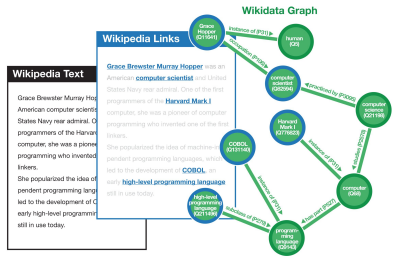
# Using Context Clues to Improve Entity Disambiguation in Text

Will Seaton, Hardik Gupta, Ruochen Zhao, Johannes K. Kolberg

## GOAL

Named Entity Disambiguation is an NLP task that involves identifying and linking an entity mention in text to a node in a knowledge base.

Incorporating "Congruence", or relatedness between entities in the same sentence, can provide added confidence in node prediction.




## DATA

**AIDA CoNLL-YAGO**
Testing dataset. Manually labeled specifically for entity disambiguation with true Wikipedia page assignments.

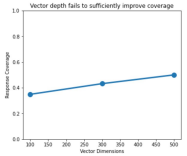
**Kensho-Derived Wikimedia Dataset**
Structured for NLP research by Kensho Tech. Provides raw English Wikipedia text, structured anchor links and an associated knowledge graph.

**Wikipedia2Vec**
Provides word and entity vector embeddings trained directly on the Wikipedia corpus of text, pages, and links.

## PIPELINE

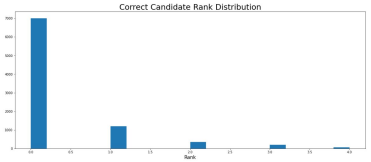
### 1. Candidate Pool Generation

*Text<>Entity Vector Similarity*
Generating candidates using vector similarity failed to contain the correct answer in more than half of the cases.



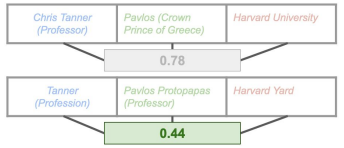

*Anchor Link Statistics*
An "Anchor Link" is a text within a Wikipedia page that hyperlinks to a different page, and represents how we might colloquially refer to that page.

Generating candidates using anchor links contained the correct answer in 87-90% of cases, and in the top 3 rank positions >95% of named entities.

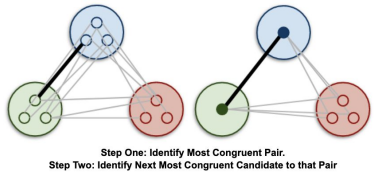



### 2. Congruent Prediction

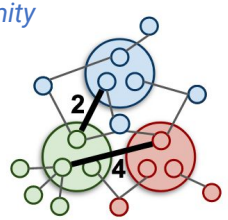
*Combination Centrality*
Calculate average embedding distances between candidates for every unique combination of candidate entity links.




*Entity<>Entity Vector Similarity*
Recursively find most related pairs of candidates by comparing embeddings that encompass text and graph data.




*Knowledge Graph Proximity*
Recursively select candidate pair with shortest path along filtered "dendrite" graph

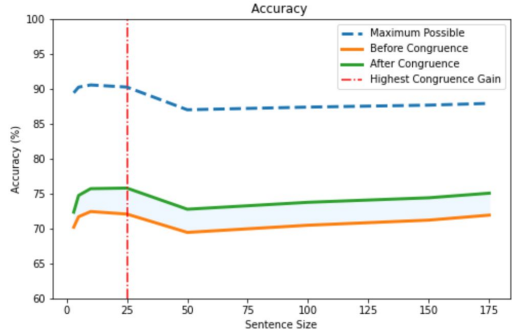



## RESULTS & LEARNINGS

Entity<>Entity Vector Similarity as congruence method improved predictive accuracy by a range of 2.4–3.7% across a range of named entities in a single sentence.

Embeddings must be sufficiently informative, including encompassing information on semantic or knowledge base relatedness, as well as textual similarity.

More granular knowledge graph use requires overcoming significant computational challenges.




## REFERENCES

Pair-Linking for Collective Entity Disambiguation: Two Could Be Better Than All (Phan et al, 2018)

Wikipedia2Vec: An Efficient Toolkit for Learning and Visualizing the Embeddings of Words and Entities from Wikipedia (Yamada et al, 2020)