Strong Heuristics for Named Entity Linking

Marko Čuljak,[‡*] **Andreas Spitz,**[§] **Robert West,**[¶] **Akhil Arora**[¶†]
[‡]University of Zagreb, [§]University of Konstanz, [¶]EPFL
marko.culjak@fer.hr, andreas.spitz@uni-konstanz.de
robert.west@epfl.ch, akhil.arora@epfl.ch

Abstract

Named entity linking (NEL) in news is a challenging endeavour due to the frequency of unseen and emerging entities, which necessitates the use of unsupervised or zero-shot methods. However, such methods tend to come with caveats, such as no integration of suitable knowledge bases (like Wikidata) for emerging entities, a lack of scalability, and poor interpretability. Here, we consider person disambiguation in QUOTEBANK, a massive corpus of speaker-attributed quotations from the news, and investigate the suitability of intuitive, lightweight, and scalable heuristics for NEL in web-scale corpora. Our best performing heuristic disambiguates 94% and 63% of the mentions on QUOTEBANK and the AIDA-CoNLL benchmark, respectively. Additionally, the proposed heuristics compare favourably to the state-of-the-art unsupervised and zero-shot methods, EIGENTHEMES and mGENRE, respectively, thereby serving as strong baselines for unsupervised and zero-shot entity linking.

1 Introduction

While many of the most famous historic quotes are wise irrespective of their origin, this is less true for the majority of contemporary quotes in the news, which require speaker attribution to be useful in journalism or the social and political sciences. This observation is the motivation behind the construction of QUOTEBANK, a corpus of 178 million unique quotations that are attributed to speaker mentions and were extracted from 162 million news articles published between 2008 and 2020 (Vaucher et al., 2021). However, given the ambiguity of names, attributing quotes to mentions is insufficient for proper attribution, and thus, named entity disambiguation is required, a feature which QUOTEBANK lacks.

To tackle this shortcoming and investigate the disambiguation of person mentions in QUOTEBANK as a prototypical example of a web-scale corpus, we explore the suitability of scalable named entity linking (NEL) heuristics, which map mentions of entity names in the text to a unique identifier in a referent knowledge base (KB) and thus, resolve the ambiguity. NEL is an established task and solutions have been used for a variety of applications such as KB population (Dredze et al., 2010) or information extraction (Hoffart et al., 2011), yet the frequency of emerging and unseen entities in news data renders the adaptation of supervised NEL approaches difficult and tends to require unsupervised or zero-shot methods.

While such unsupervised methods (Le and Titov, 2019; Arora et al., 2021) and zero-short methods (Logeswaran et al., 2019; Cao et al., 2021) have been developed in recent years, scalability is an issue. For example, fully disambiguating QUOTEBANK with the state-of-the-art zero-shot NEL method, mGENRE (De Cao et al., 2022), would require approximately 37 years on a single GPU according to our experimental estimates. Therefore, we investigate the suitability of heuristic NEL methods that rely on signals that are simple to extract from mention contexts or entity entries in a KB. In contrast to mGENRE, we find that our best-performing heuristics can solve the same task in 108 days on a single CPU core, i.e., orders of magnitude faster and on cheaper hardware, while achieving comparable performance.

Contributions. To address the need for NEL in web-scale corpora, we investigate the disambiguation performance of simple, interpretable, scalable, and lightweight heuristics and compare them to state-of-the-art zero-shot and unsupervised NEL methods. Our experiments on QUOTEBANK and the AIDA-CoNLL benchmark demonstrate the competitiveness of these heuristics.

[*]Research done while at EPFL.
[†]Corresponding author.

Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop, pages 235 - 246

2 Related Work

Viable learning-based methods for NEL in settings without available training data can be classified into *zero-shot* and *unsupervised* learning.

Zero-shot NEL was introduced by Logeswaran et al. (2019) with the objective of linking mentions to entities that were unseen during training. Later, Wu et al. (2020) proposed a BERT-based model for this task. Finally, Cao et al. (2021) proposed GENRE, a supervised NEL method that leverages BART to retrieve entities by generating their unique names autoregressively, conditioned on the context by employing beam search. While GENRE uses Wikipedia as its referent KB and is not directly compatible with our setting, we compare our methods to mGENRE (De Cao et al., 2022), a multilingual adaptation of GENRE using Wikidata.

Unsupervised NEL. Le and Titov (2019) proposed τMIL-ND, a BiLSTM model trained on noisy labels, which are generated via a heuristic that ranks the candidate entities of a mention based on matching words in a mention and candidate labels. Similarly, Fan et al. (2015) experiment with distant learning for NEL and create training data by merging Freebase with Wikipedia. Recently, Arora et al. (2021) proposed EIGENTHEMES, which is based on the observation that vector representations of gold entities lie in a low-rank subspace of the full embedding space. These low-rank subspaces are used to perform collective entity disambiguation.

While powerful, the aforementioned methods are designed for general domains and multiple entity types, and thus, cannot capitalize on domain- and entity-specific signals. In the following, we investigate the suitability of unsupervised NEL heuristics for person disambiguation in the domain of news quotes in comparison to these methods.

3 Problem Formalization

The input to our NEL system are articles $a \in \mathscr{A}$ from the set $\mathscr{A}$ of all articles in QUOTEBANK. In each article a, a set of entity mentions $\mathscr{M}_a$ is annotated. Each such mention $m \in \mathscr{M}_a$ can be mapped to a set of candidate Wikidata entities $\mathscr{E}_m$, which are uniquely identified by their Wikidata QID identifier (for further details regarding Wikidata, see Appendix D). If multiple entity candidates are available for a mention, we refer to this mention as *ambiguous*. Conversely, *unambiguous* mentions have only a single candidate entity. Given an article $a \in \mathscr{A}$, an ambiguous mention $m \in \mathscr{M}_a$, and all candidate entities $\mathscr{E}_m$, the task of NEL is to identify the entity $e \in \mathscr{E}_m$ to which m refers.

We assume that NEL methods assign a rank $r(e,m)$ to each candidate entity $e \in \mathscr{E}_m$ by ranking candidates according to the score provided by the method, which corresponds to the likelihood that e is the correct entity for m. Consequently, we assume that methods cannot identify cases in which the entity does not exist in the KB or is not contained in the list of candidates (i.e., out-of-KB or NIL predictions). Thus, our focus is on the evaluation of methods in cases where at least one candidate is available.

4 Scoring Methods

We consider three main signals for entity candidate ranking methods: *entity popularity*, *entity-content similarity*, and *entity-entity similarity*. Implementation details are provided in Appendix B.

4.1 Entity Popularity

Entity popularity is an important signal for disambiguating entities in news articles as popular entities are more likely to appear in the news (Shen et al., 2015). Since popularity cannot be measured directly, we utilize 4 proxies derived from Wikidata, some of which have also been used previously as features for supervised NEL (Delpeuch, 2020).

Number of properties (NP). Based on the assumption that Wikidata contains more information for popular entities, we use the number of Wikidata properties to approximate entity popularity.

Number of site links (NS). Similar to NP, a more popular entity is likely connected to more Wikimedia pages. We thus use the number of site links to estimate entity popularity.

PageRank (PR) is a graph centrality metric that was originally developed for web search as a part of Google's search engine (Page et al., 1999). We experiment with two PageRank scores computed on the Wikidata graph (PR_{WD}) and the Wikipedia graph (PR_{WP}) and report their results separately.

Lowest QID (LQID). The Wikidata QID is an auto-incremented integer identifier. Intuitively, well-known entities are added to Wikidata early and their QIDs are low. Therefore, we simply select the candidate with the lowest QID value.

4.2 Entity-Content Similarity

In addition to entity-centric information, we consider the mention context and attempt to match it to the attributes of candidate entities in the KB. Consider the following example from QUOTEBANK:

*"Professor **Tim Wheeler, Vice-Chancellor of the University of Chester**, said: "The university is dedicated to educating the very best nurses [...]"*

Tim Wheeler's title, *Vice-Chancellor of the University of Chester*, exactly matches the short description of a Wikidata entity with QID Q2434362. Therefore, it stands to reason that we can leverage content similarity metrics for entity linking.

Intersection score (IScore). The IScore captures word overlap between mention context and entity descriptions. Let $\mathscr{W}_a$ be a set of lowercased words occurring in article a, let $\mathscr{W}_e$ be a set of words occurring in the textual representation of an entity in Wikidata, and let $\mathscr{W}_{sw}$ be a set of English stopwords. We then compute the IScore of an entity e with respect to article a as

$$\text{IScore}(a,e) = |(\mathscr{W}_a \cap \mathscr{W}_e) \setminus \mathscr{W}_{sw}| \qquad (1)$$

While we could normalize the score by $|\mathscr{W}_a \cup \mathscr{W}_e|$ to obtain a Jaccard similarity, we intentionally bias the IScore towards entities with more substantial descriptions, thereby implicitly incorporating entity popularity information. We use the Porter stemmer (Porter, 1980) for stemming words before matching (please see Appendix E for experiments with IScore using raw input words or lemmatization).

Narrow IScore (NIScore). For a more focused context representation, we also compute a version of the IScore with a narrow context that only contains the sentences in which a mention of the given entity occurs. For further experiments with the selection of mention contexts, see Appendix E.

Cosine similarity of embeddings (CSE). Following a baseline from Arora et al. (2021), to capitalize on the effectiveness of transformer models for NLP tasks, we leverage contextualized language models to create embeddings of article contents and candidate entity descriptions, which are then compared. We employ $\text{BART}_{\text{BASE}}$ (Lewis et al., 2020) to generate embeddings and then compute cosine similarity scores. For details, see Appendix B.

Narrow CSE (NCSE). Similar to the NIScore, we consider a narrow context around entity mentions for computing the CSE by restricting the context that is used for the creation of embeddings to sentences in which the entity occurs.

4.3 Entity-Entity Similarity

Since many mentions of entities can be expected to be unambiguous, we may use such mentions as anchors and leverage their relations to ambiguous mentions for the purpose of disambiguation. Similar to the entity-content similarity methods described above, we experiment with metrics that use intersections of entity occurrences and embedding similarities of attribute values from Wikidata.

Entity-entity IScore (EEIScore). Following the above intuition, the EEIScore utilizes the information that is contained in relations between ambiguous and unambiguous mentions. Let $\mathscr{U}_a$ be the set of all entities that can be mapped to unambiguous mentions in an article a (i.e., mentions that can be trivially disambiguated). Let $\mathscr{S}_e$ be the set of all statements that occur in the Wikidata entry corresponding to an entity e. We define $\mathscr{S}_{\mathscr{U}_a} := \bigcup_{e \in \mathscr{U}_a} \mathscr{S}_e$. Using this set of all statements of unambiguous entities, we then compute the EEIScore of a candidate entity e for an ambiguous mention as:

$$\text{EEIScore}(e, \mathscr{U}_a) = |\mathscr{S}_e \cap \mathscr{S}_{\mathscr{U}_a}| \qquad (2)$$

Cosine similarity of statement value embeddings (CSSVE). We refine the idea behind the intersection score of entity relations by using embeddings of Wikidata statement values and property types (i.e., relations in Wikidata). For each entity e, Wikidata contains a set of statements $s_e = (p_e, v_e)$, consisting of a property p_e and a value v_e. Using this data, we first create embeddings $\varepsilon(v)$ of the values for all statements $s \in \mathscr{S}_{\mathscr{U}_a} \cup \mathscr{S}_e$. We then compute CSSVE as the sum of cosine similarities of statement value embeddings between all pairs of statements of the candidate entity and statements of unambiguous mentions in the article that have matching property types (i.e., describe the same type of relation):

$$\text{CSSVE}(e, \mathscr{U}_a) = \sum_{\substack{(s_u, s_e) \in (\mathscr{S}_{\mathscr{U}_a} \times \mathscr{S}_e) \\ p_u = p_e}} \frac{\varepsilon(v_u) \cdot \varepsilon(v_e)}{\|\varepsilon(v_u)\| \|\varepsilon(v_e)\|} \qquad (3)$$

4.4 Composite Scores

We also use two composite scores in our evaluation: **UIScore** refers to the weighted sum of IScore, NIScore, and EEIScore, while **UCSE** refers to the weighted sum of CSE, NCSE, and CSSVE. Since

CSE and NCSE are cosine similarities, their outputs are constrained to the $[-1,1]$ interval, while CSSVE is unbounded. To ensure similar magnitudes we map all scores to the $[0,1]$ interval by applying the transformation $f(x) = \frac{1}{2}(x+1)$ to CSE and NCSE, and additive smoothing to CSSVE.

5 Data

We focus on QUOTEBANK data, but also investigate the performance on AIDA-CoNLL as a benchmark. Similar to Arora et al. 2021, Raiman and Raiman 2018, and Guo and Barbosa 2018 we label the mentions as either *'easy'* or *'hard'*. In QUOTEBANK, we deem a mention easy if it can be correctly disambiguated using NS and hard otherwise, while in AIDA-CoNLL we use the definition proposed by Arora et al. 2021. In Table 1 we present the statistics for easy and hard mentions in the datasets.

QUOTEBANK is a collection of quotes that were extracted from 127 million news articles and attributed to one of 575 million speaker mentions (Vaucher et al., 2021), out of which 75% are unambiguous. For our evaluation, we use a randomly sampled subset of 300 articles that are manually annotated with 1,866 disambiguated person mentions. 70% of these mentions are unambiguous. Out of the ambiguous mentions, it was possible to determine ground truth labels for 310 (57%), which we use in our evaluation. We split the ground truth into 245 mentions (79%) for evaluation and 65 mentions (21%) for parameter tuning. For a more thorough description of the QUOTEBANK ground truth, see Appendix A.

AIDA-CoNLL. To assess whether the proposed methods can be used for unsupervised NEL in general, we also evaluate their performance on the AIDA-CoNLL benchmark (Hoffart et al., 2011), which is based on the CoNLL 2003 shared task (Sang and Meulder, 2003). We use the same setup as Arora et al. 2021 and use the validation set for hyperparameter optimization. The differences between the evaluation setups of QUOTEBANK and AIDA-CoNLL are explained in Appendix C.

6 Evaluation

All the resources (code, datasets, etc.) required to reproduce the experiments in this paper are available at `https://github.com/epfl-dlab/nelight`.

Table 1: The number of mentions in different difficulty categories. The definitions of *Easy* and *Hard* mentions are presented in § 6.2. On AIDA-CoNLL, #Easy + #Hard ≠ #Overall because for some mentions, the gold-entity was not contained in the candidate set.

Dataset	#Easy	#Hard	#Overall
QUOTEBANK	203	42	245
AIDA-CoNLL	2555	1136	4478

6.1 Evaluation Setup

We use *micro* precision at one (P@1) and mean reciprocal rank (MRR) as the evaluation metrics. The metrics are aggregated over all ambiguous mentions for which ground truth data is available. Performance is reported with 95% bootstrapped confidence intervals (CIs) over 10,000 bootstrap samples. To identify optimal weight parameters for the composite metrics, we perform a grid search over the range $[0,1]$. For the QUOTEBANK data, the best performance is obtained for weights $(1,1,1)$ for UIScore and $(0.45,0.9,0.2)$ for UCSE. For the AIDA-CoNLL data, we perform the parameter optimization on the official validation set, where the best performance is obtained for weights $(0.9,0,1)$ and $(0,1,1)$ for UIScore and UCSE, respectively.

Tie breaking. Several ranking methods introduce ties, which we break by using popularity heuristics. Among the popularity heuristics, only LQID is injective and always outputs distinct scores for different entities. In our experiments, we, therefore, use LQID to break ties if they remain after using other tie-breakers. A full breakdown of the tie-breaking performance for all popularity-based methods can be found in Appendix E.3.

6.2 Results

We report P@1 for all the methods in Table 2, and MRR in Appendix G. For comparison, we present the analytically computed performance of a random baseline, which picks one of the entity candidates uniformly at random.

QUOTEBANK. Among the popularity-based metrics, the best results are achieved by NS. However, considering the confidence intervals, the performance gains of NS over PR_{WP} and NP are not significant. LQID and PR_{WD} perform poorly in comparison to the other methods. All popularity methods outperform the random baseline, confirming their usefulness as a prior for NEL.

Table 2: P@1 of the methods on QUOTEBANK and AIDA-CoNLL. Eigen (IScore) refers to EIGENTHEMES weighted by IScore. Eigen on QUOTEBANK is weighted by NS, while On AIDA, it denotes the results obtained by Arora et al. 2021. The best obtained P@1 in each column is highlighted **bold**.

	QUOTEBANK			AIDA-CoNLL		
Method	**Easy**	**Hard**	**Overall**	**Easy**	**Hard**	**Overall**
Random	0.374 ± 0.017	0.260 ± 0.045	0.354 ± 0.024	0.267 ± 0.014	0.066 ± 0.004	0.169 ± 0.009
LQID	0.828 ± 0.054	0.238 ± 0.140	0.727 ± 0.056	0.856 ± 0.014	0.259 ± 0.029	0.554 ± 0.016
NP	0.921 ± 0.040	0.143 ± 0.120	0.788 ± 0.052	0.856 ± 0.014	0.190 ± 0.023	0.536 ± 0.015
NS	$\mathbf{1.000 \pm 0.000}$	0.000 ± 0.000	0.829 ± 0.048	0.908 ± 0.012	0.275 ± 0.026	0.588 ± 0.014
PR_{WD}	0.768 ± 0.059	0.214 ± 0.132	0.673 ± 0.061	0.838 ± 0.014	0.155 ± 0.021	0.517 ± 0.015
PR_{WP}	0.926 ± 0.040	0.333 ± 0.140	0.824 ± 0.048	$\mathbf{0.938 \pm 0.010}$	0.282 ± 0.027	0.607 ± 0.014
IScore	0.956 ± 0.030	0.762 ± 0.134	0.922 ± 0.034	0.863 ± 0.014	0.549 ± 0.029	0.632 ± 0.015
NIScore	0.966 ± 0.030	0.571 ± 0.151	0.851 ± 0.014	0.851 ± 0.014	0.407 ± 0.028	0.562 ± 0.015
CSE	0.901 ± 0.044	0.500 ± 0.159	0.833 ± 0.047	0.386 ± 0.019	0.276 ± 0.026	0.290 ± 0.014
EEIScore	0.951 ± 0.034	0.690 ± 0.143	0.906 ± 0.036	0.815 ± 0.016	0.382 ± 0.031	0.562 ± 0.015
CSSVE	0.872 ± 0.049	0.357 ± 0.155	0.784 ± 0.051	0.712 ± 0.017	0.256 ± 0.026	0.471 ± 0.015
UIScore	0.966 ± 0.030	$\mathbf{0.833 \pm 0.123}$	0.943 ± 0.029	0.833 ± 0.014	0.577 ± 0.028	0.621 ± 0.014
UCSE	0.941 ± 0.034	0.595 ± 0.156	0.882 ± 0.042	0.465 ± 0.019	0.386 ± 0.029	0.363 ± 0.014
Eigen	0.995 ± 0.010	0.238 ± 0.134	0.865 ± 0.044	0.859 ± 0.014	0.500 ± 0.030	0.617 ± 0.015
Eigen (IScore)	0.956 ± 0.030	0.714 ± 0.147	0.914 ± 0.037	0.794 ± 0.015	$\mathbf{0.702 \pm 0.029}^{\dagger}$	0.631 ± 0.014
mGENRE	0.995 ± 0.010	0.810 ± 0.143	$\mathbf{0.963 \pm 0.025}$	0.925 ± 0.011	0.610 ± 0.028	$\mathbf{0.682 \pm 0.014}^{\dagger}$

† Indicates statistical significance ($p < 0.05$) between the best and the second-best method using bootstrapped 95% CIs.

Table 3: P@1 of representative methods on various entity types in the AIDA-CoNLL dataset. In the evaluation dataset, there are 1016 PER, 1345 ORG, 1575 LOC, and 542 MISC mentions. The best P@1 in each column is highlighted **bold**.

Method	**PER**	**ORG**	**LOC**	**MISC**
NS	0.687 ± 0.030	0.410 ± 0.027	0.777 ± 0.021	0.292 ± 0.039
PR_{WP}	0.719 ± 0.029	0.477 ± 0.026	0.752 ± 0.023	$\mathbf{0.293 \pm 0.042}$
IScore	0.786 ± 0.026	0.597 ± 0.026	0.694 ± 0.022	0.245 ± 0.035
UIScore	$\mathbf{0.789 \pm 0.026}$	0.601 ± 0.026	0.664 ± 0.023	0.232 ± 0.035
mGENRE	0.720 ± 0.027	0.608 ± 0.027	$\mathbf{0.858 \pm 0.018}$	0.284 ± 0.039
Eigen (IScore)	0.760 ± 0.026	$\mathbf{0.732 \pm 0.025}$	0.608 ± 0.024	0.205 ± 0.035
Eigen	0.696 ± 0.028	0.671 ± 0.026	0.655 ± 0.023	0.223 ± 0.035

The performances of entity-entity similarity methods are similar to their entity-content similarity counterparts. This is in line with the hypothesis that the gold entities mentioned in the same article are more closely related than the other subsets of entity candidates (Arora et al., 2021). Generally, combining the entity-content similarity methods with their entity-entity similarity counterparts leads to performance gain, as seen from the example of UIScore and UCSE. Considering the overall performance, UIScore outperforms CSE and all entity popularity methods. The performance of CSE is similar to the performance of NS, which is considerably simpler. Finally, the performance of UIScore is comparable to mGENRE, achieving a slightly higher P@1 on hard mentions.

AIDA-CoNLL. In the AIDA-CoNLL data, IScore and UIScore achieve a comparable performance to the current state-of-the-art in unsupervised entity linking, EIGENTHEMES (Arora et al., 2021), but lag slightly behind mGENRE (De Cao et al., 2022), the state-of-the-art zero-shot method. In contrast to QUOTEBANK, we do not observe performance gains as a result of combining different heuristics as UIScore fails to outperform IScore. EIGENTHEMES weighted by IScore achieves by far the strongest performance on the hard mentions, despite a relatively poor performance on easy mentions. Overall, the performance makes an encouraging case for the heuristics to be used as strong baselines for entity linking in general, and on large data sets in particular.

AIDA-CoNLL entity type analysis. The results of the analysis with respect to the entity types available in the original CoNLL 2003 dataset (Sang and Meulder, 2003) are shown in Table 3. In CoNLL 2003, there are four entity types: person (PER), organization (ORG), location (LOC), and miscellaneous (MISC).

The UIScore heuristic achieves the best performance on PER mentions, outperforming even mGENRE. As described in Subsection 4.2, persons that are mentioned in the news are usually introduced by a simple description of their background or current occupation even if they are well known. Since the heuristics proposed for person disambiguation in QUOTEBANK are based on this assumption, this explains a relatively strong performance of the UIScore heuristic on PER type entities in AIDA-CoNLL.

Despite superior performance on PER mentions, UIScore lags behind mGENRE and EIGENTHEMES on other types. We attribute this to a lack of introductory context in comparison to PER mentions (e.g., a mention of "China" in an article would typically not be followed by "a state in East Asia"). Furthermore, non-person named entities are frequently used as metonyms (e.g., "Kremlin" is a frequent metonym for the Russian government, but it can also refer to the Kremlin building). Depending on the context, a simple heuristic such as IScore may thus struggle to properly link candidates.

Computational performance. While mGENRE achieves the best performance on both QUOTEBANK and on AIDA-CoNLL, it is a transformer model and takes *substantially* longer to run in comparison to UIScore. Disambiguating a single mention with mGENRE takes approximately 533 times longer than with UIScore, and approximately 533K times longer than with NS, thereby rendering it infeasible for speaker disambiguation in QUOTEBANK, which contains millions of news articles. For a detailed breakdown of inference times per mention, see Appendix F.

7 Discussion

Overall, the results highlight the practicality of the proposed heuristics. Our simple heuristics outperform those based on word embeddings and are competitive in comparison to mGENRE.

7.1 Error analysis

To take a closer at avenues for improvement, we show a manual error analysis for UIScore in Table 4. In 6 cases, the predicted entity and the gold entity have a matching domain (e.g., both are sportsmen). In 4 cases, the key property by which a human could determine the correct entity was only implicitly mentioned in the context, which caused a failure in string matching. For 3 articles, a key property of the gold entity was not listed in Wikidata, even though it could be found in external sources such as Wikipedia. The remaining error stems from the presence of a "decoy" entity, i.e, an influential but unrelated entity that induced spurious matches. For a thorough description and illustration of the error categories, see Appendix H.

7.2 Limitations

Since UIScore is the most promising of our heuristics, we focus on it and its components.

Table 4: Error sources for UIScore.

Error source	#Mentions
Similar domain	6 (42.9%)
Key property implicit in the text	4 (28.6%)
Key property not in Wikidata	3 (21.4%)
Decoy mention	1 (7.1%)

The biggest limitation of IScore is imposed by the equal importance that is assigned to words in the context, which could be improved by re-ranking important words for given entities. Similarly, Wikidata properties for EEIScore and CSSVE could be ranked or filtered (for example, the property *date of birth* is likely to cause spurious matches, while *occupation* is likely useful).

Regarding tie-breaking, the use of LQID is intuitive for persons in the news domain, but may fail for other entity types and other domains, and is dependent on Wikidata. Finally, in our focus on QUOTEBANK data, we are reliant on the authors' method for candidate generation, which could be improved for better performance in the future.

8 Conclusions and Future Work

We tackled the problem of entity linking in QUOTEBANK by employing heuristics that rely on simple signals in the context of mentions and the referent KB. The solid overall performance of the proposed heuristics on QUOTEBANK, their low computational complexity, and competitive performance on the AIDA-CoNLL benchmark suggest that they can be used as strong baselines for unsupervised entity linking in large datasets.

Future work. We plan to experiment with weighting schemes that account for word importance, utilize additional signals from the KB, and include improved candidate generation methods. Finally, we aim to provide a disambiguated version of QUOTEBANK to the community.

Acknowledgements

We would like to thank Vincent Ng for providing insightful feedback during the pre-submission mentorship phase. This project was partly funded by the Swiss National Science Foundation (grant 200021_185043), the European Union (TAILOR, grant 952215), the Microsoft Swiss Joint Research Center, and the University of Konstanz Zukunftskolleg. We also acknowledge generous gifts from Facebook and Google supporting West's lab.

References

Akhil Arora, Alberto García-Durán, and Robert West. 2021. Low-rank subspaces for unsupervised entity linking. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.

Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2021. Autoregressive entity retrieval. In *9th International Conference on Learning Representations*.

Nicola De Cao, Ledell Wu, Kashyap Popat, Mikel Artetxe, Naman Goyal, Mikhail Plekhanov, Luke Zettlemoyer, Nicola Cancedda, Sebastian Riedel, and Fabio Petroni. 2022. Multilingual Autoregressive Entity Linking. *Transactions of the Association for Computational Linguistics*, 10:274–290.

Antonin Delpeuch. 2020. Opentapioca: Lightweight entity linking for wikidata. In *Proceedings of the 1st Wikidata Workshop (Wikidata 2020) co-located with the 19th International Semantic Web Conference*.

Mark Dredze, Paul McNamee, Delip Rao, Adam Gerber, and Tim Finin. 2010. Entity disambiguation for knowledge base population. In *Proceedings of the 23rd International Conference on Computational Linguistics*.

Miao Fan, Qiang Zhou, and Thomas Fang Zheng. 2015. Distant supervision for entity linking. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation*.

Zhaochen Guo and Denilson Barbosa. 2018. Robust named entity disambiguation with random walks. *Semantic Web*, 9:459–479.

Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. 2011. Robust disambiguation of named entities in text. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*.

Phong Le and Ivan Titov. 2019. Distant learning for entity linking with automatic noise detection. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Lajanugen Logeswaran, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, Jacob Devlin, and Honglak Lee. 2019. Zero-shot entity linking by reading entity descriptions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

George A. Miller. 1995. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab. Previous number = SIDL-WP-1999-0120.

Dario Pavllo, Tiziano Piccardi, and Robert West. 2018. Quootstrap: Scalable unsupervised extraction of quotation-speaker pairs from large news corpora via bootstrapping. In *Proceedings of the Twelfth International Conference on Web and Social Media*.

Martin F Porter. 1980. An algorithm for suffix stripping. *Program*, 14(3):130–137.

Jonathan Raiman and Olivier Raiman. 2018. Deeptype: Multilingual entity linking by neural type system evolution. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning*.

Wei Shen, Jianyong Wang, and Jiawei Han. 2015. Entity linking with a knowledge base: Issues, techniques, and solutions. *TKDE*, 27(2).

Timoté Vaucher, Andreas Spitz, Michele Catasta, and Robert West. 2021. Quotebank: A corpus of quotations from a decade of news. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*.

Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. 2020. Scalable zero-shot entity linking with dense entity retrieval. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*.

Appendix

A Ground Truth Data

For the method evaluation, we randomly sample 300 articles from QUOTEBANK. The ground truth for 160 articles is determined by the author, while the remaining 140 articles are annotated by the author's colleagues. The annotators were provided with article content, article title, publication date, article URL, a list of ambiguous named entity mentions, and for each ambiguous mention, a candidate set of QIDs as listed in QUOTEBANK. The annotators had to either select the correct QID from the candidate set or select one of the following categories if the correct QID is not listed:

- *The mention does not refer to a person.* Sometimes, buildings and other artifacts named after some person are identified as a person. We ignore such mentions in the evaluation.
- *The correct QID does not exist in Wikidata.* This means that a person is likely not significant enough to have a Wikidata item. For example, sometimes a journalist or a photographer of a newspaper where the article is published shares the name of a famous person and is therefore listed as a speaker candidate.
- *The correct QID exists in Wikidata but is not listed.* This can happen if the correct QID is added to Wikidata after the candidate entities were generated.
- *Impossible to determine.* Some articles are either too noisy or do not contain enough information for disambiguation to be feasible.

In Table 5, we present the distribution of person mentions in the evaluation data with respect to different categories. We observe that more than 70% of the 1866 mentions are unambiguous. For 310 (57%) of the ambiguous mentions, it was possible to determine the ground truth based on the given candidate sets. For the majority of the remaining 43% of ambiguous mentions no correct entity was available in Wikidata.

The main drawback of the QUOTEBANK evaluation dataset is its small size. Since all articles were annotated by only one annotator, there is no data on the inter-annotator agreement. In the future, we aim to create a more sophisticated benchmark dataset via crowdsourcing.

Table 5: Distribution of mentions in the ground truth data with respect to ambiguity and availability of ground truth.

Category		**#Mentions**
Unambiguous		1322 (70.8%)
Ambiguous	Gold entity exists	310 (16.6%)
	No correct QID in Wikidata	151 (8.1%)
	Impossible	37 (2.0%)
	Correct QID not listed	24 (1.3%)
	Not a person	22 (1.2%)
Total		1866

B Implementation Details of the Scoring Methods

B.1 IScore

To calculate the IScore, we first obtain labels of Wikidata statement values listed for e. We then tokenize the content of a using the tagset of the Penn Treebank Tokenizer. We use the computed tokens to create sets $\mathcal{W}_a$ and $\mathcal{W}_e$. Then, we apply the formula given in equation 1 and compute the IScore based on $\mathcal{W}_a$, $\mathcal{W}_e$, and a predefined set of English stopwords $\mathcal{W}_{sw}$[1].

B.2 CSE

To embed an article, we follow the standard transformer model preprocessing procedure. We tokenize the article content using the model-specific tokenizer, respecting BART's 1024 token limit by simply truncating the input if the limit is exceeded. We then feed the obtained tokens to BART and average the last hidden state of the model output. Since truncation leads to loss of information in comparison to other methods, we experimented with chunking the input into chunks of at most 1024 tokens, computing token embeddings in each chunk separately, and aggregating the obtained token embeddings. However, this did not improve performance on QUOTEBANK (0.698 P@1 and 0.818 MRR), while all articles from AIDA-CoNLL are within the token limit so we report the results of the first approach.

Embedding the entity is slightly more challenging. Following the same procedure as for the computation of the article content embeddings, we compute the embedding the the first paragraph in an entity's Wikipedia page if such a page is available. Otherwise, we compute the embeddings of the short description, and each statement value la-

[1] https://gist.github.com/sebleier/554280

Table 6: Results of the IScore ablation study with respect to word normalization and inclusion of different Wikidata features. In each row, we report P@1 and MRR of IScore method for the combinations of the following Wikidata features: short description (D), Wikipedia first paragraph (P), statement value labels (S), and statement value labels and aliases (S_A) for a setting without word normalization, as well as for settings with stemming and lemmatization. The best results in each column are in bold. Since S_A is essentially a superset of S, we omit the combinations where both S and S_A appear. All the experiments were run with NS as a tie-breaker.

	No normalization		Lemmatization		Stemming	
Combination	**P@1**	**MRR**	**P@1**	**MRR**	**P@1**	**MRR**
D	0.869 ± 0.044	0.921 ± 0.027	0.890 ± 0.040	0.930 ± 0.026	0.894 ± 0.039	0.934 ± 0.026
P	0.832 ± 0.049	0.903 ± 0.030	0.816 ± 0.051	0.895 ± 0.029	0.832 ± 0.047	0.902 ± 0.029
S	0.894 ± 0.040	0.936 ± 0.026	0.898 ± 0.039	0.940 ± 0.024	0.906 ± 0.038	0.944 ± 0.024
S_A	0.886 ± 0.041	0.932 ± 0.025	0.890 ± 0.042	0.935 ± 0.025	0.898 ± 0.039	0.939 ± 0.024
D + P	0.841 ± 0.046	0.907 ± 0.028	0.820 ± 0.050	0.898 ± 0.030	0.841 ± 0.047	0.906 ± 0.028
D + S	**0.902 ± 0.039**	**0.943 ± 0.024**	**0.906 ± 0.038**	0.945 ± 0.022	**0.918 ± 0.035**	**0.952 ± 0.021**
D + S_A	0.890 ± 0.041	0.937 ± 0.023	**0.906 ± 0.038**	**0.947 ± 0.022**	0.914 ± 0.037	0.950 ± 0.022
P + S	0.861 ± 0.044	0.919 ± 0.028	0.861 ± 0.045	0.920 ± 0.028	0.873 ± 0.044	0.925 ± 0.026
P + S_A	0.878 ± 0.042	0.928 ± 0.025	0.882 ± 0.041	0.931 ± 0.025	0.882 ± 0.042	0.930 ± 0.025
D + P + S	0.861 ± 0.045	0.921 ± 0.026	0.861 ± 0.045	0.920 ± 0.027	0.873 ± 0.042	0.926 ± 0.026
D + P + S_A	0.886 ± 0.042	0.934 ± 0.025	0.886 ± 0.041	0.934 ± 0.025	0.882 ± 0.042	0.930 ± 0.026

Table 7: Comparison of performances of CSE and IScore when considering different context sizes. Ensemble refers to the sum of the scores obtained considering the narrow and entire context of the article, respectively. The best results for each scoring method are in bold. All the experiments were run with NS as a tie-breaker.

Method	**Context**	**P@1**	**MRR**
CSE	Narrow	0.751 ± 0.055	0.857 ± 0.033
	Entire	0.833 ± 0.050	0.902 ± 0.029
	Ensemble	**0.857 ± 0.044**	**0.921 ± 0.025**
IScore	Narrow	0.898 ± 0.039	0.941 ± 0.023
	Entire	0.918 ± 0.035	0.952 ± 0.021
	Ensemble	**0.922 ± 0.036**	**0.954 ± 0.022**

bel listed for an entity in Wikidata, and aggregate them via arithmetic mean.

B.3 mGENRE

We use mGENRE in a similar setup as De Cao et al. (2022). Suppose that we want to disambiguate entity mention m occurring in an article a. We first enclose m with special tokens `[START]` and `[END]` that correspond to the start and the end of a mention span. We then take at most t mBART (Liu et al., 2020) tokens from either side. As the input for mGENRE, we use a string consisting of the left context, the mention enclosed with the special tokens, and the right context. mGENRE then outputs the top k entity QIDs and their respective scores, where k is the beam size. For entities in $\mathcal{Q}_m$ that are not retrieved by mGENRE, we simply assign 0 as a score. Note that mGENRE outputs the scores corresponding to the negative log-likelihood of the resulting sequence. Thus, in order for 0 to be the smallest possible score, we exponentiate the scores obtained from mGENRE. In the QUOTEBANK setup, we also perform one additional step: since each speaker candidate can be mentioned multiple times in the text, we run mGENRE for each of the speaker candidate mentions and sum the scores obtained for each of the candidate Wikidata entities.

Table 8: Performances of mGENRE for different context sizes. The best result in each column is highlighted **bold**.

	QUOTEBANK		AIDA-CoNLL	
t	**P@1**	**MRR**	**P@1**	**MRR**
64	0.951 ± 0.029	0.968 ± 0.018	0.664 ± 0.013	0.713 ± 0.012
128	**0.963 ± 0.025**	**0.976 ± 0.017**	0.675 ± 0.014	0.723 ± 0.013
256	0.959 ± 0.026	0.972 ± 0.021	**0.682 ± 0.014**	**0.730 ± 0.013**

In Table 8, we present the performances of mGENRE on both QUOTEBANK and AIDA-CoNLL for different values of t, while in Table 2 we report only the best obtained P@1. In all our experiments with mGENRE, we set the beam size k to 10.

C Evaluation Setup Details

QUOTEBANK. The QUOTEBANK data exclusively contains annotations of person mentions. Before training a model that attributes the quotations to their respective speakers, the quotations and speaker candidates are identified in the article text

(Vaucher et al., 2021). The extraction of speaker candidates is explained in detail by Pavllo et al. (2018). Although a speaker candidate can appear in an article multiple times, the quotations are not attributed to specific mentions but rather to the most likely speaker candidate. Thus, we evaluate our methods on QUOTEBANK on a speaker candidate level and refer to speaker candidates as mentions to ensure that our method and result descriptions are consistent with the standard nomenclature.

AIDA-CoNLL. When evaluating our methods on the AIDA-CoNLL benchmark, we do not ignore the mentions for which the gold entity either cannot be determined or is not retrieved by the candidate generator. As a consequence, the resulting P@1 and MRR reported on AIDA-CoNLL are significantly lower in comparison to the QUOTEBANK results as they are bounded by the recall of the candidate generator. We use the same candidate generator as Arora et al. (2021), which imposes an upper bound of 0.824 to P@1 and MRR. Additionally, to ensure a fair comparison with Arora et al. (2021), we break ties by selecting the first speaker candidate with the same score and use the same definition of the easy and hard mentions when reporting the method performances.

D Wikidata

Wikidata is a large community-driven KB. It boasts more than 96 million data items as of January 2022, out of which 6 million are humans. Each Wikidata item is identified by a unique positive integer prefixed with the upper-case letter Q, also known as QID (e.g. Earth (Q2), Mahatma Gandhi (Q1001)). Obligatory data fields of items are a label and a description. Labels and descriptions need not be unique, but each item is uniquely identified by a combination of a label and a short description. Therefore, each QID is linked to the label-description combination. Optionally, some items consist of aliases (alternative names for an entity) and statements. Statements provide additional information about an item and they consist of at least one property-value pair. A property is a pre-defined data type, identified by a unique positive integer, but unlike items, it is prefixed with the upper-case letter P (e.g. occupation (P106), sex or gender (P21)). The value of a statement may take on many types, such as Wikidata items, strings, numbers, or media files. Some items also have a list of site links that connect them to the corresponding page of the entity in other Wikimedia projects, such as Wikipedia or Wikibooks. The methods we propose in Section 4 leverage the described information to link the named entity mentions in the news articles to their respective Wikidata entities.

E Additional Experiments

E.1 Wikidata Features and Word Normalization Ablation for IScore

In Table 6, we show the results of an ablation study that aims to assess the effect of the inclusion of different Wikidata entity features on the performance of IScore and word normalization methods. The features we consider are short descriptions, statement value labels with and without aliases, and Wikipedia first paragraphs. We obtain the best results by leveraging short descriptions and Wikidata statement values. When using only Wikipedia first paragraphs, we obtain a performance similar to NS, a simple entity popularity metric. Seemingly, the inclusion of aliases does not improve the performance. Additionally, we observe that lemmatization (using the WordNet lemmatizer (Miller, 1995)) and stemming (using the Porter stemmer (Porter, 1980)) improve IScore performance by a small margin. Furthermore, we observe a slight performance gain of stemming over lemmatization. This is especially important considering the volume of the data and the inefficiency of lemmatization when compared to stemming.

E.2 Context Size

As shown in Table 7, narrowing down the context has a negative impact on the performances of both the CSE and IScore scoring methods. However, we hypothesize that the words that occur close to the entity mention are more important than those in a broader context. Therefore, we also experiment with the linear combination of the respective scores for each context size. In both cases, the optimal weights obtained through grid search optimization are $(1, 1)$. We observe a slight performance gain for the ensemble of both scoring methods.

E.3 Tie breakers

In Table 6, we present the results of the experiment with various tiebreakers. Seemingly, all the tie-breakers are a reasonable choice since no tie-breaker clearly outperforms the others.

Table 9: P@1 of different popularity metrics as tiebreakers. Rows correspond to scoring methods and columns to tiebreakers. CSE and UCSE are omitted from the table because their performance remains the same irrespective of the tiebreaker. The best P@1 in each row is highlighted **bold**.

	NS	NP	PR WP	PR WD	LQID
IScore	0.918 ± 0.036	$\mathbf{0.922 \pm 0.035}$	0.918 ± 0.036	0.918 ± 0.036	0.906 ± 0.038
EEIScore	0.898 ± 0.039	0.894 ± 0.039	$\mathbf{0.906 \pm 0.037}$	0.878 ± 0.042	0.873 ± 0.042
CSSVE	$\mathbf{0.784 \pm 0.052}$	0.780 ± 0.054	$\mathbf{0.784 \pm 0.053}$	$\mathbf{0.784 \pm 0.051}$	$\mathbf{0.784 \pm 0.052}$
UIScore	0.939 ± 0.032	0.939 ± 0.032	$\mathbf{0.942 \pm 0.031}$	0.935 ± 0.033	0.931 ± 0.033

Table 10: Estimated per-mention inference times of the selected methods. mGENRE is run on Nvidia GeForce GTX TITAN X, while UIScore and NS were executed on a single 2.5 GHz core of Intel Xeon E5-2680 processor.

Method	Inference time	
	QUOTEBANK	AIDA-CoNLL
mGENRE	8.0 s	1.9 s
NS	15 μs	26 μs
IScore	7.9 ms	67 ms
UIScore	15 ms	135 ms
Eigen	11 ms	39 ms

F Inference Time

In Table 10, we present the inference times of mGENRE, EIGENTHEMES, our best-performing methods on QUOTEBANK and AIDA-CoNLL: UIScore and IScore, respectively, and the well-performing entity popularity metric NS. EIGENTHEMES and the selected heuristics are significantly more efficient than mGENRE. The differences in inference times on Quotebank and AIDA-CoNLL are due to the setup differences (see C). Additionally, the inference times of NS, IScore, UIScore, and EIGENTHEMES largely depend on the number of candidates per mention. Thus, since on average, the number of candidate entities per mention on AIDA-CoNLL (approx. 18) is substantially larger than in QUOTEBANK (approx. 5), their inference times on AIDA-CoNLL are longer. Note that our best methods do not require GPU, making them easily parallelizable on CPU cores.

G Mean reciprocal rank of the methods

As an extension of Table 2, in Table 11 we present the MRR of the methods. MRR follows similar trends as P@1.

H Error Source Descriptions

Similar domain. If the gold entity and the system output have similar backgrounds or occupations, their Wikidata items tend to contain similar statements. For example, in one of the articles, the gold entity for Shawn Williams was Q7491485 (lacrosse player), while the output of the model was Q13064143 (American football player, defensive back). Shawn Williams first appears in the following sentence:

*Canada head coach Randy Mearns kept his No. 51 warm-up shirt - honoring Tucker Williams, the son of NLL star **Shawn Williams** of the Buffalo Bandits who is currently undergoing the treatment for Burkitt's Lymphoma - on throughout the game.*

Earlier in the article, *lacrosse* was mentioned directly, which in addition to the mention of *NLL* (National Lacrosse League) made it clear that Q7491485 is the gold entity. However, the UIScore of Q13064143 was just 1 point higher than the UIScore of Q7491485, which led to the erroneous prediction.

Key property not in Wikidata. In some cases, the Wikidata item does not contain the key information that is used to describe the entity in the article. Such cases are difficult even for humans as they require background knowledge stored in multiple sources. An example of this is John Prendergast (Q6253345), who was described in one article as the *co-founder of Enough*. This property is not listed in the Wikidata item of Q6253345 but can be found in external sources. The output of the model was Q6253343, a late British Army officer who served in World War II. The article in which Prendergast was mentioned was about violent events in Congo and was thus rich in war-related terms. Most importantly, World War II was mentioned in the article, leading to three spuriously matched words in Q6253343's Wikidata item. The final scores of Q6253345 and Q6253343 were 8 and 12

Table 11: MRR of the methods on QUOTEBANK and AIDA-CoNLL. Eigen and Eigen (IScore) have the same definition as in Table 2. The best obtained MRR in each column is highlighted **bold**.

	QUOTEBANK			AIDA-CoNLL		
	Easy	**Hard**	**Overall**	**Easy**	**Hard**	**Overall**
Random	0.622 ± 0.022	0.484 ± 0.058	0.597 ± 0.030	0.387 ± 0.013	0.205 ± 0.006	0.273 ± 0.009
LQID	0.904 ± 0.030	0.505 ± 0.094	0.836 ± 0.036	0.912 ± 0.009	0.451 ± 0.021	0.635 ± 0.013
NP	0.959 ± 0.021	0.457 ± 0.082	0.873 ± 0.034	0.901 ± 0.010	0.352 ± 0.021	0.603 ± 0.013
NS	$\mathbf{1.000 \pm 0.000}$	0.389 ± 0.044	0.895 ± 0.031	0.943 ± 0.007	0.485 ± 0.020	0.661 ± 0.012
PR_{WD}	0.873 ± 0.032	0.453 ± 0.098	0.801 ± 0.039	0.903 ± 0.009	0.336 ± 0.019	0.601 ± 0.013
PR_{WP}	0.962 ± 0.020	0.561 ± 0.101	0.893 ± 0.031	$\mathbf{0.966 \pm 0.005}$	0.491 ± 0.020	0.676 ± 0.012
IScore	0.977 ± 0.016	0.842 ± 0.096	0.954 ± 0.022	0.908 ± 0.009	0.686 ± 0.021	0.692 ± 0.013
NIScore	0.980 ± 0.016	0.750 ± 0.093	0.941 ± 0.023	0.903 ± 0.010	0.538 ± 0.024	0.651 ± 0.013
CSE	0.947 ± 0.023	0.682 ± 0.099	0.902 ± 0.029	0.871 ± 0.011	0.455 ± 0.024	0.612 ± 0.013
EEIScore	0.972 ± 0.018	0.801 ± 0.097	0.943 ± 0.023	0.555 ± 0.016	0.467 ± 0.022	0.435 ± 0.011
CSSVE	0.930 ± 0.027	0.586 ± 0.100	0.871 ± 0.033	0.796 ± 0.013	0.412 ± 0.023	0.559 ± 0.013
UIScore	0.980 ± 0.015	$\mathbf{0.891 \pm 0.080}$	0.965 ± 0.019	0.888 ± 0.010	0.718 ± 0.020	0.689 ± 0.013
UCSE	0.970 ± 0.018	0.743 ± 0.099	0.931 ± 0.025	0.874 ± 0.011	0.630 ± 0.021	0.659 ± 0.013
Eigen (IScore)	0.974 ± 0.018	0.817 ± 0.092	0.947 ± 0.024	0.864 ± 0.011	$\mathbf{0.804 \pm 0.020}^{\dagger}$	0.697 ± 0.013
Eigen	0.998 ± 0.005	0.529 ± 0.090	0.917 ± 0.027	0.910 ± 0.009	0.674 ± 0.019	0.690 ± 0.012
mGENRE	0.998 ± 0.005	0.869 ± 0.089	$\mathbf{0.976 \pm 0.017}$	0.959 ± 0.006	0.720 ± 0.022	$\mathbf{0.730 \pm 0.012}^{\dagger}$

† Indicates statistical significance ($p < 0.05$) between the best and the second-best method using bootstrapped 95% CIs.

respectively. If *co-founder of Enough* was listed in Wikidata and if *World War II* was treated as a single noun phrase, the UIScore of the gold entity, Q6253345, would beat the score of Q6253343.

Key property implicit in text. Some errors occur when enough information is provided in the article and in Wikidata, but the key properties are not mentioned in the text explicitly. For example, professional golfer Will Mackenzie (Q8002946) was mentioned in an article that was clearly about golf. However, golf was not mentioned at all in the article, yet Mackenzie's profession could be inferred from other terms related to golf, such as PGA Tour, which does not appear in the Wikidata item of Q8002946. The output of the method was Q4019878 (actor and director). Although there were other golfers mentioned in the article (leading to an EEIScore of 4 for Q8002946), its item matched no stems in text, while Q4019878 matched two stems that were completely unrelated to the article: *provid* (He was born in Providence which shares the same stem as provide) and *televis* (he was a television actor). Furthermore, Q4019878 matched citizenship, spoken language, and gender with other unambiguous mentions in the article. As a result, Q4019878 was the predicted label. This indicates the need for assigning weights to Wikidata properties to avoid irrelevant matches.

Decoy mention. To illustrate the decoy mention error source, we consider the following example:

*"Amazon will debut five new comedy drama pilots in 2014, including "The After", from **Chris Carter** ("The X-Files"); "Bosch", based on book series by Michael Conelly; "Mozart in the Jungle", from Roman Coppola ("The Darjeeling Limited"); "The Rebels" from former New York Giants football player **Michael Strahan**; and "Transparent" from Jill Soloway ("Six Feet Under")."*

Suppose that we want to disambiguate Chris Carter. Clearly, the correct entity corresponding to Chris Carter is the movie producer who created the science-fiction drama "The X-Files" (Q437267). However, the appearance of Michael Strahan increased the IScore of sportsmen named Chris Carter that played for a New York team (due to the appearance of the words "player", "New", and "York"). Note that a limitation of IScore is that it treats the words New and York separately, although they should be treated as a single noun phrase.