Discriminative Information Retrieval for Question Answering Sentence Selection

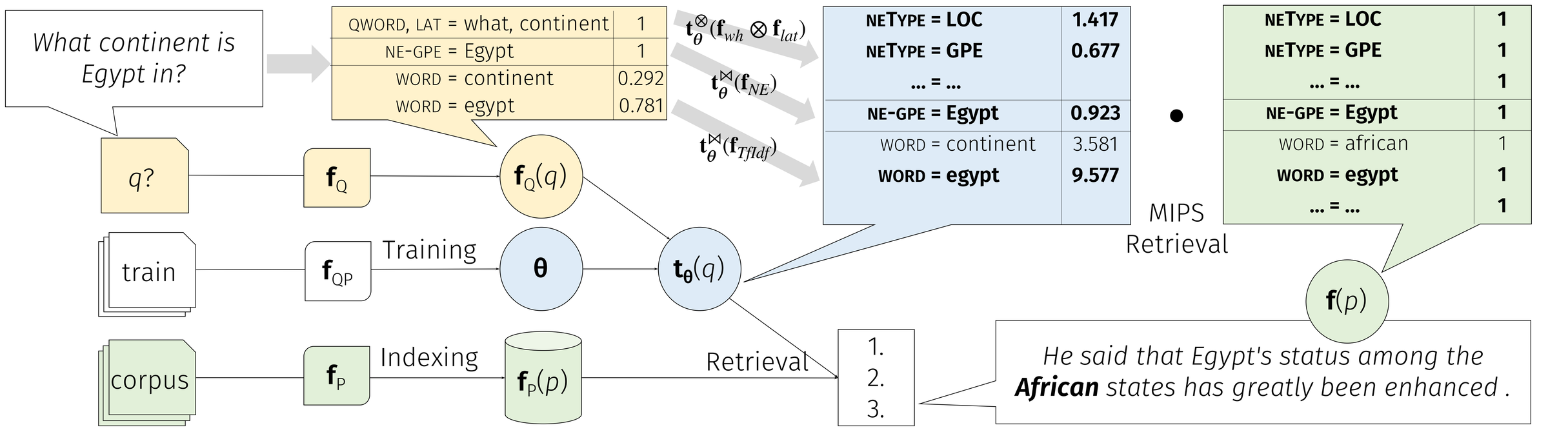
Tongfei Chen, Benjamin Van Durme
Johns Hopkins University
{tongfei, vandurme}@cs.jhu.edu

Problem

Given a question q, rank the candidate sentences in a corpus w.r.t. a scoring function $s(q, p)$ that measures how likely p answers q.

Motivation

Vanilla IR – efficient, results not good
Neural reranking – good results, slow (linear)
Is there a better way to triage the set?



Discriminative IR

$$\text{argmax}_{p \in P}\, s(q, p)$$

We want s to be

- Trainable by question/answer pairs
- Decomposable into inner products of sparse vectors $\mathbf{g}(q) \cdot \mathbf{f}(p)$

Under these conditions IR algorithms can be reused! Given q, we compute the feature vector that is *expected* from answers: $\mathbf{t}_{\boldsymbol{\theta}}(\mathbf{f}_Q(q))$.

Given feature functions $\mathbf{f}_Q$ / $\mathbf{f}_P$ / $\mathbf{f}_{QP}$:

$$\mathbf{f}_{QP}(q, p) = \mathbf{f}_Q(q) \circledcirc \mathbf{f}_P(p)$$
$$s(q, p) = \boldsymbol{\theta} \cdot \mathbf{f}_{QP}(q, p) = \mathbf{t}_{\boldsymbol{\theta}}(\mathbf{f}_Q(q)) \cdot \mathbf{f}_P(p)$$

Modeled as linear feature-based IR (trained using log loss)

Trainable $\boldsymbol{\theta}$

Decoupled

Compositional operators ($\circledcirc$):

$$\mathbf{f} \otimes \mathbf{g} = \{((k_f, k_g) = (v_f, v_g), w_f w_g)\}$$
$$\mathbf{f} \bowtie \mathbf{g} = \{((k_f = k_g) = 1, w_f w_g)\}$$

Projection:

$$\mathbf{t}^{\otimes}_{\boldsymbol{\theta}}(\mathbf{f}) = \{(k = v, w_f \theta((k_f, k) = (v_f, v))\}$$
$$\mathbf{t}^{\bowtie}_{\boldsymbol{\theta}}(\mathbf{f}) = \{(k = v_f, w_f \theta((k_f = k) = 1))\}$$

It can be proven that

$$\mathbf{t}^{\otimes}_{\boldsymbol{\theta}}(\mathbf{f}) \cdot \mathbf{g} = \boldsymbol{\theta} \cdot (\mathbf{f} \otimes \mathbf{g})$$
$$\mathbf{t}^{\bowtie}_{\boldsymbol{\theta}}(\mathbf{f}) \cdot \mathbf{g} = \boldsymbol{\theta} \cdot (\mathbf{f} \bowtie \mathbf{g})$$

* Feature vectors are represented as a set of (key = value, weight) tuples.
* For all $(k_f = v_f, w_f) \in \mathbf{f}$, $(k_g = v_g, w_g) \in \mathbf{g}$.

Paper

Software

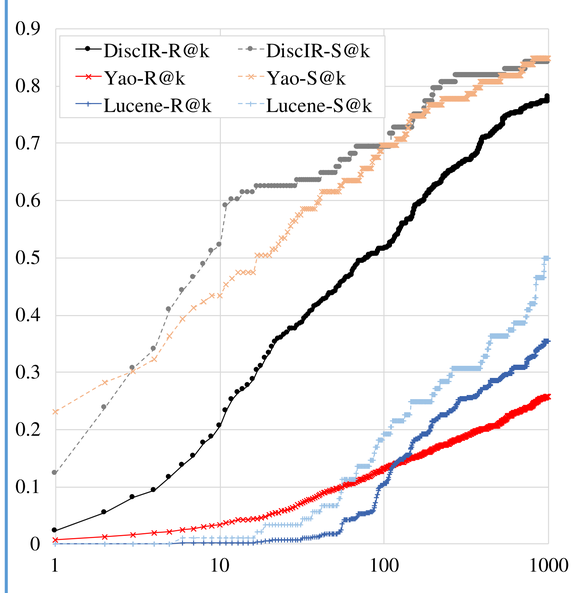
Feature set

```
((QWord ⊗ LexAnsType) ⊗ NamedEntityTypes) +
((QWord ⊗ LexAnsType) ⊗ BagOfWords      ) +
( NamedEntities       ⋈ NamedEntities   ) +
( NormalizedTfIdf     ⋈ BagOfWords      )
```

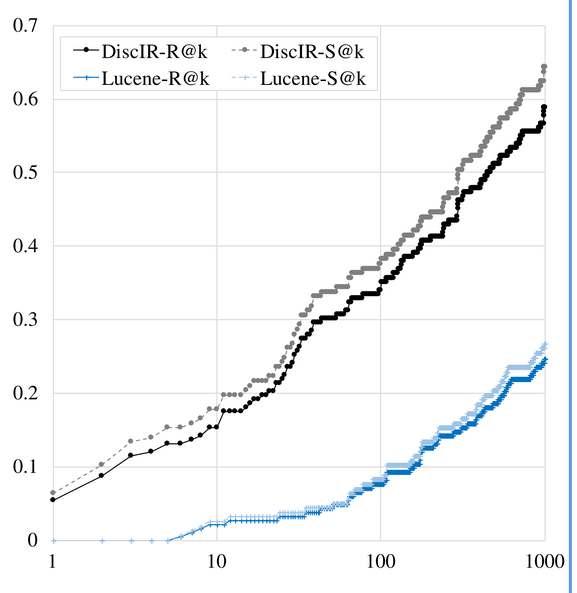
Experiments

Dataset	# of questions			# of
	train	dev	test	sentences
TREC/AQUAINT	2150	53	99	23,398,842
WikiQA/Wikipedia	2118	77	157	20,368,761

	R@1k	b-pref	MAP	MRR
TREC/AQUAINT				
Lucene	35.47%	38.22%	9.78%	15.06%
Yao et al. (2013)	25.88%	45.41%	13.75%	**29.87%**
DiscIR	**78.20%**	**75.15%**	**17.84%**	25.30%
WikiQA/Wikipedia				
Lucene	24.73%	25.69%	0.58%	0.72%
DiscIR	**58.79%**	**60.88%**	**10.26%**	**11.42%**



TrecQA



WikiQA