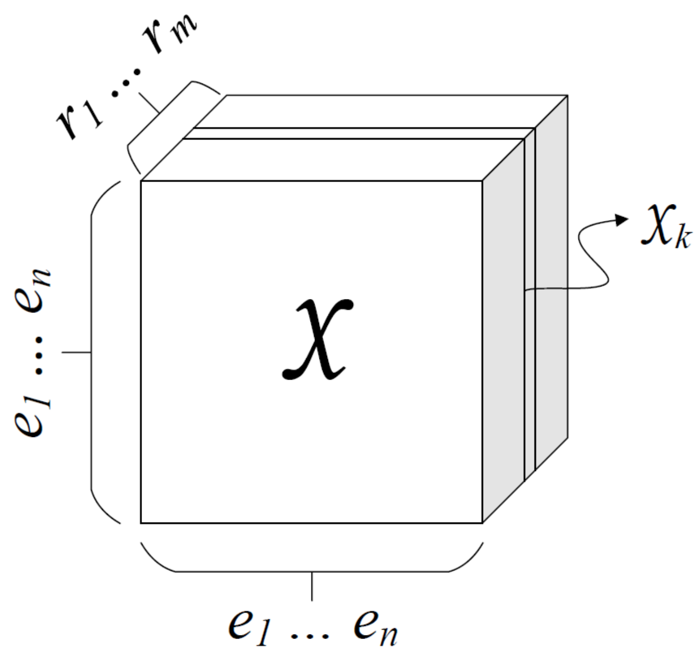
$Cos\ (\quad \times \quad , \quad \times \quad)$



Knowledge Graphs (1/2)

- Collection of subj-pred-obj triples – (e_1, r, e_2)

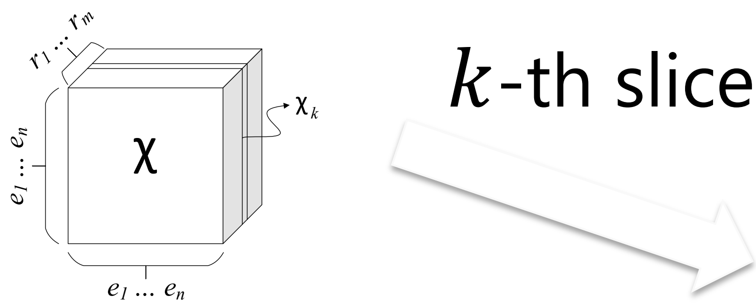
Subject	Predicate	Object
Obama	Born-in	Hawaii
Bill Gates	Nationality	USA
Bill Clinton	Spouse-of	Hillary Clinton
Satya Nadella	Work-at	Microsoft
…	…	…

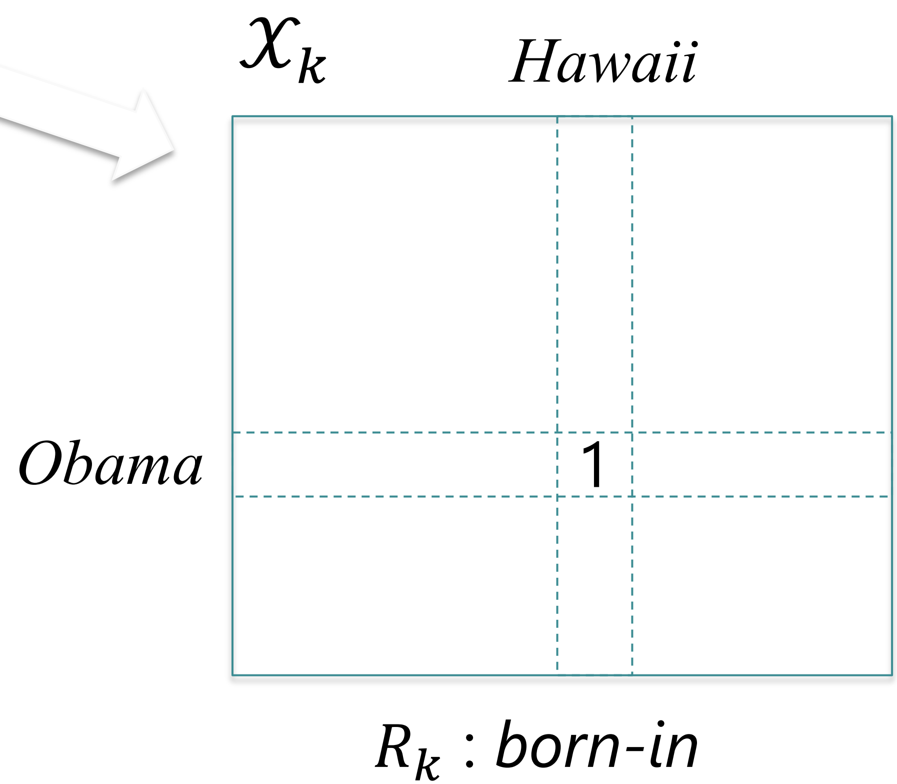


n: # entities, m: # relations

M. Nickel, V. Tresp, and H.-P. Kriegel. A three-way model for collective learning on multi-relational data. In Proceedings of the 28th International Conference on International Conference on Machine Learning, ICML'11, pages 809–816, **2011**.

Knowledge Graphs (2/2)

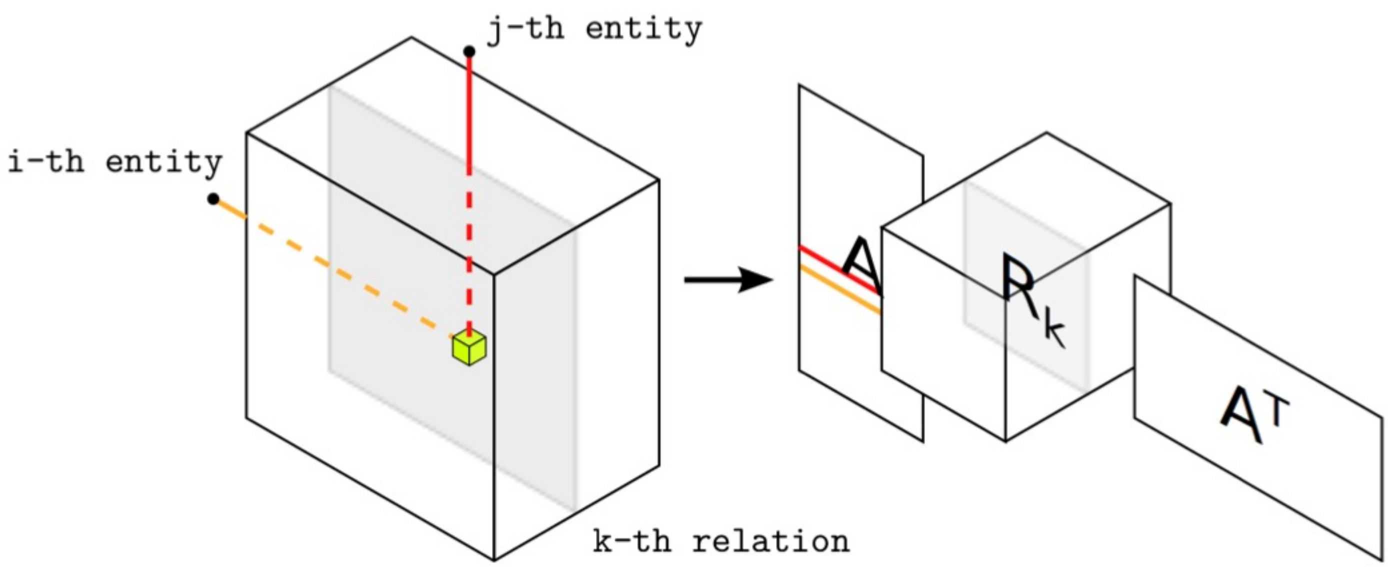




A 0 entry means:
- Incorrect (*false*)
- Unknown

M. Nickel, V. Tresp, and H.-P. Kriegel. A three-way model for collective learning on multi-relational data. In Proceedings of the 28th International Conference on International Conference on Machine Learning, ICML'11, pages 809–816, **2011**.

Factorization



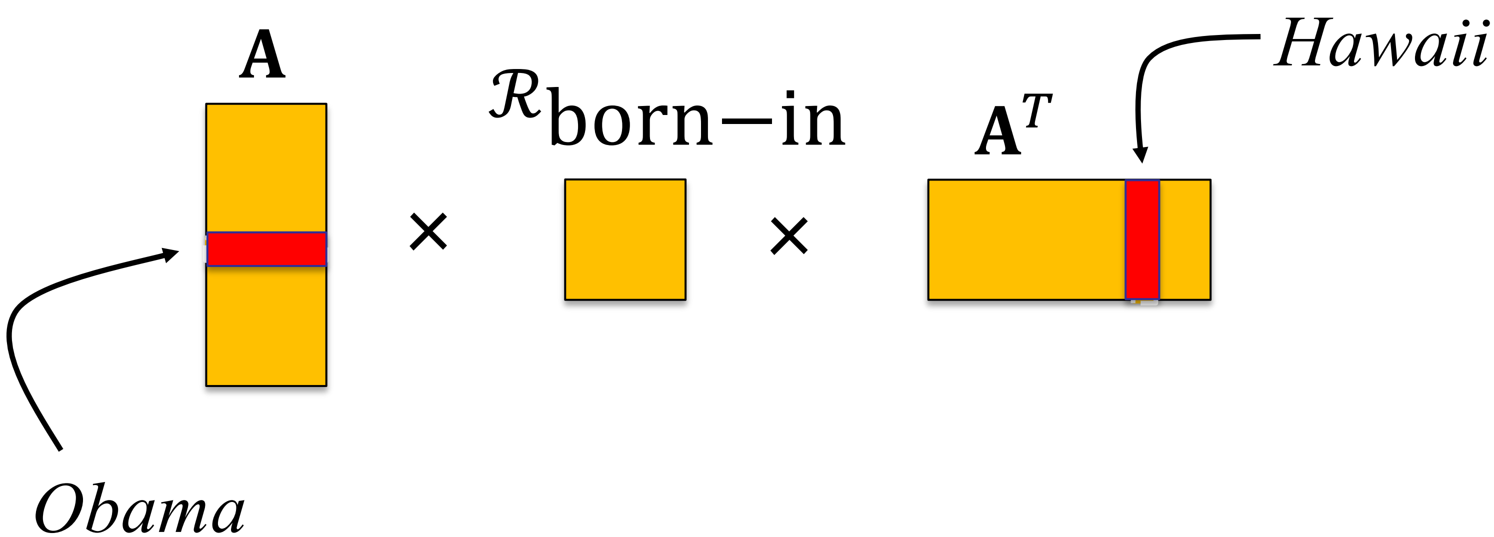
M. Nickel, V. Tresp, and H.-P. Kriegel. A three-way model for collective learning on multi-relational data. In Proceedings of the 28th International Conference on International Conference on Machine Learning, ICML'11, pages 809–816, **2011**.

Measure the Degree of a Relationship

$$f_{\text{born-in}}(\text{Obama}, \text{Hawaii})$$

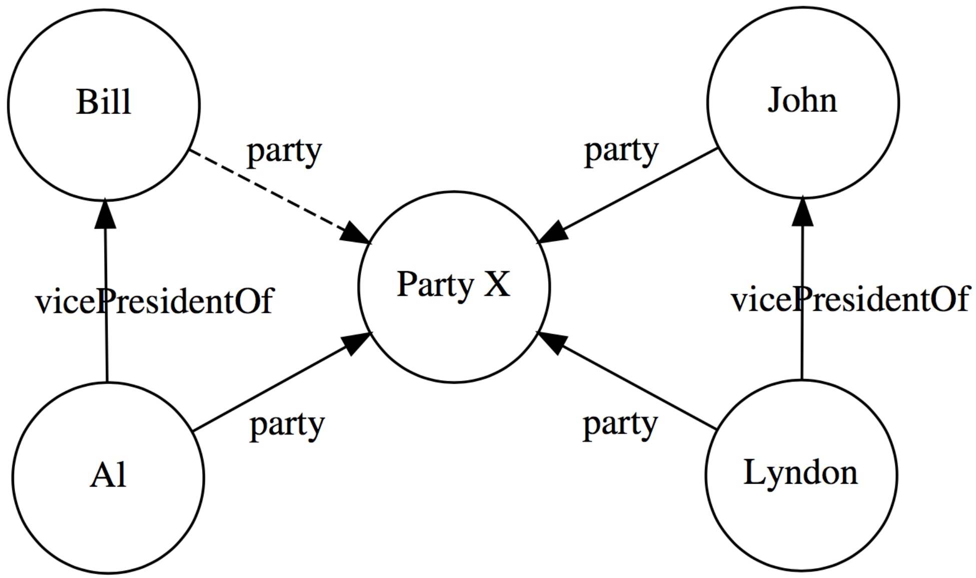
$$=$$

$$\mathbf{A}_{\text{Obama},:}\ \mathcal{R}_{\text{born-in}}\ \mathbf{A}^{\mathrm{T}}_{\text{Hawaii},:}$$



Prediction of Unknown Facts

- Predict party membership of US (vice) presidents



Prediction of unknown fact party(Bill, Party X)

Problem: Relational Domain Knowledge

- Relational data – the entity type
  - Relation can only hold between the right types of entities
    - Words having *is-a* relation have the same part-of-speech
    - For relation *born-in*, the entity types are: (person, location)
- Leverage type information to improve multi-relational LSA

Chang, Kai-Wei & Yih, Wen-tau & Yang, Bishan & Meek, Chris. Typed Tensor Decomposition of Knowledge Bases for Relation Extraction. In: Proc. EMNLP-14. **2014**.

Knowledge Graph?

🤔 Where is the knowledge in a knowledge graph (KG) ?

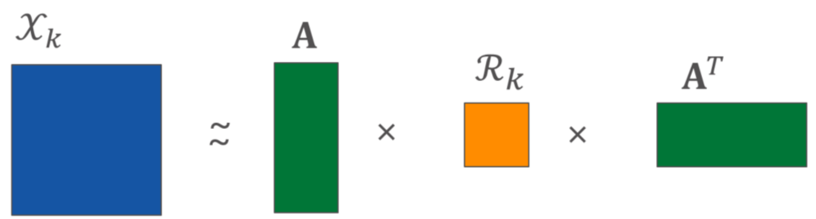
- Queries as with SQL database
- Embedding approaches rank existence of tuples
- Thresholds difficult to specify
- Use top-k queries with ranking w.r.t. score to establish existence of relations (or links)
- Want as many “true” tuples as possible in the answer set
  - Standard evaluation measures: Precision and Recall
- But applications may treat all query answers as true answers
- No uncertainty about answers to queries

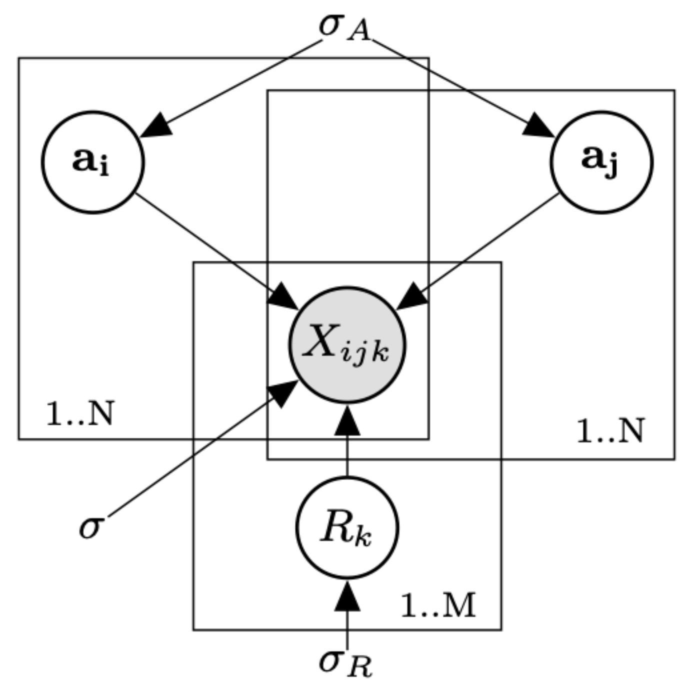
😉 Probabilistic database with open-world assumption ?

- ➢ Course “Non-standard Databases and Data Mining”
- ➢ But: Want sparsity (or “tuples computed on demand”)

RESCAL: Graphical Model in Plate Notation

- Tensor factorization can be seen as a probabilistic model
  - Specified here in plate notation
- With appropriate CPTs, queries for the distribution $\mathrm{P}(R(e_i, e_j))$ can be answered
- Can be used for prediction of unknown facts



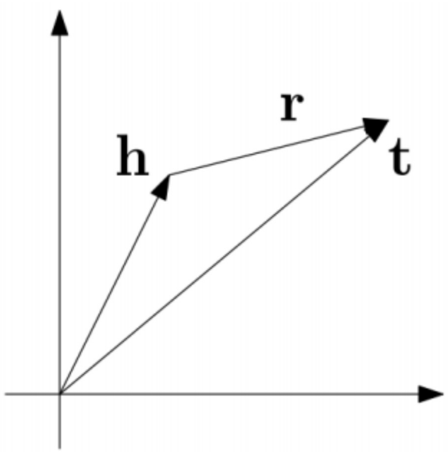


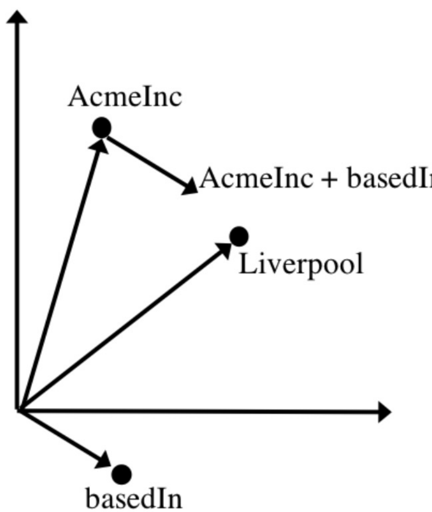
Nickel, M, Tresp, V, Kriegel, HP: Factorizing YAGO. Scalable Machine Learning for Linked Data. In Proceedings of the 21st International World Wide Web Conference, **2012**.

TransE: KG-Completion

- Inspired by word2vec

$$\text{score}(\mathcal{R}_p(\mathrm{e}_s, \mathrm{e}_o)) = -\|\boldsymbol{e}_s + \boldsymbol{r}_p - \boldsymbol{e}_o\|_1$$





Learning objective: **h + r = t**

Loss Function

- Closed-world assumption: square loss

$$L = \sum_{h,t \in E, r \in R} (y_{h,r,t} - f(h,r,t))^2$$

Triple, triplet

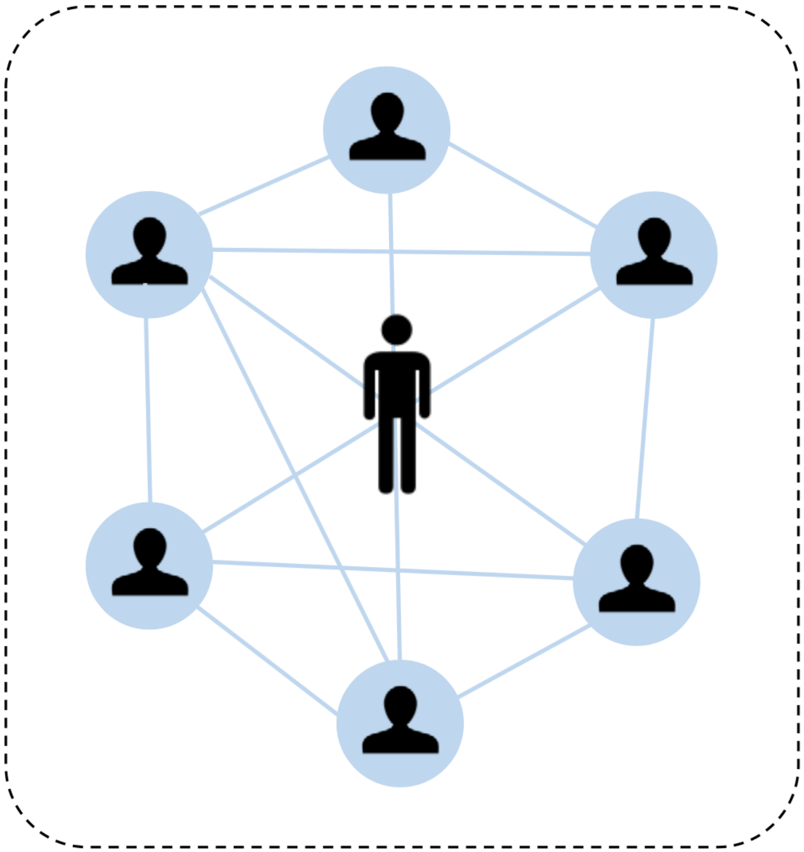
- Open-world assumption: triplet loss

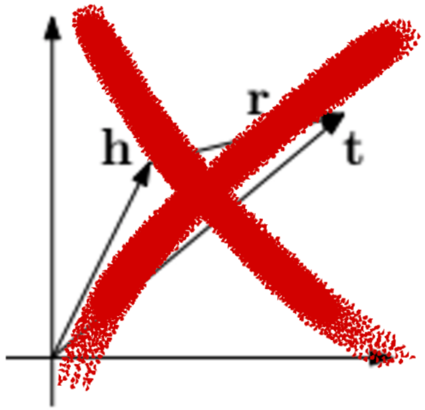
$$L = \sum_{T+} \sum_{T-} max(0, \gamma - f(h,r,t) + f(h',r',t'))$$

  - Need negative sampling

TransE: KG-Completion

However…





- In real world, we construct many relationships with many subjects.
- TransE can't represent more than one relationship between entities.

Overview

Théo Trouillon, Christopher R. Dance, Éric Gaussier, Johannes Welbl, Sebastian Riedel, and Guillaume Bouchard. 2017. Knowledge graph completion via complex tensor factorization. J. Mach. Learn. Res. 18, 1, 4735–4772. **2017**.

Model	**Scoring Function** ϕ	**Relation Parameters**	$\mathcal{O}_{time}$	$\mathcal{O}_{space}$
CP (Hitchcock, 1927)	$\langle w_r, u_s, v_o \rangle$	$w_r \in \mathbb{R}^K$	$\mathcal{O}(K)$	$\mathcal{O}(K)$
RESCAL (Nickel et al., 2011)	$e_s^T W_r e_o$	$W_r \in \mathbb{R}^{K^2}$	$\mathcal{O}(K^2)$	$\mathcal{O}(K^2)$
TRANSE (Bordes et al., 2013b)	$-\lvert\lvert(e_s + w_r) - e_o\rvert\rvert_p$	$w_r \in \mathbb{R}^K$	$\mathcal{O}(K)$	$\mathcal{O}(K)$
NTN (Socher et al., 2013)	$u_r^\top f(e_s W_r^{[1..D]} e_o + V_r \begin{bmatrix} e_s \\ e_o \end{bmatrix} + b_r)$	$W_r \in \mathbb{R}^{K^2 D}, b_r \in \mathbb{R}^K$ $V_r \in \mathbb{R}^{2KD}, u_r \in \mathbb{R}^K$	$\mathcal{O}(K^2 D)$	$\mathcal{O}(K^2 D)$
DISTMULT (Yang et al., 2015)	$\langle w_r, e_s, e_o \rangle$	$w_r \in \mathbb{R}^K$	$\mathcal{O}(K)$	$\mathcal{O}(K)$
HOLE (Nickel et al., 2016b)	$w_r^T(\mathcal{F}^{-1}[\overline{\mathcal{F}[e_s]} \odot \mathcal{F}[e_o]]))$	$w_r \in \mathbb{R}^K$	$\mathcal{O}(K \log K)$	$\mathcal{O}(K)$
COMPLEX (this paper)	$\mathrm{Re}(\langle w_r, e_s, \bar{e}_o \rangle)$	$w_r \in \mathbb{C}^K$	$\mathcal{O}(K)$	$\mathcal{O}(K)$

Table 1: Scoring functions of state-of-the-art latent factor models for a given fact $r(s, o)$, along with the representation of their relation parameters, and time and space (memory) complexity. K is the dimensionality of the embeddings. The entity embeddings e_s and e_o of subject s and object o are in $\mathbb{R}^K$ for each model, except for COMPLEX, where $e_s, e_o \in \mathbb{C}^K$. $\bar{x}$ is the complex conjugate, and D is an additional latent dimension of the NTN model. $\mathcal{F}$ and $\mathcal{F}^{-1}$ denote respectively the Fourier transform and its inverse, $\odot$ is the element-wise product between two vectors, Re(.) denotes the real part of a complex vector, and $\langle \cdot, \cdot, \cdot \rangle$ denotes the trilinear product.

F. L. Hitchcock. The expression of a tensor or a polyadic as a sum of products. J. Math. Phys, 6(1):164–189, **1927**.

Evaluation Metrics

Learning to Rank metrics

How well are positive triples ranked against their corruptions?

$$Hits@N = \frac{1}{|Q|}\sum_{i=1}^{|Q|} 1 \text{ if } rank_{(s,p,o)_i} \leq N$$

$$MR = \frac{1}{|Q|}\sum_{i=1}^{|Q|} rank_{(s,p,o)_i}$$ [Mean Rank]

$$MRR = \frac{1}{|Q|}\sum_{i=1}^{|Q|} \frac{1}{rank_{(s,p,o)_i}}$$ [Mean Reciprocal Rank]

Evaluation

	FB15k					WN18				
	MR	MRR	H@1	H@3	H@10	MR	MRR	H@1	H@3	H@10
TransE	-	.463	.297	.578	.749	-	.495	.113	.888	.943
DistMult	42	.798	-	-	**.893**	655	.797	-	-	.946
HolE	-	.524	.402	.613	.739	-	.938	.930	.945	.949
ComplEx	-	.692	.599	.759	.840	-	.941	.936	.945	.947
ConvE	51	.657	.558	.723	.831	374	.943	.935	.946	.956
pRotatE	43	**.799**	**.750**	.829	.884	**254**	.947	.942	.950	.957
RotatE	**40**	.797	.746	**.830**	.884	309	**.949**	**.944**	**.952**	**.959**

	FB15k-237					WN18RR				
	MR	MRR	H@1	H@3	H@10	MR	MRR	H@1	H@3	H@10
TransE	357	.294	-	-	.465	3384	.226	-	-	.501
DistMult	254	.241	.155	.263	.419	5110	.43	.39	.44	.49
ComplEx	339	.247	.158	.275	.428	5261	.44	.41	.46	.51
ConvE	244	.325	.237	.356	.501	4187	.43	.40	.44	.52
pRotatE	178	.328	.230	.365	.524	**2923**	.462	.417	.479	.552
RotatE	**177**	**.338**	**.241**	**.375**	**.533**	3340	**.476**	**.428**	**.492**	**.571**

Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, Jian Tang:
RotatE: Knowledge Graph Embedding by Relational Rotation
in Complex Space. In Proc. ICLR 2019.

Summary

- Very many RESCAL- and TransE-like approaches for handcrafted embeddings of relational data
- None of the many approaches covers what's in a text
- There no knowledge in a knowledge graph ….
  - but a formal subjective content description (SCD) of certain aspects
  - SCDs might indeed be very helpful for certain applications
- Forget about handcrafted approaches