Contrastive Entity Coreference and Disambiguation for Historical Texts

Abhishek Arora[1†], **Emily Silcock**[1†], **Leander Heldring**[2,3], **Melissa Dell**[1,2*]
[1]Harvard University; Cambridge, MA, USA.
[2]National Bureau of Economic Research; Cambridge, MA, USA.
[3]Kellogg School of Management, Northwestern University, Evanston, IL, USA.
[†] These authors contributed equally. *Corresponding author: melissadell@fas.harvard.edu.

Abstract

Massive-scale historical document collections are crucial for social science research. Despite increasing digitization, these documents typically lack unique cross-document identifiers for individuals mentioned within the texts, as well as individual identifiers from external knowledge bases like Wikipedia/Wikidata. Existing entity disambiguation methods often fall short in accuracy for historical documents, which are replete with individuals not remembered in contemporary knowledge bases. This study makes three key contributions to improve cross-document coreference resolution and disambiguation in historical texts: a massive-scale training dataset replete with hard negatives - that sources over 190 million entity pairs from Wikipedia contexts and disambiguation pages - high-quality evaluation data from hand-labeled historical newswire articles, and trained models evaluated on this historical benchmark. We contrastively train bi-encoder models for coreferencing and disambiguating individuals in historical texts, achieving accurate, scalable performance that identifies out-of-knowledge base individuals. Our approach significantly surpasses other entity disambiguation models on our historical newswire benchmark. Our models also demonstrate competitive performance on modern entity disambiguation benchmarks, particularly on certain news disambiguation datasets.

1 Introduction

Massive scale historical document collections - such as historical newspapers or the 14 billion documents in the U.S. National Archives - are central source materials for social science research. While historical documents are increasingly being digitized, they are not typically tagged with unique cross-document identifiers for individuals mentioned in the texts, or with individual identifiers from an external knowledge base such as English Wikipedia/Wikidata, which provides structured data for over a million individuals.

While there is a large literature on entity disambiguation to a knowledge base (but to a lesser extent, entity coreference across documents in a corpus), we found that existing methods did not meet our accuracy requirements when applied to historical documents. Some widely-used methods require the entity to be in the knowledge base, whereas historical documents are replete with individuals not in Wikipedia. Indeed, one motivation for disambiguating entities in historical texts is to understand why some people are remembered and others are not. Historical texts have a different distribution of entities than modern texts; for example, there may be fewer hyperlinks in crawl corpora - a common source of training data - to these individuals' Wikipedia pages. Moreover, historical texts often have OCR noise, and language can evolve across time. To feasibly run entity disambiguation over massive-scale historical texts, the method also needs to be highly computationally efficient, as academic and archival budgets are typically highly constrained, and the amount of potential historical material to disambiguate is vast.

This study makes three central contributions designed to improve and encourage further research on cross-document coreference resolution and disambiguation of individuals in historical texts: a massive-scale training dataset replete with hard negatives, high quality evaluation data drawn from historical documents, and entity coreference and disambiguation models trained and evaluated on these data.

Our first contribution is to train coreference and disambiguation models, with a focus on historical texts, and news in particular. Our aims are: 1) accurate performance, 2) highly scalable, 3) allowing out-of-knowledge base mentions, and 4) trainable with simple recipes and limited compute.

A contrastively trained bi-encoder retrieval architecture, widely used for open-domain retrieval (*e.g.*, Karpukhin et al. (2020)), is an excellent fit

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 6174–6186
November 12-16, 2024

for these requirements. In a contrastively trained bi-encoder, the neural network encodes mentions (in the document corpus or a knowledge base) that refer to the same entity nearby in embedding space and encodes mentions referring to different entities further apart. At inference time, a given entity mention is disambiguated to the nearest entry of the knowledge base if their encodings are within some threshold similarity. If the mention encoding is sufficiently dissimilar to all entries in the knowledge base, the entity is marked as out-of-knowledge base. Analogously, an arbitrary number of entity mentions across documents in a corpus (*e.g.,*, different newspaper articles) can be coreferenced by clustering their embeddings.

This approach is highly scalable because each entity mention and entry in the knowledge base is embedded only once, no matter how many entities are to be disambiguated. Moreover, a Facebook AI Similarity Search (FAISS) (Johnson et al., 2019) backend can be used to retrieve the most similar knowledge base embedding for each mention in the document corpus. FAISS is extremely optimized, scaling up to billion-scale datasets on relatively modest hardware. This architecture also easily handles out-of-knowledge-base entities. Assuming an appropriate training dataset, it is straightforward to train, making it feasible, even on a highly constrained academic compute budget, to update the models as the deep learning literature advances or to tune them for specific document collections.

A second central contribution of the paper is the creation of a massive-scale dataset for contrastive training of entity coreference and entity disambiguation models. `WikiConfusables` constructs over 190 million entity pairs for contrastive training. Positive pairs come from contexts (paragraphs) in Wikipedia that contain hyperlinks to the same page (for coreference), or from a context and the first paragraph of the relevant entity that it links to (for disambiguation).

We obtain hard negative entity pairs - *e.g.,* pairs that are highly confusable - at scale from Wikipedia disambiguation pages, which list entities that have confusable names or aliases. The hard negative pairs come from contexts that link to different pages on a given disambiguation page. For example, the disambiguation page "John Kennedy" includes John F. Kennedy the president, John Kennedy (Louisiana politician), John F. Kennedy Jr, and a variety of other John Kennedys. Hard negatives sample contexts mentioning John F. Kennedy (*e.g.,* with hyperlinks to John F. Kennedy's page) and pair them with contexts mentioning other entities from the John Kennedy disambiguation page. We overrepresent hard negatives from within families (*e.g.*, paired mentions of Henry Ford Jr. and Henry Ford Sr.) by mining family members from Wikidata pages. These challenging cases are very common in historical texts (*e.g.,* fathers and sons with the same name and profession). Constructing our open-source `WikiConfusables` required substantial wrangling of Wikipedia dumps, and we have made it publicly available (CC-BY) to allow other researchers to more easily exploit this rich source of information.

The extensive hard negative pairs in `WikiConfusables` allow us to contrastively train our entity coreference and disambiguation models on four Nvidia A6000 GPU cards, a modest setup by deep learning standards. In contrast, contrastive training with random negative pairs requires massive batch sizes to achieve strong performance (He et al., 2020).

We train our `LinkMentions` coreference model on paired contexts around Wikipedia mention hyperlinks. We further tune this coreference model for disambiguation, creating `LinkWikipedia`, by training on paired contexts and first paragraphs. If desired, this model can be further tuned on target data. We do so for historical newspapers, creating the `LinkNewsWikipedia` model by tuning on a hand-labeled dataset linking individuals in newspapers to Wikipedia.

A third contribution is the development of a high quality benchmark that coreferences and disambiguates individuals in historical newswire articles from the 1950s and 1960s, that appear in the Newswire dataset (Silcock et al., 2024). We hand disambiguate entities - or mark them as not in the knowledge base - creating the `Entities of the Union` historical benchmark.

We document performance on `Entities of the Union` that significantly exceeds that of other widely used entity disambiguation models. We are also competitive in disambiguating individuals in modern benchmarks, especially on the MSNBC and ACE2004 benchmarks (both of which disambiguate modern news), where we outperform other state-of-the-art models. This suggests that our models and training data have broader applications beyond historical documents, especially to modern news.

Cross-document coreference is moreover highly

accurate. Cross-document coreference resolution is often central to cataloging historical document collections, and can be applied as a first step to creating a knowledge base when the individuals in a corpus are not covered in existing knowledge bases.

Finally, we briefly illustrate some of the facts that can be gleaned from tagging unique individuals in historical document collections, by applying the `LinkNewsWikipedia` model to a large-scale historical news dataset. We are enthusiastic about the promise of coreferenced and disambiguated historical documents to lead to many informative insights. Our datasets and models are open-source, with a CC-BY license, and we hope that they encourage further engagement with historical cross-document coreference resolution and disambiguation.

The rest of this study is organized as follows: Section 2 discusses the related literature, and Section 3 introduces the novel massive scale `WikiConfusables` training dataset and historical benchmark. Section 4 develops models for entity coreference and disambiguation, and Section 5 evaluates model performance. Section 6 applies our models to a massive-scale historical news corpus. Finally, Section 8 considers limitations, and Section 9 discusses ethical considerations.

2 Literature

Entity disambiguation has inspired a variety of architectures, including a masked language model (LUKE) (Yamada et al., 2022) and a neural translation model (GENRE) (De Cao et al., 2020). While these architectures work well for some problems, they are not well-suited for disambiguation of large-scale historical corpora. The masked language model approach limits to the top 50K Wikipedia entries, many of whom are not people, due to computational constraints in computing the softmax. It also does not allow out-of-knowledge base entities and requires sparse entity priors. The neural translation approach's sequence-to-sequence architecture is slow at inference time, requiring around 60 times longer to run than other models considered in our comparisons.

This paper follows the most scalable entity disambiguation approaches in employing a bi-encoder architecture. One of the inspirations for the current study is BLINK (Wu et al., 2019), which models entity disambiguation as a text retrieval problem, using a contrastively trained BERT (Devlin et al., 2018) bi-encoder and a re-ranking cross-encoder. The model assumes all entities are in the knowledge base. This study updates the bi-encoder architecture with advances made over the past five years (such as using mean pooling rather than a [CLS] token for the representation (Reimers and Gurevych, 2019) and advances in training efficiency). It also develops an expansive training dataset, replete with hard negatives. We develop a coreference model and incorporate it into the disambiguation pipeline and allow for out-of-knowledge base individuals.

Another well-known model that uses a bi-encoder is ReFinED (Ayoola et al., 2022), an entity linking model that performs mention detection and disambiguation for all mentions within a document in a single pass. Like BLINK, it uses a bi-encoder architecture but allows entities to be out-of-knowledge base. This study compares the performance of our models to GENRE, BLINK, and ReFinED, widely-used models with well-maintained codebases.

The cross-document coreference resolution literature is less dense, but a closely related study is Hsu and Horwood (2022), which uses a contrastively trained RoBERTa (Liu et al., 2019) bi-encoder and clustering for cross-document resolution of entities and events. They do not consider disambiguation to an external knowledge base.

A key distinction between this study and most existing entity disambiguation benchmarks is its focus on real-world contexts with lesser known entities - many lost to history except in the contexts of the documents being considered. Most benchmarks only contain in-knowledge base entities. An exception is Kassner et al. (2022). They detect out-of-knowledge base mentions by clustering representations, with a cluster defined as out-of-knowledge base if there are no Wikipedia embeddings in the cluster. They then add the mean embeddings of the out-of-knowledge base clusters to the knowledge base index and run entity linking with this fully comprehensive embedding index. They use BLINK as the encoder. To create a dataset with out-of-knowledge base entities, they link a large crawl corpus (OSCAR) to two Wikipedia dumps taken at time t_0 and t_1. Links to pages added between t_0 and t_1 are then out of knowledge base when disambiguating to the knowledge base in t_0. This type of out-of-knowledge base entity is different to the type we encounter in historical texts, as they are entities that were prominent enough at the time of mention to link to Wikipedia but are

not in an earlier snapshot. In contrast, in historical applications, many individuals are simply not very prominent. We do not compare on this benchmark, as all these entities were in the late-2022 Wikipedia snapshot we used for training. Hence, they were seen in training by our models (but not by some of the older comparisons) and no longer approximate truly out-of-knowledge base entities. We do not compare on the dataset introduced by Zaporojets et al. (2022) for similar reasons.

3 Datasets

3.1 Training Data

High-quality hard negatives, as well as paired positive data, are needed to train contrastive models for entity coreference and disambiguation. We create a novel, massive-scale training dataset - `WikiConfusables` - by mining entity pairs from Wikipedia disambiguation pages and Wikidata family relationships.

Entity contexts are drawn from a Wikipedia XML dump[1] from November 11, 2022, with mentions of each entity appearing as a hyperlink to their page. We split the entities into train, test, and validation sets, pairing mentions of the same entity along with their context (defined by the paragraph containing the entity mention) to create positive pairs. We create 'easy negatives' by pairing an entity mention with that of a different entity. We obtain 'hard' negatives using Wikipedia's disambiguation pages, *e.g.*, John Fitzgerald Kennedy and John Kennedy (Louisiana senator). We further enhance our training data with in-context negatives, other entities that appear in the context window of the entity under consideration. Table 1 describes the resulting dataset.

	Train	Val	Test
`Wiki Coreference`	179,069,981	5,819,525	5,132,56
`Wiki Disambiguation`	4,202,145	522,385	528,709
`NewsConfusables`	5046	666	666

Table 1: Statistics on dataset size.

We also create data for disambiguation by linking contexts with entity mentions to their associated template, forming positive pairs. To create the template, we use Wikidata names, aliases, and occupations/positions held by individuals. For example, for President Kennedy: "John F. Kennedy is of type human. Also known as Kennedy, Jack Kennedy, President Kennedy, John Fitzgerald Kennedy, J. F. Kennedy, JFK, John Kennedy, John Fitzgerald 'Jack' Kennedy, and JF Kennedy. Has worked as a politician, journalist, and statesperson." We then append this text with the first paragraph of the associated Wikipedia page.

Easy negatives are created by linking contexts with random entity templates. Similar to our coreference training, we use Wikipedia disambiguation pages to associate entity contexts with hard negative templates. We also create negative pairs using family relationships in Wikidata - *e.g.*, John F. Kennedy and Jacqueline Kennedy Onassis (who could often be referred to as "Mrs. John Kennedy"). We split the entities into an 80-10-10 train-validation-test split.

Finally, we further adapt the training domain to newspapers. An advantage of our bi-encoder architecture is that it is straightforward to tune to specific domains, plausibly helpful given historical document collections can be quite idiosyncratic.

We prepare a hand-labeled dataset, `NewsConfusables`, to tune our `LinkWikipedia` model to the historical news domain. First, we obtain names and aliases of individuals from Wikidata, then do a sparse search for them in a newspaper corpus spanning a century. We hand label whether the article refers to the anchor (*e.g.,* John F. Kennedy) or someone with the same name or alias (*e.g.,* City Councilman Jack Kennedy). When they refer to different individuals, these form hard negatives. We create extra hard negatives by matching an individual with another individual mentioned in the same context, and Wikipedia hard negatives by matching an individual with another individual mentioned in the same Wikipedia disambiguation dictionary. Easy negatives are created by matching with a random individual. Further information on this labeling process is given in the supplementary materials.

3.2 Evaluation Data

An important contribution of the study is to create high quality evaluation data for entity coreference resolution and disambiguation with historical documents. Our `Entities of the Union` benchmark labels historical, off-copyright U.S. newswire articles (Silcock et al., 2024). We double label 157 newswire articles, from 4 different days from 4 years in the 1950s and 60s, totaling 1,137 person mentions. The articles were labeled by highly moti-

[1] https://dumps.wikimedia.org/

vated North American undergraduate students[2] and all discrepancies were resolved by hand. We label days on which State of the Union addresses took place, as there are modestly more coreferences to resolve, providing more power for evaluating this task.

We split `Entities of the Union` into a 50-50 evaluation-test split, so the coreference clustering threshold can be chosen on the val split. Table 2 compares the size of the `Entities of the Union` test split to other widely-used modern benchmarks for entity disambiguation.

Benchmark	Total Mentions	Mentions in Wikipedia	People Mentions	People in Wikipedia
AIDA-CoNLL (test)	1,824	1,824	248	248
ACE2004	257	257	20	20
AQAINT	727	727	64	64
MSNBC	656	656	228	228
WNED-WIKI	6,821	6,821	625	625
WNED-CWEB	11,154	11,154	1,361	1,361
EotU (test)	569	446	569	446

Table 2: Entity and people mentions across different benchmarks.

Note that our full dataset is twice the size of the test set described here. The larger WNED datasets are generated automatically from web texts - e.g., using links from elsewhere on the web to Wikipedia - and hence are more akin to `WikiConfusables`.

Note that these widely used-benchmarks have all entities in the knowledge base. In `Entities of the Union` (validation and test splits), there are 220 unique individuals that are in Wikipedia, totaling 898 mentions. There are 239 mentions that are not in Wikipedia.

4 Methods

`LinkMentions` and `LinkWikipedia` coreference mentions across documents in a corpus and disambiguate person mentions to a knowledge base, respectively. An overview of the model architecture is shown in Figure 1.

We separate named entity recognition - which tags the tokens in a text that refer to named entities - from entity disambiguation, rather than doing them end-to-end as in entity linking. Even in noisy historical news articles, we are able to achieve 94% accuracy tagging people with named entity recognition as a token classification task. Errors tend to occur when OCR noise is so severe that there is little hope of disambiguating the entity. Hence, there is not much scope for errors to propagate. Separating named entity recognition and coreference/disambiguation simplifies the architecture, making it easier for the social science community to implement or customize to individualized applications.

We moreover focus on [PER] (person) tags from named entity recognition, as these are of primary interest for many historical applications. We found that locations could be disambiguated very well using non-neural methods and Geonames, a larger structured database of georeferenced locations.

Coreference Resolution Model: Our coreference resolution model links mentions of a given person across documents in a corpus. Depending on the question at hand, coreference resolution can produce the final output or can be used to create a prototype entity for disambiguation to an external knowledge base.

We choose a bi-encoder infrastructure, as bi-encoders are relatively straightforward for researchers with limited exposure to deep learning to customize to novel settings. This is essential for academic applications, which tend to be highly diverse and hence often require customization. They train with relatively little supervised data and fine-tuning is not very sensitive to hyperparameter selection. Training a bi-encoder is feasible on a small compute budget. Pre-trained bi-encoder models are lightweight and offer efficient inference for large datasets. Other popular architectures often have limitations on the size of the knowledge base that can be used, cannot handle out-of-knowledge base entities, or are very slow to run. Out-of-knowledge base entities and a large set of potential entities to disambiguate to are particularly common in historical applications. Finally, bi-encoders are easily implementable using the sentence transformers package, which continues to have an active user community. This will hopefully add to the longevity of a bi-encoder approach to disambiguation.

To contrastively train `LinkMentions`, using the 179,069,981 coreference training pairs in our novel `WikiConfusables` coreference training set, we employ Online Contrastive Loss as implemented in Reimers and Gurevych (2019), with cosine similarity and margin of 0.4, and utilize AdamW as the optimizer with a linear warm up scheduler set to 18.2%. Our training setup includes 4 Nvidia A6000 GPUs, a batch size of 512, and a learning rate of 1e-5. We train for a single epoch, processing each pair in the training split only once. The best model is chosen based on the pair-wise classifica-

[2]They were paid at the rate set by our department.

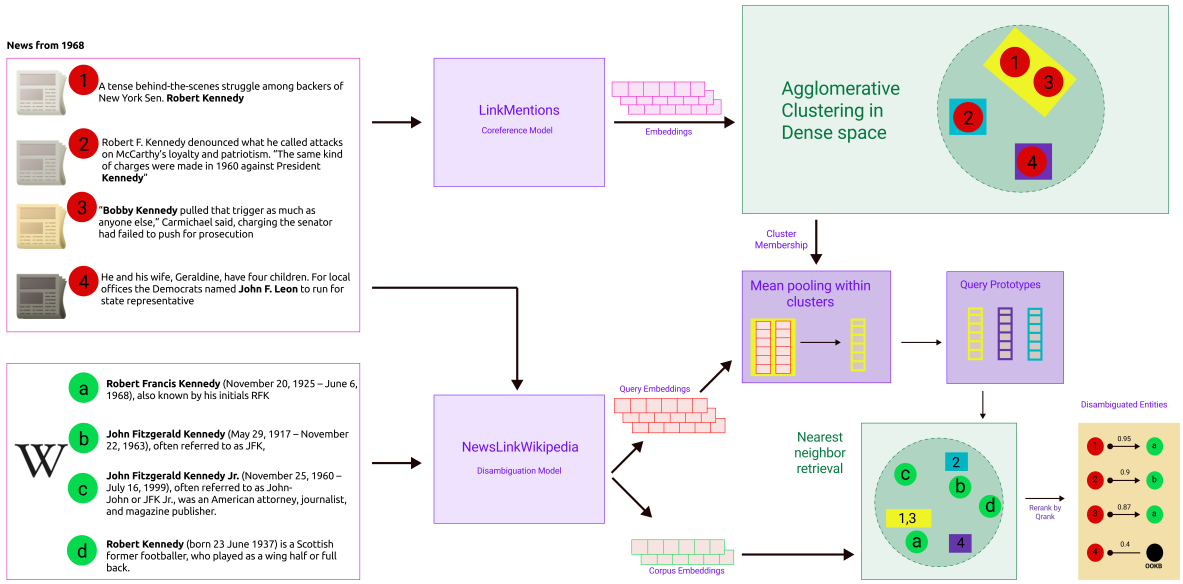

Figure 1: Entity Disambiguation Pipeline. Newspaper articles with pre-identified entities are embedded using `LinkMentions` and then clustered to group mentions of the same entity. Templates constructed from Wikidata and Wikipedia are then embedded using the `LinkNewsWikipedia` to create a lookup FAISS index. The news articles corresponding to the same cluster are embedded using `LinkNewsWikipedia` and mean-pooled to create a prototype embedding to query the lookup index. The entity of the nearest Wikipedia template to the query is assigned to each article in the entity cluster.

tion F1 score on the validation set, with the highest validation F1 being 92.75%.

We use a sequence length of 256, and initialize with an *all-mpnet-base-v2* Sentence-BERT model (Reimers and Gurevych, 2019) from the Hugging Face hub. This is a lightweight model, making it more feasible for those in the academic community with limited resources to train and deploy it at scale. The model is trained in Pytorch (Paszke et al., 2019) with hyperparameters tuned using hyperband implemented in Weights and Biases (Weights&Biases, 2023). Given the large training dataset, we found it beneficial to divide it into 10 chunks before training. After completing each chunk (1/10 of an epoch), we resumed training from an intermediate checkpoint and lowered the learning rate to 2e-6 after the first chunk to minimize the risk of the optimizer overshooting the minima. Since training each chunk began with a warm up, our approach effectively simulates a linear scheduler with restarts.

We observed significant performance improvement by using special tokens ($[M]$ Entity $[\backslash M]$) around an entity mention (Wu et al., 2019). For instance, "Eisenhower sharing a light moment with President-elect $[M]$ John F. Kennedy $[\backslash M]$ during their meeting in the Oval Office."

At inference time, the contrastively trained `LinkMentions` is used to embed the mentions, and then they are grouped together via hierarchical agglomerative clustering (with average linkage) using cosine similarity. A threshold of 0.175 was chosen on the validation set. (Other clustering methods are straightforward to swap in, as desired.) In the newspaper corpus, we coreference entities across articles within dates.

Disambiguation: For disambiguation, we finetune our coreference model on the disambiguation portion of `WikiConfusables`, with similar hyperparameters to those used in coreference training, but without restarts or chunking. The learning rate is 2e-6, with a 20% warm up, and batch size is 256. The model is trained for three epochs, and the best checkpoint is selected based on the classification F1, achieving a maximum validation F1 of 97%.

This model can be further tuned for specific applications. We create `LinkNewsWikipedia` by tuning on the paired disambiguation data in `NewsConfusables`. We use an identical training setup, achieving a maximum validation F1 of 85%.

Next, we prepare a lookup corpus to disambiguate entity mentions to the correct entity using semantic information from both the context around the mention and information from a template we create from Wikipedia and Wikidata pages

of individuals, as described above. We prune our knowledge base to remove extraneous entities. We include only entities of instance type human, who have a birth or death date, as we found most that did not were instance type errors. We remove pages of individuals born after the conclusion of the corpus and remove entities with no overlap and a high edit distance between the Wikidata label and the associated Wikipedia page title. We found that those pages overwhelmingly were not a person page for that Wikidata entry.

The resulting knowledge base has 1.12 million person pages. We embed these templates using our disambiguation model and store them in a FAISS IndexFlatIP index (Johnson et al., 2019).

To run disambiguation, we embed mentions using the disambiguation model. Using the clusters obtained from coreferencing, we mean pool within each cluster to create the entity prototype embeddings, and use these to query the nearest neighbor(s) in the knowledge base. If there is no embedding in the knowledge base within a threshold cosine similarity of the query - where this threshold is chosen on a validation set - we mark the entity as not in the knowledge base.

To choose the no-match threshold, we annotate the output of our disambiguation pipeline on a set of 6,425 pairs sampled from 13 years in a large-scale newswire dataset (Silcock et al., 2024). We then find the cut-off threshold that maximizes pairwise classification precision and use it as the no-match threshold. We chose a threshold of 0.14986.

If the returned matches are very close to each other, we find there are modest gains from disambiguating to the most popular near entity. If the nearest neighbor is within the no-match threshold distance for a match, and the second-nearest neighbor is at least 0.01 cosine distance from the nearest neighbor, we disambiguate to the nearest neighbor. Otherwise, we use Qrank[3], which ranks Wikidata entities by aggregating page views on Wikipedia, Wikispecies, Wikibooks, Wikiquote, and other Wikimedia projects. We keep all entities that are within 0.01 cosine distance of the nearest neighbor to the query, and choose the one with the highest Qrank.

We do not add a re-ranking step with a cross-encoder, as in Wu et al. (2019), in order to maximize scaleability to massive datasets. However, training a cross-encoder on `WikiConfusables` would be straightforward.

For reproducibility and ease of access, we have made our models and training/evaluation data available on the Hugging Face hub (links are redacted to maintain anonymity for review). All code is in our GitHub repository.

5 Evaluation

To measure performance on coreferencing and disambiguating individuals in historical documents, we apply our models and others from the literature to `Entities of the Union`. We made significant efforts to run existing models on our evaluation data, but not all models in the literature have maintained their codebases, or even made code available. Moreover, some models are simply not suitable for the task. For example, LUKE (Yamada et al., 2019) limits to top 50K Wikipedia entities (many of whom are not people), meaning relatively few entities in our datasets are in the knowledge base. We run BLINK (Wu et al., 2019), GENRE (De Cao et al., 2020), and ReFinED (Ayoola et al., 2022), all prominent models in the entity disambiguation literature with well-maintained codebases. We closely follow their implementations, providing details in the supplementary materials. The latter two models have a zero-shot version and an AIDA-CoNLL fine-tuned version. We report results for both.

These are disambiguation, not coreference, models. The coreference literature is thinner and disambiguation is our main focus, and so we do not have comparisons for this task. Coreferencing individuals across newswire articles is very accurate, achieving an adjusted rand index (ARI) of 96.42.

Table 3 documents that on the `Entities of the Union` dataset, both our zero-shot `LinkWikipedia` model and fine-tuned `LinkNewsWikipedia` model beat other models by a wide margin. `LinkNewsWikipedia` correctly retrieves or classifies as out-of-knowledge base 78% of individual mentions, whereas `LinkWikipedia` has an accuracy of 74%. The next best alternative is ReFinED, which correctly disambiguates around 65% of mentions. When only considering entities in Wikipedia, `LinkNewsWikipedia` correctly disambiguates 89% of entities, `LinkWikipedia` correctly disambiguates 85% of entities, and the next best alternative is GENRE, with an accuracy of 81%. Hence, while much of the advantage is with out-of-knowledge base entities, our models

[3] https://github.com/brawer/wikidata-qrank/tree/main

Benchmark	LinkWikipedia	LinkNewsWikipedia	BLINK	GENRE BLINK Data	GENRE AIDA-CoNLL	ReFinED Base	ReFinED AIDA-CoNLL
EotU (all)	74.0	78.3	59.9	63.4	62.4	65.4	64.0
EotU (in KB)	85.2	89.0	76.5	80.9	79.6	60.3	59.9
AIDA-CoNLL (test set)	70.6	71.7	79.0	58.1	62.1	99.2	99.2
ACE2004	90.0	90.0	85.0	80.0	80.0	80.0	80.0
AQAINT	92.2	95.3	98.4	95.3	95.3	93.8	93.8
MSNBC	89.4	98.2	81.6	82.5	84.2	84.2	82.9
WNED-WIKI	88.9	88.5	93.9	94.2	93.4	95.5	94.7
WNED-CWEB	70.7	71.5	69.1	70.7	71.5	73.3	72.4

Table 3: Benchmark performance comparison across different methods. The first row evaluates on all entities in Entities of the Union, whereas the second row only considers in-knowledge base entities.

also do better on in-knowledge base entities, even zero-shot.

We also compare performance on disambiguating people in existing, widely-used benchmarks. While modern data are not our main focus, our models do reasonably well. In particular, LinkNewsWikipedia achieves a near-perfect 98% accuracy on MSNBC. LinkWikipedia has 89% accuracy, as compared to the next best (GENRE) with 84% accuracy. This suggests that our models - beyond being suited to historical applications - can also be well-suited to disambiguating modern news. We also beat other models on the news dataset ACE2004, which has very few people. On other modern benchmarks, there are model(s) that perform better, but our performance is in the range of the other models.

Model	**Base MPNet**	**Our Model Only Disambig.**	**Add Coref.**	**Add Birth Date Filter**	**Add Qrank Rerank**
LinkWikipedia	26.5	62.6	73.8	73.8	74.0
NewsLinkWikipedia	26.5	69.1	77.9	77.9	78.3

Table 4: Ablations. The first column uses a Sentence-BERT MPNet model, which we use to initialize training. The second column discards the coreferencing step, as well as birthdate filtering and Qrank re-ranking. The next three columns add back coreference resolution, birthdate filtering, and Qrank re-ranking, respectively.

We conduct ablations in Table 4, to quantify the contributions of different elements in our disambiguation pipeline. Base MPNet disambiguates with a Sentence-BERT (Reimers and Gurevych, 2019) MPNet (Song et al., 2020) model (*all-mpnet-base-v2*), the base model that we initialize with. This model is not intended for entity disambiguation, but we include it to quantify how much is gained through our training. Performance is very poor. The next column reports results from our trained disambiguation models, without coreference resolution or additional processing steps: filtering entities in the knowledge base to be born prior to the end date of the corpus and re-ranking by Wikipedia QRank when the nearest entities are very close to each other. Accuracy falls relative to the baseline by around ten percentage points in both models. Adding coreference resolution restores almost all of this decline, with birthdate filtering and Qrank re-ranking contributing little. Coreference resolution plausibly combines information across mentions and reduces noise, leading to overall better quality disambiguation.

6 Exploring Entities in Historical News

To give a flavor of how our framework can be combined with historical documents, we apply our disambiguation pipeline to a large-scale corpus of historical newswire articles - sent out over newswires such as the Associated Press between 1878 and 1977 (Silcock et al., 2024). The dataset contains 2.7 million unique articles, reproduced in a corpus of local news over 32 million times.

We disambiguate 15,323,463 person mentions, encompassing 61,933 unique individuals. Only 4.6% of disambiguated entity mentions refer to women, with Golda Meir being the most mentioned woman. The most mentioned entity is Dwight D. Eisenhower, appearing in 9,530 unique articles which are reproduced an average of 33.7 times. Richard Nixon, Harry S. Truman, and Adolf Hitler are the next most mentioned in unique newswire articles. Figure 2 plots their mention counts. For U.S. presidents, electoral cycles are clearly visible.

Entity disambiguation also allows us to see individuals' occupations in Wikidata. The most common occupations of disambiguated entities are politician, military officer and lawyer.

This dataset, also publicly available, provides fascinating data that researchers can use to study who appeared in historical news and which of these

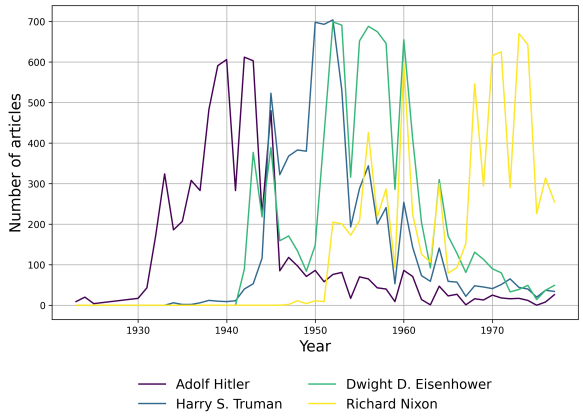

Figure 2: Mentions over time of entities that appeared most commonly in newswire articles.

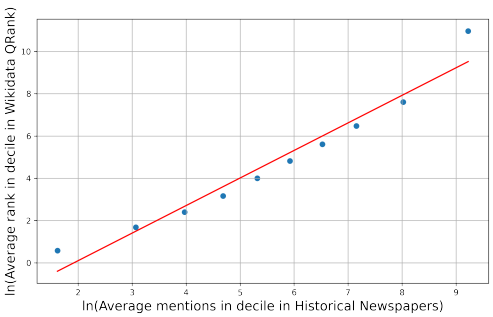

Figure 3: Mentions against Wikipedia Qrank.

individuals are remembered via Wikipedia today.

To examine whether we are able to detect less prominent entities as well as prominent entities, we plot log mentions by decile in the large scale newswire dataset (Silcock et al., 2024) - a measure of how prominent individuals were in the news historically - against log Wikidata Qrank by decile, showing the correlation between these two measures.

We find a virtually linear relationship, showing that we do not underdetect less prominent entities, or overdetect prominent entities.

7 Conclusion

We provide new data for training and evaluating entity coreference and disambiguation. We propose bi-encoder models trained on this data, which achieve high accuracy for disambiguating and coreferencing entities in historical documents, as well as on existing entity disambiguation benchmarks. Our models are able to handle out-of-knowledge base individuals.

Our data provide high-quality resources for other researchers developing methods for entity coreference and disambiguation, while our models provide an off-the-shelf solution for social scientists wanting to explore who is mentioned in historical texts.

8 Limitations

The present paper focuses on individuals and disambiguation to Wikipedia/Wikidata. We found disambiguating locations to Wikipedia to be unproductive, as we could achieve very strong performance using sparse methods and Geonames, a more comprehensive database of locations. In the future, we hope to extend the model to organizations (though in practice, many historical organizations end up being out-of-knowledge base).

While Wikipedia is extensive, there are many people who never entered this knowledge base, and various biases may influence which historical figures are remembered in Wikipedia. Part of the objective of applying `LinkNewsWikipedia` to a massive scale corpus of historical newswires is to understand more about which individuals were considered broadly newsworthy at the time but have since been forgotten. This information could be used to expand databases such as Wikipedia in the future. Nevertheless, we cannot disambiguate an individual who is not in the knowledge base, no matter how noteworthy they were historically.

9 Ethical Considerations

This study presents no major ethical concerns. Its methods are entirely open source, and its training data are entirely in the public domain. We disambiguate individuals in widely reproduced historical newspaper articles, which are in the public domain and hence do not pose privacy concerns.

It is possible that some applications could raise concerns. Historical news, government publications, and other historical documents reflect the biases of their time and may contain factual inaccuracies or offensive content. Moreover, while our models are reasonably accurate, they are not perfect, and depending on the usage of the output, human revision of the match - potentially bringing in additional information - may be required. It is important to interpret the disambiguated texts critically, as is the norm in rigorous historical research.

Acknowledgements

We received excellent research assistance from Arpit Bhate, Vania Cheung, William Cox, Dennis Du, Eyad Elsafoury, Connor Fogal, Jude Ha,

Zachary Lee, Alice Liu, Catherine Liu, Shiloh Liu, Andrew Lu, Omer Mujawar, Sethu Odayappan, Domenick Regina, Cristopher Rosas, and Ryan Xia. Harvard Data Science Initiative and Microsoft Azure provided compute credits to support article digitization.

Supplemental Materials

10 Model hyperparameters

10.1 Entity coreference

At training time, we split the large training set into 10 chunks and we use the following hyperparameters:

- Online contrastive loss, with cosine distance metric and a margin of 0.4
- Batch size: 256
- Max sequence length: 256
- Epochs: 1
- Learning rate 1e-5, increased to 2e-06 after first chunk
- Optimiser: AdamW
- Warm-up rate: 0.182, increased to 1 after the first chunk

At inference time, we cluster entities using Hierarchical Agglomerative clustering, with average linking and cosine distance metric, with a threshold of 0.175, which was validated on the validation split.

10.2 Entity disambiguation

At training time we use the following hyperparameters:

- Online contrastive loss, with cosine distance metric and a margin of 0.4
- Batch size: 256
- Max sequence length: 256
- Epochs: 3
- Learning rate 2e-6
- Optimiser: AdamW
- Warm-up rate: 0.2

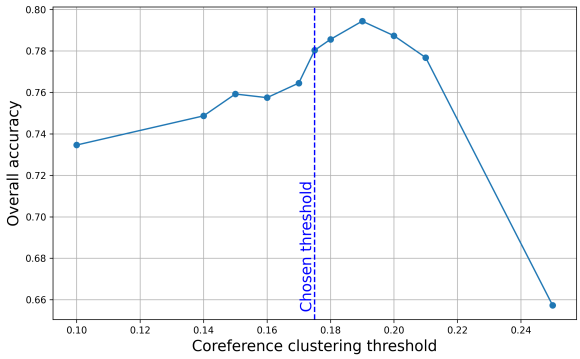

Figure 4: Sensitivity of disambiguation results to choice of coreference clustering threshold.

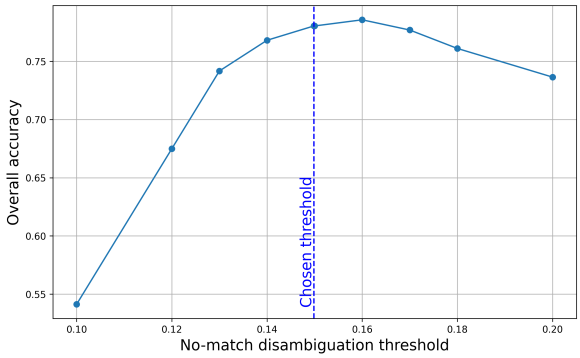

Figure 5: Sensitivity of disambiguation results to choice of no-match threshold.

10.3 Sensitivity to inference hyperparameters

The two main hyperparameters at inference time are the clustering threshold used in the coreference step, and the no-match threshold used in the disambiguation step. In figure 4 in the main text, we demonstrate that our results are not particularly sensitive to other choices.

Here we examine sensitivity of our results to choices of clustering threshold and no-match threshold, with our fine-tuned disambiguation model.

Figure 4 graphs the disambiguation accuracy on the test set of `Entities of the Union` for different choices of clustering threshold. The results are quite flat around the threshold that we chose, based on the validation set.

Figure 4 shows sensitivity of these same results to choice of no-match threshold. The results are not especially sensitive. The threshold can be changed ± 2 percentage points with little change to the results. In fact, we can see here that the threshold that we chose was not in fact the optimal threshold, but this makes minimal difference to our overall results.

11 Comparison to other entity disambiguation models

We compare our method against existing entity disambiguation architectures. Results from this are given in the text. Details of the implementations of other architectures follow.

11.1 BLINK

To run BLINK (Wu et al., 2019), we use the implementation from their github repository directly (https://github.com/facebookresearch/BLINK). Specifically we use `main_dense.run()` with default hyperparameters.

11.2 GENRE

For GENRE (De Cao et al., 2020), we used their huggingface implementation (https://huggingface.co/facebook/genre-linking-blink). We used 10 beams, with truncation around the entities to a max token length of 384 as suggested in the appendix of the original paper. Otherwise default hyperparameters were used. We evaluate using both their base model (`facebook/genre-linking-blink`) and their model finetuned on Aida-YAGO.

11.3 ReFinED

To evaluate ReFinED, we use entity disambiguation implementation from their github repository (https://github.com/amazon-science/ReFinED). We use `refined.process_text()` with default hyperparameters. For each mention, we specify that the mention type is PER using `coarse_mention_type = "PERSON"`.

12 Annotator instructions

12.1 `NewsConfusables` annotator instructions

`NewsConfusables` was labeled by The articles were labeled by paid North American undergraduate students. Annotators were given the following instructions:

- "The box at the top shows the entity in question, the years in which they held some kind of important position, and their aliases (see more detail at the end of this email).
- It also shows a code starting with Q. This is the unique reference for the person in Wikidata. You can query search this there (https://www.wikidata.org/) and pull up more detail about the person, including information on positions they've held, what else they've done etc. Most of these people also appear on Wikipedia, so this is also a good source for finding out more info about the person.
- Then there are (up to) 32 passages of text. In each of these there is a highlighted term. Your job is to label each of these passages for whether the highlighted term is the same as the person in the box at the top ('positive') or not ('negative'). In some cases this will be pretty clear, and in some cases it might need a bit of digging.
- In some cases we don't have 32 passages, so you'll see some turn up as "Empty" at the end - no need to label these "

12.2 `Entities of the Union` annotator instructions

To create `Entities of the Union`, 1,137 entity mentions across 157 newswire articles were double-annotated by undergraduate research assistants. Annotator labeling instructions were as follows:

"We've pulled out lots of articles from the day of the State of the Union speech in 1958, 1959, 1960 and 1961. A team of RAs over this semester has been labeling the spans in these texts that refer to an entity. What we want you to focus on is working out which unique entity these spans refer to. The database of unique identifiers that we will work from is Wikidata (https://www.wikidata.org/). If you search for entities on here, you will see they have a unique identifier beginning with Q (eg. Melissa's is Q58009782). For each entity in the articles we'd like you to find the unique identifier in wikidata (if it exists). It might also be useful to use wikipedia in difficult cases - all pages on wikipedia will have an entry in wikidata, but wikidata is bigger than wikipedia, so there might be some entities in wikidata that you don't find in wikipedia [...] there's 14 entities to label on average per article. The entity in question should be highlighted in the text."

Annotators were encouraged to reach out with questions and clarifications.

References

Tom Ayoola, Shubhi Tyagi, Joseph Fisher, Christos Christodoulopoulos, and Andrea Pierleoni. 2022. ReFinED: An efficient zero-shot-capable approach to end-to-end entity linking. In *NAACL*.

Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2020. Autoregressive entity retrieval. *arXiv preprint arXiv:2010.00904*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738.

Benjamin Hsu and Graham Horwood. 2022. Contrastive representation learning for cross-document coreference resolution of events and entities. *arXiv preprint arXiv:2205.11438*.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*.

Nora Kassner, Fabio Petroni, Mikhail Plekhanov, Sebastian Riedel, and Nicola Cancedda. 2022. Edin: An end-to-end benchmark and pipeline for unknown entity discovery and indexing. *arXiv preprint arXiv:2205.12570*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Emily Silcock, Abhishek Arora, Luca D'Amico-Wong, and Melissa Dell. 2024. Newswire: A large-scale structured database of a century of historical news. *arXiv preprint arXiv:2406.09490*.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. *Advances in Neural Information Processing Systems*, 33:16857–16867.

Weights&Biases. 2023. Tune hyperparameters.

Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. 2019. Scalable zero-shot entity linking with dense entity retrieval. *arXiv preprint arXiv:1911.03814*.

Ikuya Yamada, Koki Washio, Hiroyuki Shindo, and Yuji Matsumoto. 2019. Global entity disambiguation with pretrained contextualized embeddings of words and entities. *arXiv preprint arXiv:1909.00426*.

Ikuya Yamada, Koki Washio, Hiroyuki Shindo, and Yuji Matsumoto. 2022. Global entity disambiguation with bert. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3264–3271.

Klim Zaporojets, Lucie-Aimée Kaffee, Johannes Deleu, Thomas Demeester, Chris Develder, and Isabelle Augenstein. 2022. Tempel: Linking dynamically evolving and newly emerging entities. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.