A Graphical Model for Entity-based Document Retrieval

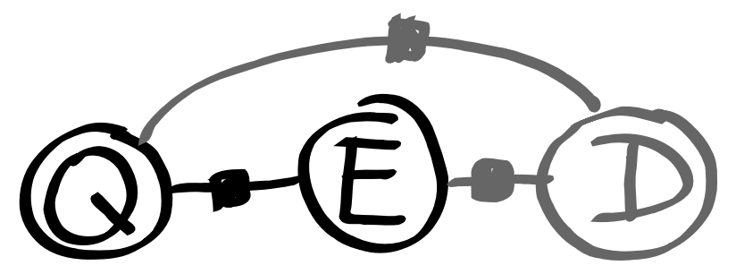
Laura Dietz, Jeffrey Dalton, W. Bruce Croft {dietz, jdalton, croft}@cs.umass.edu

Abstract

With the availability of entity linking tools information retrieval has a new source to leverage. We study the utility of entity links for the well-known problem of document retrieval given a keyword query. We devise a graphical model to jointly reason about the compatibility of the query, and latent documents and entities.

Indicators for relevant entities are available after entity linking the query, but these links are very rare. One way is to to use probabilistic retrieval from the knowledge base. We can further complement this information with entity links found in relevant documents. We exploit the jointness of the model to improve the prediction of document relevance with relevant entities. To induce a document ranking, the marginal model likelihood is used as a score for document relevance. Furthermore, we assume that depending on the context of the query, different aspects of the entity become important and the way it is referred to changes. Therefore, each entity is associated with a set of structured attributes such as names, related entities, and types, we estimate along with other variables in the model.

Entities Relevant for Query?



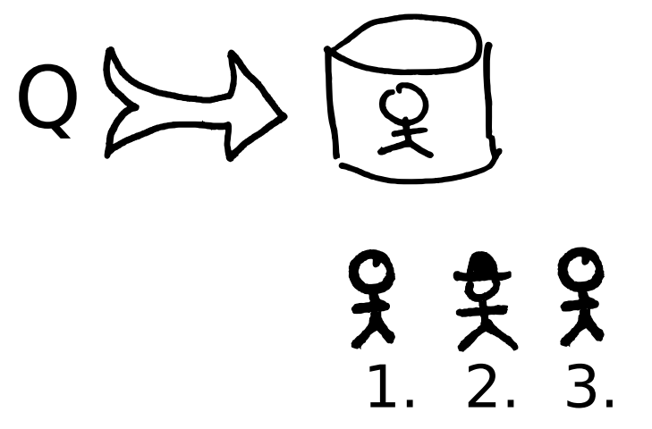
Different indicators for relevant Entities E:

a) Q annotated with linked entities E

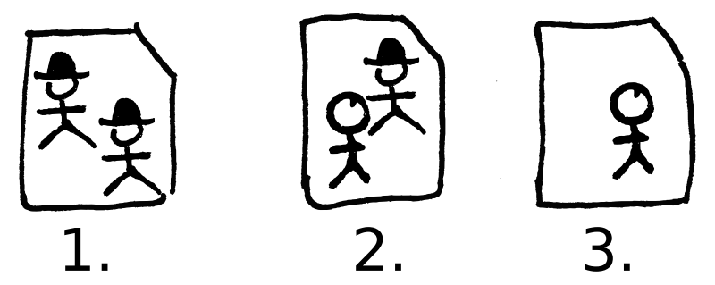
Query: obama family tree

```
id: Barack_Obama
Cat: US President
Cat: Politician
Words: 44th President office
Words: African American
```

b) run query Q against Wikipedia;
get a ranking of entities E (with score)



c) run query Q against corpus index;
get a 1st pass ranking of docs D;
entity link D to get a multi-bag of E
use "Relevance Model" to derive distr over E

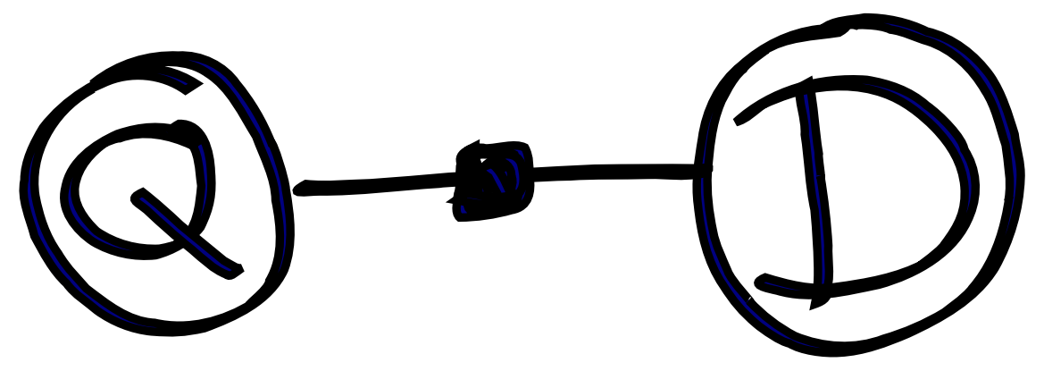


$$\underbrace{\int p(D|Q)}_{\approx \exp(score_q D)} \underbrace{p(E|D)}_{LM} dD$$

Problem: ad hoc Document Retrieval

Model indicating relevance

keyword query

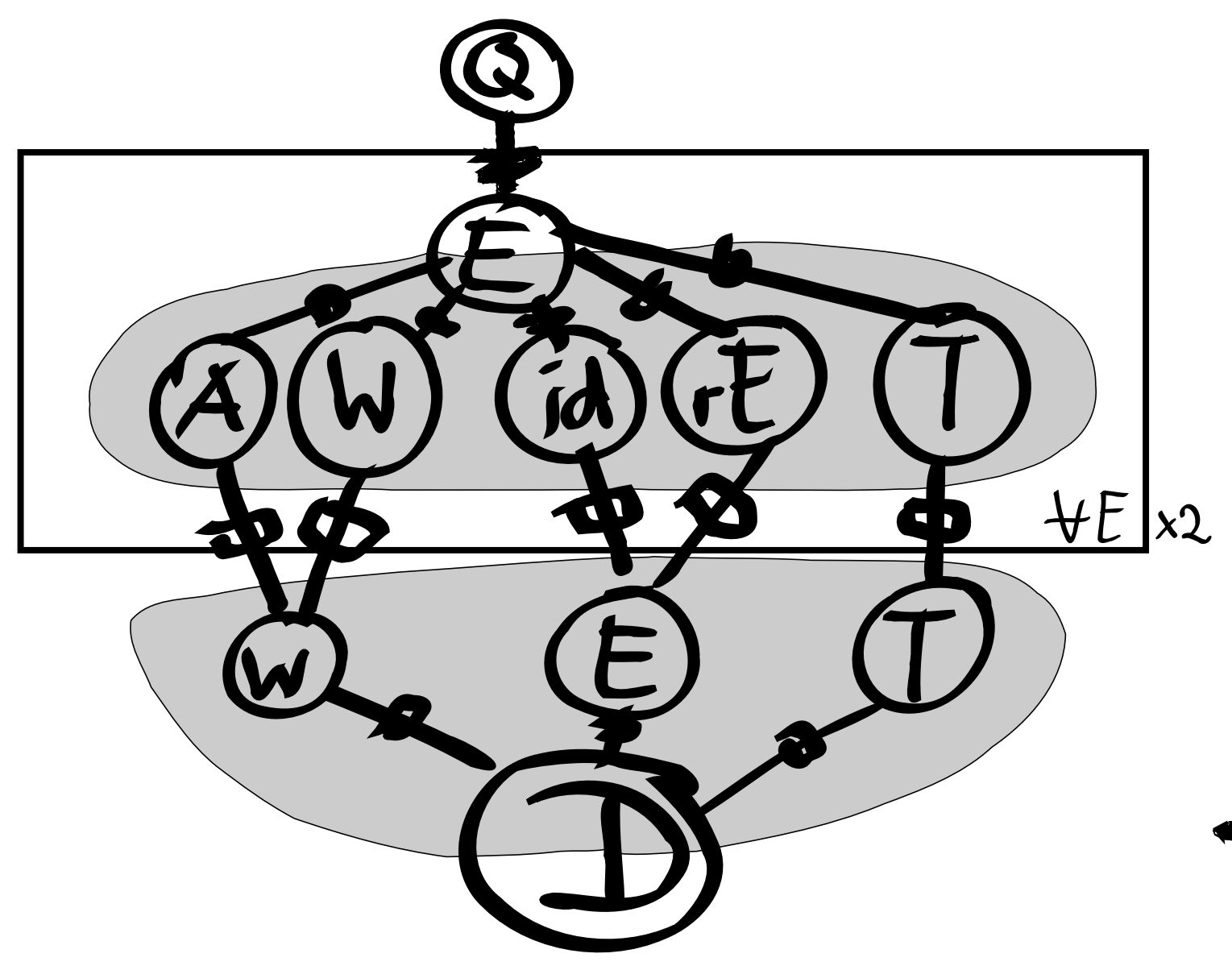


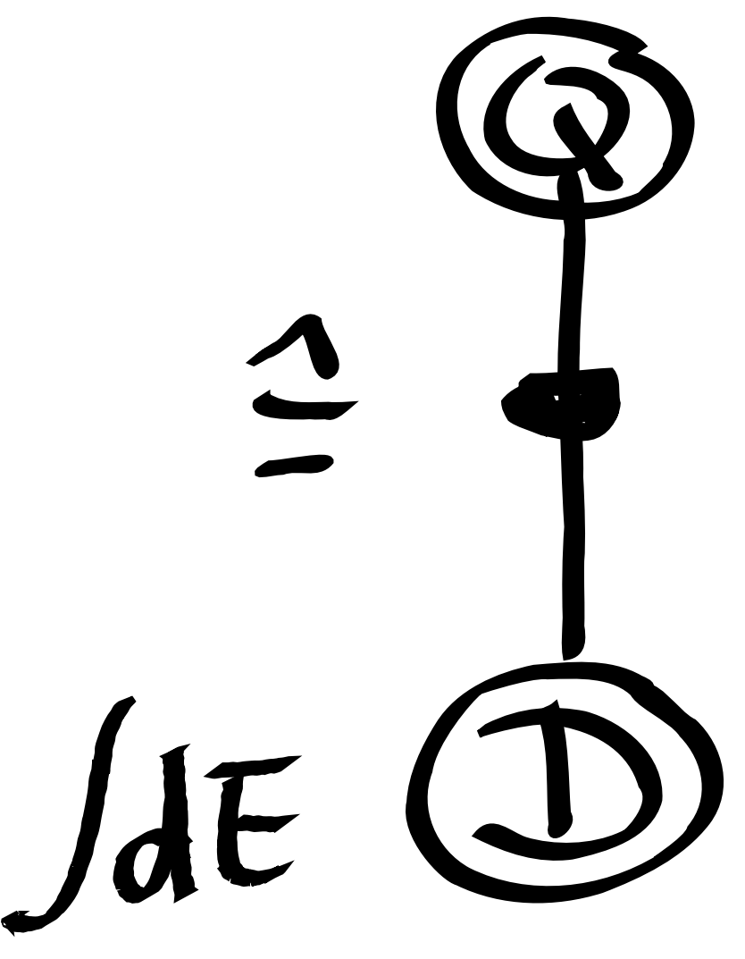
sequence of terms

Rank all D in the corpus according to the likelihood under the model.

Retrieval Models for keyword queries
- BM25, TF-IDF
- Language Model
  p(q1, D) * p(q2, D) * p(q3, D) ...
- Language Model with Dirichlet Smoothing
- Bigram Model, windowed Bigram Model
- Sequential Dependence Model

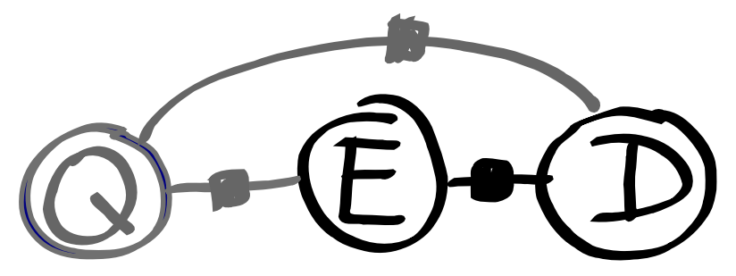
Graphical model: E-based Retrieval



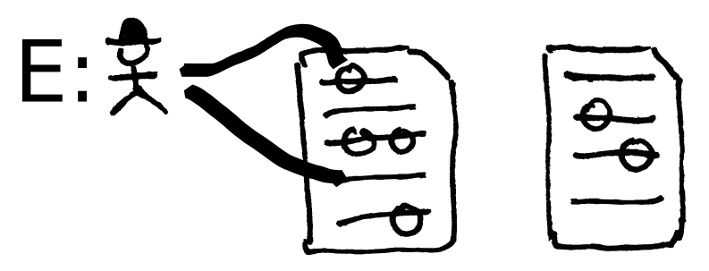


Entities integrated out
Merging distributions over A, Q, ids, rE and T

Document Relevant given Entities?

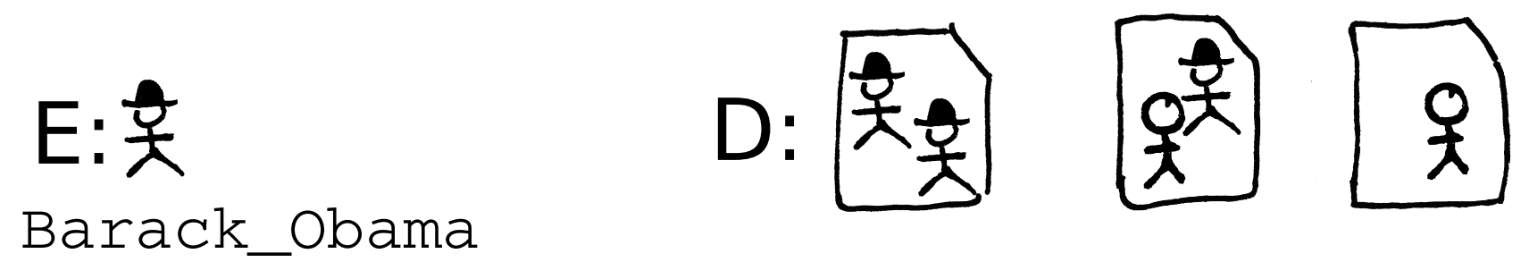


Assume we know relevant Entities
Different indicators for relevant Documents:
a) D contains name alias of E
b) D contains words that are also charactistic for E



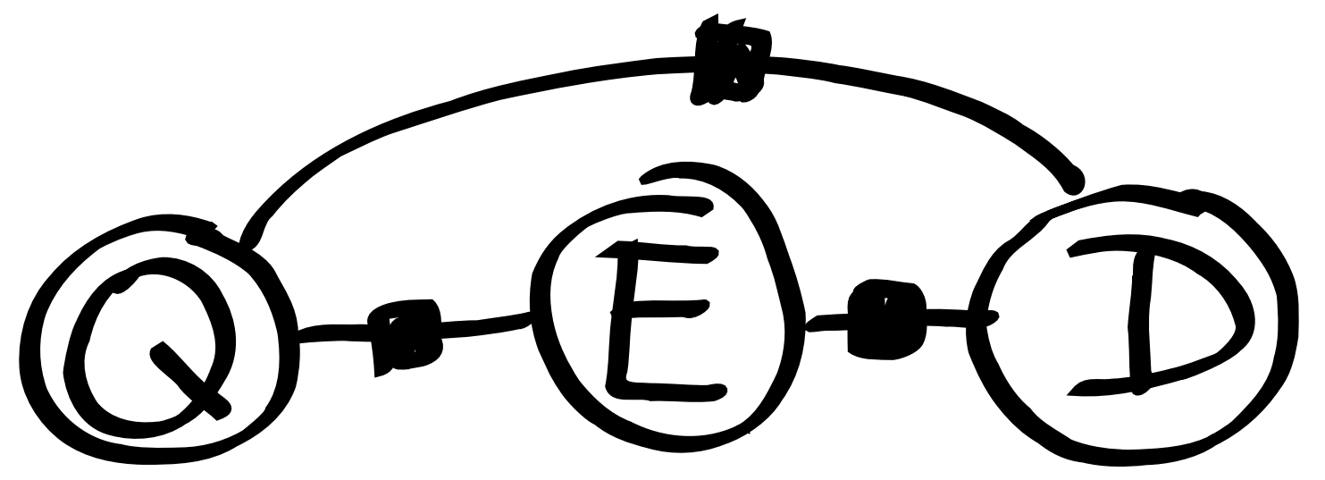
(only require D as text)

c) D contains link to E
d) D contains link to rel entities of E (neighbor)
e) D contains links to entities with same type



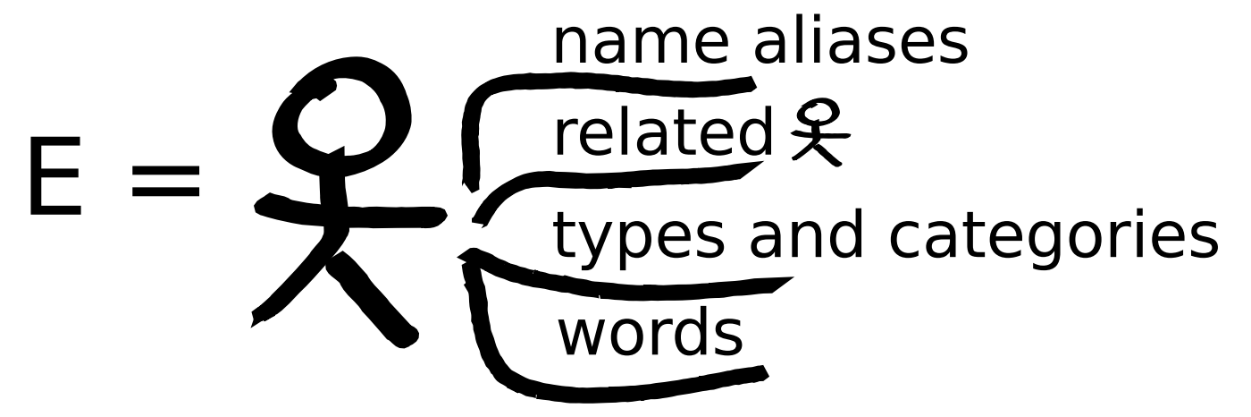
```
id: Barack_Obama
Cat: US President
```

Retrieval with Entities



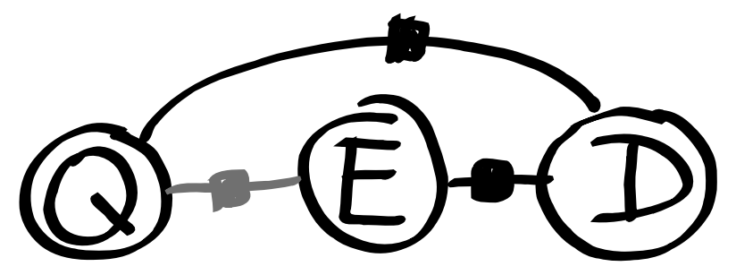
E rel. for Q D rel. for E

Retrieval score for each D:
Iterate over all E to rank all D
according to Likelihood



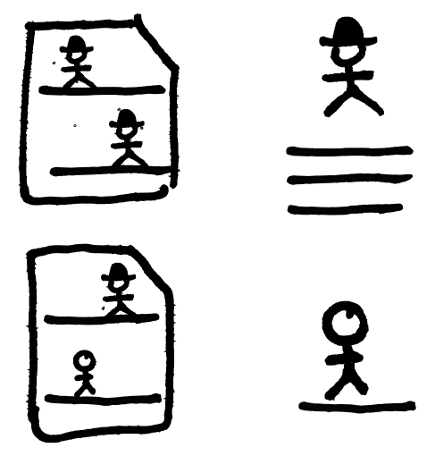
name aliases
related
types and categories
words

Entity Context Model



Entity representation from Knowledge Base
But can be updated in the light of the query

- Use 1st pass retrieved Docs
- For every unique E
- collect snippets surrounding E
  (8 words, 50 words, 1 sentence)



Given retrieved Documents (1st pass)
Different indicators for Entity attributes:

a) surface forms of links (--> name aliases)

b) co-mentioned entities (--> related E)

Mr Obama said the sanctions were not aimed at Russian President Vladimir Putin personally.

c) types/categories of entities in snippets

Mr Obama said the sanctions were not aimed at Russian President Vladimir Putin personally.

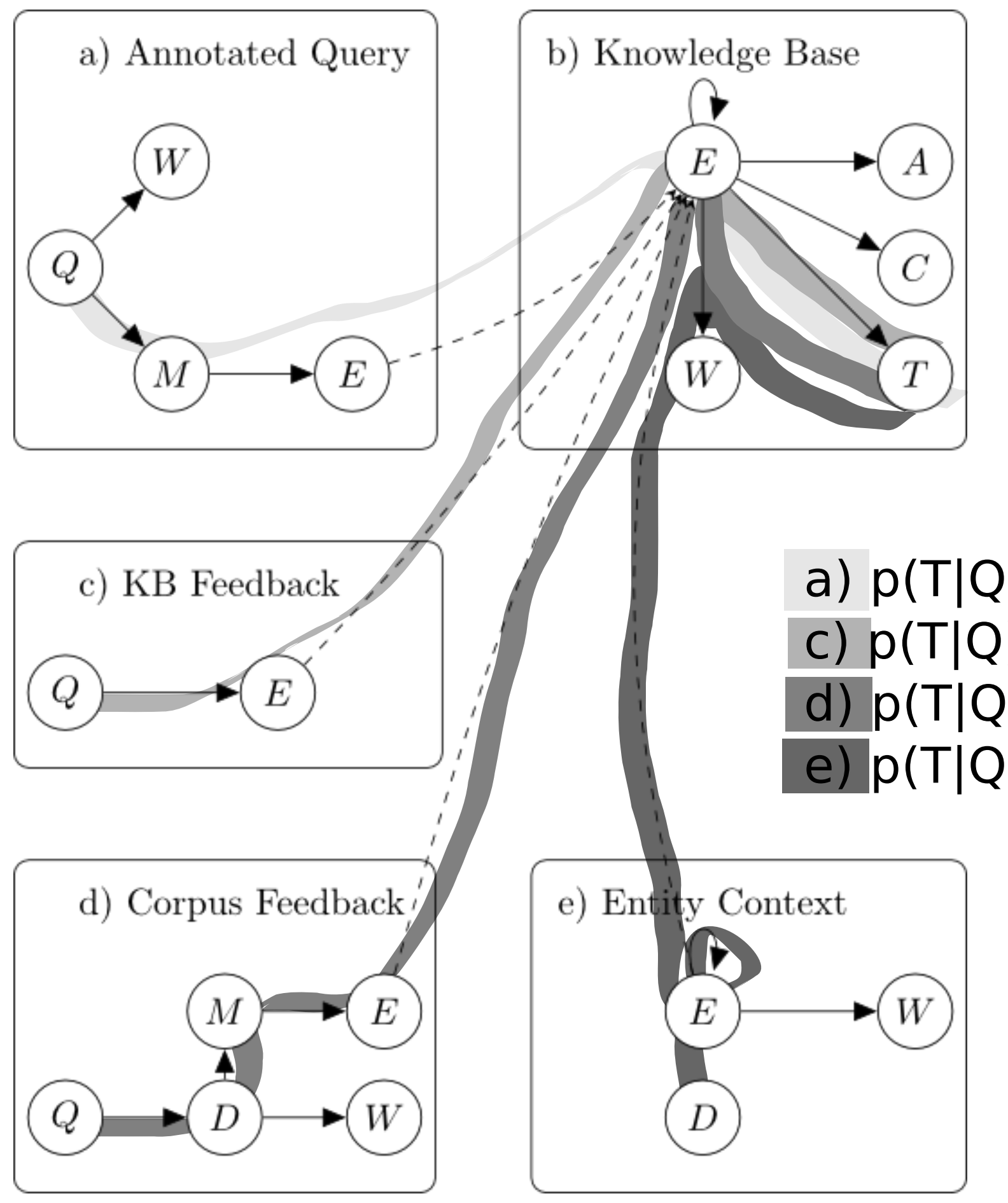
```
Cat: Politician
                     Cat: Politician
```

d) word distribution of snippets (--> words)

Mr Obama said the sanctions were not aimed at Russian President Vladimir Putin personally.

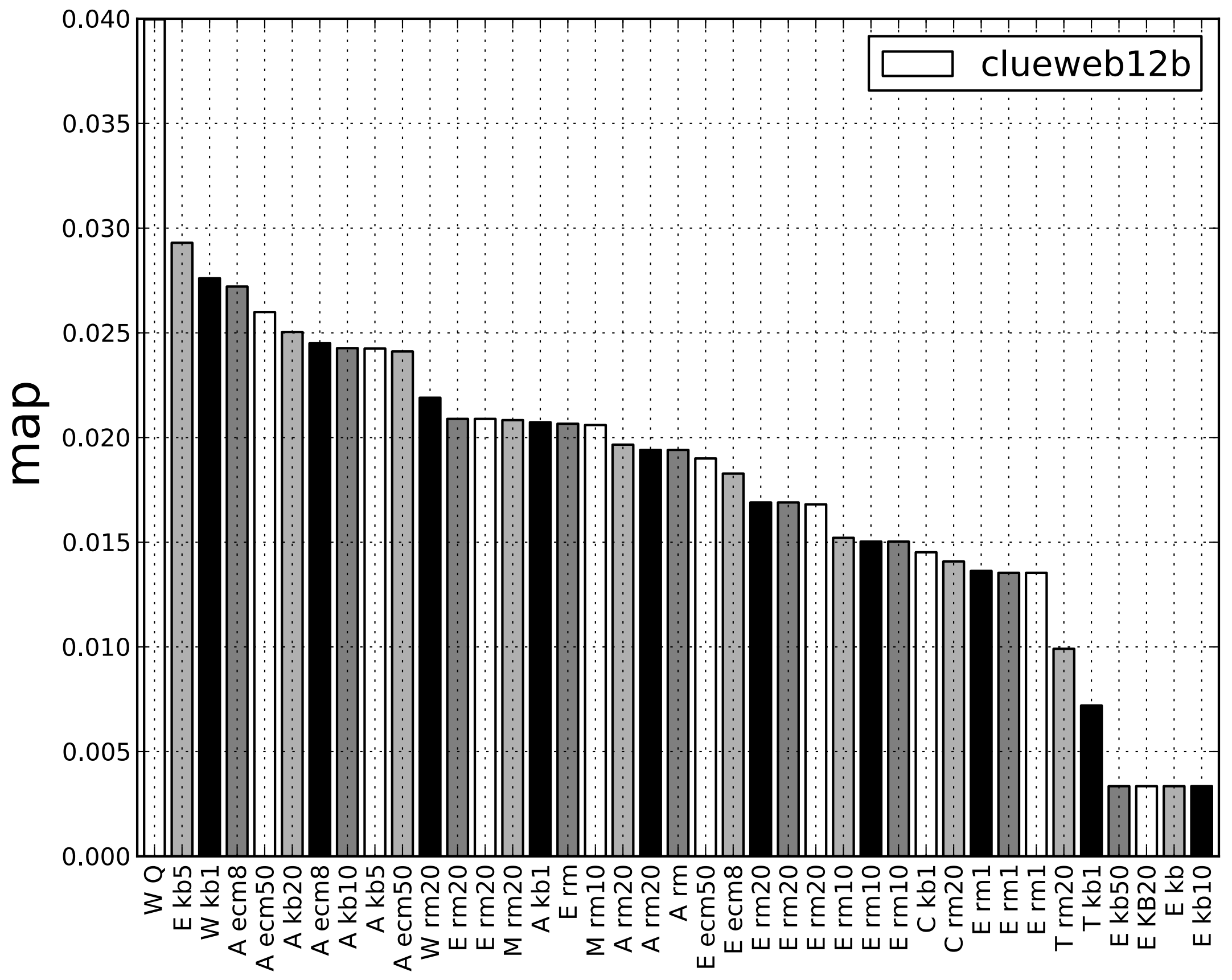
```
Words: sanctions, russian, president,
Words: aimed, personally
```

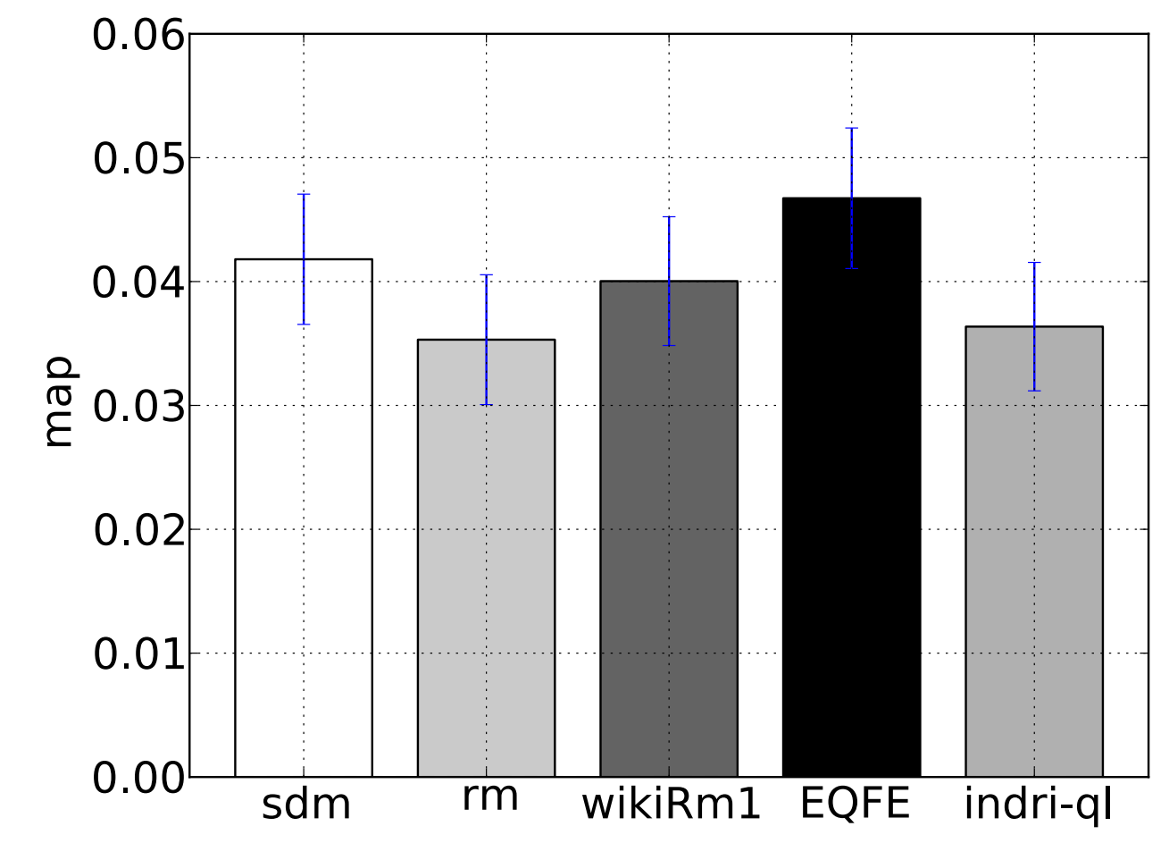
Different Paths and Different Types as Features



a) p(T|Q) = p(M|Q) p(E|M) p(T|E)
c) p(T|Q) = p(E|Q) p(T|E)
d) p(T|Q) = p(D|Q) p(M|D) p(E|M) p(T|E)
e) p(T|Q) = p(D|Q) p(E',E|D) p(T|E')

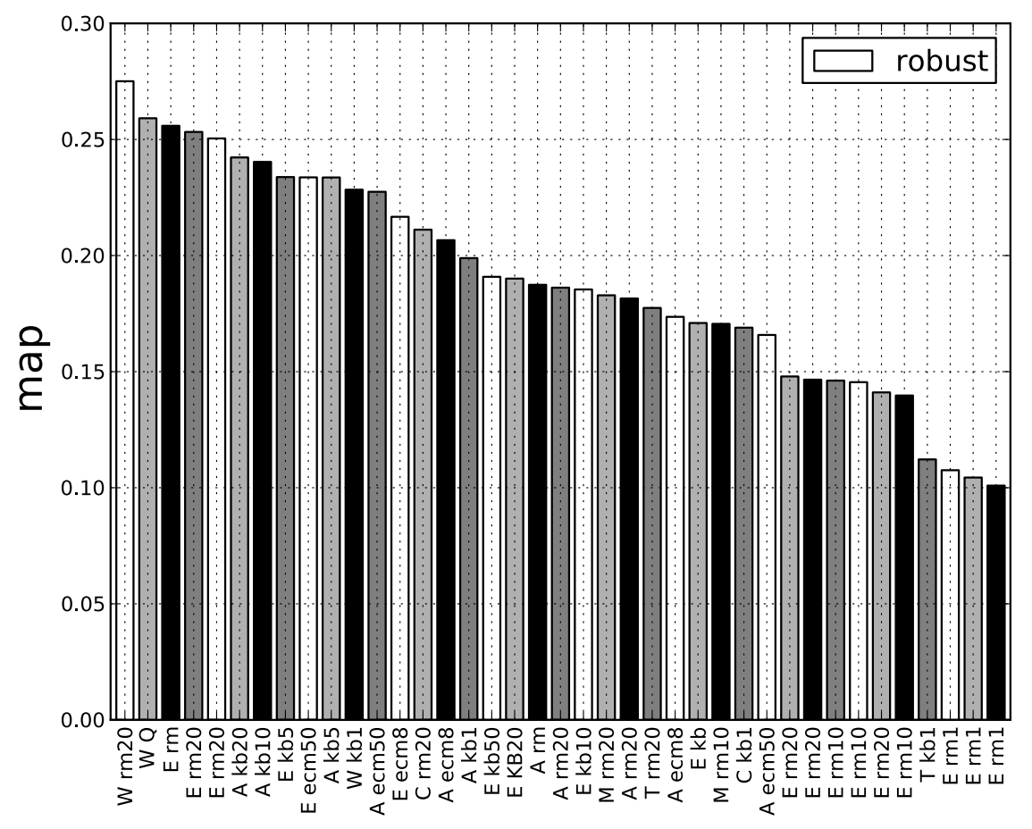
ClueWeb12b Results

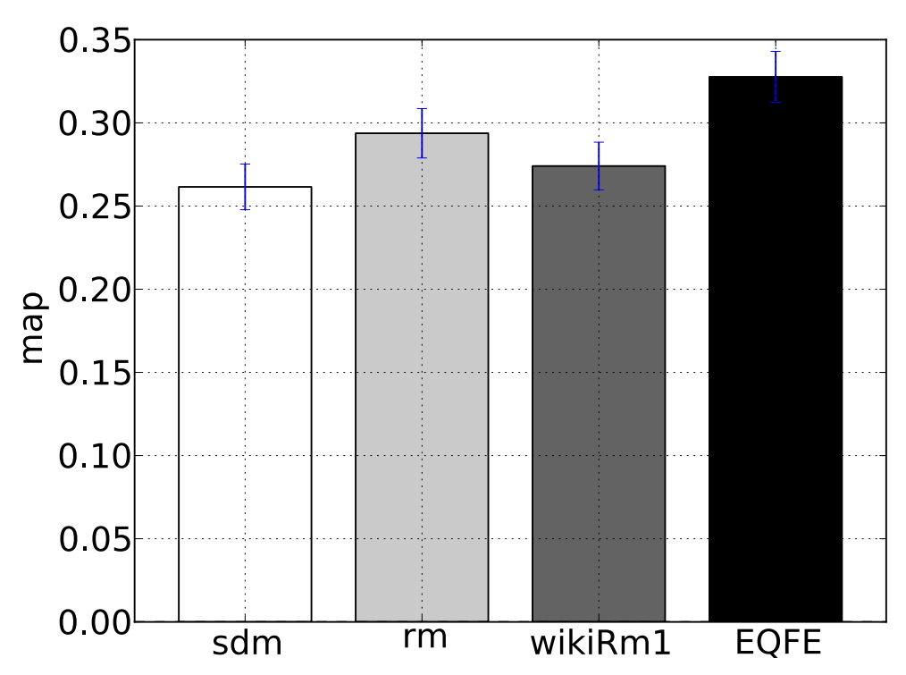




Entity Links from Google FACC1 Dataset.

Robust04 Results





Entity Links generated with KB Bridge