Modeling Topics and Knowledge Bases with Embeddings

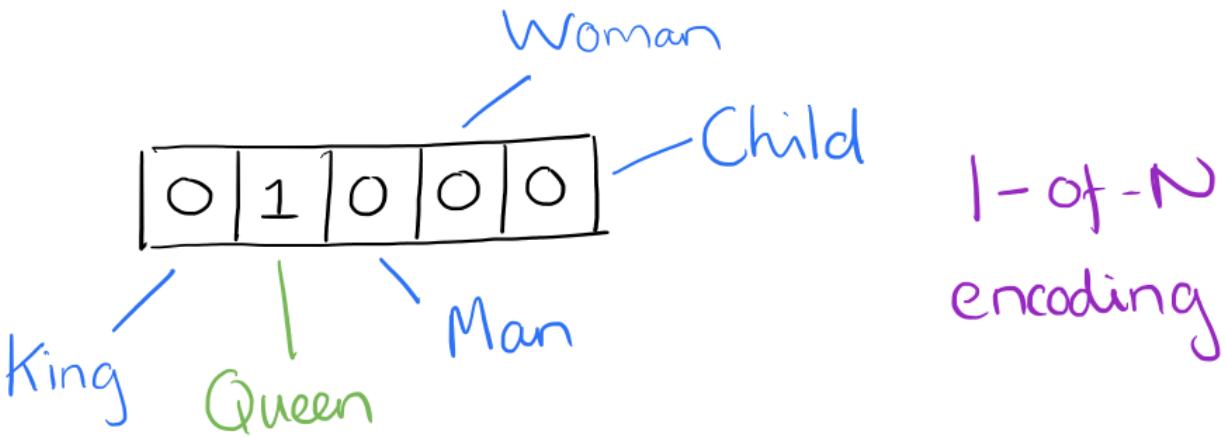
Dat Quoc Nguyen and Mark Johnson

Department of Computing
Macquarie University
Sydney, Australia

December 2016

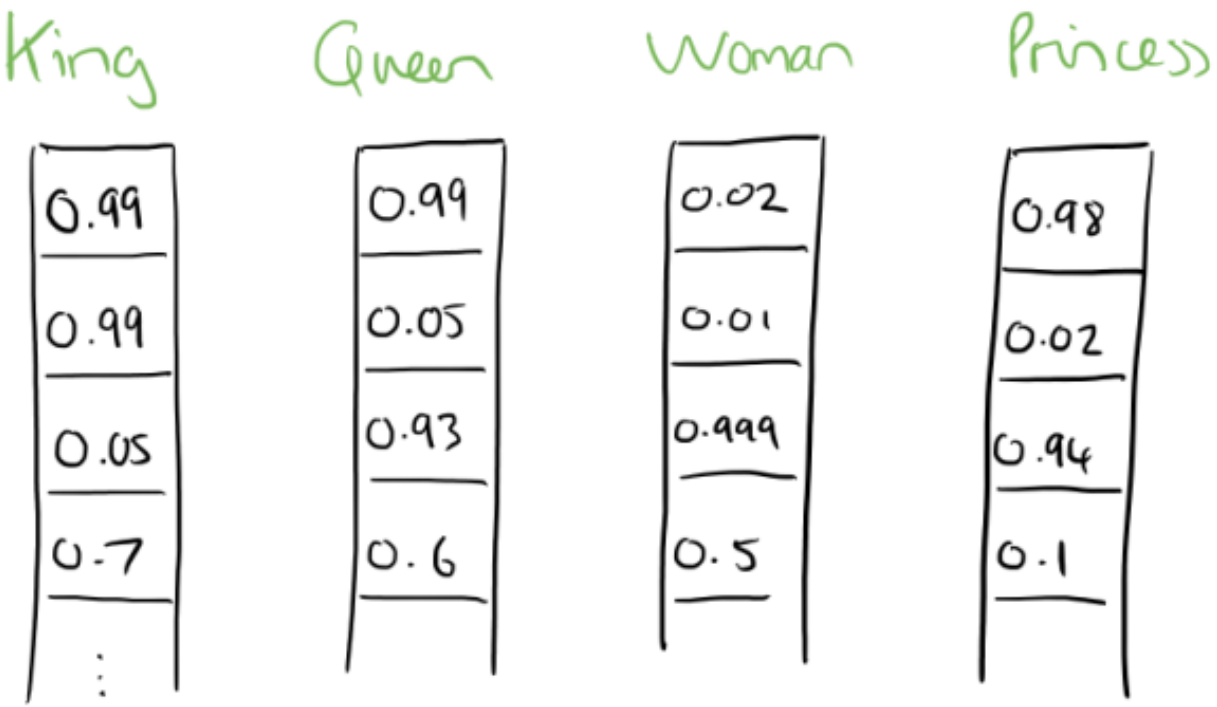
Vector representations/embeddings of words

- One-hot representation: high-dimensional and sparse vector
  - Vocabulary size: N
  - N-dimensional vector: filled with 0s, except for a 1 at the position associated with word index



Vector representations/embeddings of words

- Deep learning evolution: most neural network toolkits do not play well with one-hot representations
  - Dense vector: low-dimensional distributed vector representation
  - Vector size $k \ll N$, for example: $k = 100$, Vocabulary size $N = 100000$



Vector representations/embeddings of words

- Unsupervised learning models are proposed to learn low-dimensional vectors of words efficiently, e.g W2V Skip-gram
- Word embeddings learned from large external corpora capture various aspects of word meanings
  - Possibly assign topics (i.e. labels) to clusters of "similar" meaning words
  - Topic **1** for $\{$*banking*, *bank*, *transaction*, *finance*, $_{money}$*laundering*$\}$

$\Rightarrow$Can we incorporate word embeddings to topic models?

Improving topic models with word embeddings

- *Topic models* take a corpus of documents as input, and
  - Learn a set of latent *topics* for the corpus
  - Infer *document-to-topic* and *topic-to-word* distributions from co-occurrence of words within documents

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

- If the corpus is small and/or the documents are short, the topics will be noisy due to the limited information of word co-occurrence
- IDEA: Use the word embeddings learned on a large external corpus to improve the topic-word distributions in a topic model

Improving topic models with word embeddings

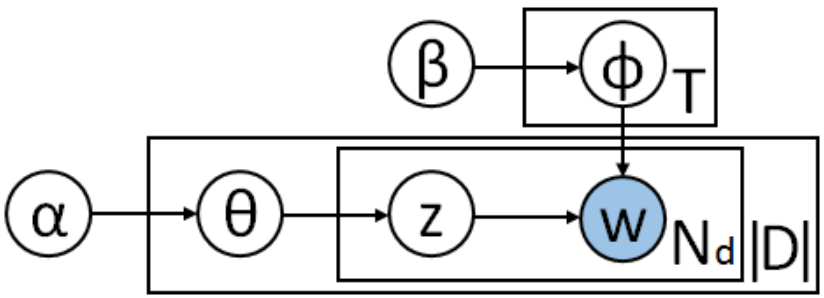
- Each word w is associated with a pre-trained *word embedding* $\boldsymbol{\omega}_w$
- Each topic t is associated with a *topic embedding* $\boldsymbol{\tau}_t$
- We define a latent feature topic-to-word distribution $\text{CatE}(w)$ over words:

$$\text{CatE}(w \mid t) = \frac{\exp(\boldsymbol{\omega}_w \cdot \boldsymbol{\tau}_t)}{\sum_{w' \in V} \exp(\boldsymbol{\omega}_{w'} \cdot \boldsymbol{\tau}_t)}$$

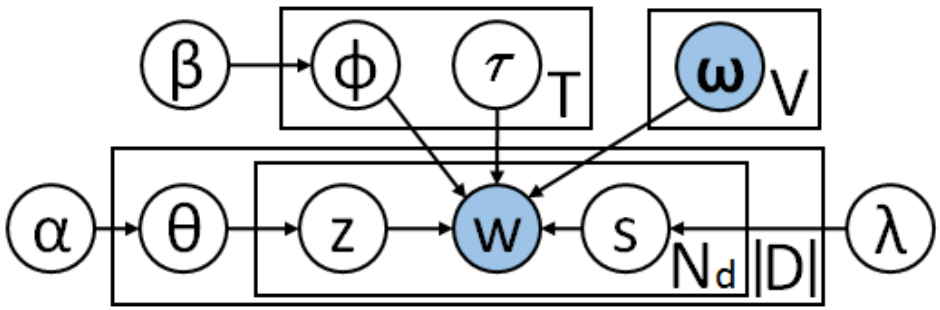
  - Optimize the log-loss to learn $\boldsymbol{\tau}_t$
- Our new topic models *mix the* CatE *distribution* with a multinomial distribution over words
  - Combine information from a large, general corpus (via the CatE distribution) and a smaller but more specific corpus (via the multinomial distribution)

Improving topic models with word embeddings

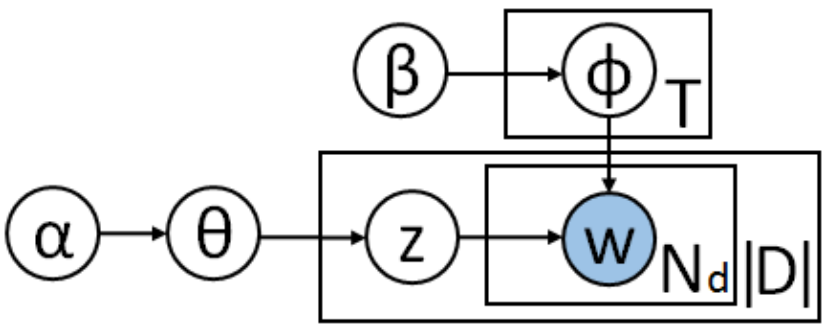
- Combine Latent Dirichlet Allocation (LDA) and Dirichlet Multinomial Mixture (DMM) with the word embeddings: LF-LDA & LF-DMM



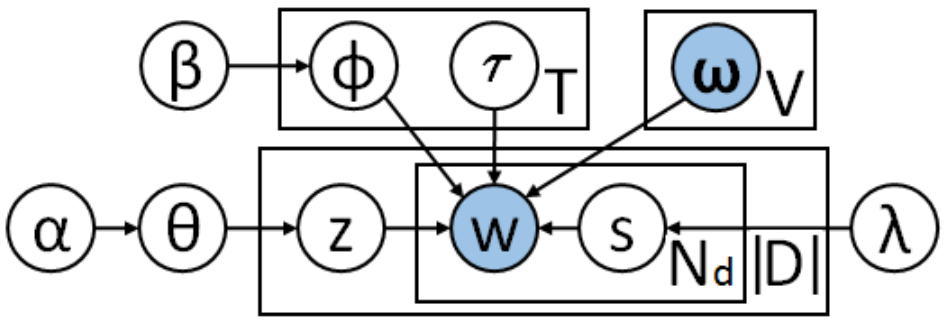
(LDA)



(LF-LDA)



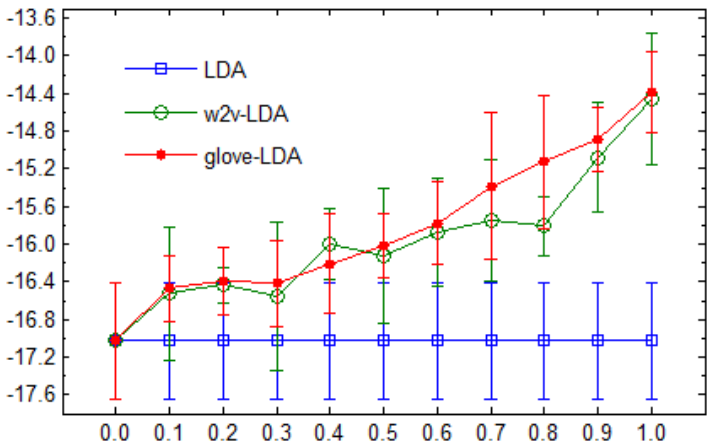
(DMM)



(LF-DMM)

Improving topic models with word embeddings

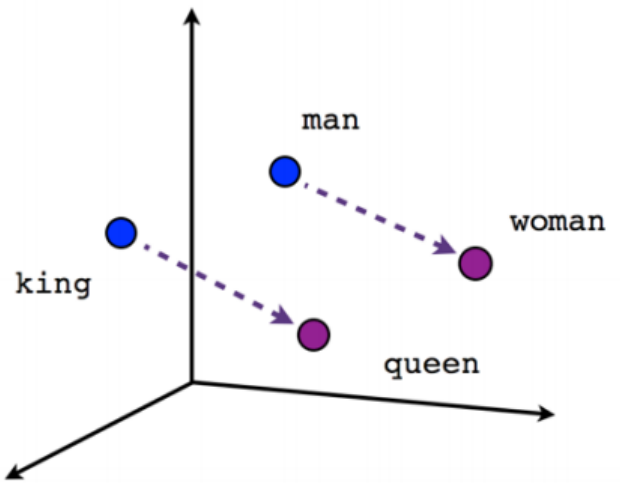
Topic 1		Topic 3		Topic 4	
DMM	LF-DMM	DMM	LF-DMM	DMM	LF-DMM
japan	japan	u.s.	prices	egypt	libya
nuclear	nuclear	oil	sales	china	egypt
u.s.	u.s.	japan	oil	u.s	**iran**
crisis	plant	prices	u.s.	mubarak	**mideast**
plant	**quake**	stocks	profit	bin	**opposition**
china	radiation	sales	stocks	libya	**protests**
libya	**earthquake**	profit	japan	laden	**leader**
radiation	**tsunami**	fed	rise	france	**syria**
u.n.	**nuke**	rise	**gas**	bahrain	u.n.
vote	crisis	growth	growth	air	**tunisia**
korea	**disaster**	wall	shares	report	**chief**
europe	**power**	street	**price**	rights	**protesters**
government	**oil**	china	**profits**	court	mubarak
election	**japanese**	fall	**rises**	u.n.	**crackdown**
deal	**plants**	shares	**earnings**	war	bahrain

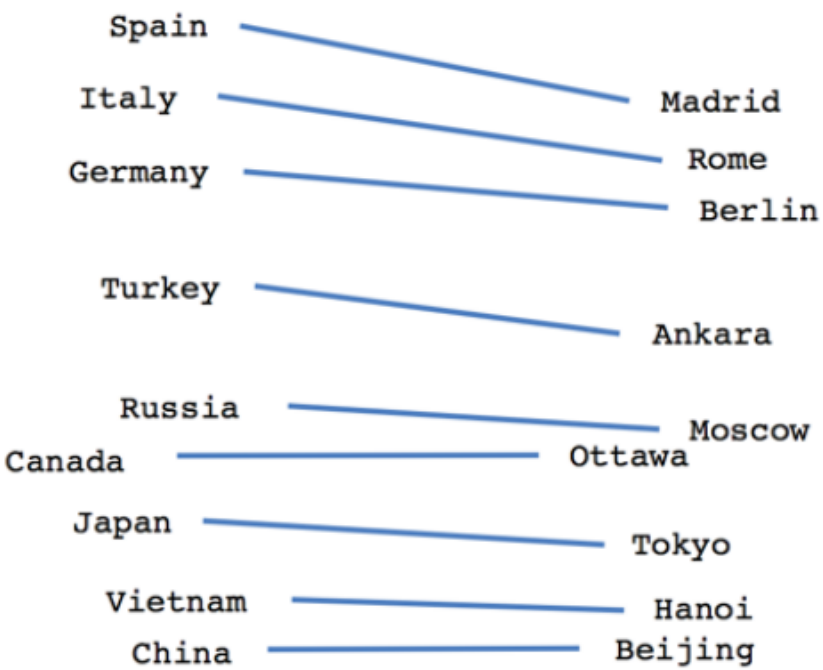


(Topic coherence)

- Significant improvement of topic coherence scores on all models and experimental corpora
- Obtain **5+%** absolute improvements in clustering and classification evaluation scores on the small or short datasets
- No reliable difference between pre-trained Word2Vec and Glove vectors

Vector representations/embeddings: "distance" results





$$\boldsymbol{v}_{king} - \boldsymbol{v}_{queen} \approx \boldsymbol{v}_{man} - \boldsymbol{v}_{woman}$$

$$\boldsymbol{v}_{Vietnam} - \boldsymbol{v}_{Hanoi} \approx \boldsymbol{v}_{Japan} - \boldsymbol{v}_{Tokyo} \approx \boldsymbol{v}_{Germany} - \boldsymbol{v}_{Berlin} \approx \boldsymbol{v}_{some\_relationship,\ saying:\ is\_capital\_of}$$

$$\boldsymbol{v}_{Hanoi} + \boldsymbol{v}_{is\_capital\_of} \approx \boldsymbol{v}_{Vietnam}$$

$$\boldsymbol{v}_{Tokyo} + \boldsymbol{v}_{is\_capital\_of} \approx \boldsymbol{v}_{Japan}$$

$$\boldsymbol{v}_{Berlin} + \boldsymbol{v}_{is\_capital\_of} \approx \boldsymbol{v}_{Germany}$$

Triple $(head\ entity, relation, tail\ entity)$

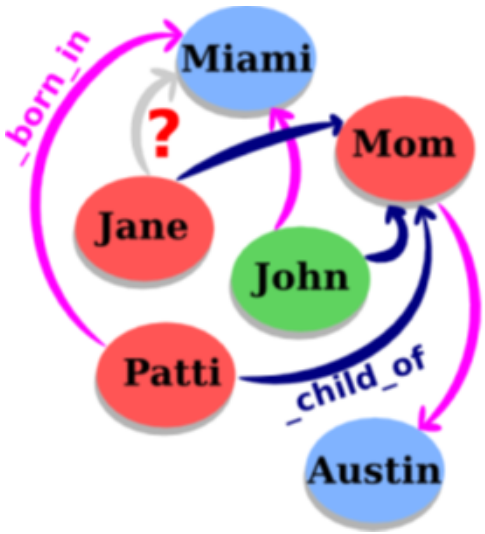
$$\boldsymbol{v}_h + \boldsymbol{v}_r \approx \boldsymbol{v}_t$$

$$\Rightarrow \|\boldsymbol{v}_h + \boldsymbol{v}_r - \boldsymbol{v}_t\|_{\ell_{1/2}} \approx 0$$

Link prediction in knowledge bases?

STransE: a new embedding model for link prediction

- Knowledge bases (KBs) of real-world triple facts (head entity, relation, tail entity) are useful resources for NLP tasks
- **Issue**: large KBs are still far from complete
- So it is useful to perform *link prediction in KBs* or *knowledge base completion*: predict which triples not in a knowledge base are likely to be true
- Embedding models for link prediction in KBs:
  - Associate entities and/or relations with dense feature vectors or matrices
  - Obtain SOTA performance and generalize to large KBs



STransE: a new embedding model for link prediction

- The TransE model (Bordes et al., 2013) represents each relation r by a translation vector $\boldsymbol{v}_r$, which is chosen so that $\|\boldsymbol{v}_h + \boldsymbol{v}_r - \boldsymbol{v}_t\|_{\ell_{1/2}} \approx 0$
  - Good for 1-to-1 relationships, e.g: *is_capital_of*
  - Not good for 1-to-Many, Many-to-1 and Many-to-Many, e.g: *gender*
- STransE: a novel embedding model of entities and relationships in KBs
  - Our STransE represents each entity as a low dimensional vector, and each relation by two matrices and a translation vector
  - STransE choose matrices $\mathbf{W}_{r,1}$ and $\mathbf{W}_{r,2}$, and vector $\boldsymbol{v}_r$ so that: $\|\mathbf{W}_{r,1}\boldsymbol{v}_h + \boldsymbol{v}_r - \mathbf{W}_{r,2}\boldsymbol{v}_t\|_{\ell_{1/2}} \approx 0$
  - Optimize a margin-based objective function to learn the vectors and matrices

STransE results for link prediction in KBs

- We conducted experiments on two benchmark datasets WN18 and FB15k (Bordes et al., 2013)

Dataset	#E	#R	#Train	#Valid	#Test
WN18	40,943	18	141,442	5,000	5,000
FB15k	14,951	1,345	483,142	50,000	59,071

- Link prediction task:
  - Predict h given $(?, r, t)$ or predict t given $(h, r, ?)$ where ? denotes the missing element
  - Evaluation metrics: mean rank (MR) and Hits@10 (H10)

Method	WN18		FB15k	
	MR	H10	MR	H10
TransE	251	89.2	125	47.1
TransH	303	86.7	87	64.4
TransR	225	92.0	77	68.7
CTransR	218	92.3	75	70.2
KG2E	348	93.2	59	74.0
TransD	212	92.2	91	77.3
TATEC	-	-	**58**	76.7
STransE	**206**	**93.4**	69	**79.7**
RTransE	-	-	**50**	76.2
PTransE	-	-	58	84.6

- Find a new relation-path based embedding model in our CoNLL paper!

STransE results for search personalization

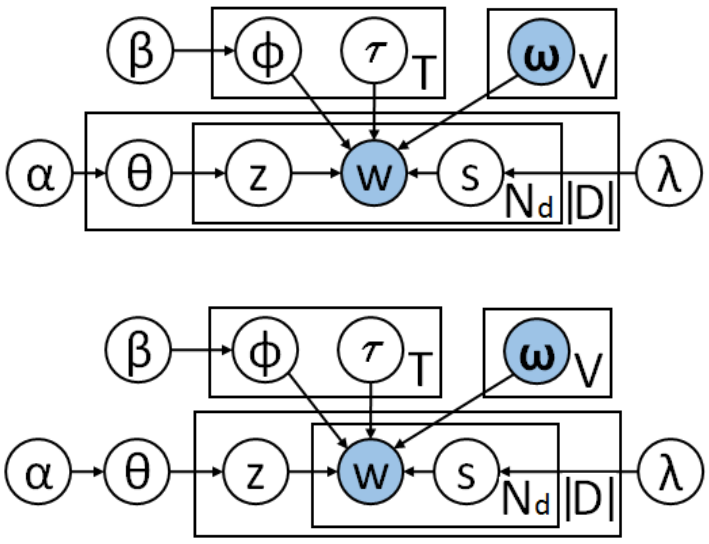
- Two users search using the same keywords, they are often looking for different information (i.e. difference due to the users' interests)
- Personalized search customizes results based on user's search history (i.e. submitted queries and clicked documents)
- Let (q, u, d) represent a triple (query, user, document)
- Represent the user u by two matrices $\mathbf{W}_{u,1}$ and $\mathbf{W}_{u,2}$ and a vector $\boldsymbol{v}_u$, which represents the user's topical interests, so that:

$$\|\mathbf{W}_{u,1}\boldsymbol{v}_q + \boldsymbol{v}_u - \mathbf{W}_{u,2}\boldsymbol{v}_d\|_{\ell_{1/2}} \approx 0$$

- $\boldsymbol{v}_d$ and $\boldsymbol{v}_q$ are pre-determined by employing the LDA topic model
- Optimize a margin-based objective function to learn the user embeddings and matrices

Metric	**Search Eng.**	**L2R SP**	**STransE**	**TransE**
MRR	0.559	$0.631_{+12.9\%}$	$\mathbf{0.656}_{+17.3\%}$	$0.645_{+15.4\%}$
P@1	0.385	$0.452_{+17.4\%}$	$\mathbf{0.501}_{+30.3\%}$	$0.481_{+24.9\%}$

Conclusions



$$\|\mathbf{W}_{r,1}\boldsymbol{v}_h + \boldsymbol{v}_r - \mathbf{W}_{r,2}\boldsymbol{v}_t\|_{\ell_{1/2}}$$

- Latent feature vector representations induced from large external corpora can be used to improve topic modeling on smaller datasets
- Our new embedding model STransE for link prediction in KBs
  - Applying to a search personalization task, STransE helps to significantly improve the ranking quality
- https://github.com/datquocnguyen

Thank you for your attention!

- Dat Quoc Nguyen, Richard Billingsley, Lan Du and Mark Johnson. Improving Topic Models with Latent Feature Word Representations. *Transactions of the ACL*, 2015.
- Dat Quoc Nguyen, Kairit Sirts, Lizhen Qu and Mark Johnson. STransE: a novel embedding model of entities and relationships in knowledge bases. In *Proceedings of NAACL-HLT*, 2016.
- Dat Quoc Nguyen, Kairit Sirts, Lizhen Qu and Mark Johnson. Neighborhood Mixture Model for Knowledge Base Completion. In *Proceedings of CoNLL*, 2016.
- Thanh Vu*, Dat Quoc Nguyen*, Mark Johnson, Dawei Song and Alistair Willis. Search Personalization with Embeddings. In *Proceedings of ECIR 2017*, to appear.