Learning Discriminative Projections for Text Similarity Measures

Scott Wen-tau Yih

Joint work with
Kristina Toutanova, John Platt, Chris Meek

Microsoft Research

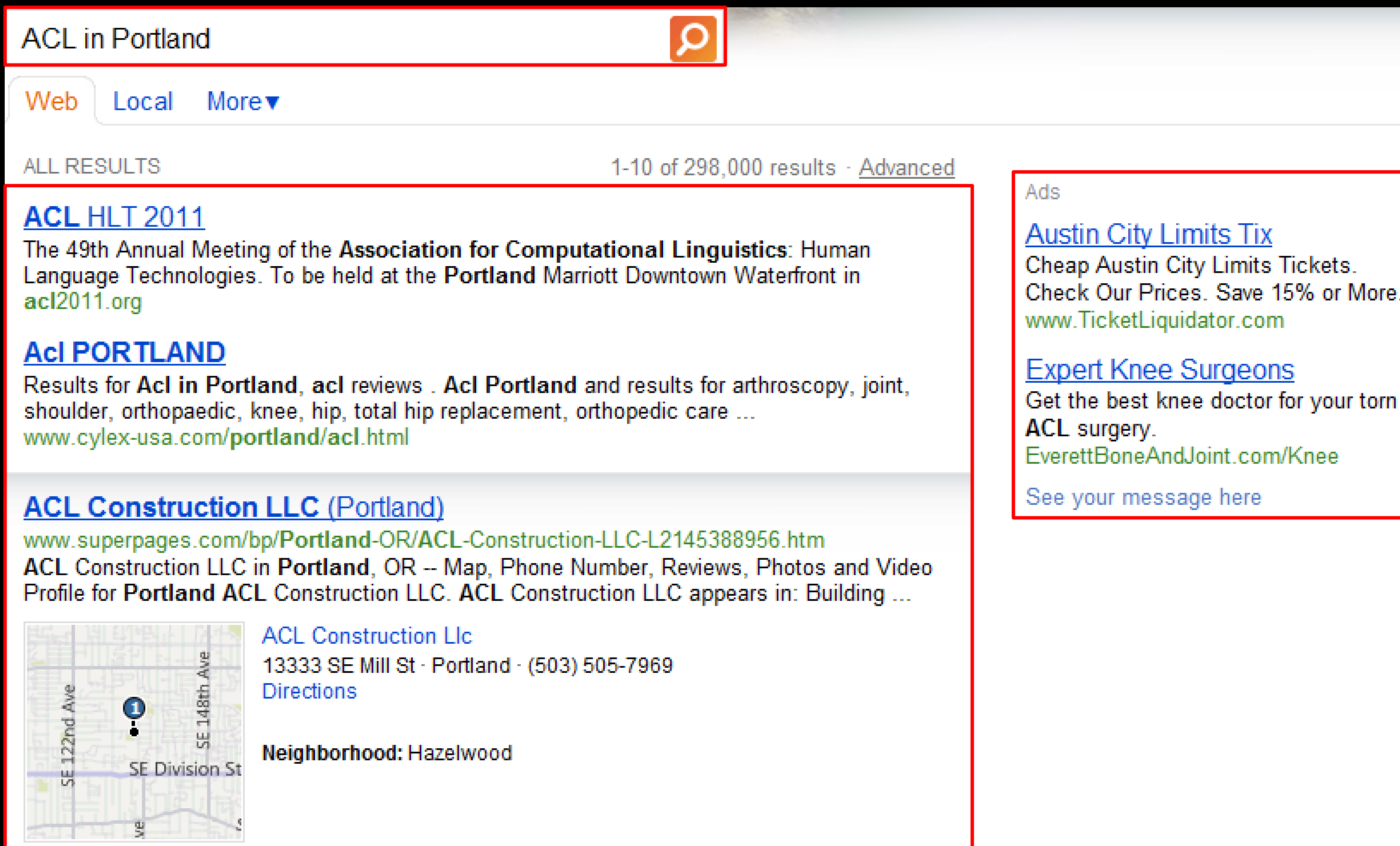
Cross-language Document Retrieval

English Query Doc

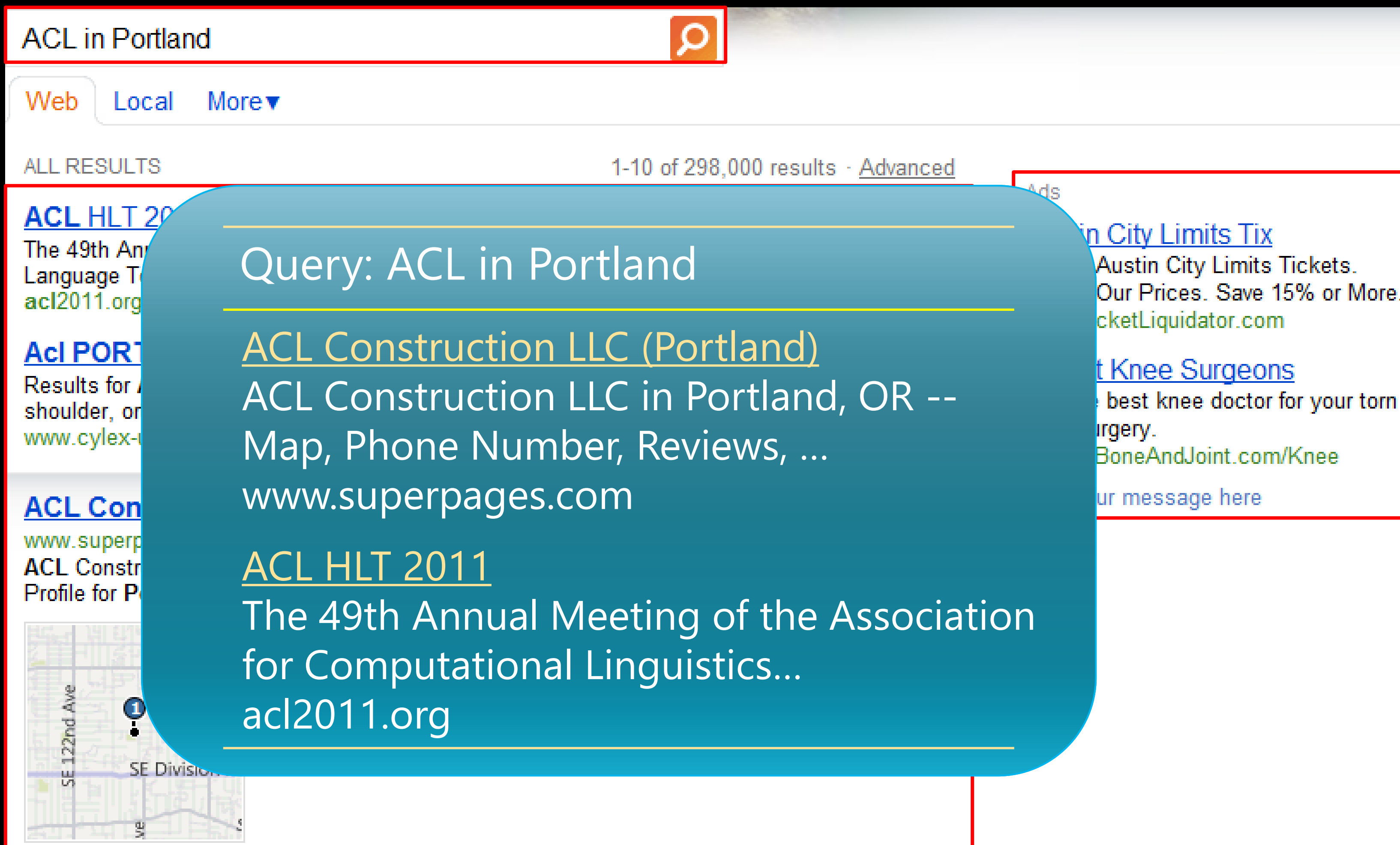

Article Discussion Read View source
Penguin
From Wikipedia, the free encyclopedia
(Redirected from Spheniscidae)
For other uses, see Penguin (disambiguation).
Penguins (order **Sphenisciformes**, family **Spheniscidae**) are a group of aquatic, flightless birds living almost exclusively in the southern hemisphere, especially in Antarctica. Highly adapted for life in the water, penguins have countershaded dark and white plumage, and their

Spanish Document Set

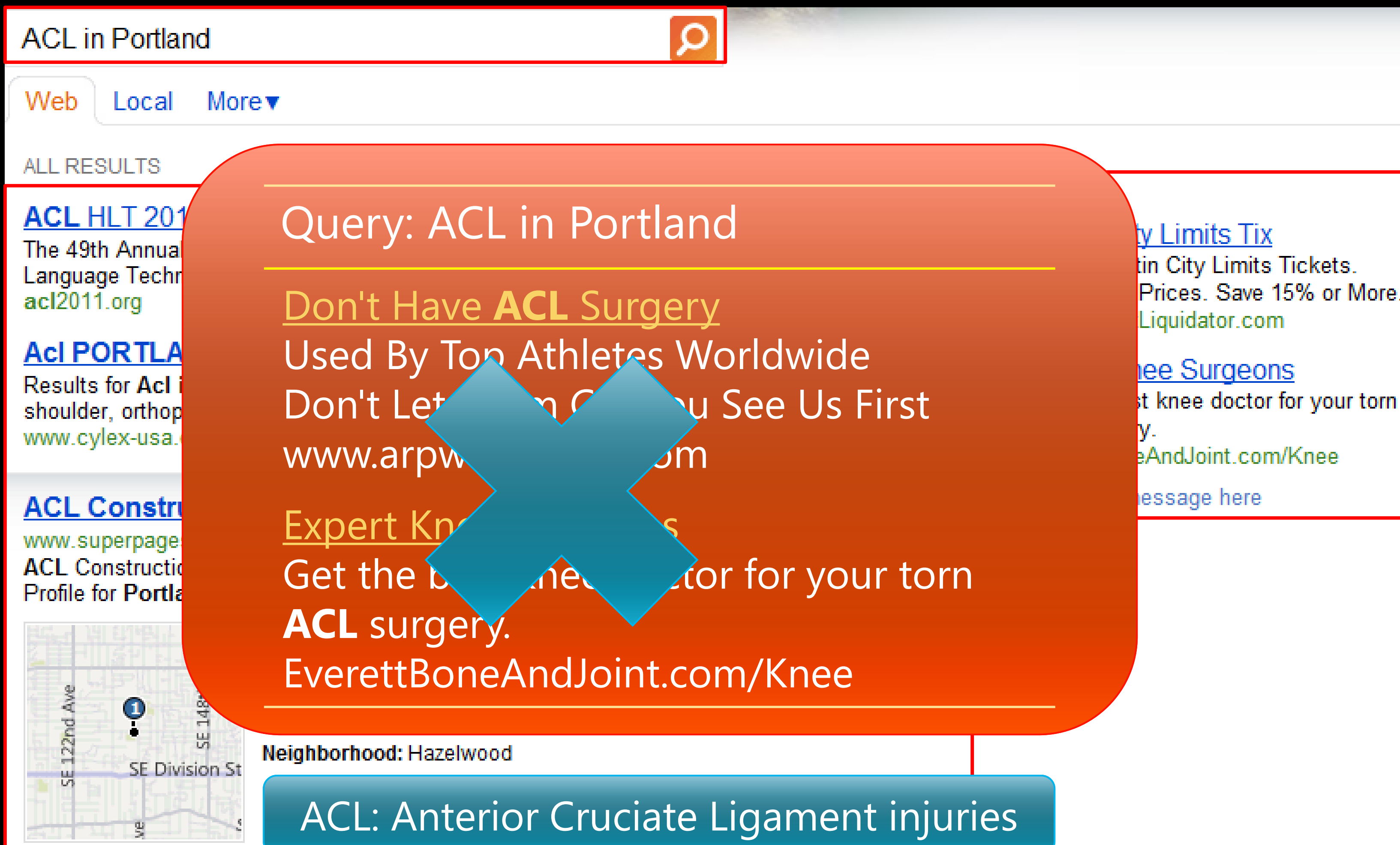
Web Search & Advertising



Web Search & Advertising

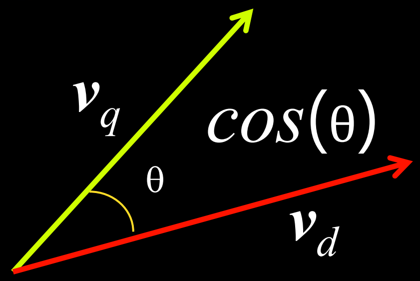


Web Search & Advertising



ACL: Anterior Cruciate Ligament injuries

Vector Space Model

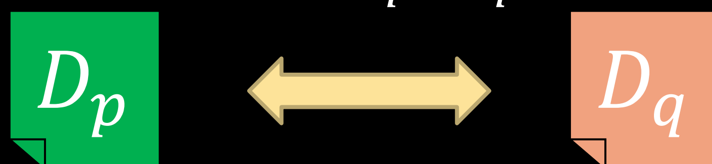


- Represent text objects as vectors
  - Word/Phrase: term co-occurrences
  - Document: term vectors with TFIDF/BM25 weighting
  - Similarity is determined using functions like cosine of the corresponding vectors
- Weaknesses
  - Different but related terms cannot be matched
    - e.g., (buy, used, car) vs. (purchase, pre-owned, vehicle)
  - Not suitable for cross-lingual settings

Learning Concept Vector Representation

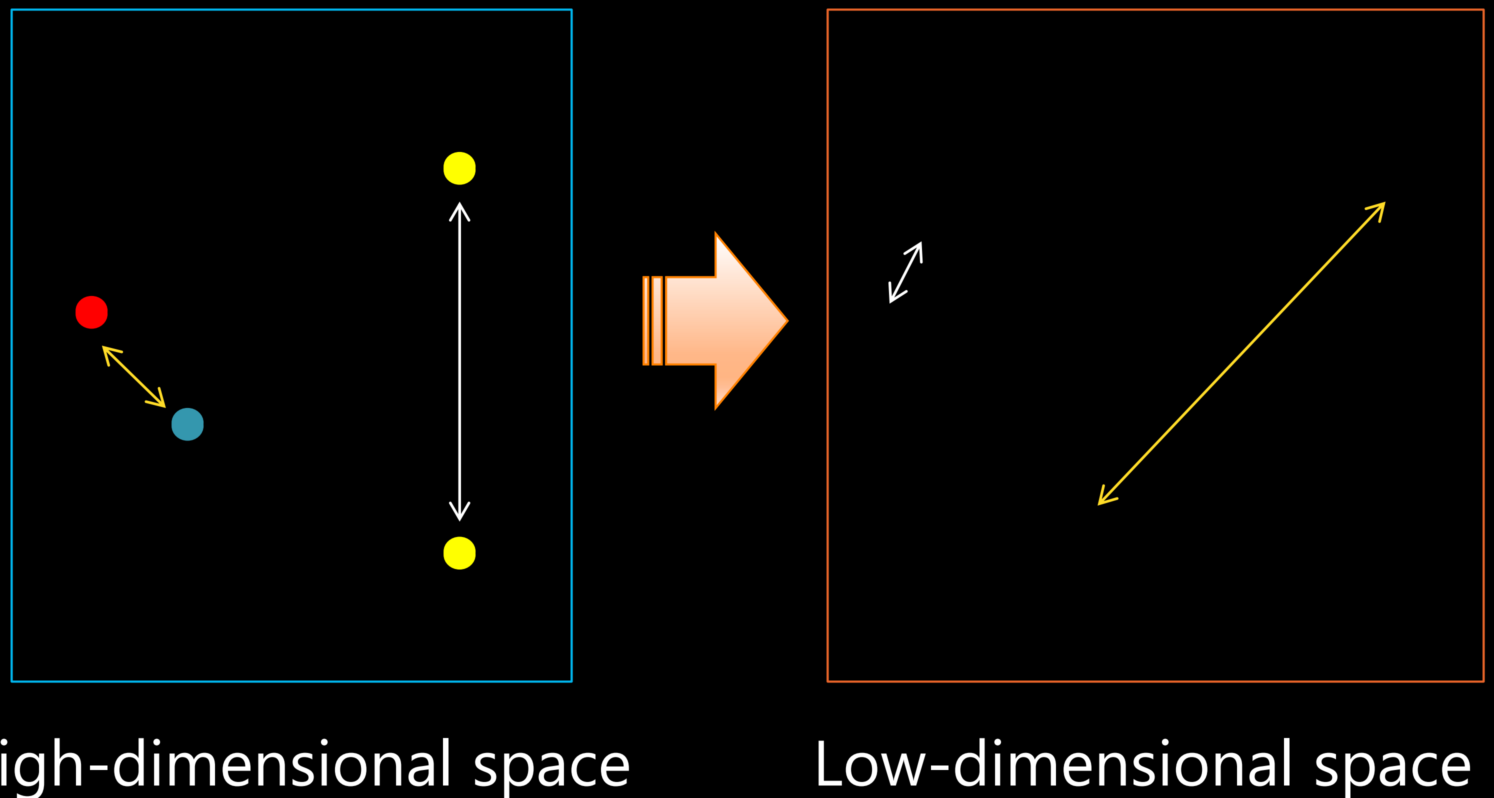
- Are D_p and D_q relevant or semantically similar?

$$sim(D_p, D_q)$$



- Input: high-dimensional, sparse term vectors
- Output: low-dimensional, dense concept vectors
- Model requirements
  - Transformation is easy to compute
  - Provide good similarity measures

Ideal Mapping



Microsoft Research

Dimensionality Reduction Methods

	Probabilistic	Projection
Supervised	CPLSA, PLTM, JPLSA	OPCA, S2Net, HDLR, CCA, CL-LSI
Unsupervised	PLSA, LDA	PCA, LSA

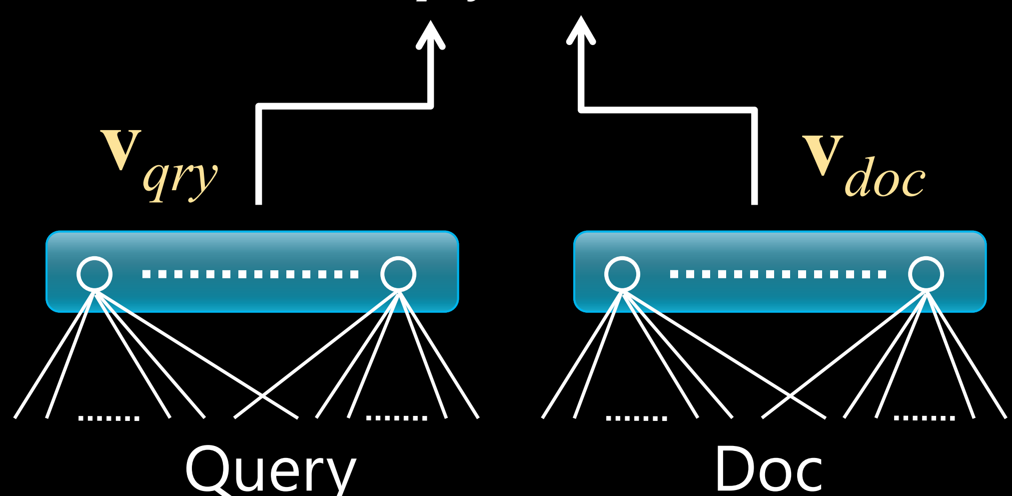
Outline

- Introduction
- Problem & Approach
- Experiments
  - Cross-language document retrieval
  - Ad relevance measures
  - Web search ranking
- Discussion & Conclusions

Goal – Learn Vector Representation

- Approach: Siamese neural network architecture
  - Train the model using labeled (query, doc)
  - Optimize for pre-selected similarity function (cosine)

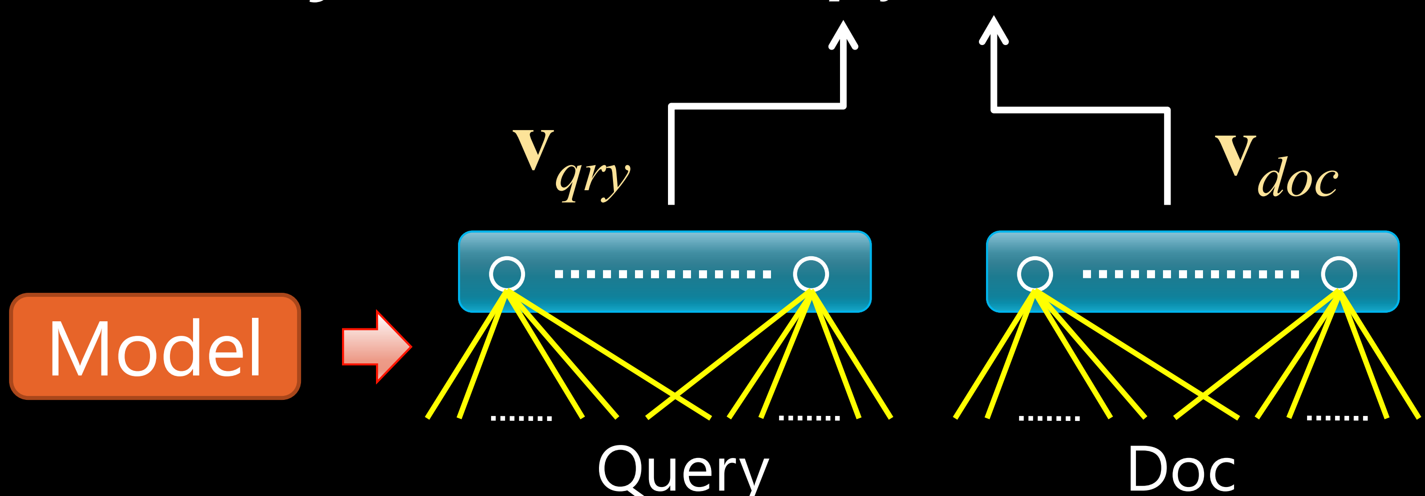
$$y \longleftrightarrow f_{\text{sim}}\left(\mathbf{v}_{qry}, \mathbf{v}_{doc}\right)$$



Goal – Learn Vector Representation

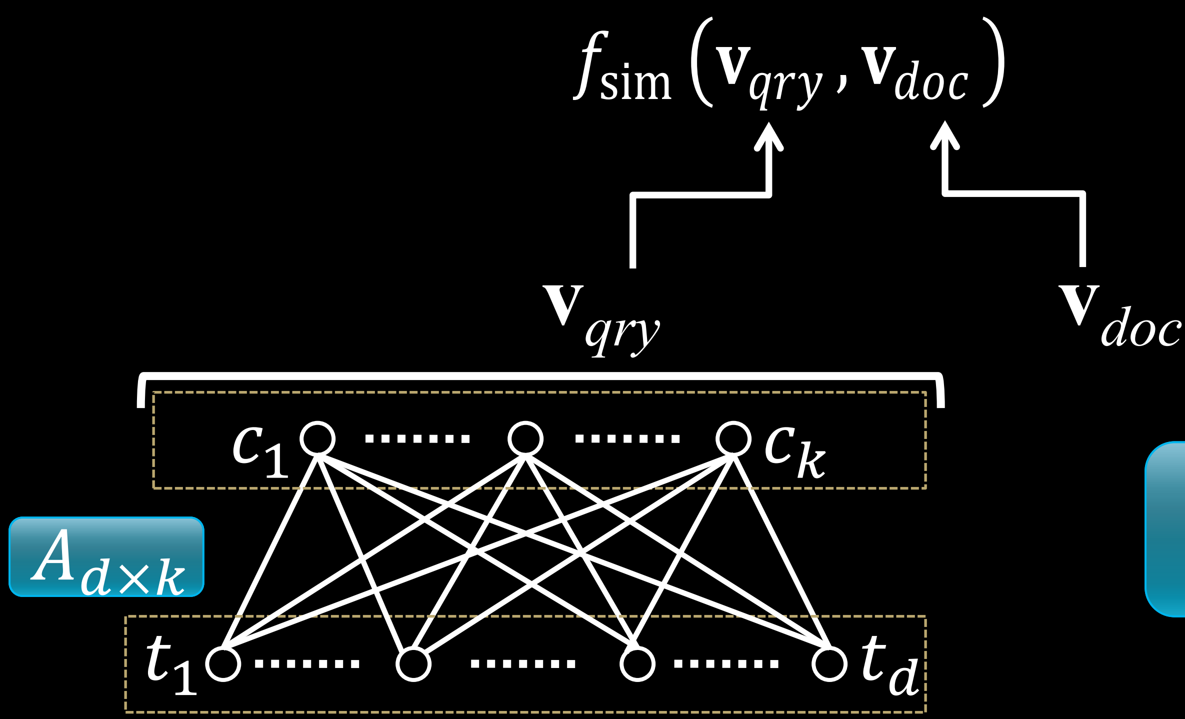
- Approach: Siamese neural network architecture
  - Train the model using labeled (query, doc)
  - Optimize for pre-selected similarity function (cosine)

$$y \longleftrightarrow f_{\text{sim}}\left(\mathbf{v}_{qry}, \mathbf{v}_{doc}\right)$$



S2Net – Similarity via Siamese NN

- Model form is the same as LSA/PCA
- Learning the projection matrix discriminatively



$$V_{qry} = A^T F_{qry}$$

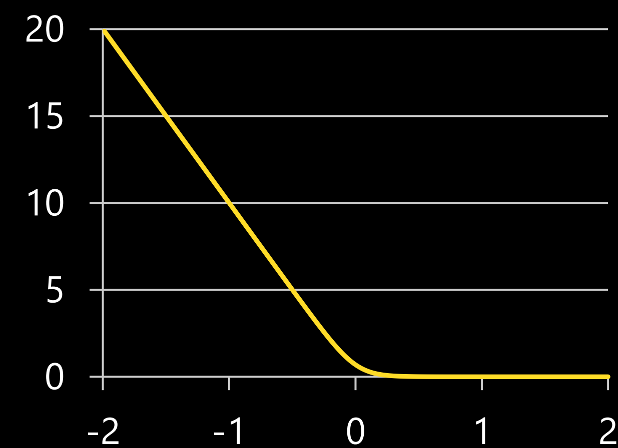
Pairwise Loss – Motivation

- In principle, we can use a simple loss function like mean-squared error: $\frac{1}{2}\left(y - f_{sim}(\mathrm{v}_{\mathrm{qry}}, \mathrm{v}_{\mathrm{doc}})\right)^2$. But…

	Doc_1	$-$	0.95
	Doc_2	$+$	0.89
$Query$	Doc_3	$-$	0.54
	Doc_4	$+$	0.43
	Doc_5	$-$	0.21

Pairwise Loss

- Consider a query q and two documents d_1 and d_2
  - Assume d_1 is more related to q, compared to d_2
- F_q, F_{d_1}, F_{d_2}: original term vectors of q, d_1 and d_2
- $\Delta = f_{sim}\left(A^T F_q, A^T F_{d_1}\right) - f_{sim}\left(A^T F_q, A^T F_{d_2}\right)$
- $Loss(\Delta; A) = \log(1 + \exp(-\gamma\Delta))$
  - γ: scaling factor, as $\Delta \in [-2,2]$
  - $\gamma = 10$ in the experiments



Microsoft Research

Model Training

- Minimizing the loss function can be done using standard gradient-based methods
  - Derive batch gradient and apply L-BFGS
- Non-convex loss
  - Starting from a good initial matrix helps reduce training time and converge to a better local minimum
- Regularization
  - Model parameters can be regularized by adding a smoothing term $\frac{\beta}{2}||A - A_0||^2$ in the loss function
  - Early stopping can be effective in practice

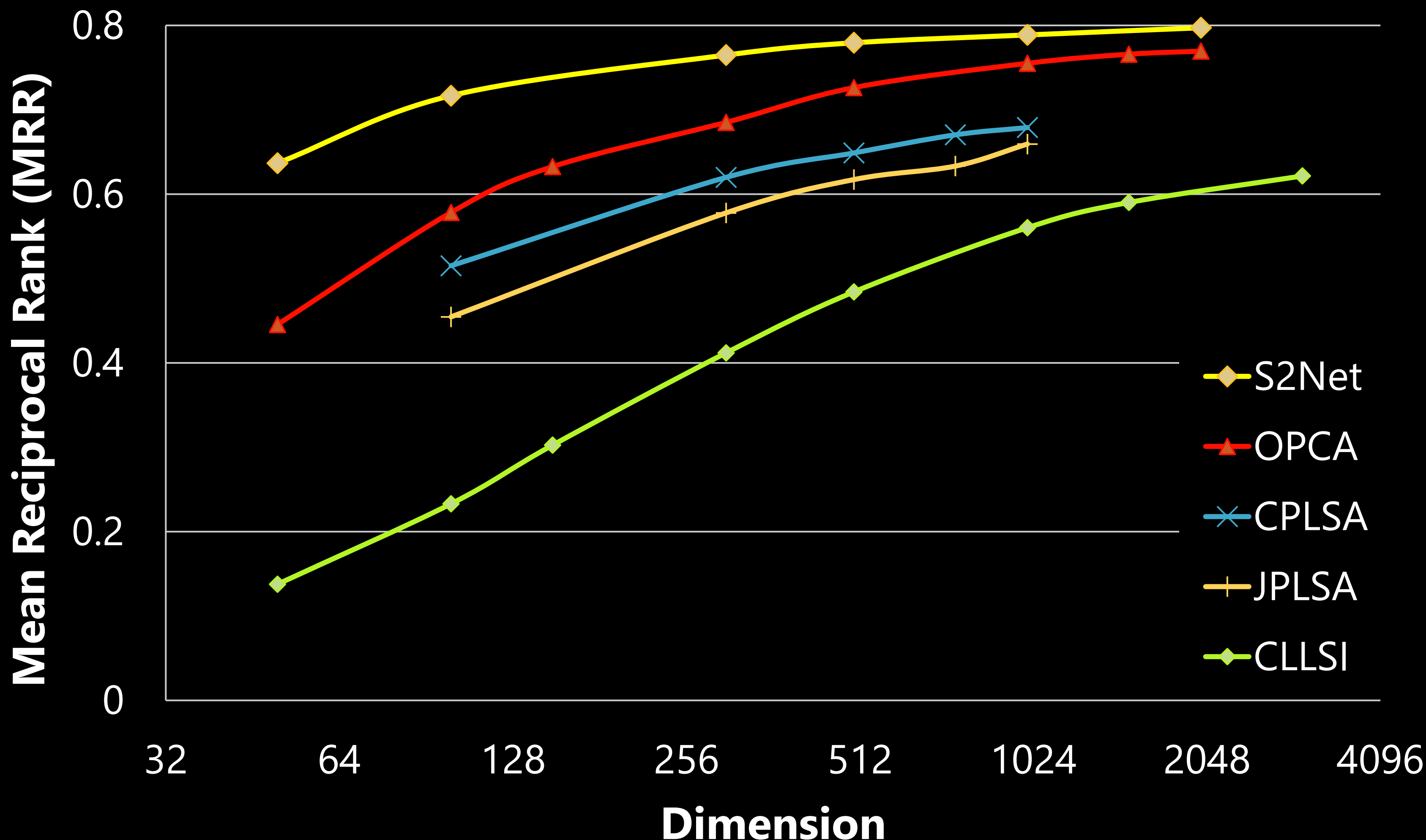
Outline

- Introduction
- Problem & Approach
- Experiments
  - Cross-language document retrieval
  - Ad relevance measures
  - Web search ranking
- Discussion & Conclusions

Cross-language Document Retrieval

- Dataset: pairs of Wiki documents in EN and ES
  - Same setting as in [Platt et al. EMNLP-10]
  - #document in each language
    - Training: 43,380, Validation: 8,675, Test: 8,675
  - Effectively, $43380^2 = 1.88$ billion training examples
    - Positive: EN-ES documents in the same pair
    - Negative: All other pairs
- Evaluation: find the comparable document in the different language for each query document

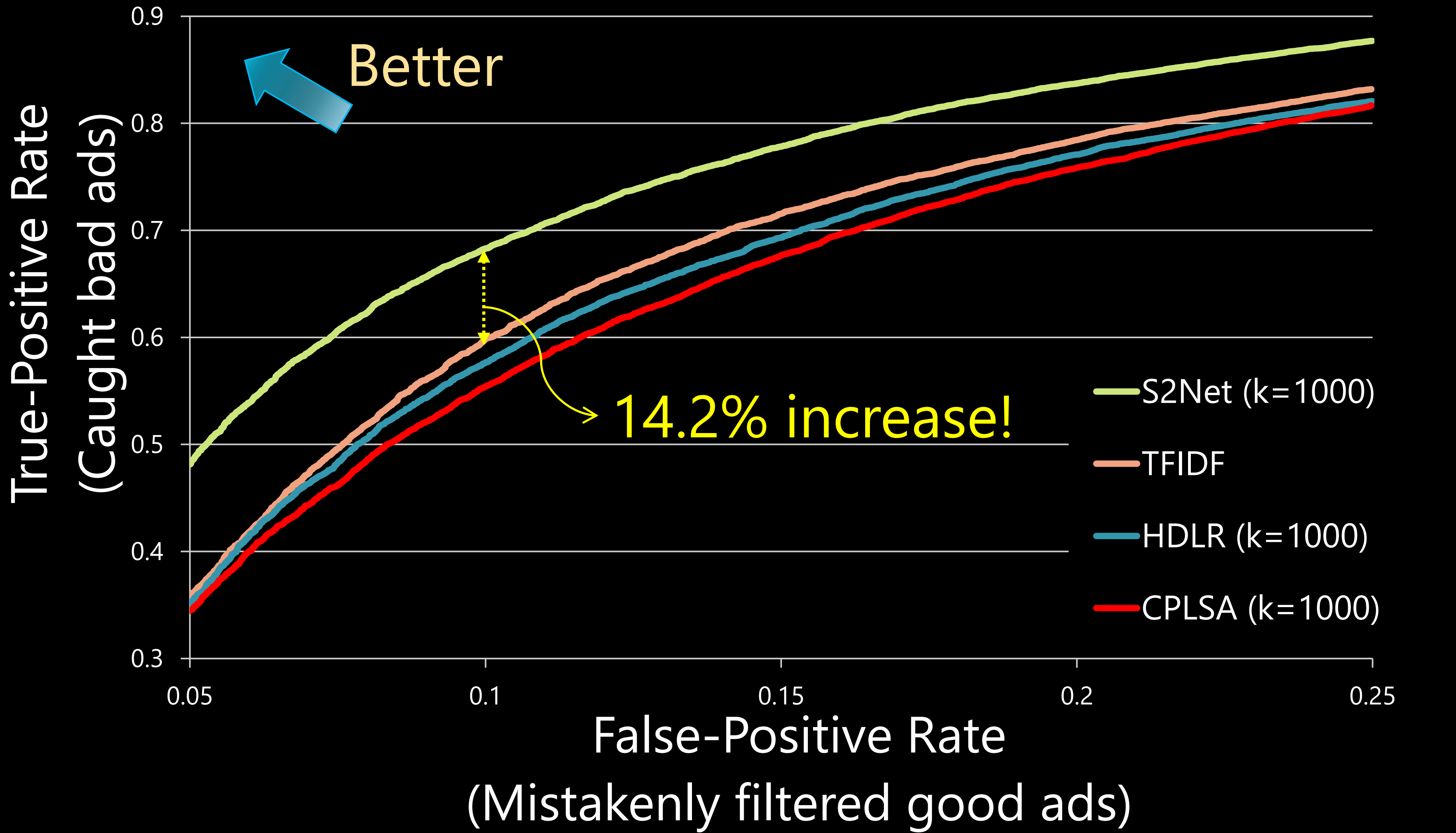
Results on Wikipedia Documents



Ad Relevance Measures

- Task: Decide whether a paid-search ad is relevant to the query
  - Filter irrelevant ads to ensure positive search experience
  - F_{qry}: pseudo-document from Web relevance feedback
  - F_{doc}: ad landing page
- Data: query-ad human relevance judgment
  - Training: 226k pairs
  - Validation: 169k pairs
  - Testing: 169k pairs

The ROC Curves of the Ad Filters



Web Search Ranking [Gao et al., SIGIR-11]

Train latent semantic models

- *Parallel corpus* from clicks
- 82,834,648 query-doc pairs

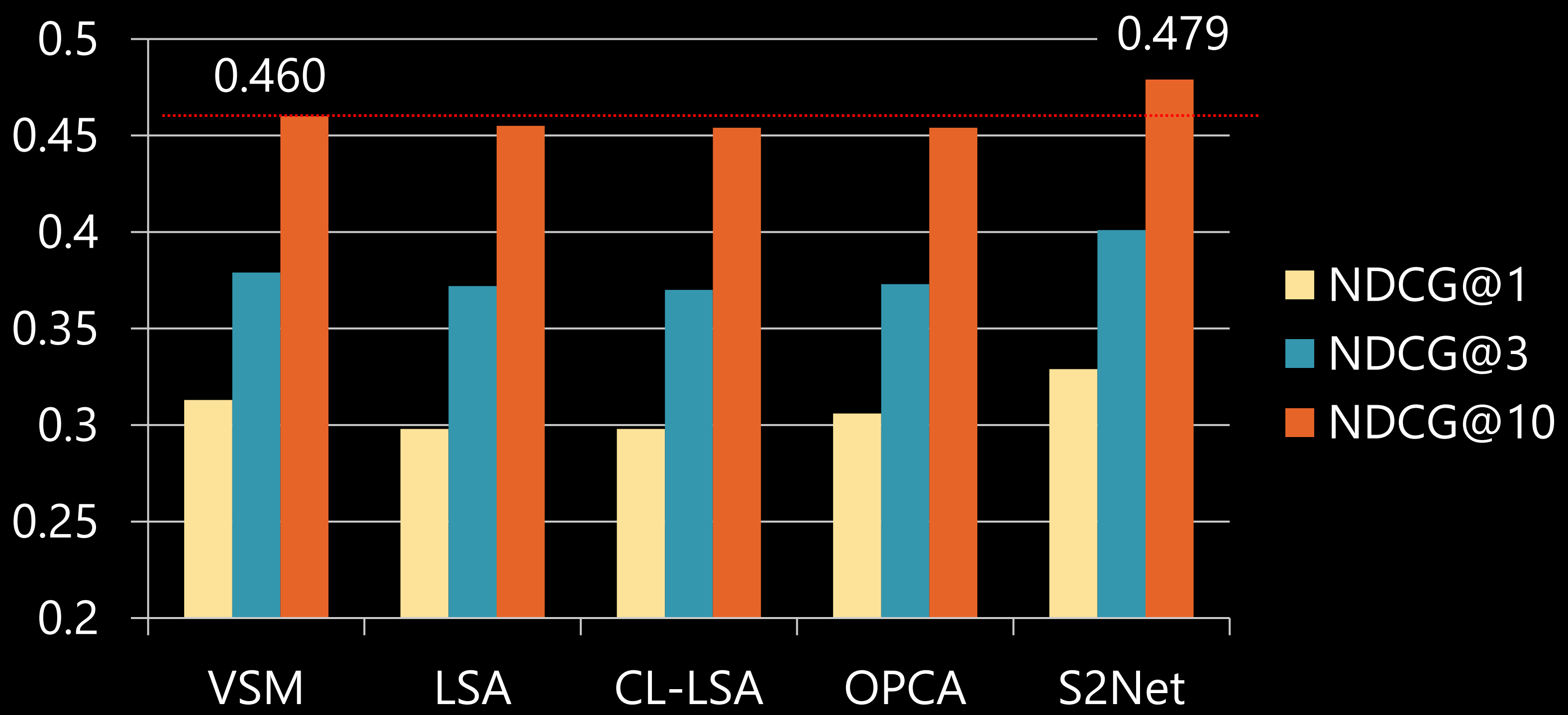
query 1	—	doc 1
query 2	—	doc 2
query 3	—	doc 3
…	—	…

Evaluate using labeled data

- Human relevance judgment
- 16,510 queries
- 15 doc per query in average

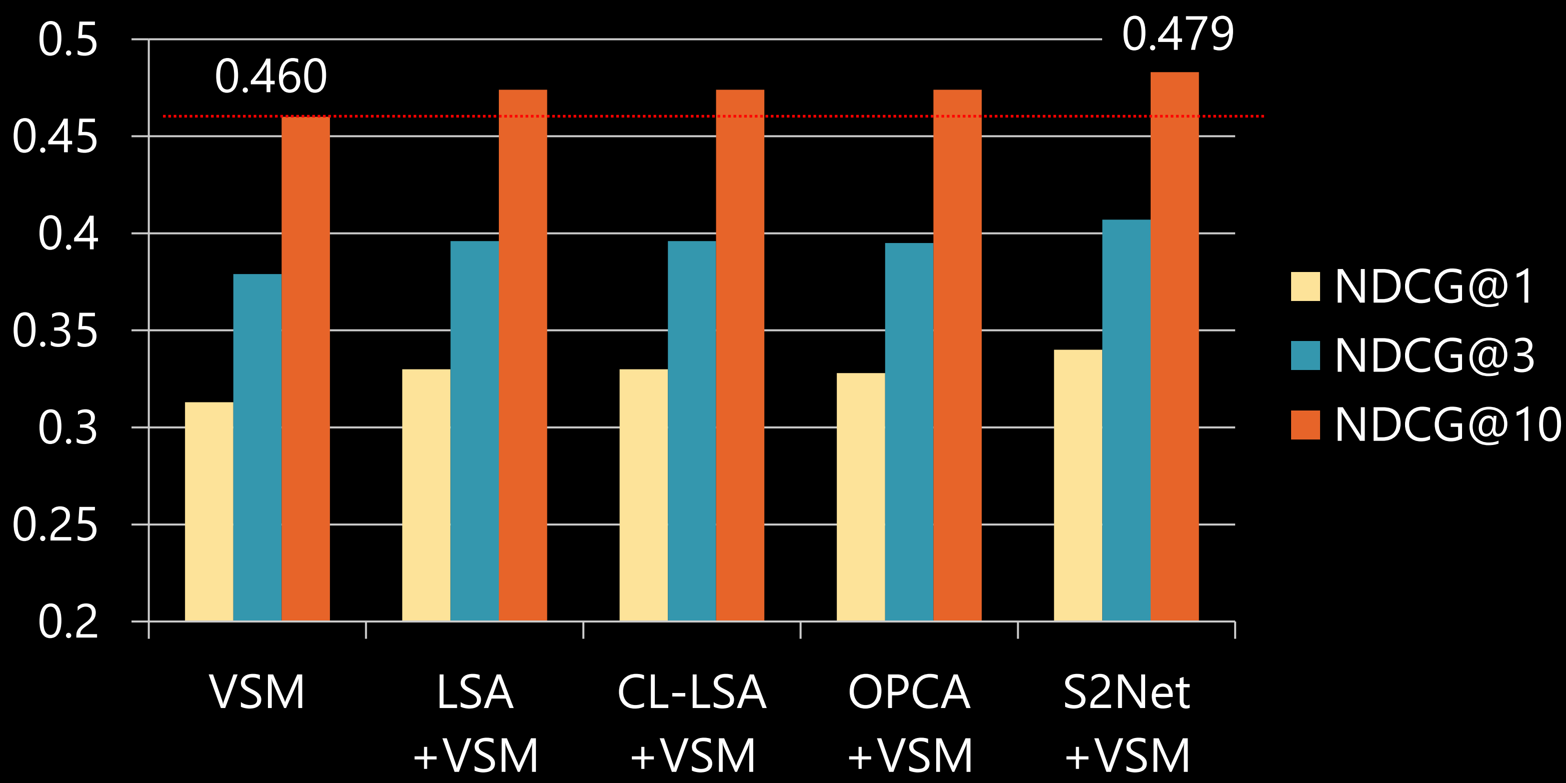
query	doc 1	Good
	doc 2	Fair
	doc 3	Bad

Results on Web Search Ranking



- Only S2Net outperforms VSM compared to other projection models

Results on Web Search Ranking



- After combined with VSM, results are all improved
- More details and interesting results of generative topic models can be found in [SIGIR-11]

Outline

- Introduction
- Problem & Approach
- Experiments
  - Cross-language document retrieval
  - Ad relevance measures
  - Web search ranking
- Discussion & Conclusions

Model Comparisons

- S2Net vs. generative topic models
  - Can handle explicit negative examples
  - No special constraints on input vectors
- S2Net vs. linear projection methods
  - Loss function designed to closely match the true objective
  - Computationally more expensive
- S2Net vs. metric learning
  - Target high-dimensional input space
  - Scale well as the number of examples increases

Why Does S2Net Outperform Other Methods?

- Loss function
  - Closer to the true evaluation objective
- Slight nonlinearity
  - Cosine instead of inner-product
- Leverage a large amount of training data
  - Easily parallelizable: distributed gradient computation

Microsoft Research

Conclusions

- S2Net: Discriminative learning framework for dimensionality reduction
  - Learns a good projection matrix that leads to robust text similarity measures
  - Strong empirical results on different tasks
- Future work
  - Model improvement
    - Handle Web-scale parallel corpus more efficiently
    - Convex loss function
  - Explore more applications
    - e.g., word/phrase similarity