

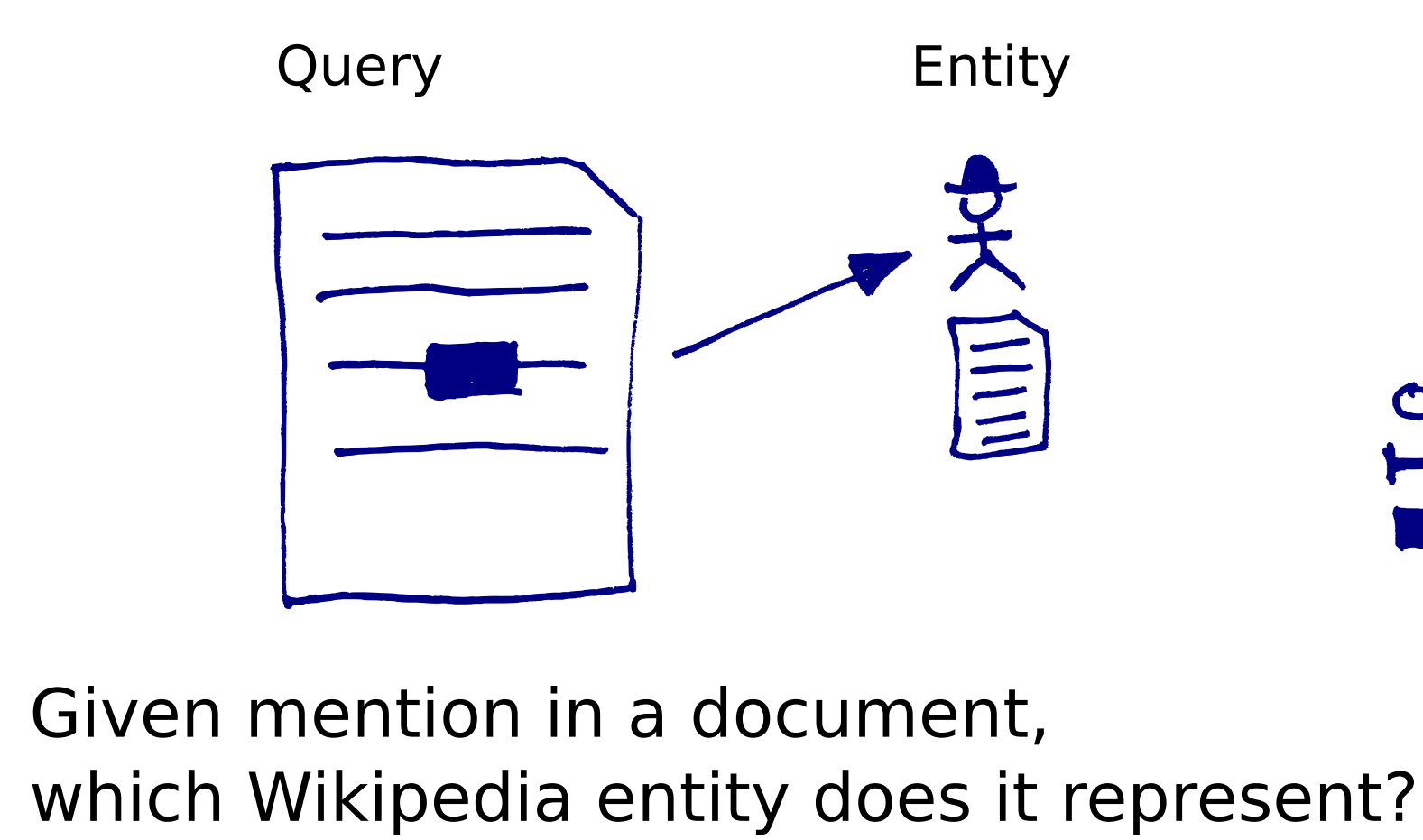
Across-Document Neighborhood Expansion for Candidate Retrieval

Laura Dietz dietz@cs.umass.edu and Jeffrey Dalton jdalton@cs.umass.edu

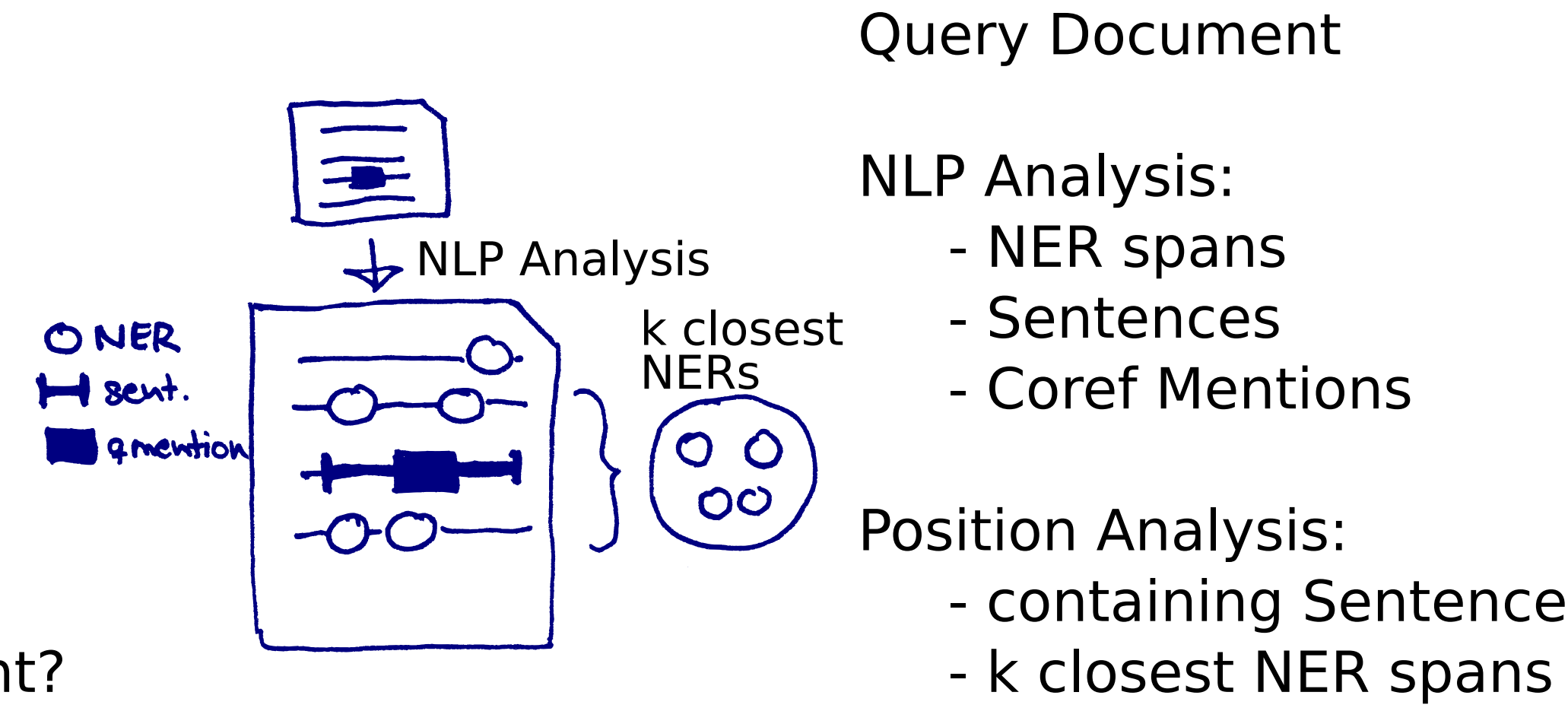
Abstract

Last year's competition demonstrated that the NER context contains important information that should not be ignored in entity linking. State-of-the-art approaches anchor on unambiguous entities, look for overlap in categories, or approximate a joint model of candidate assignments, after Wikipedia candidates have been selected. Current candidate approaches, such as anchor text maps, are effective but may lead to very large candidate sets to be examined. UMass has two objectives for our TAC submission. First, we use cross-document context information to perform entity neighborhood expansion and estimate the importance of entity context using corpus-wide information. Second, we use probabilistic information retrieval that incorporates the neighborhood information to generate a ranked candidate set in a single step. The result is a small candidate set that even for less than 50 candidates contains the true answer in 95% of the cases, allowing for computationally intensive inference in the next phase. It turns out that our best performing run simply predicts the top candidate of the unsupervised candidate ranking, outperforming more than half of the contestants.

Entity Linking Problem



NLP Preprocessing



Candidate Retrieval Model

Mention t , name variants v , sentences s , NER spans e
 component weights λ , relevance weights ϕ

$$\#combine:0=(\lambda_T + \lambda_V):1=\lambda_S:2=\lambda_E$$

$$\#combine:0=\lambda_T:1=\lambda_V$$

$$\#seqdep(t)$$

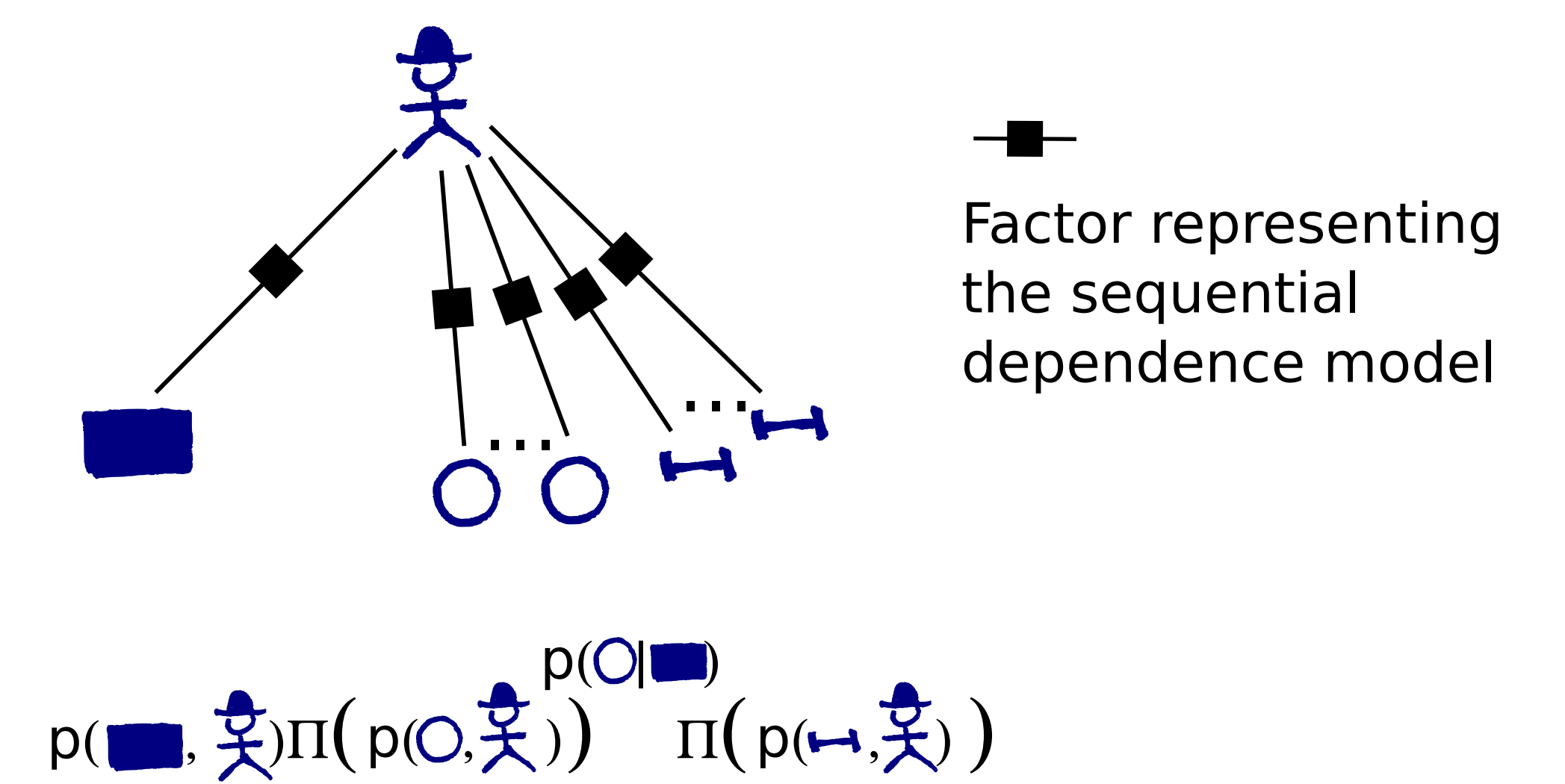
$$\#combine(\#seqdep(v_0) \dots \#seqdep(v_V))$$

$$\#combine(\#seqdep(s_0), \dots, \#seqdep(s_S))$$

$$\#combine:0 = \phi_0^E : \dots : k : \phi_k^E$$

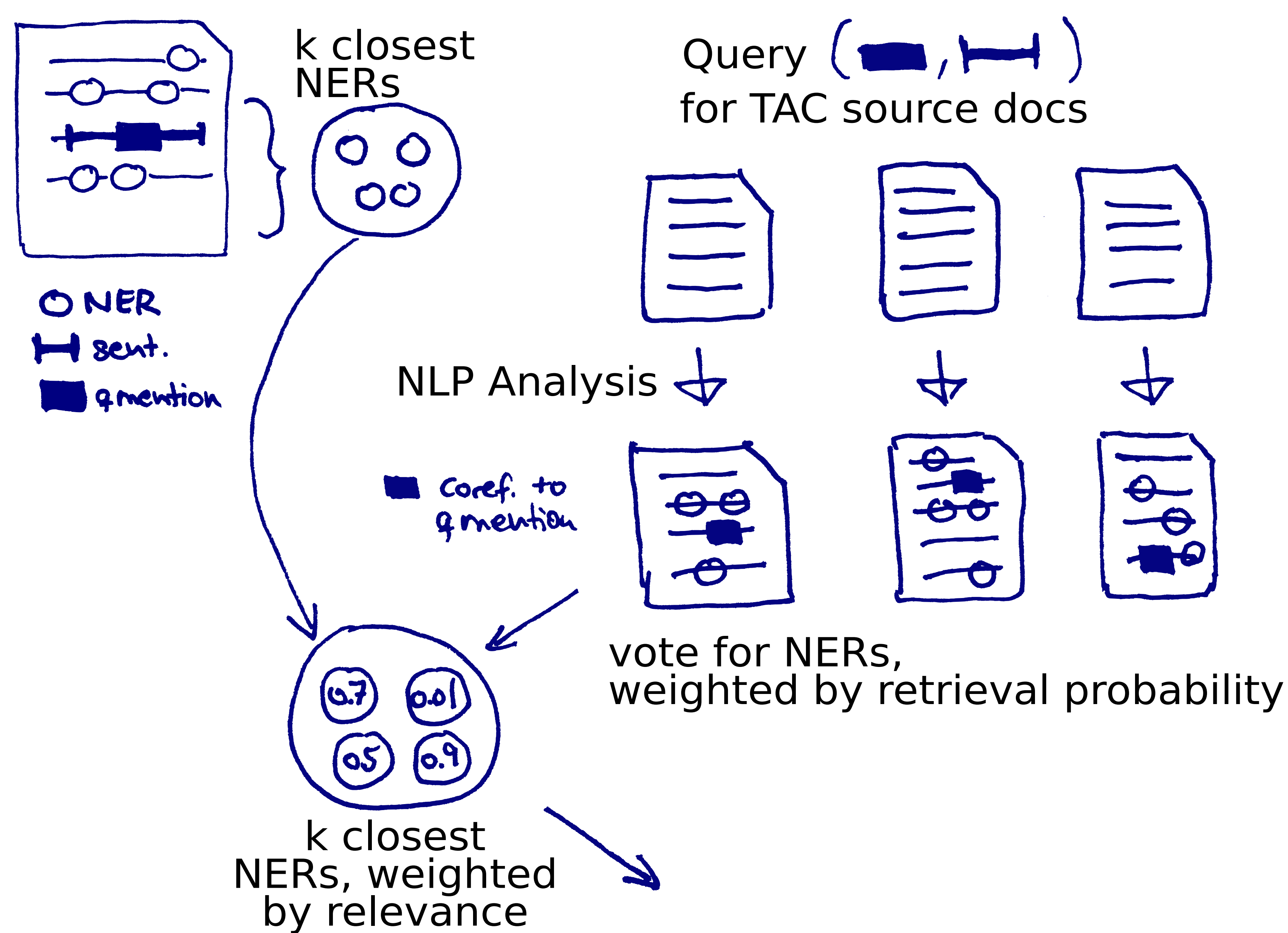
$$\#seqdep(e_0), \dots, \#seqdep(e_k)$$

Retrieval model based on Markov Random Fields

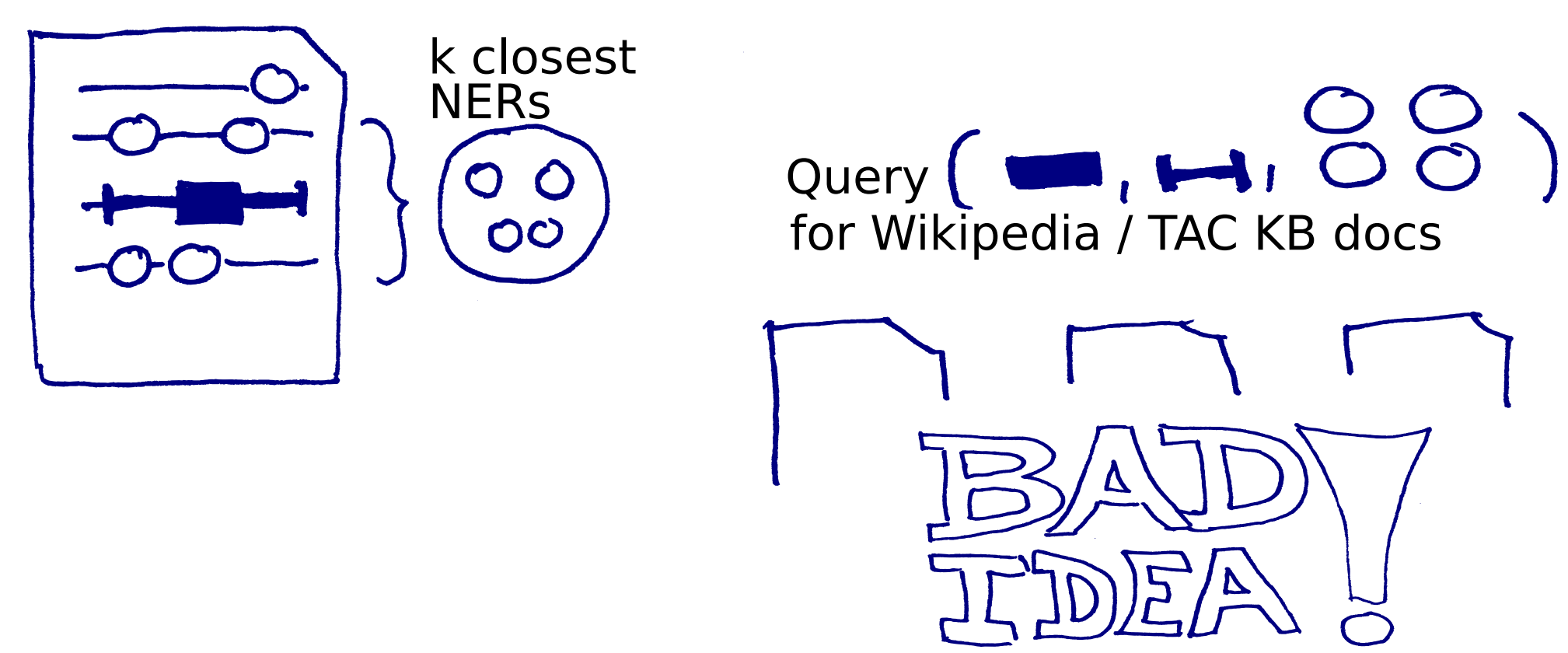


Relevance of NER:
 NERs that help disambiguation
 Which NERs occur near Pseudo-Coreferent Mentions?

Neighborhood Expansion



Motivating Example: Relevance of NERs



Example Query:

ABC shot the TV drama "Lost" in Australia.

Candidates:

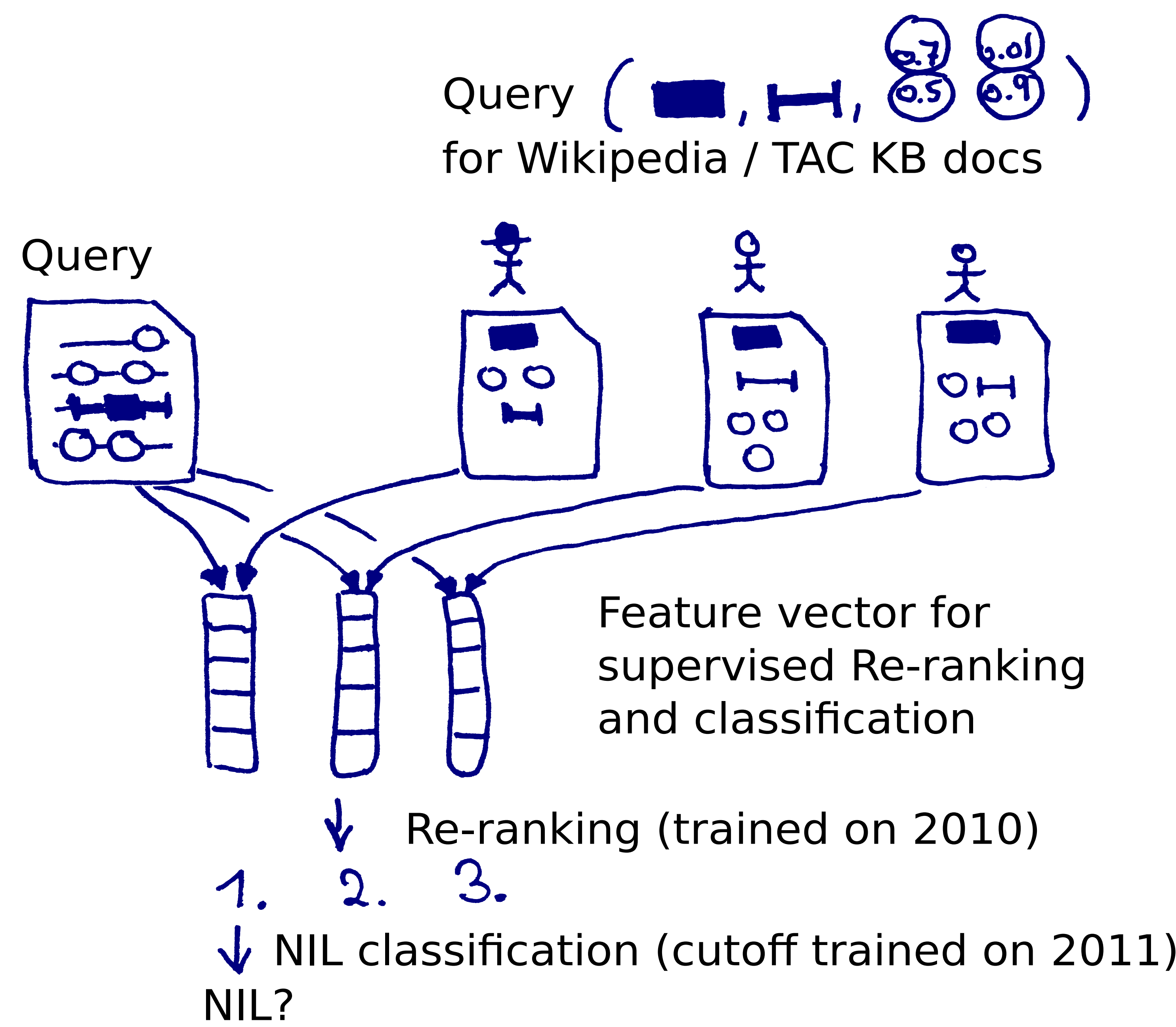
- Australian Broadcasting Corporation Television
- American Broadcast Central
- ...

"Australia" is an unambiguous entity

But: "Australia" is not really relevant for American Broadcast Central.

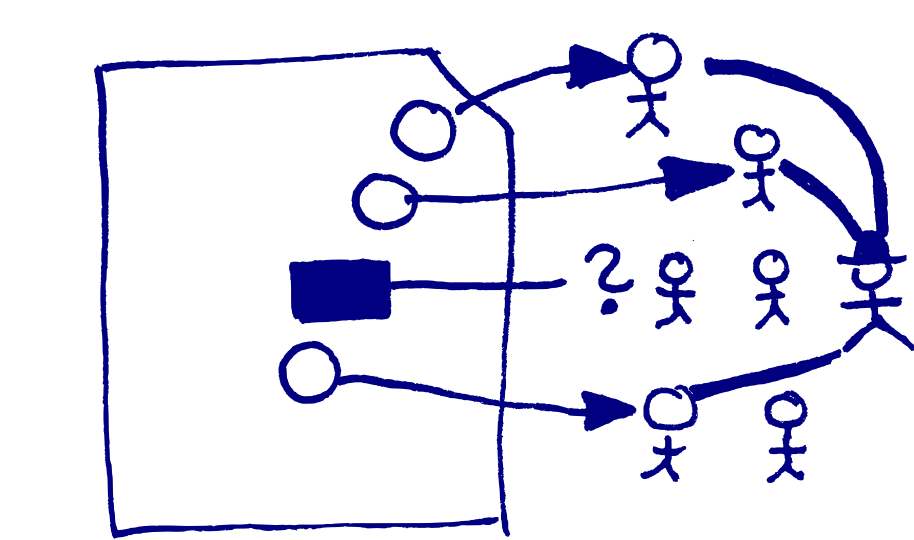
Danger to lead to the wrong conclusion.

Candidate Retrieval and Entity Linking



Joint Neighborhood Assignment Models

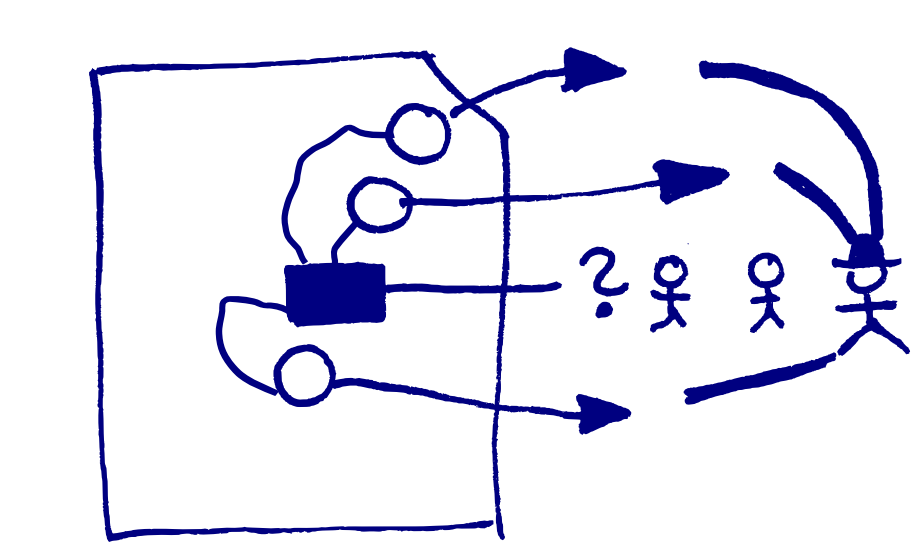
For each NER span:
 Assuming candidate set is retrieved
 Goal: find joint assignment that maximize likelihood



$p(\mathbf{O}, \mathbf{A}) \prod (p(\mathbf{O}, \mathbf{A})) \prod (p(\mathbf{A}, \mathbf{A}))$ Pair-wise model

Candidate Retrieval with Neighborhood Expansion

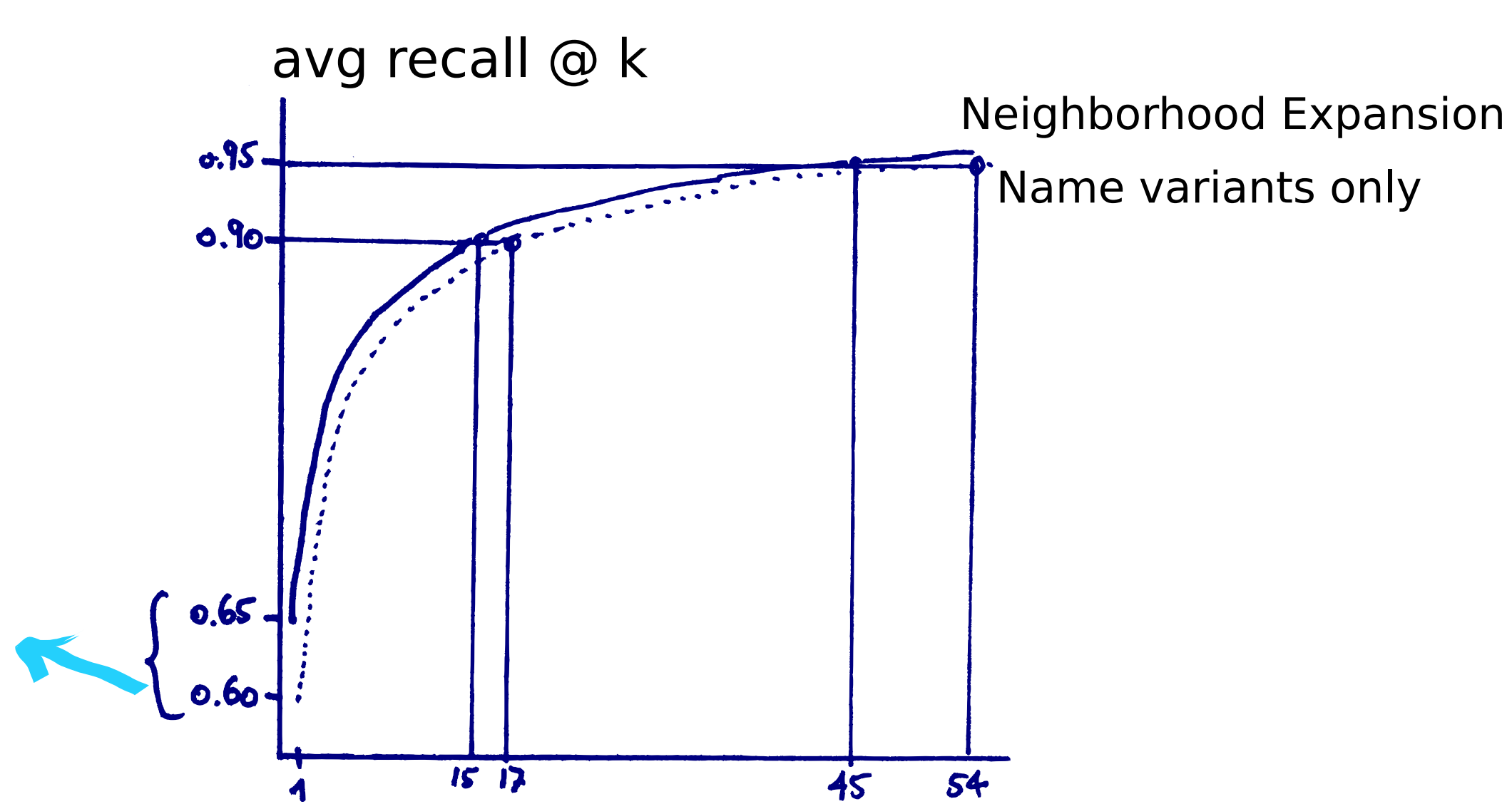
Neighborhood expansion estimates reliability for disambiguating the query mention.



No candidate set necessary!
 Joint assignment model is optimized during candidate retrieval!

$p(\mathbf{O}, \mathbf{A}) \prod (p(\mathbf{O}, \mathbf{A})) \prod (p(\mathbf{A}, \mathbf{A}))$ where $p(\mathbf{O}, \mathbf{A}) =$ Relevance of NER for disambiguation

Results as Candidate Retrieval

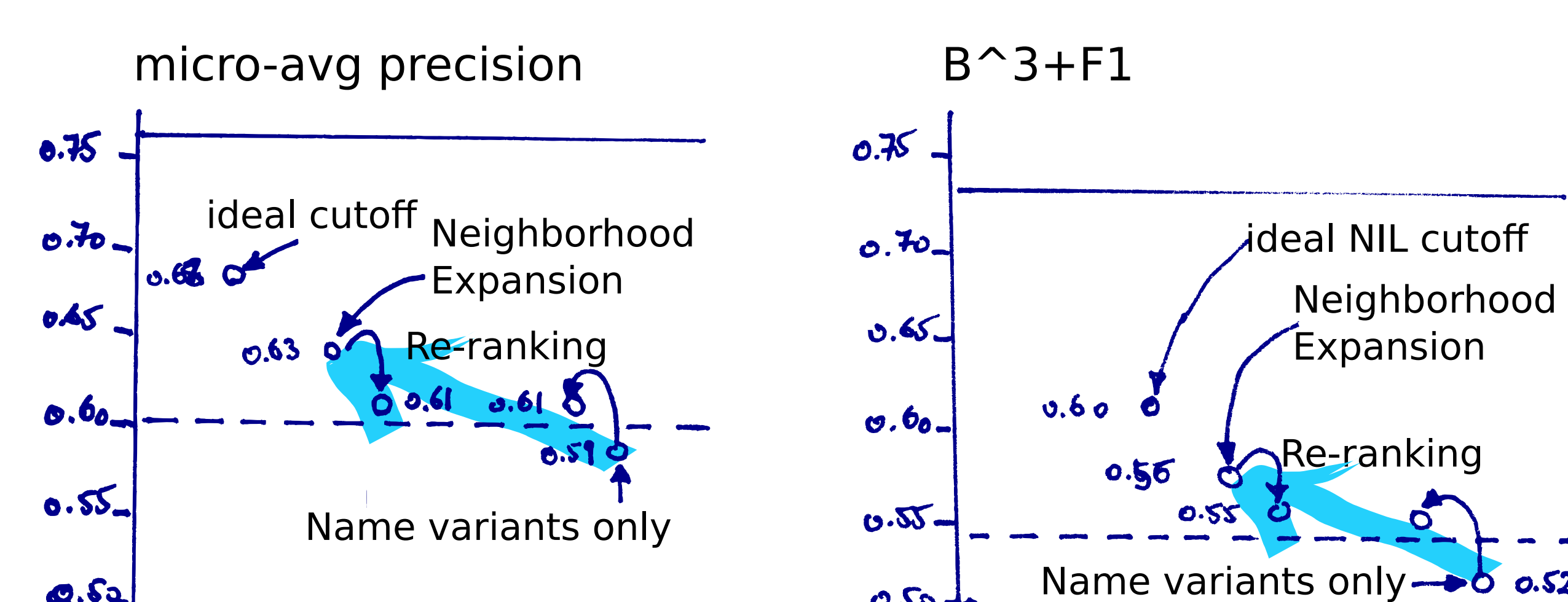


Neighborhood Expansion retrieves the true entity at high cutoff rates

MRR 0.75 (versus 0.72) 95% recall at rank 45

Small candidate set allows for time intensive re-ranking methods!

Results as Entity Linking System 2012



Results as Entity Linking System 2011

