Intelligent Agents

Multi-Relational Latent Semantic Analysis

Ralf Möller
Universität zu Lübeck
Institut für Informationssysteme

Open-Domain Question Answering

Fulfill user's information need with direct answers

- Answer Sentence Selection

Q: Who won the best actor Oscar in 1973?
S_1: Jack Lemmon was awarded the Best Actor Oscar for Save the Tiger (1973).
S_2: Academy award winner Kevin Spacey said that Jack Lemmon is remembered as always making time for others.

- Word-embedding based approaches
- Combined with models for enhanced lexical semantic

Continuous Semantic Representations

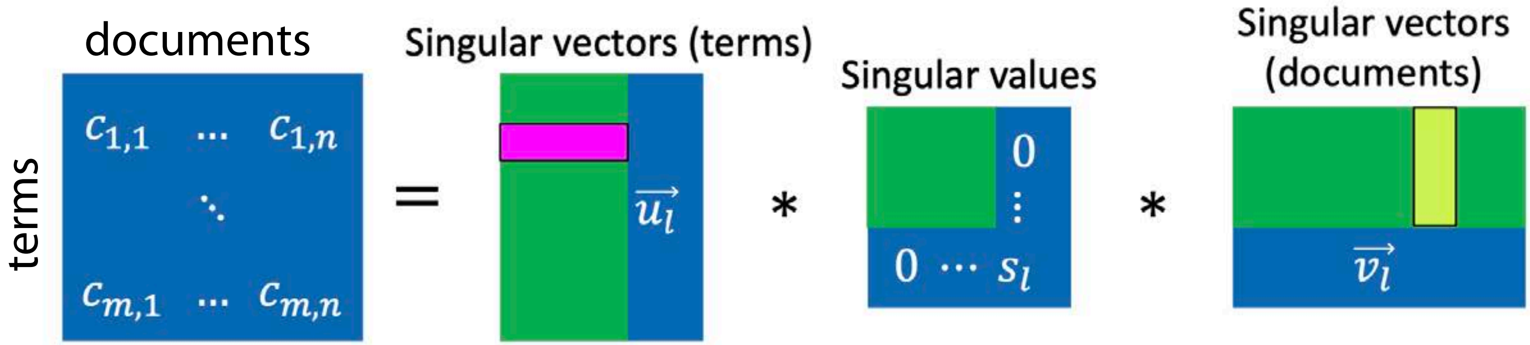
- A lot of popular methods for creating representations
  - Word/Document embedding: Vector Space Model (BoW)
  - Dimension reduction: Latent Semantic Analysis

Recap: Latent Semantic Analysis (LSA, ca. 1990)

Rank reduction of $A_k = US_kV^T$ with singular values set to 0

Matrix factorization method

- Compute a term-document matrix A
- Use singular value decomposition: $A = U * S * V^T$



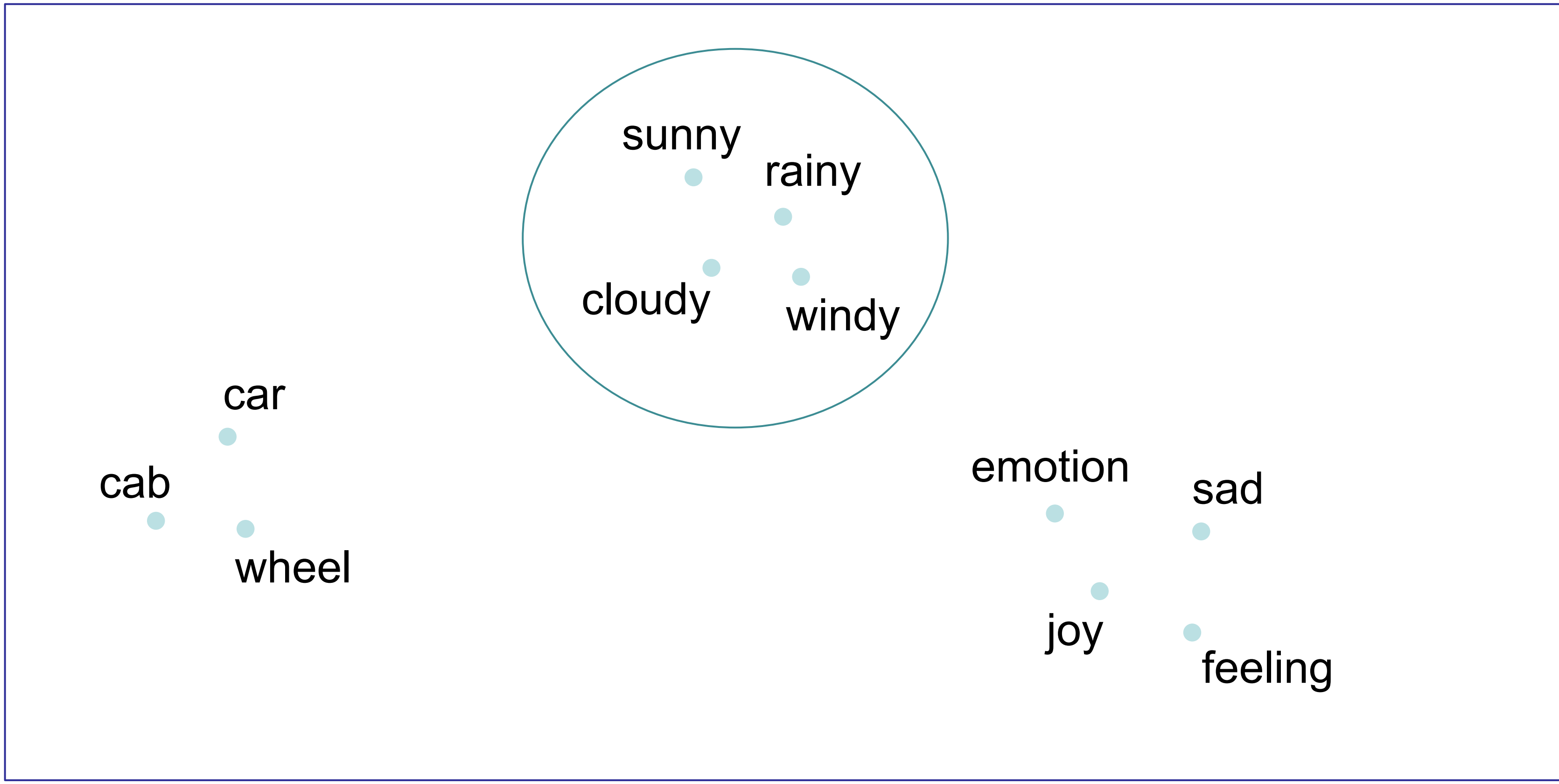
- Dimensional reduction by omitting singular values
- Term and document vectors can be treated as semantic spaces

corresponds to a term
corresponds to a document

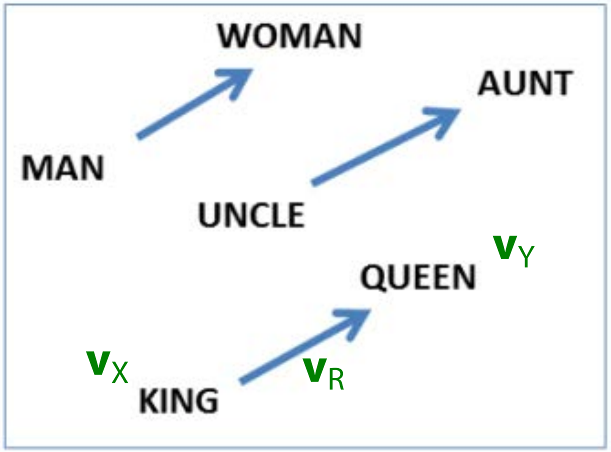
Continuous Semantic Representations

- A lot of popular methods for creating representations
  - Word/Document embedding: Vector Space Model (BoW)
  - Dimension reduction: Latent Semantic Analysis
  - Encoding of term co-occurrence information: PMI
  - Shallow parsing: HMMs, MRFs, Deep Network Learning
    - POS tagging, Phase chunking, NER, SRL
  - Topic models: Latent Dirichlet Allocation
  - Word embedding w.r.t. contexts: Word2Vec, GloVe, Paragraph Vector (doc2vec)…

Continuous Semantic Representations



Using Learned Relational Networks for IR



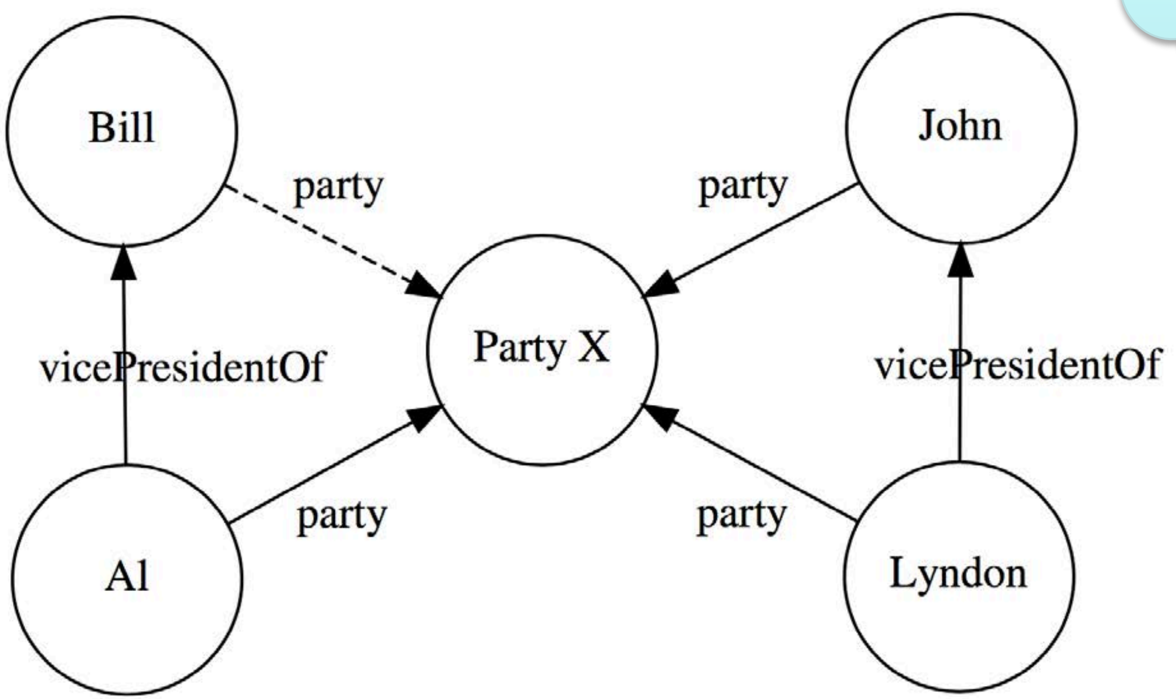
Instead of words, also named entities (represented by words) can be embedded

Two entities can be associated automatically

Shallow parsing and/or semantic role labeling might help

Goal: Predict relation membership

Databases from text data are inherently incomplete

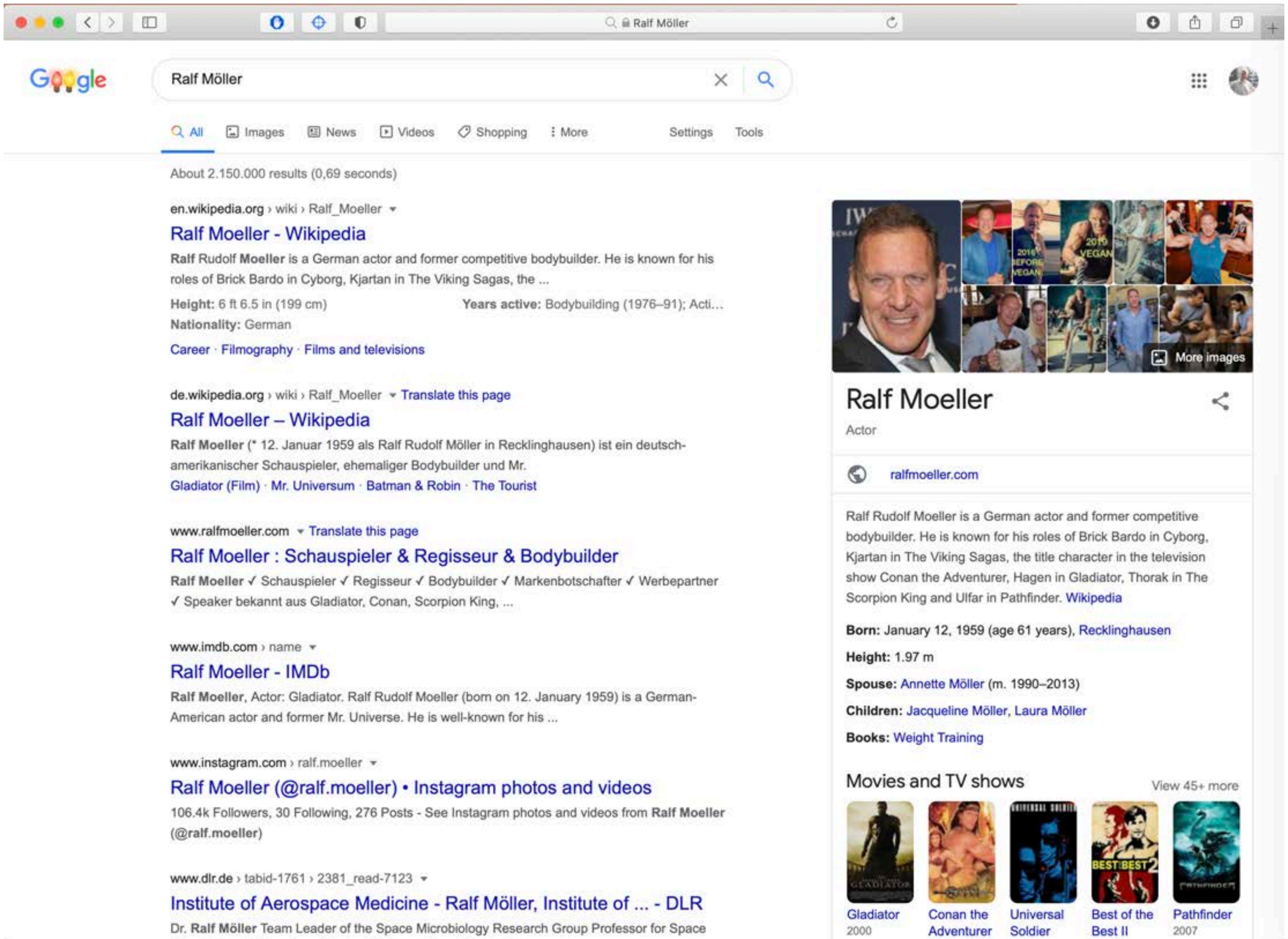


Structure embedding due to similarities

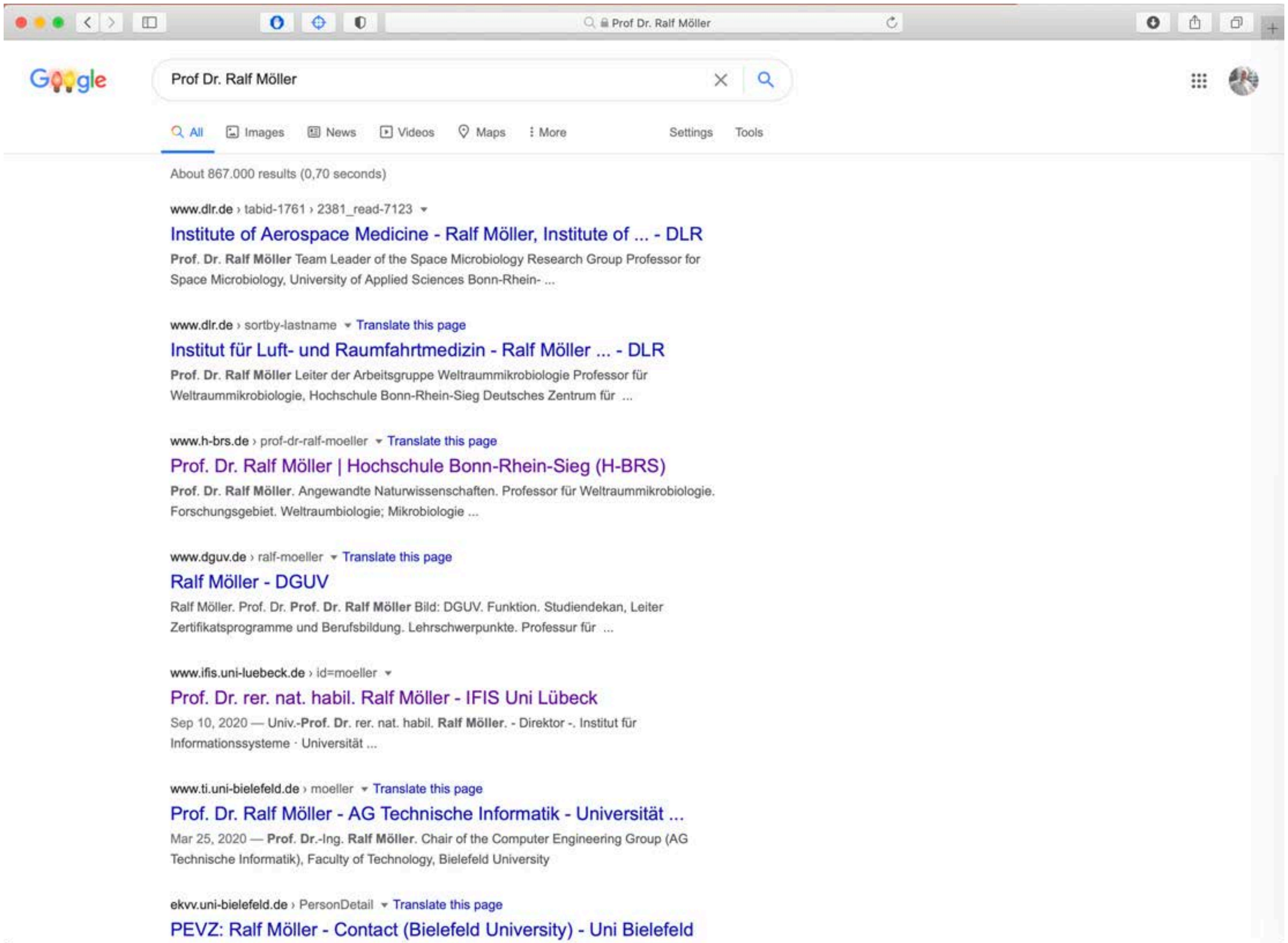
"Knowledge graph" completion

Actually, we deal with database completion

Relational Database Search vs. String Search



Relational Data only for Important People



Negative Sampling of Relational Structures?

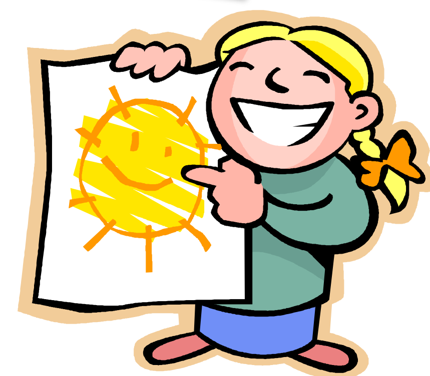
- Naively selecting arbitrary tuples of instances as negative examples by sampling?
- Does not impose much control on the problem of learning embeddings
- Open-world vs. closed-world assumption
- Generative adversarial networks to generate negative examples of relational structures (not discussed in detail)

Liwei Cai and William Yang Wang, "KBGAN: Adversarial Learning for Knowledge Graph Embeddings", in Proceedings NAACL 2018.

Semantics Needs More Than Similarity

Tomorrow will be rainy.

Tomorrow will be sunny.



~~*similar*(r, sunny)?~~

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
  - 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

Target word: row-vector

Inducing polarity

	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

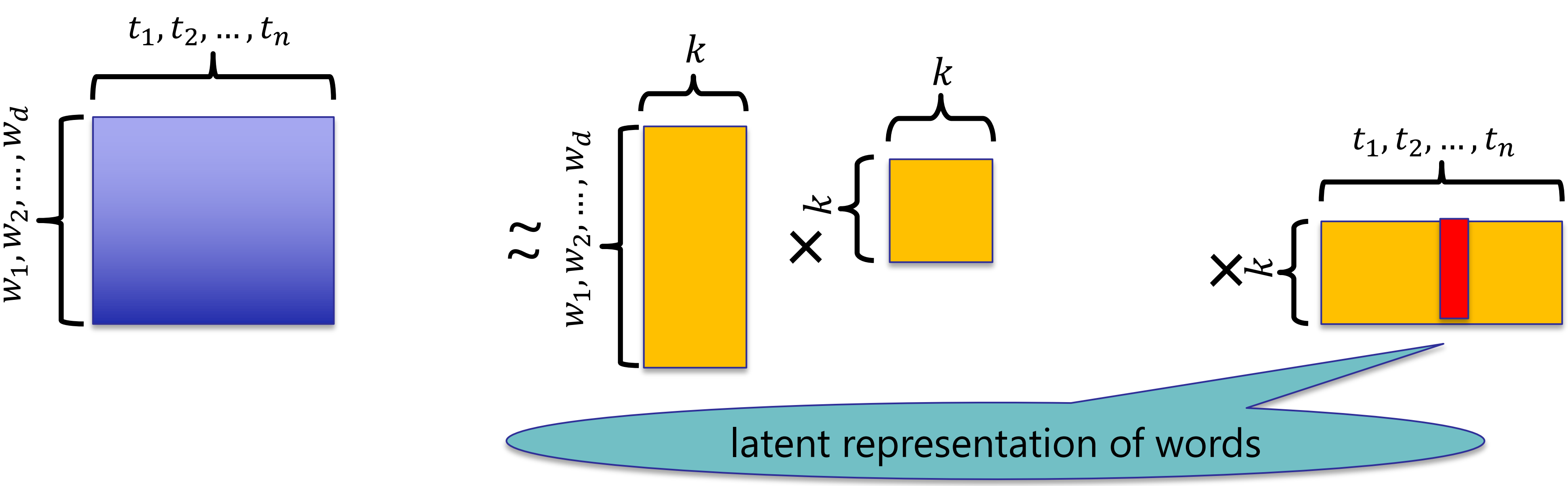
	joy	gladden	sadden	feeling
joyfulness	0	0	0	1
gladden	0	0	0	0
sad	0	0	0	1
anger	0	0	0	1

Hyponym layer

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

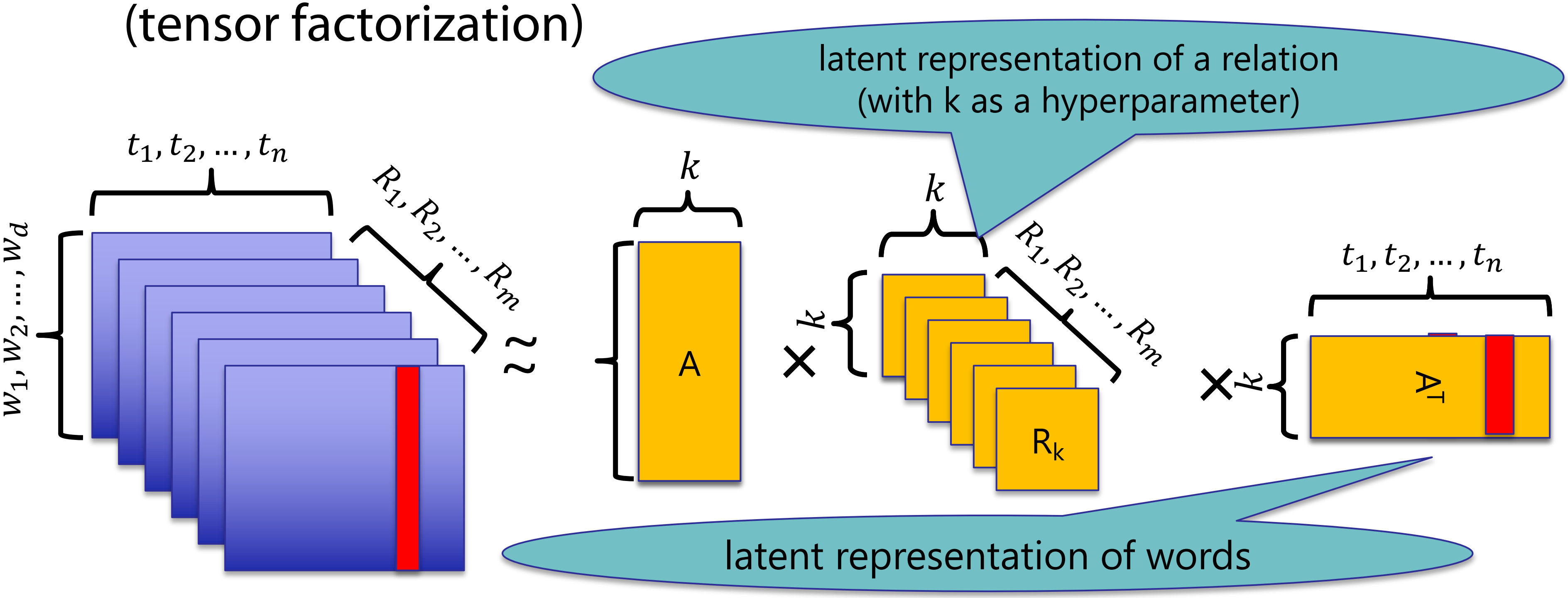
Tensor Decomposition – Analogy to SVD

- 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

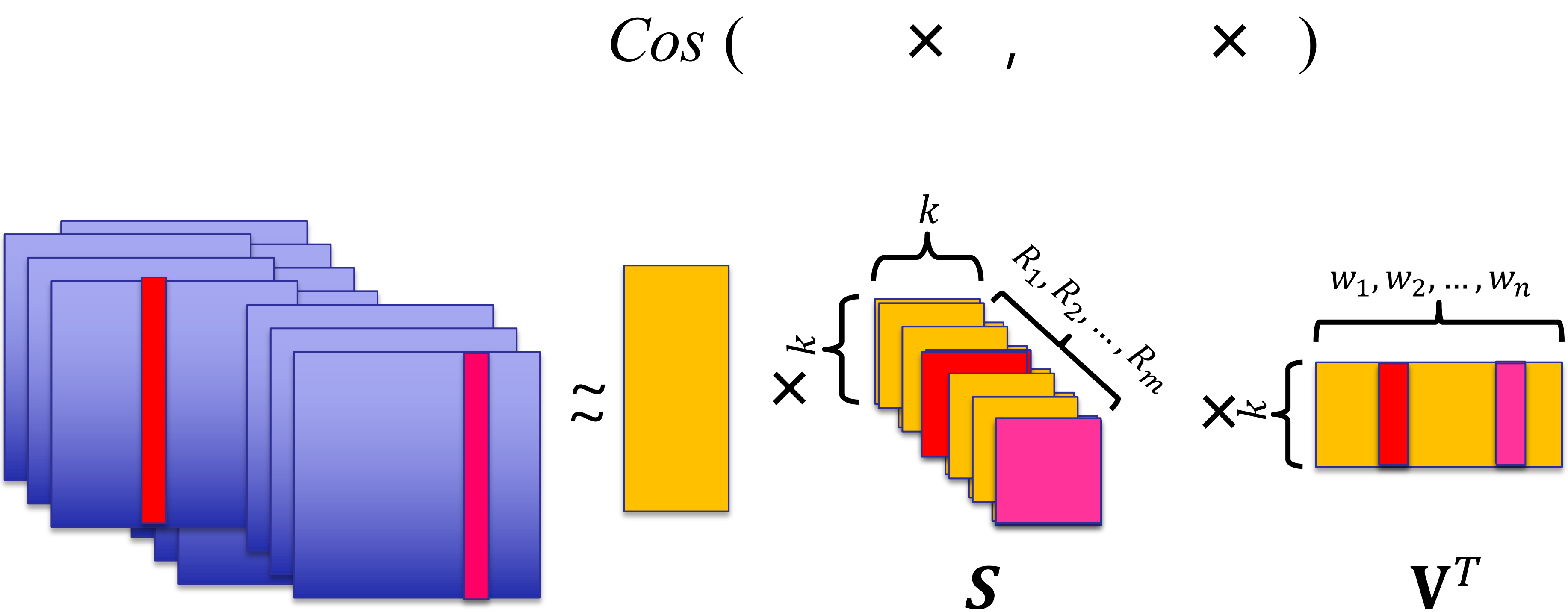
Synonym layer

	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

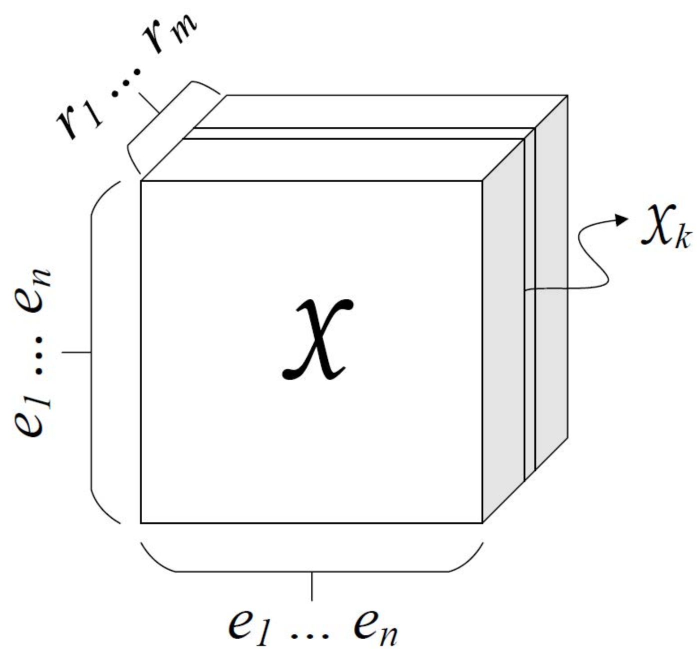
- $rel(\mathrm{w}_i, \mathrm{w}_j) = \cos(\boldsymbol{S}_{:,:,syn}\mathbf{V}_{i,:}^T, \boldsymbol{S}_{:,:,rel}\mathbf{V}_{j,:}^T)$



Representation of Facts (1/2)

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

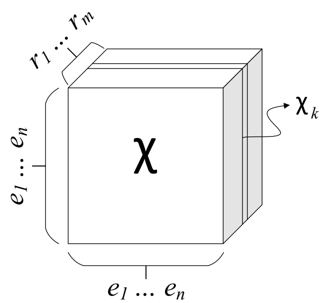
Subject	Predicate	Object
Obama	Born-in	Hawaii
Bill Gates	Nationality	USA
Bill Clinton	Spouse-of	Hillary Clinton
Satya Nadella	Work-at	Microsoft
…	…	…



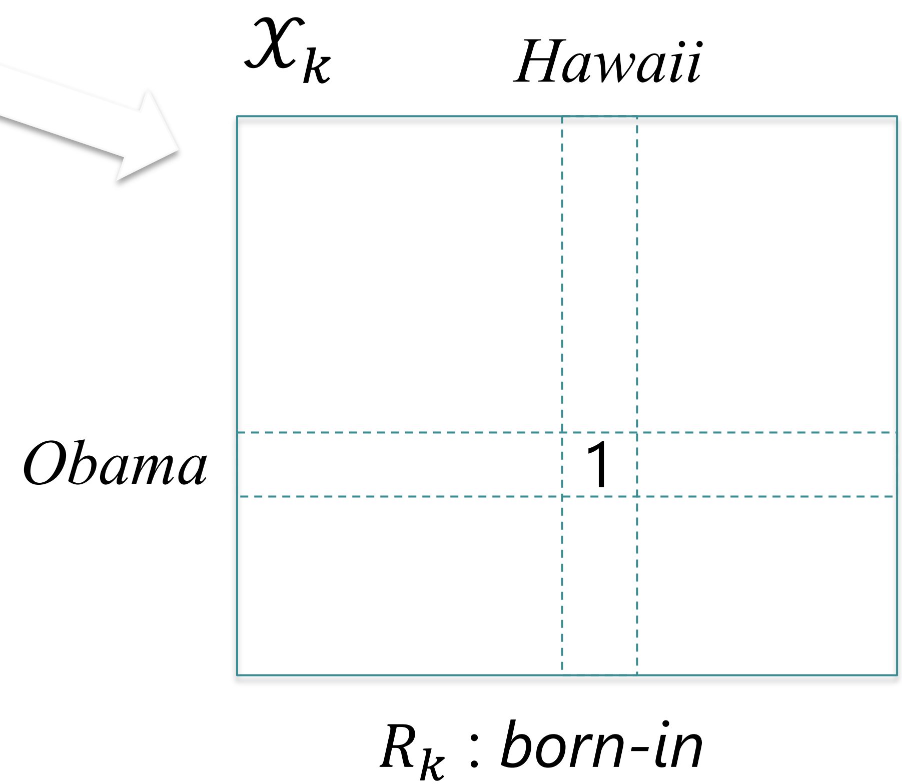
n: # entities, m: # relations

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**.

Database Representation (2/2)



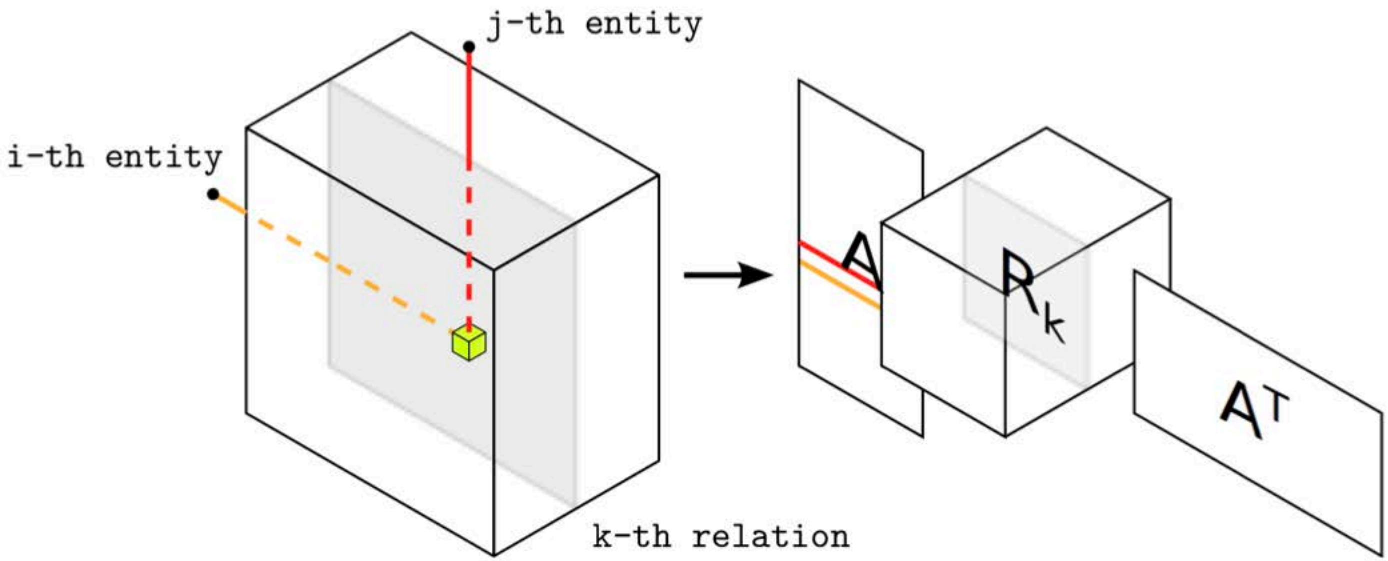
k-th slice



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**.

Tensor Factorization

- Values a_i and a_i are representations of the i-th and j-th entity by latent components (rows of A, columns of A^T)
- Claim: Entities' similarity in this space reflects entities' similarity in relational domain
- R_k models the interactions of the latent components in the k-th relation
- Could even invent new layers R_k
- Expressing data in terms of newly invented latent components (new layers R_k) is often referred to as predicate invention in statistical relational learning

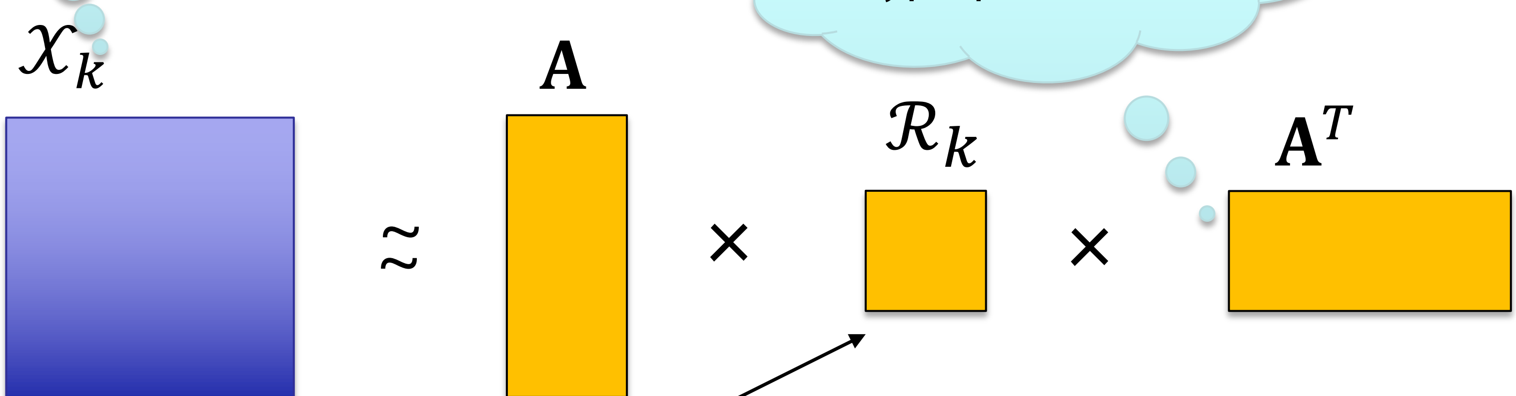
Nickel, M, Tresp, V, Kriegel, HP: Factorizing YAGO. Scalable Machine Learning for Linked Data. In Proceedings of the 21st International World Wide Web Conference, **2012**.

Tensor Decomposition Objective

$$argmin_{AR} \quad \underbrace{\frac{1}{2}\left(\sum_k \|\mathcal{X}_k - \mathbf{A}\mathcal{R}_k\mathbf{A}^T\|_F^2\right)}_{\text{Reconstruction Error (Loss)}} + \underbrace{\frac{1}{2}\lambda\left(\|A\|_F^2 + \sum_k \|\mathcal{R}_k\|_F^2\right)}_{\text{Regularization}}$$

Rank reduction?

Dimension is a hyperparameter



k-th relation

Frobenius Norm: $\|A\|_F \equiv \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} |a_{ij}|^2}.$

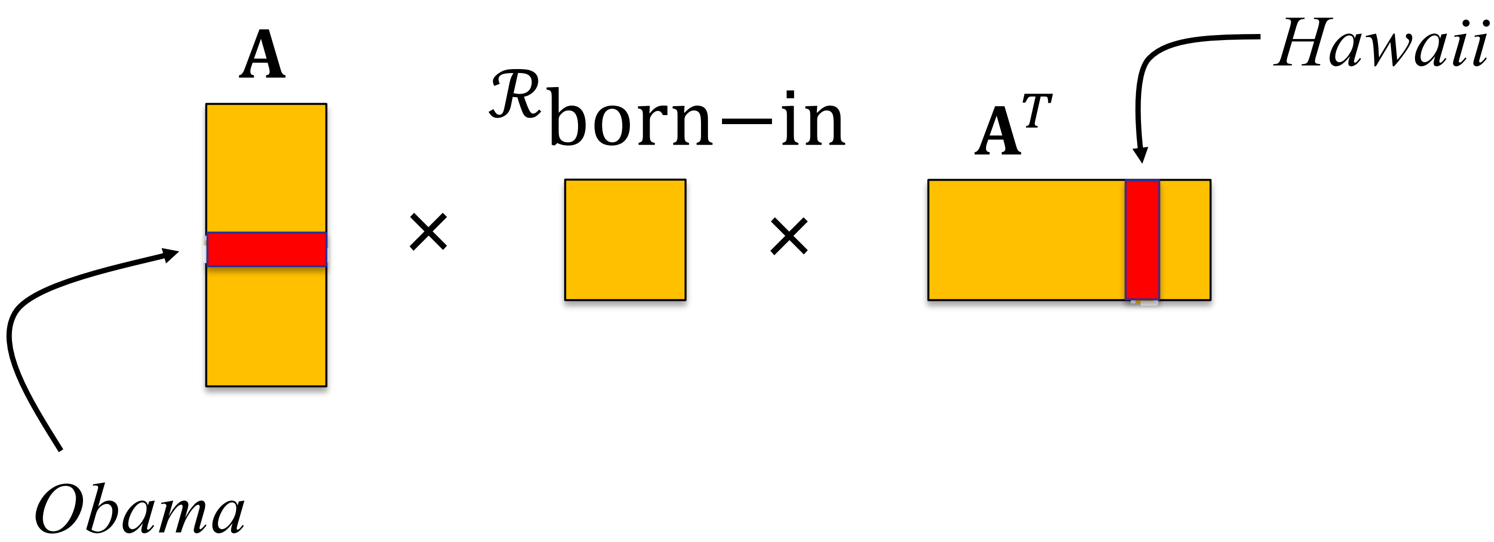
Brett W. Bader, Richard A. Harshman, and Tamara G. Kolda. 2007. Temporal Analysis of Semantic Graphs Using ASALSAN. In Proc. ICDM '07. 2007

Measure the Degree of a Relationship

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

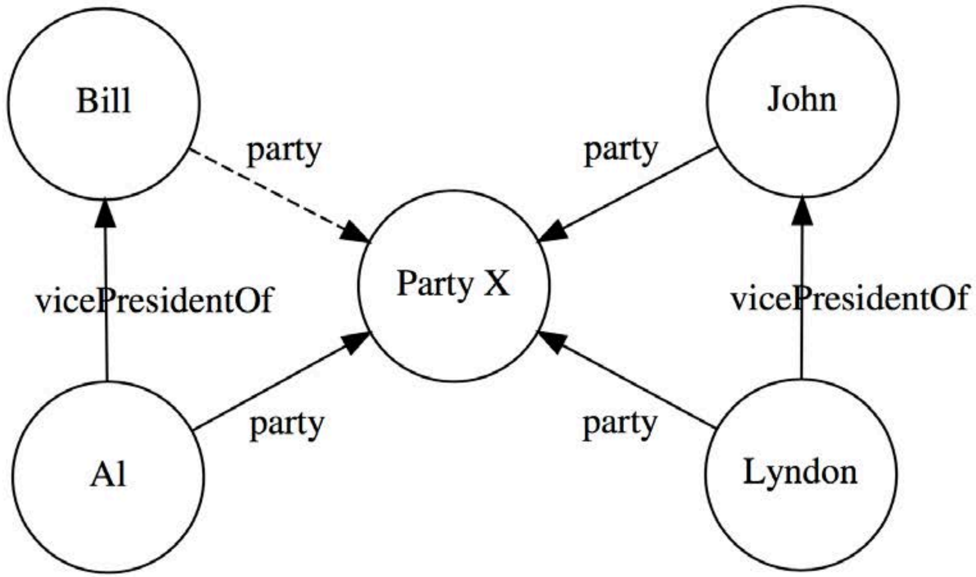
$$=$$

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



Prediction of Unknown Facts

- Predict party membership of US (vice) presidents



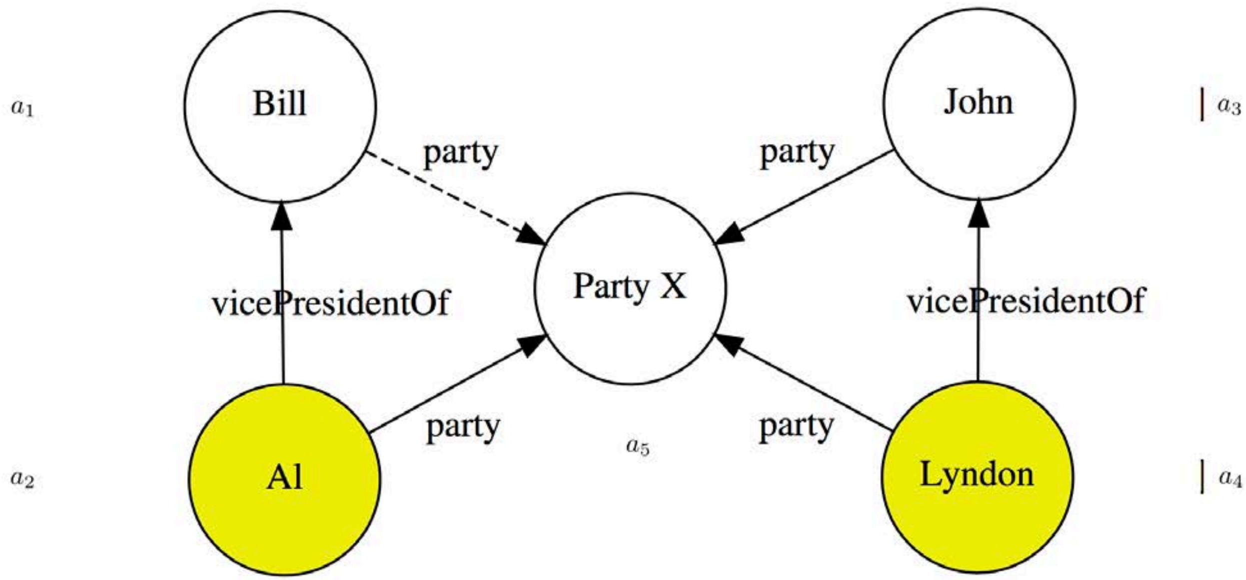
Prediction of unknown fact party(Bill, Party X)

Link prediction

- $\hat{X}_{ijk} = a_i R_k a_j^T$ is the score that the model assigns to the truth of $R_k(e_i, e_j)$ or the existence of $(e_i, e_j) \in R_k$.
- If $\hat{X}_{ijk} > \theta$, where θ is a threshold, then $(e_i, e_j) \in R_k$
- Due to regularization, sparseness of R_k is enforced
  - Difficult to determine useful threshold
  - Better use ranking and top-k queries
- To determine which entities most likely have a specific link to entity e_i, it is sufficient to compute the matrix product $A\,R_k a_i$, where a_i is the latent vector for e_i
- Exact inference is indeed tractable with this approach

Link Prediction

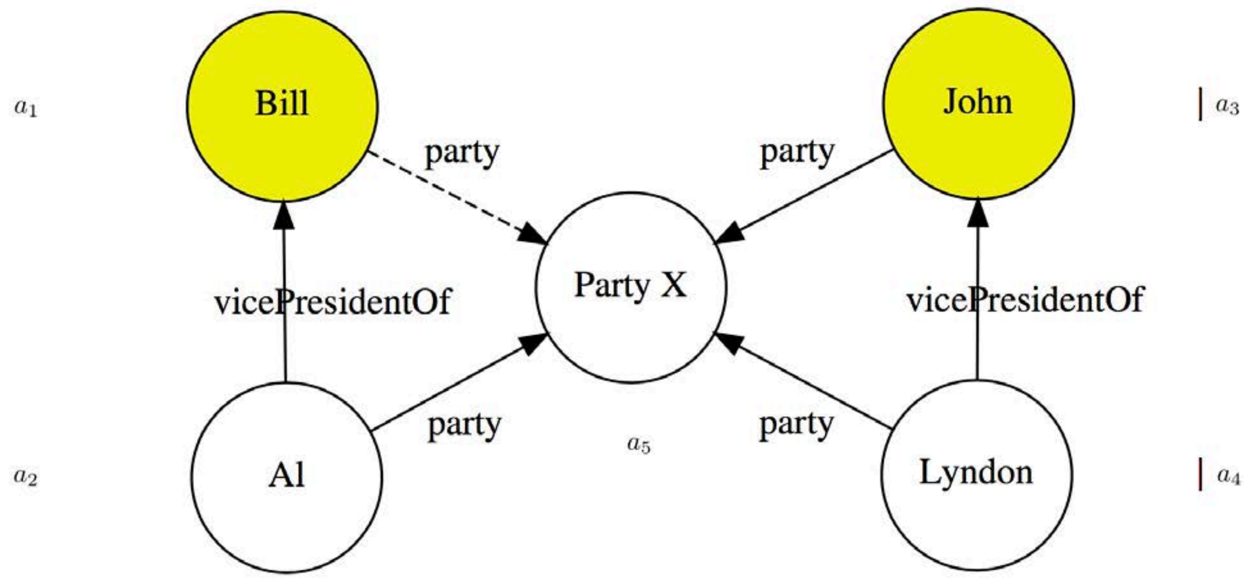
- Predict party membership of US (vice) presidents



- Latent component representation of Al and $Lyndon$ will be similar to each other
  - Both representations reflect that the corresponding entities are related to the Object $Party\_X$

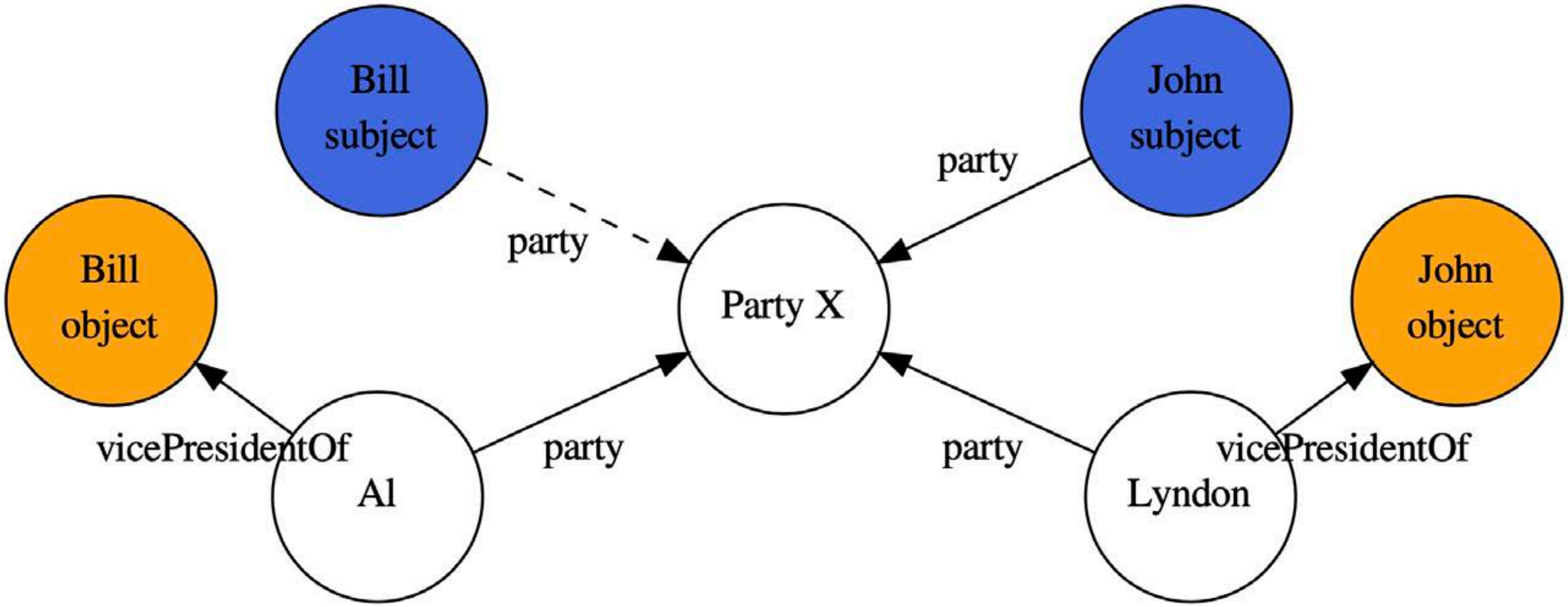
Link Prediction

- Predict party membership of US (vice) presidents



- Therefore, *Bill* and *John* will also have a similar latent-component representation
- Consequently $a_{Bill} R_{party}\, {a^T}_{Party\_X}$ will yield a similar value to $a_{John} R_{party}\, {a^T}_{Party\_X}$
- Missing relation can be predicted correctly

Collective Learning on Multi-Relational Data



- This would break if $Bill$ and $John$ had different representations as subjects and objects

$$f(A, R_k) = \frac{1}{2} \sum_{i,j,k} \left(\mathcal{X}_{ijk} - \boldsymbol{a}_i^T R_k \boldsymbol{b}_j \right)^2$$

Benefits of the Latent Representation

- Retrieve similar documents
- Retrieve similar entities via latent representations
- Matrix A can be interpreted as an embedding into a latent component space that reflects their similarity over all relations in the domain of discourse
- For similarity retrieval use ranking (and top-k), which is effectively computable

Problem: Relational Domain Knowledge

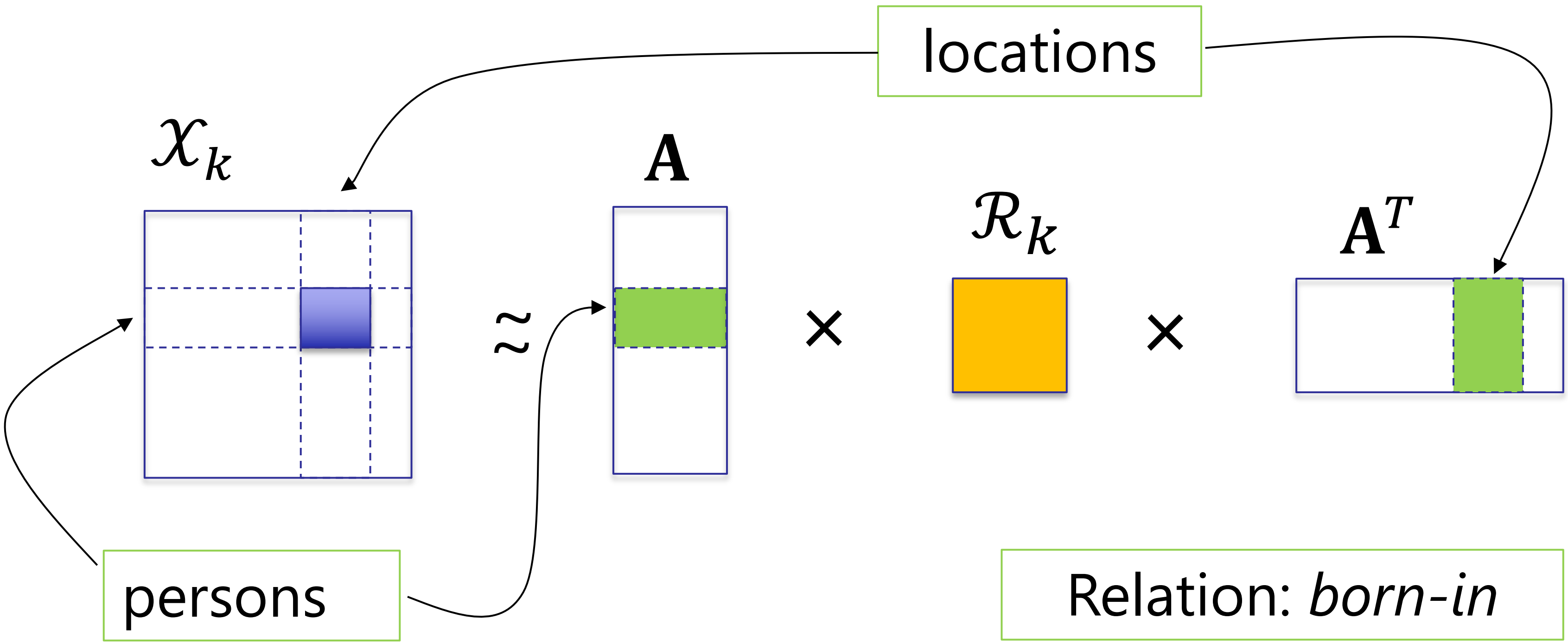
- Relational data – the entity type
  - Relation can only hold between the right types of entities
    - Words having *is-a* relation have the same part-of-speech
    - For relation *born-in*, the entity types are: (person, location)
- Leverage type information to improve multi-relational LSA
- Idea #3: Change the objective function

Typed Multi-Relational LSA (TRESCAL)

- Only legitimate entities are included in the objective function of tensor decomposition
- Benefits of leveraging the type information
  - Faster model training time
  - Higher prediction accuracy
- Experiments conducted using database
  - Application to Relation Extraction

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

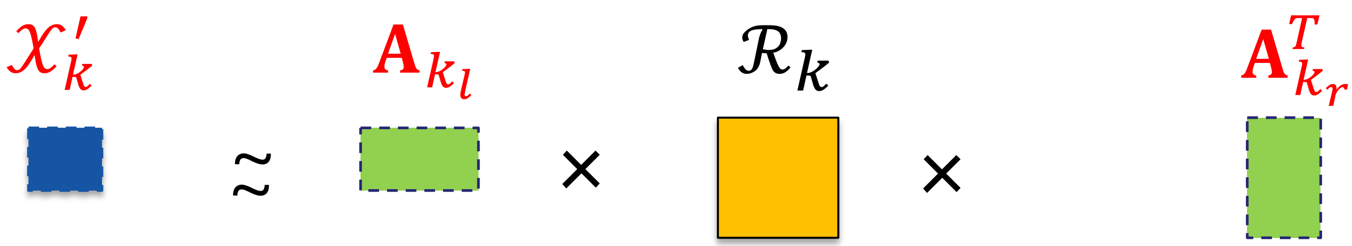
Typed Tensor Decomposition Objective



- Impose constraints on embeddings
- How to reprensent instance-of relation?
  - X_i = ISA? Types/classes are sets not entities

Typed Tensor Decomposition Objective

- Reconstruction error: $\frac{1}{2}\sum_{k}\left\|\mathcal{X}'_k - \mathbf{A}_{k_l}\mathcal{R}_k\mathbf{A}^T_{k_r}\right\|_F^2$



- Improvements in learning time
- Performance improvements

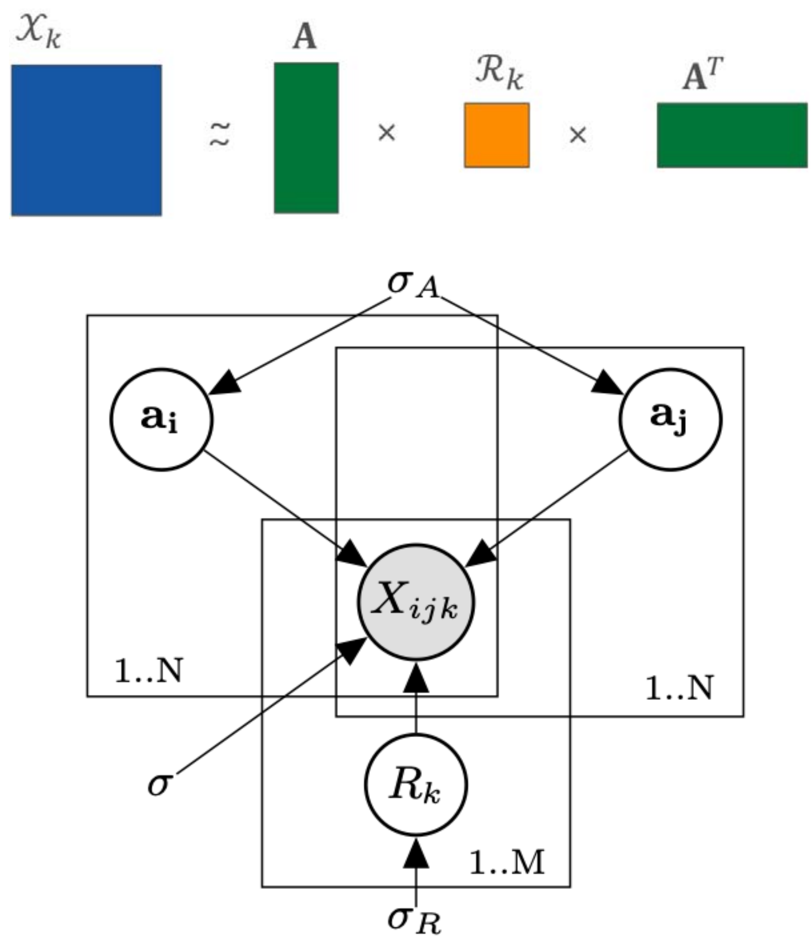
Chang, Kai-Wei & Yih, Wen-tau & Yang, Bishan & Meek, Chris. Typed Tensor Decomposition of Knowledge Bases for Relation Extraction. In: Proc. EMNLP-14. **2014**.

Tasks

- Entity Retrieval: $(e_i, r_k, ?\,)$
  - One positive entity with 100 negative entities
- Relation Retrieval: $(e_i, ?\,, e_j)$
  - Positive entity pairs with equal number of negative pairs

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



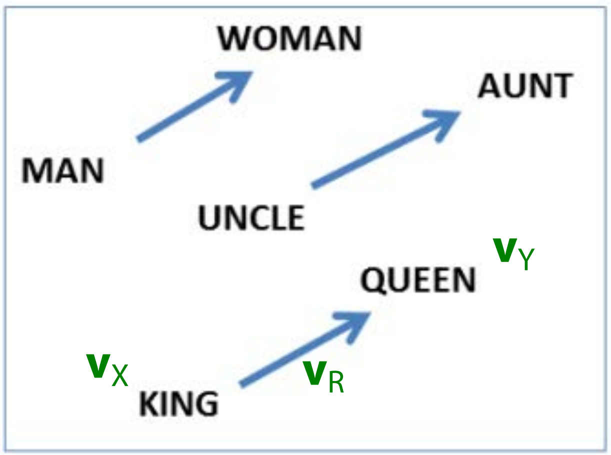
Nickel, M, Tresp, V, Kriegel, HP: Factorizing YAGO. Scalable Machine Learning for Linked Data. In Proceedings of the 21st International World Wide Web Conference, **2012**.

Types of Relations

- **Symmetric/Antisymmetric** Relations
  - Symmetric: e.g., Marriage
  - Antisymmetric: e.g., Filiation if $R(a, b)$ with $a \neq b$, then $R(b, a)$ must not hold,
- **Inverse** Relations
  - Husband and wife
- **Composition** Relations
  - My mother's husband is my father

Asymmetric = antisymmetric and irreflexive

Using Learned Relational Networks for IR



TransE [Bordes *et al.*, 2013]

TransE: Translating Embeddings:
Find an embedding for entities and relations so that R(X,Y) iff $\mathbf{v}_Y - \mathbf{v}_X \sim= \mathbf{v}_R$

Instead of words, also named entities (represented by words) can be embedded

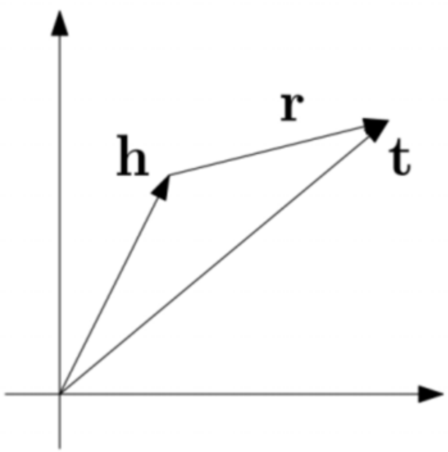
Two entities can be associated automatically

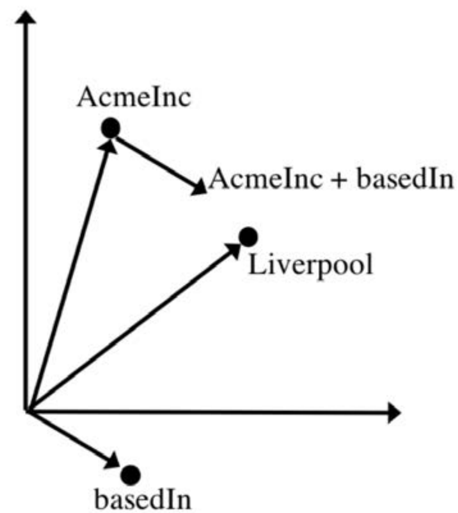
Antoine Bordes, Nicolas Usunier, Alberto Garcia-Durán, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In Proceedings NIPS **2013**.

TransE: Additive Scoring Function

- 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

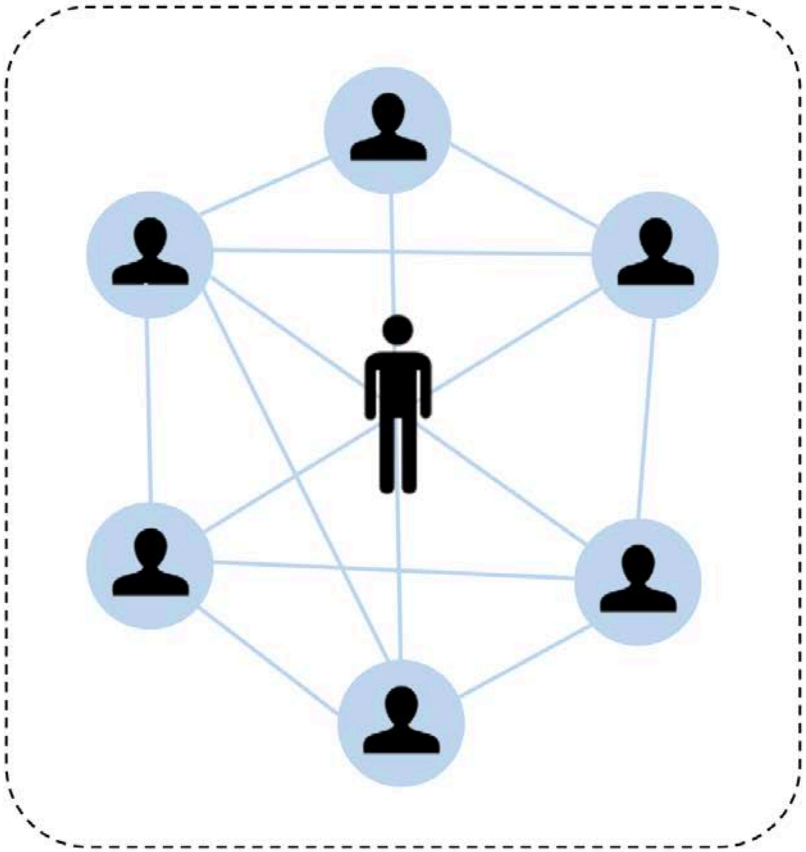
- Open-world assumption: triplet loss

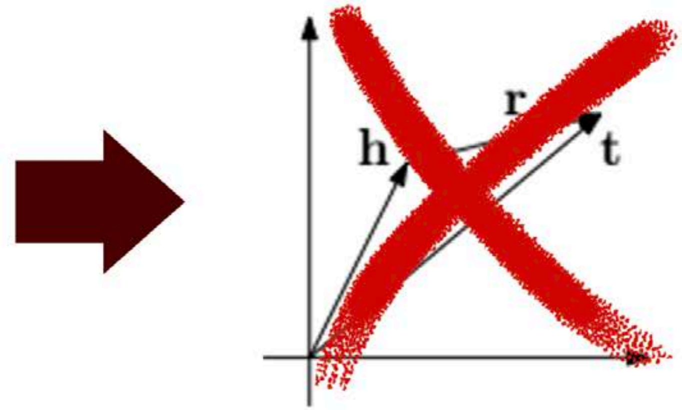
$$L = \sum_{T+} \sum_{T-} max(0, \gamma - f(h,r,t) + f(h',r',t'))$$

  - Need negative sampling

TransE

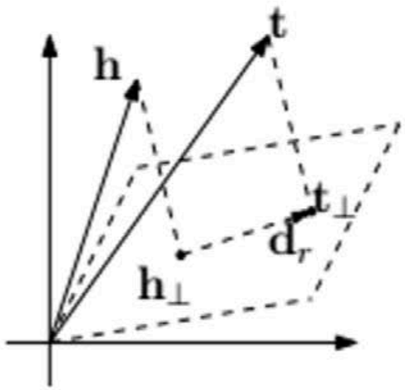
However...





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

TransH: Relation-specific Embeddings



TransH

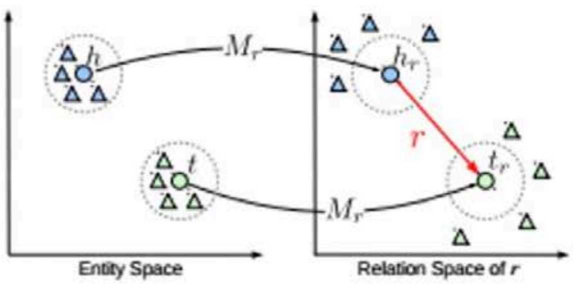
- Representing **projected** relations with differences between entity vectors
- Able to represent M-to-M relations

reference : TransH (Wang et al. AAAI2014)

- **Entities** and **relations** have different characteristics
- However, they are represented in the same space

Wang, et al. Knowledge graph embedding by translating on hyperplanes. In: Proc. AAAI-14. **2014**.

TransR/CTransR: Relation-specific Embeddings



TransR

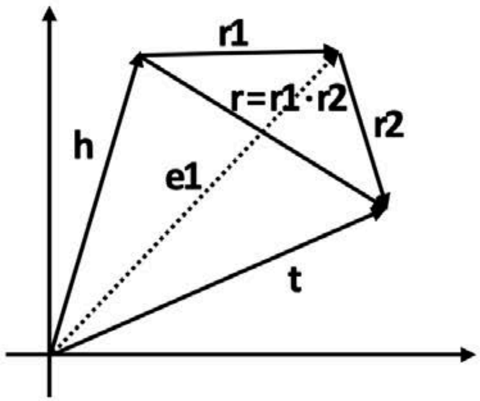
- Representing **projected** entities and **mapping** them into relation space ‘r’
- Able to represent different characteristics between entities and relations

reference : TransR (AAAI2015)

However, these were all not good enough for detecting and addressing misrelations!!

Lin, et al. Learning entity and relation embeddings for knowledge graph completion. In: Proc. AAAI **2015**.

PTransE

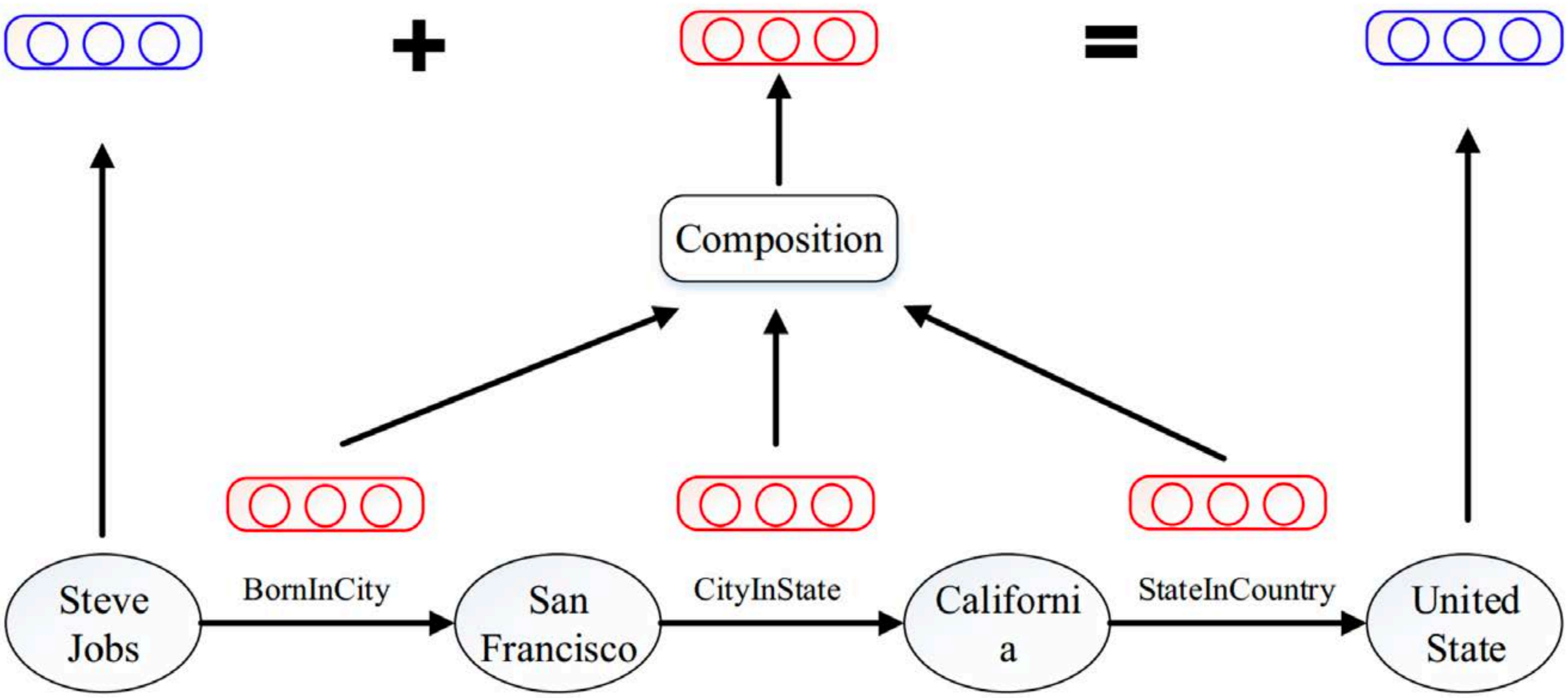


PTransE

- Representing relations through **composition of relations** between entity vectors

Lin, et al. Modeling Relation Paths for Representation Learning of Knowledge Bases. In: Proc. EMNLP **2015**.

PTransE


+
=
Composition
Steve Jobs
BornInCity
San Francisco
CityInState
Californi a
StateInCountry
United State

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 application 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")

Benchmark Datasets

- **FB15K**: a subset of Freebase. The main relation types are **symmetry/antisymmetry** and **inversion** patterns.
- **WN18**: a subset of WordNet. The main relation types are **symmetry/antisymmetry** and **inversion** patterns.
- **FB15K-237**: a subset of FB15K, where inversion relations are deleted. The main relation types are **symmetry/antisymmetry** and **composition** patterns.
- **WN18RR**: a subset of WN18, where inversion relations are deleted. The main relation types are **symmetry/antisymmetry** and **composition** patterns.

Dataset	#entity	#relation	#training	#validation	#test
FB15k	14,951	1,345	483,142	50,000	59,071
WN18	40,943	18	141,442	5,000	5,000
FB15k-237	14,541	237	272,115	17,535	20,466
WN18RR	40,943	11	86,835	3,034	3,134

Using Learned Relational Networks for IR

- Query answering: indirect queries requiring chains of reasoning
- DB Completion: exploits redundancy in the KB + chains to infer missing facts

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

- Fraction of positives that rank in the top N among their negatives in validation set
- Rank of a positive example is determined by the rank of its score against the scores of a certain number of negative examples
- A rank of 1 is the "best" outcome as it means that the positive example had a higher score than all the negatives.

Freebase 15k benchmark

Methods	Hits@10	
Unstructured [Bordes *et al.*, 2014]	4.5	baseline method
RESCAL [Nickel *et al.*, 2011]	28.4	tensor factorization
SE [Bordes *et al.*, 2011]	28.8	Translation-based embedding
SME [Bordes *et al.*, 2014]	31.3	
LFM [Jenatton *et al.*, 2012]	26.0	
TransE [Bordes *et al.*, 2013]	34.9	
ConvNets [Shi and Zhu, 2015]	37.7	
TransH [Wang *et al.*, 2014b]	45.7	
TransR [Lin *et al.*, 2015b]	48.4	
PTransE [Lin *et al.*, 2015a]	51.8	

Using Learned Relational Networks for IR

	Entity Retrieval			Relation Retrieval		
	TransE	RESCAL	TRESCAL	TransE	RESCAL	TRESCAL
w/o type checking	51.41%‡	51.59%	54.79%	75.88%	73.15%†	76.12%
w/ type checking	67.56%	62.91%‡	69.26%	70.71%‡	73.08%†	75.70%

Table 2: Model performance in mean average precision (MAP) on *entity retrieval* and *relation retrieval*. † and ‡ indicate the comparison to TRESCAL in the same setting is statistically significant using a paired-t test on average precision of each query, with $p < 0.01$ and $p < 0.05$, respectively. Enforcing type constraints during test time improves entity retrieval substantially, but does not help in relation retrieval.

$$\text{AveP} = \frac{\sum_{k=1}^{n} (P(k) \times \text{rel}(k))}{\text{number of relevant documents}}$$

$P(k)$ = Precision @ Rank k (ranked precision)

where $\text{rel}(k)$ is an indicator function equaling 1 if the item at rank k is a relevant document, zero otherwise.

$$\text{MAP} = \frac{\sum_{q=1}^{Q} \text{AveP(q)}}{Q}$$

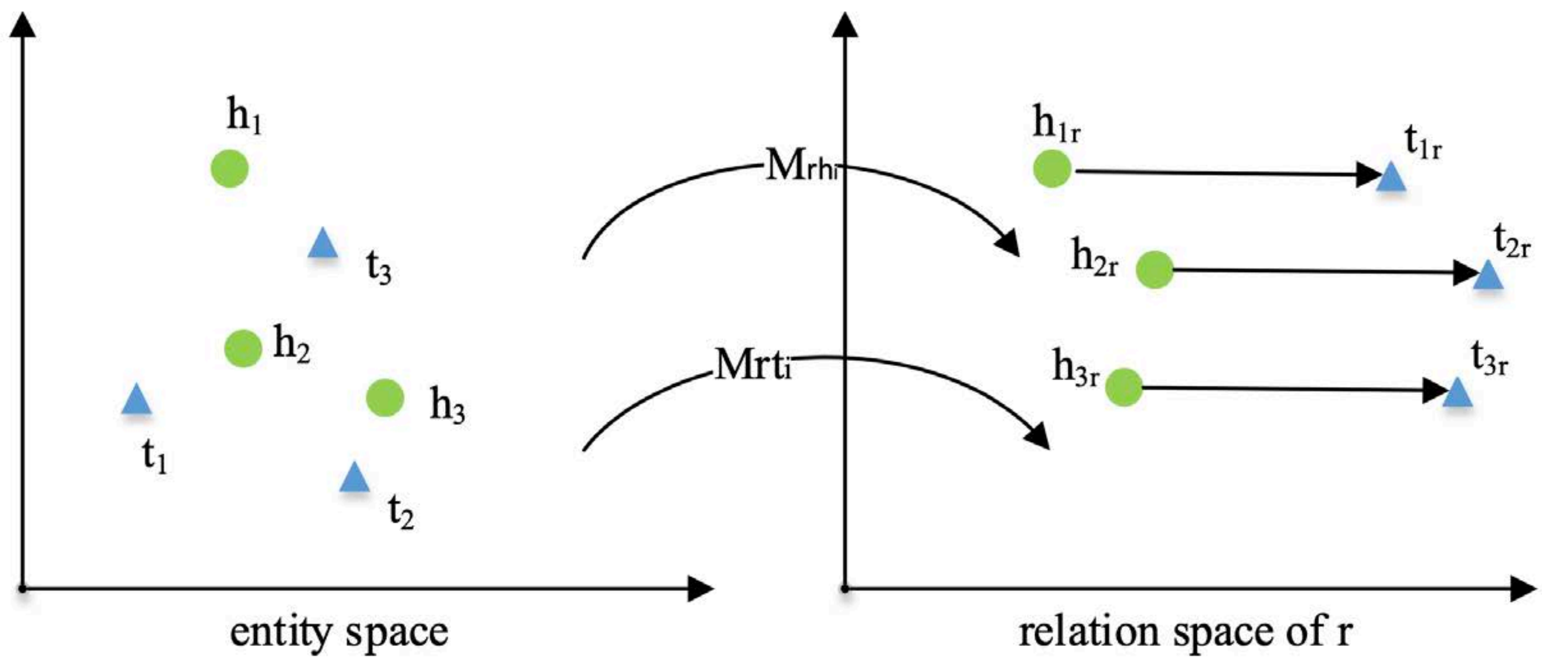
where Q is the number of queries.

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

Maximilian Nickel, Kevin Murphy, Volker Tresp, and Evgeniy Gabrilovich. A Review of Relational Machine Learning for Knowledge Graphs. Proc. IEEE 104, 1 (2016), 11–33. **2016**.

TransD

- Projection matrices not only related to relation but also head/tail entities



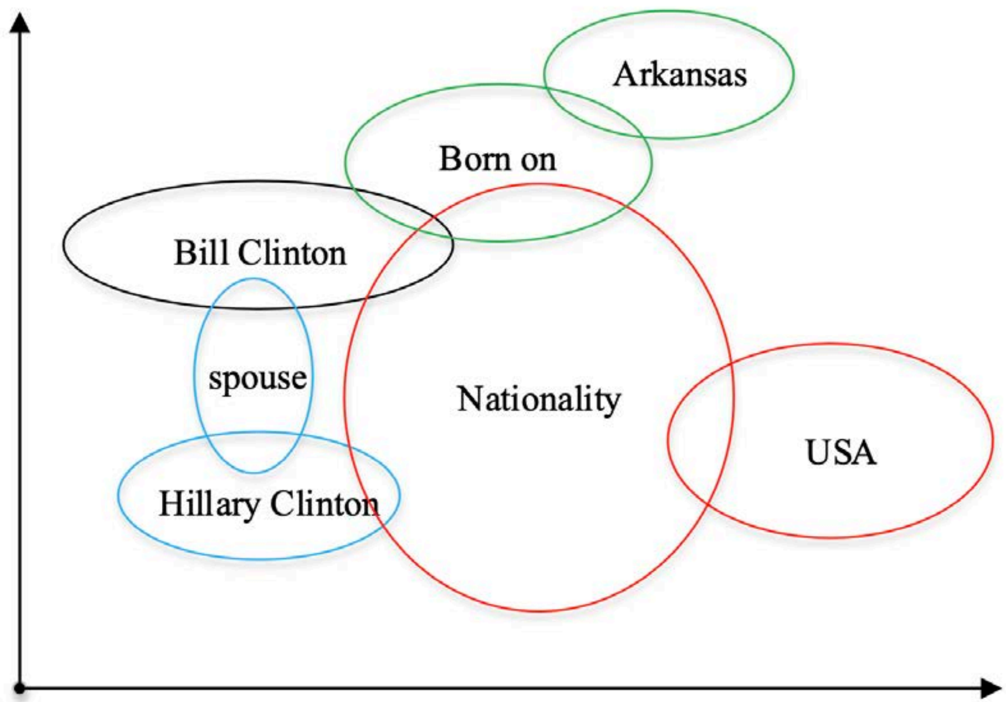
Ji, et al. Knowledge Graph Embedding via Dynamic Mapping Matrix. In: Proc. ACL-IJCNLP-15. **2015.**

TransD

- Uses two vectors to represent a named symbol object (entity and relation)
  - The first one represents the meaning of a(n) entity (relation),
  - the other one is used to construct mapping matrix dynamically
- Compared with TransR/CTransR, TransD not only considers the diversity of relations, but also entities
- TransD has less parameters and has no matrix-vector multiplication operations, which makes it applicable to large scale graphs

KG2E

- Represent relations / entities with Gaussian distribution
- Consider (un)certainties of entities and relations



He, et al. Learning to Represent Knowledge Graphs with Gaussian Embedding. In: Proc. CIKM **2015**.

Further Developments

- Stochastic Neighbor Embeddings (SNE)
  - Non-linear dimensionality reduction

    Geoffrey Hinton and Sam Roweis. Stochastic neighbor embedding. In Proceedings of the 15th International Conference on Neural Information Processing Systems (NIPS'02). **2002**.

- Holographic Embeddings (HolE)

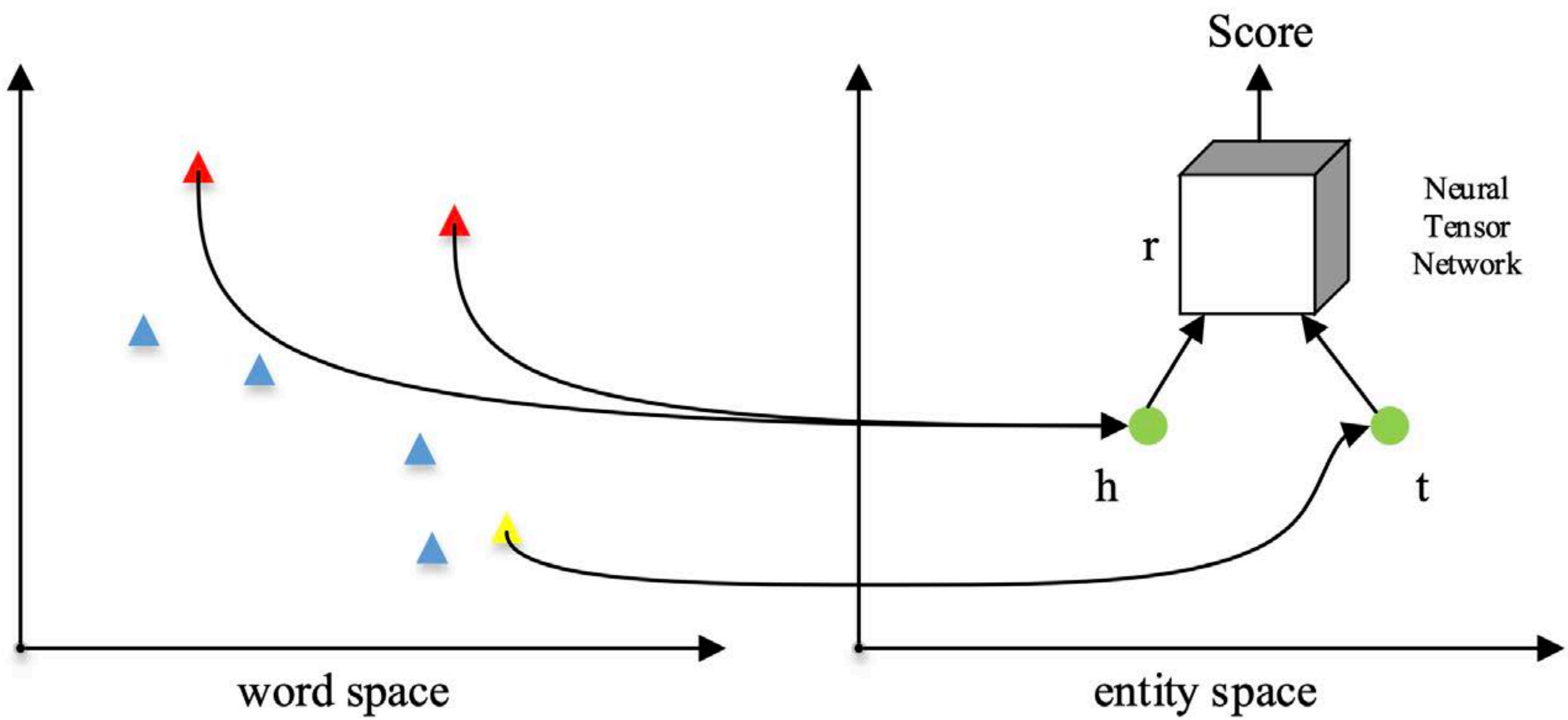
    Maximilian Nickel, Lorenzo Rosasco, and Tomaso Poggio. Holographic embeddings of knowledge graphs. In Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI'16). AAAI Press, 1955–1961. **2016**.

- Poincaré Embeddings

    Maximilian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17). **2017**.

KB Completion with Tensor Networks

- Model a KG with a Neural Tensor Network (NTN) and represent entities via word vectors

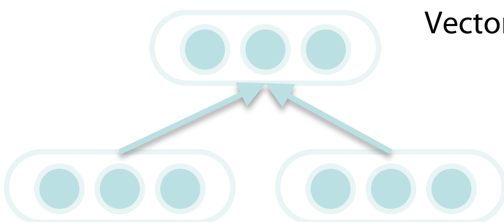


Socher, et al. Reasoning with neural tensor networks for knowledge base completion. In: Proc. NIPS **2013**.

KB Completion with Tensor Networks

- It represents each entity as the average of its word vectors

$\text{Vector}_{\text{Bank\_of\_China}}$



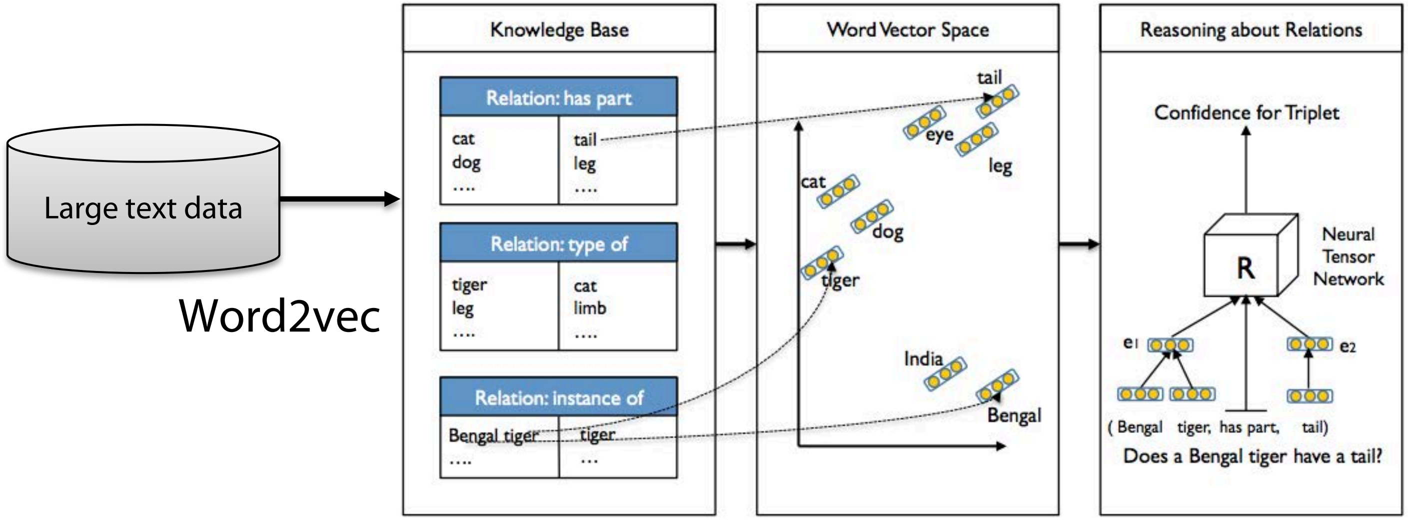
$\text{Vector}_{\text{Bank}}$ $\text{Vector}_{\text{China}}$

$\text{Vector}_{\text{Bank\_of\_China}} = 0.5 * (\text{Vector}_{\text{Bank}} + \text{Vector}_{\text{China}})$

- The goal of NTN is to be able to state whether two entities (e_1, e_2) are in a certain relationship R
  - For instance, whether the relationship (e_1, R, e_2) = (Bengal tiger, has part, tail) is true and with what certainty.

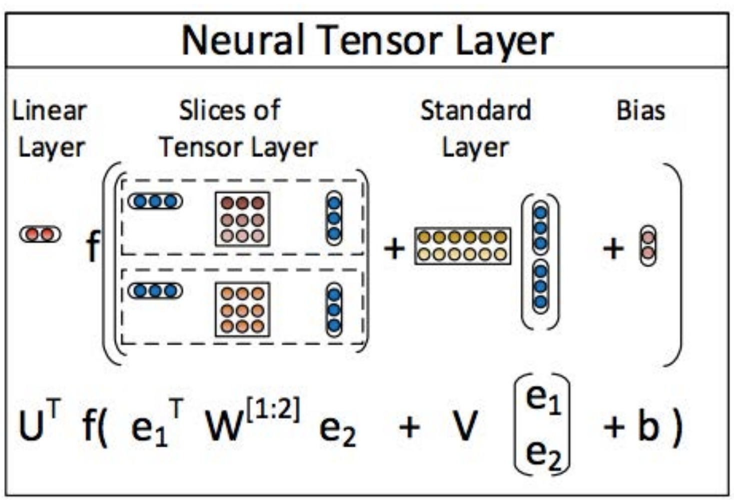
KB Completion with Tensor Networks

- Each relation is described by a so-called tensor network and pairs of entities are given as input to the model
- Each entity has a vector representation, which can be constructed by its word vectors
- The model returns a high score if they are in that relationship and a low one otherwise
- This allows any fact, whether implicitly or explicitly mentioned in the database, to be answered with a certainty score

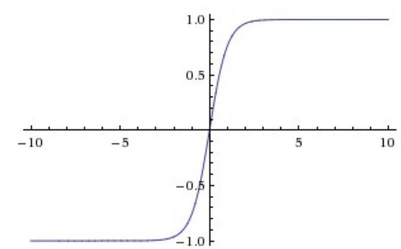


KB Completion with Tensor Networks

- The Neural Tensor Network (NTN) replaces a standard linear neural network layer with a bilinear tensor layer
  - that directly relates the two entity vectors across multiple dimensions
  - The model computes a score of how likely it is that two entities are in a certain relationship by the following NTN-based function $g(e_1, R, e_2)$:



f = tanh is a standard nonlinearity applied element-wise



$W_R^{[1:k]} \in \mathcal{R}^{d \times d \times k}$ is a tensor and the bilinear tensor product

Standard layer weight $V_R \in \mathcal{R}^{k \times 2d}$

Tensor Networks

- Originally developed in the context of condensed-matter physics and based on renormalization group ideas
  - Formal apparatus that allows systematic investigation of the changes of a physical system as viewed at different scales

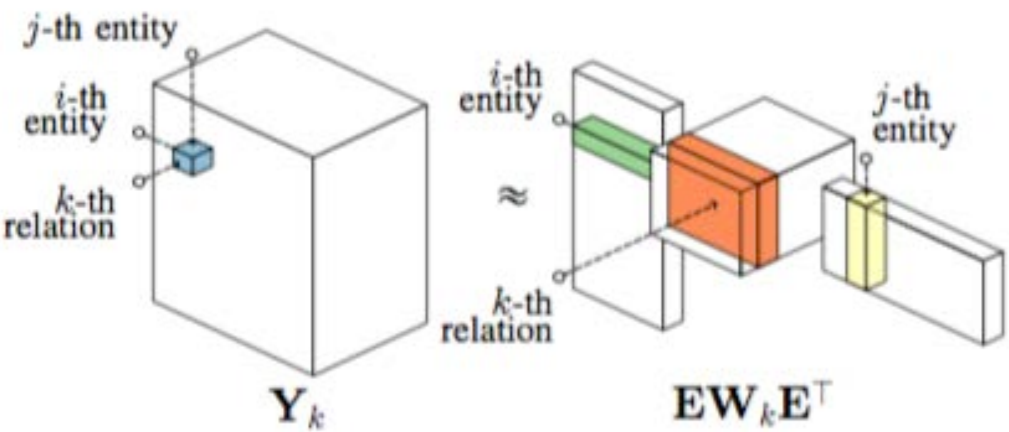
I recommend studying
tensor networks
in more detail!

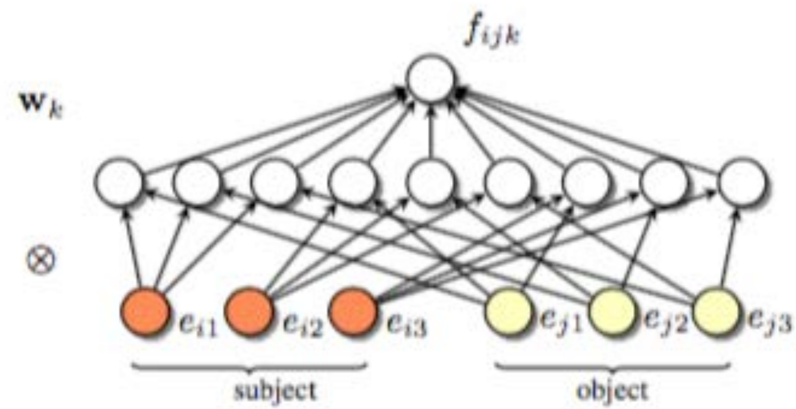
Back to Triple Scoring – Multiply instead of Add

- Multiplication: $\mathbf{h} \circ \mathbf{r}$ =?= $\mathbf{t}$
  - RESCAL: score(h,r,t) = $\mathbf{h}^\top \mathbf{W}_r \mathbf{t}$

Problem: Too many parameters!!

Pseudo neuro





Triple Scoring - Multiply

- Multiplication: $\mathbf{h} \circ \mathbf{r}$ =?= $\mathbf{t}$
  - RESCAL: score(h,r,t) = $\mathbf{h}^\top \mathbf{W}_r \mathbf{t}$
  - DistMult: score(h,r,t) = $\mathbf{h}^\top \mathrm{diag}(\mathbf{r})\mathbf{t}$
    Simplify RESCAL by using a diagonal matrix

B. Yang, W. Yih, X. He, J. Gao, and L. Deng. Embedding entities and relations for learning and inference in knowledge bases. In Proceedings of the International Conference on Learning Representations (ICLR), **2015**.

Triple Scoring - Multiply

- Multiplication: $\mathbf{h} \circ \mathbf{r}$ =?= $\mathbf{t}$
  - RESCAL: score(h,r,t) = $\mathbf{h}^\top \mathbf{W}_r \mathbf{t}$
  - DistMult: score(h,r,t) = $\mathbf{h}^\top \mathrm{diag}(\mathbf{r}) \mathbf{t}$
    Simplify RESCAL by using a diagonal matrix

  Problem: Cannot deal with asymmetric relations!!

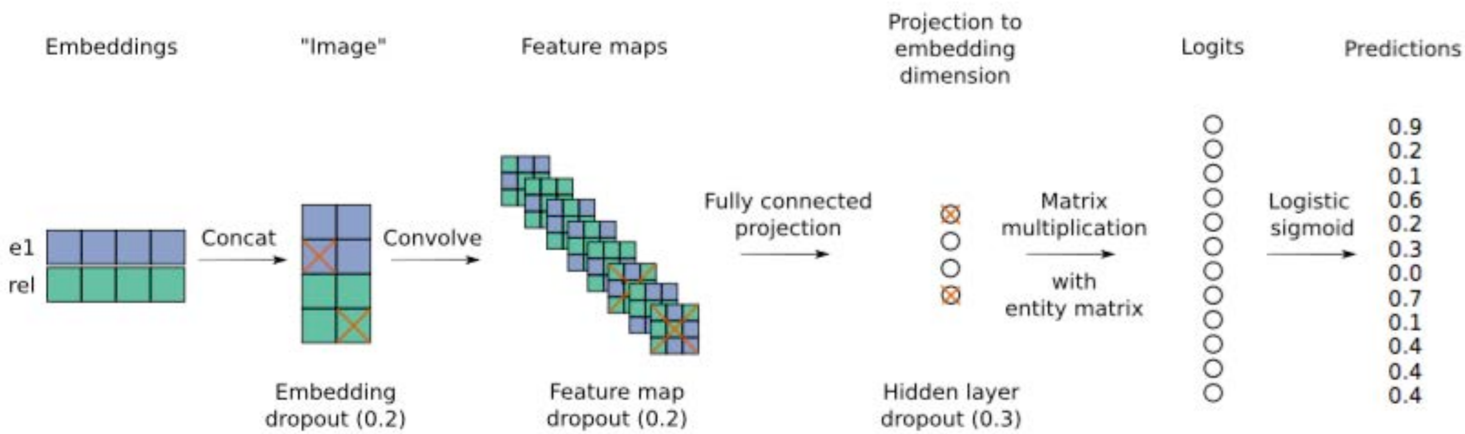
Triple Scoring - Multiply

- Multiplication: $\mathbf{h} \circ \mathbf{r}$ =?= $\mathbf{t}$
  - RESCAL: score(h,r,t) = $\mathbf{h}^\top \mathbf{W}_r \mathbf{t}$
  - DistMult: score(h,r,t) = $\mathbf{h}^\top \mathrm{diag}(\mathbf{r})\mathbf{t}$
    Simplify RESCAL by using a diagonal matrix
  - ComplEx: score(h,r,t) = $\mathrm{Re}(\mathbf{h}^\top \mathrm{diag}(\mathbf{r})\mathbf{t})$ –
    Extend DistMult by introducing complex value embedding, so can handle asymmetric relations

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**.

Triple Scoring - Multiply

- Multiplication: $\mathbf{h} \circ \mathbf{r}$ =?= $\mathbf{t}$
  - RESCAL: score(h,r,t) = $\mathbf{h}^\top \mathbf{W}_r \mathbf{t}$
  - DistMult: score(h,r,t) = $\mathbf{h}^\top \mathrm{diag}(\mathbf{r})\mathbf{t}$
  - ComplEx: score(h,r,t) = $\mathrm{Re}(\mathbf{h}^\top \mathrm{diag}(\mathbf{r})\mathbf{t})$
  - ConvE: Use convolutional network to reduce parameters



RotatE

- Representing head and tail entities in complex vector space, i.e., $\mathbf{h}, \mathbf{t} \in \mathbb{C}^{\boldsymbol{k}}$
- Define each relation $\mathbf{r}$ as an element-wise rotation from the head entity $\mathbf{h}$ to the tail entity $\mathbf{t}$, i.e.,

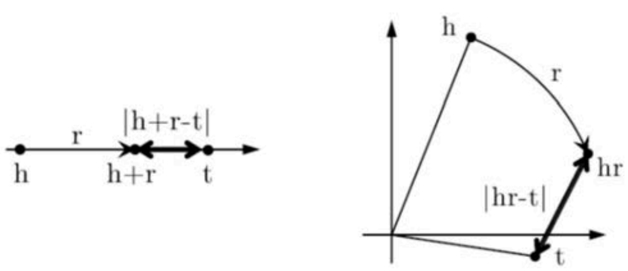
$$\mathbf{t} = \mathbf{h} \circ \boldsymbol{r}, \quad \text{where } |r_i|=1$$

- $\circ$ is the element-wise product. More specifically, we have $\mathrm{t_i} = \mathrm{h_i r_i}$, and

$$\mathrm{r_i} = e^{i\theta_{r,i}},$$

- where $\theta_{r,i}$ is the phase angle of $\mathbf{r}$ in the i-th dimension.
- Define the distance function of RotatE as

$$d_{\boldsymbol{r}}(\boldsymbol{h}, \boldsymbol{t}) = ||\mathbf{h} \circ \mathbf{r} - \mathbf{t}||$$



(a) TransE models $\mathbf{r}$ as translation in real line.

(b) RotatE models $\mathbf{r}$ as rotation in complex plane.

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

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)	$-\|\|(e_s + w_r) - e_o\|\|_p$	$w_r \in \mathbb{R}^K$	$\mathcal{O}(K)$	$\mathcal{O}(K)$
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?

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

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

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

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**

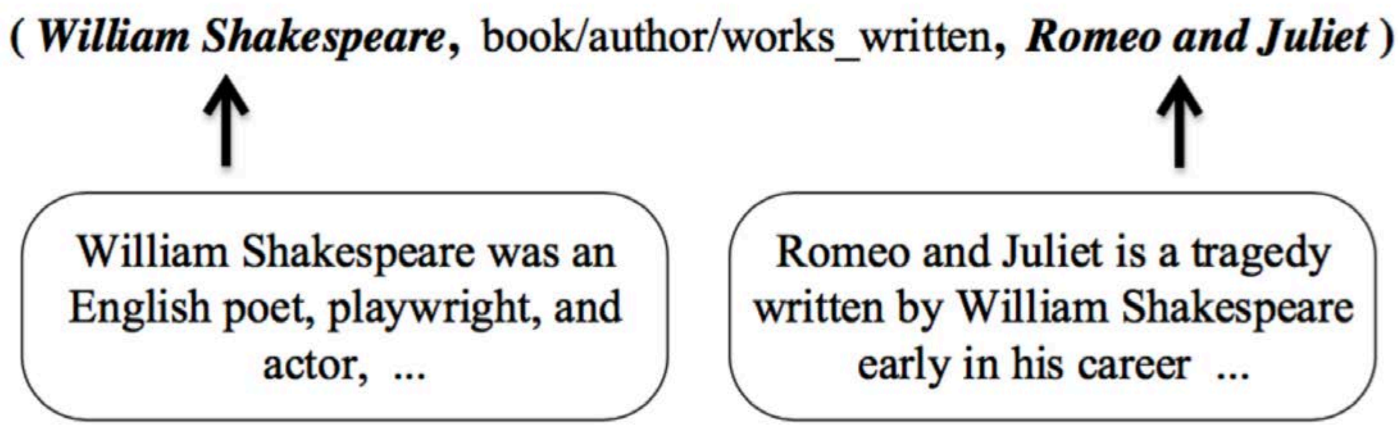
Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, Jian Tang:
RotatE: Knowledge Graph Embedding by Relational Rotation
in Complex Space. In Proc. ICLR 2019.

Fusion of Text and KG

- Relation prediction for KG – r ~ t-h
- Relation extraction from text

RL4KG with Entity Descriptions

- KG contains rich information besides network structure

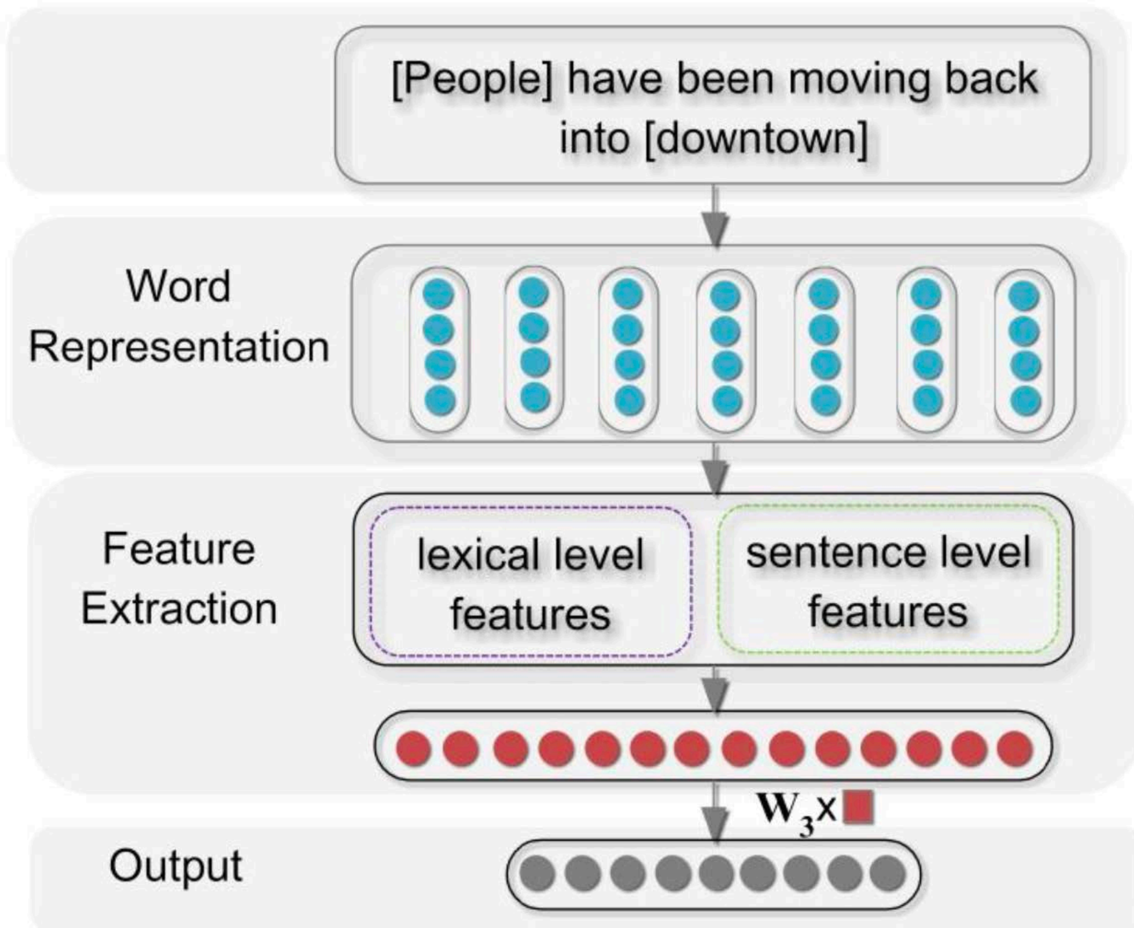


TransE+Word2Vec

- KG=>TransE, Text=>Word2Vec
- Make embeddings of the same entities in KGs and text related to each other

Wang, et al. Knowledge graph and text jointly embedding.
In: Proc. EMNLP **2014**.

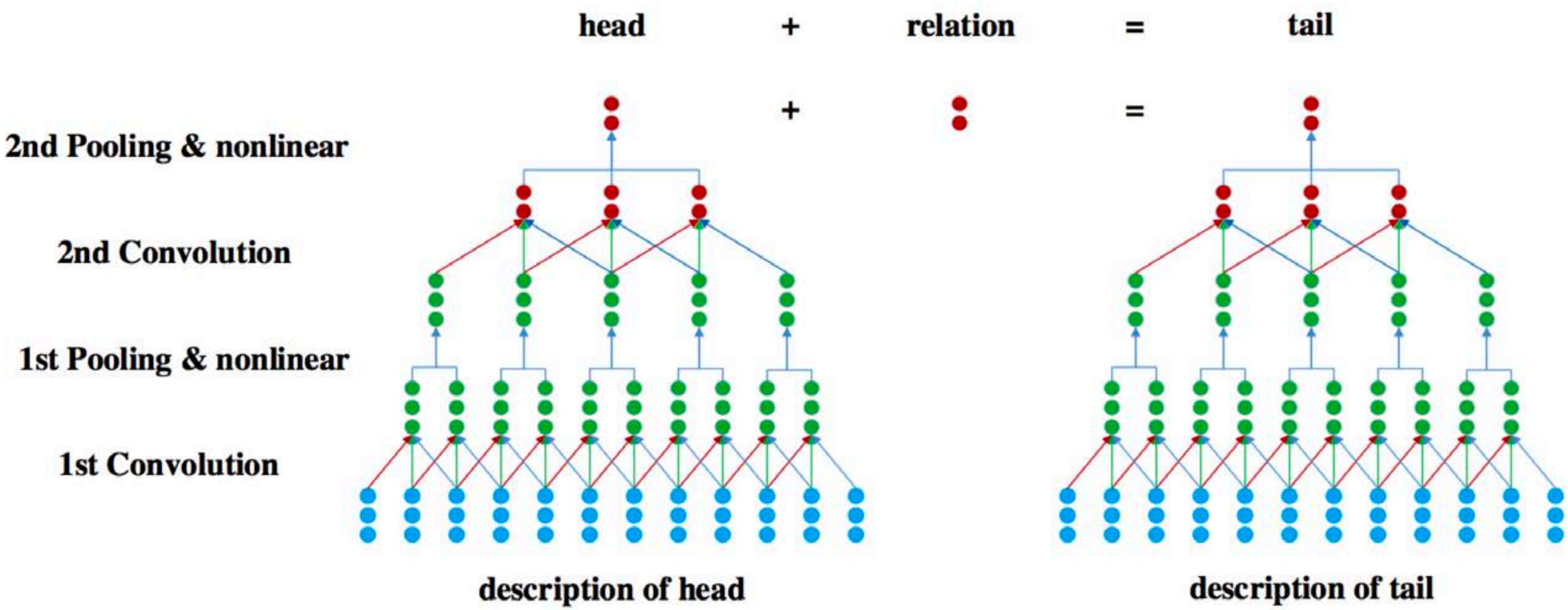
Relation Classification via CNN



Zeng, et al. Relation Classification via Convolutional Deep Neural Network. In: Proc. COLING **2014**.

Description-Embodied RL4KG

- Enhance entity representation with descriptions
- Model descriptions with convolutional network



Xie, et al. Representation Learning of Knowledge Graphs with Entity Descriptions. In: Proc. AAAI **2016**.

TransSparse

- Deals with heterogeneity …
  - Some relations link many entity pairs and others do not
- … and imbalance
  - The number of head entities and that of tail entities in a relation could be different
- Transfer matrices replaced by adaptive sparse matrices
  - Sparseness degrees determined by the number of entities (or entity pairs) linked by relations

Guoliang Ji, Kang Liu, Shizhu He, and Jun Zhao. Knowledge graph completion with adaptive sparse transfer matrix. In Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI'16). **2016**.

TransF

- Model the correlation between relations
- Translation–based method mitigating the burden of relation projection by explicitly modeling the basis subspaces of projection matrices

Method	WN18			FB15k			WN18RR			FB15k-237		
	MR	MRR	Hits@10	MR	MRR	Hits@10	MR	MRR	Hits@10	MR	MRR	Hits@10
TransE	251	-	89.2	125	-	47.1	-	-	-	-	-	-
TransH	388	-	82.3	87	-	64.4	-	-	-	-	-	-
TransR	225	-	92.0	77	-	68.7	-	-	-	-	-	-
CTransR	218	-	92.3	75	-	70.2	-	-	-	-	-	-
TransD	212	-	92.2	91	-	77.3	-	-	-	-	-	-
TransSparse (s)	221	-	92.8	82	-	79.5	-	-	-	-	-	-
TransSparse (us)	211	-	93.2	82	-	79.9	-	-	-	-	-	-
TransF	**198**	0.856	*95.3*	62	0.564	*82.3*	**3246**	**0.505**	49.8	**210**	0.286	**47.2**
PTransE	-	-	-	**58**	-	84.6	-	-	-	-	-	-
KG2E	331	-	92.8	59	-	74.0	-	-	-	-	-	-
ManifoldE	-	-	93.2	-	-	88.1	-	-	-	-	-	-
DistMult*	-	0.83	93.6	-	0.35	57.7	5110	0.425	49.1	254	0.241	41.9
ComplEx*	-	0.941	94.7	-	0.69	84.0	5261	0.444	**50.7**	248	0.240	41.9
ConvE	504	**0.942**	**95.5**	64	**0.745**	**87.3**	7323	0.342	41.1	330	**0.301**	45.8

Zichao Huang, Bo Li, and Jian Yin. Knowledge graph embedding via multiplicative interaction. In Proceedings of the 2nd International Conference on Innovation in Artificial Intelligence (ICIAI '18). **2018**.

Software

- OpenKE
  - https://github.com/thunlp/OpenKE
- KnowldgeGraphEmbedding
  - https://github.com/DeepGraphLearning/KnowledgeGraphEmbedding
- GraphVite
  - https://graphvite.io/

    Zhaocheng Zhu, Shizhen Xu, Meng Qu, Jian Tang. "GraphVite: A High-Performance CPU-GPU Hybrid System for Node Embedding". WWW'19. **2019**.
- DGL-KGE
  - https://github.com/awslabs/dgl-ke

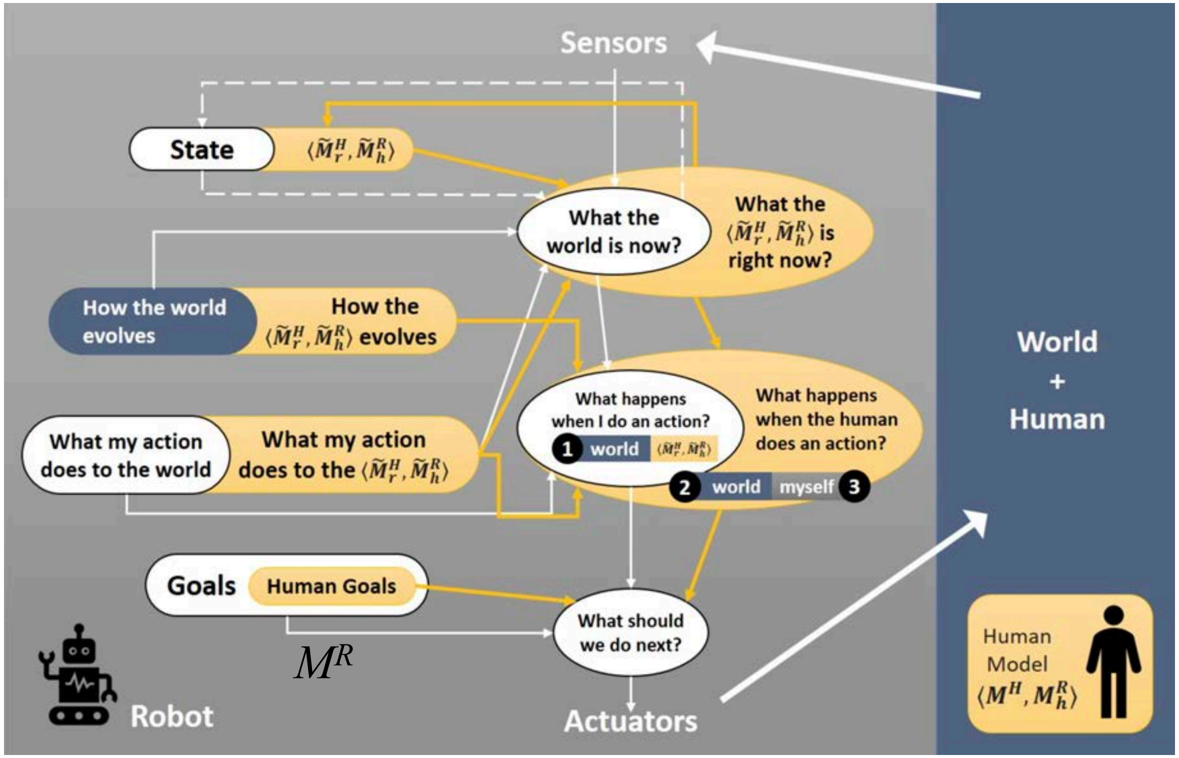
Embedding Approaches: Summary

- Continuous semantic representations that
  - Leverage existing rich linguistic resources
  - Discover new relations
  - Enable us to measure the degree of (different) relations
- Good ol' fashioned AI
  - Define embedding space
  - Implement learning approach
  - Evaluate against benchmark dataset
- Challenges
  - Modeling complex relations (explained in detail)
  - Fusion of text and KG (indicated above)
  - Combination with symbolic logic for inference on KG (next)

Back to Information Retrieval Agents

- Agents are told to fulfil IR goal
  - How exactly is the IR goal specified?
    - Search string
    - (Abstract of) example document
    - Formal query
    - Relational structure as an example
    - …

Human specifies goal: Solve a certain problem



- M^H human model of the problem to be solved
- $M^R{}_h$ is the human's understanding of the robot's M^R
- M^R robot model of the problem to be solved
- $\widetilde{M}^H{}_r$ is the robot's understanding of M^H (anticipate human behavior)
- $\widetilde{M}^R{}_h$ is the robot's understanding of $M^R{}_h$ (anticipate human's expectations)

Back to Information Retrieval Agents

- Agents are told to fulfil IR goal
  - How exactly is the IR goal specified?
    - Search string
    - (Abstract of) example document
    - Formal query
    - Relational structure as an example
    - …
- Agent develops plan to fulfil goal as fast as possible
  - How is the IR goal specified internally?
    - Match word vectors of IR goal with word vectors of documents
    - Match relational structure of IR goal with relational structures of documents
    - …