Benchmarking Scalable Methods for Streaming Cross Document Entity Coreference

Robert L. Logan IV[*♢] **Andrew McCallum**[♠♡]
Sameer Singh[♢] **Daniel Bikel**[♠]
[♢]University of California, Irvine [♠]Google Research
[♡]University of Massachusetts, Amherst
{rlogan, sameer}@uci.edu
{mccallum, dbikel}@google.com

Abstract

Streaming cross document entity coreference (CDC) systems disambiguate mentions of named entities in a scalable manner via incremental clustering. Unlike other approaches for named entity disambiguation (e.g., entity linking), streaming CDC allows for the disambiguation of entities that are unknown at inference time. Thus, it is well-suited for processing streams of data where new entities are frequently introduced. Despite these benefits, this task is currently difficult to study, as existing approaches are either evaluated on datasets that are no longer available, or omit other crucial details needed to ensure fair comparison. In this work, we address this issue by compiling a large benchmark adapted from existing free datasets, and performing a comprehensive evaluation of a number of novel and existing baseline models.[1] We investigate: how to best encode mentions, which clustering algorithms are most effective for grouping mentions, how models transfer to different domains, and how bounding the number of mentions tracked during inference impacts performance. Our results show that the relative performance of neural and feature-based mention encoders varies across different domains, and in most cases the best performance is achieved using a combination of both approaches. We also find that performance is minimally impacted by limiting the number of tracked mentions.

1 Introduction

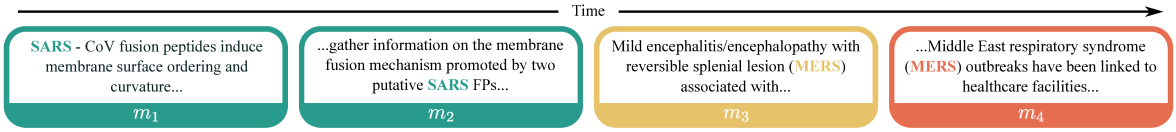
The ability to disambiguate mentions of named entities in text is a central task in the field of information extraction, and is crucial to topic tracking, knowledge base induction and question answering. Recent work on this problem has focused almost solely on entity linking–based approaches, i.e., models that link mentions to a fixed set of known entities. While significant strides have been made on this front—with systems that can be trained end-to-end (Kolitsas et al., 2018), on millions of entities (Ling et al., 2020), and link to entities using only their textual descriptions (Logeswaran et al., 2019)—all entity linking systems suffer from the significant limitation that they are restricted to linking to a curated list of entities that is fixed at inference time. Thus they are of limited use when processing data streams where new entities regularly appear, such as research publications, social media feeds, and news articles.

In contrast, the alternative approach of cross-document entity coreference (CDC) (Bagga and Baldwin, 1998; Gooi and Allan, 2004; Singh et al., 2011; Dutta and Weikum, 2015), which disambiguates mentions via clustering, does not suffer from this shortcoming. Instead most CDC algorithms suffer from a different failure mode: lack of scalability. Since they run expensive clustering routines over the entire set of mentions, they are not well suited to applications where mentions arrive one at a time. There are, however, a subset of *streaming* CDC methods that avoid this issue by clustering mentions incrementally (Figure 1). Unfortunately, despite such methods' apparent fitness for streaming data scenarios, this area of research has received little attention from the NLP community. To our knowledge there are only two existing works on the task (Rao et al., 2010; Shrimpton et al., 2015), and only the latter evaluates truly streaming systems, i.e., systems that process new mentions in constant time with constant memory.

One crucial factor limiting research on this topic is a lack of free, publicly accessible benchmark datasets; datasets used in existing works are either small and impossible to reproduce (e.g., the dataset collected by Shrimpton et al. (2015) only contains a few hundred unique entities, and many of the

[*]Work done during an internship at Google Research.
[1]Code and data available at: https://github.com/rloganiv/streaming-cdc

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pages 4717–4731
August 1–6, 2021.



(a) New mentions arrive over time.

(b) Mentions are encoded as points in a vector space and incrementally clustered. As the space grows some points are removed to ensure that the amount of memory used does not exceed a given threshold.

Figure 1: **Streaming Cross-Document Coreference**.

annotated tweets are no longer available for download) or lack the necessary canonical ordering and are expensive to procure (e.g., the ACE 2008 and TAC-KBP 2009 corpora used by Rao et al. (2010)). To remedy this, we compile a benchmark of three datasets for evaluating English streaming CDC systems along with a canonical ordering in which evaluation data should be processed. These datasets are derived from existing datasets that cover diverse subject matter: biomedical texts (Mohan and Li, 2019), news articles (Hoffart et al., 2011), and Wikia fandoms (Logeswaran et al., 2019).

We evaluate a number of novel and existing streaming CDC systems on this benchmark. Our systems utilize a two step approach where: 1) each mention is encoded using a neural or feature-based model, and 2) the mention is then clustered with existing mentions using an incremental clustering algorithm. We investigate the performance of different mention encoders (existing feature-based methods, pretrained LMs, and encoders from entity linkers such as RELIC (Ling et al., 2020) and BLINK (Wu et al., 2020)), and incremental clustering algorithms (greedy nearest-neighbors clustering, and a recently introduced online agglomerative clustering algorithm, GRINCH (Monath et al., 2019)). Since GRINCH does not use bounded memory, which is required for scalability in the streaming setting, we introduce a novel bounded memory variant that prunes nodes from the cluster tree when the number of leaves exceeds a given size, and compare its performance to existing bounded memory approaches.

Our results show that the relative performance of different mention encoders and clustering algorithms varies across different domains. We find that existing approaches for streaming CDC (e.g., feature-based mention encoding with greedy nearest-neighbors clustering) outperform neural approaches on two of three datasets (+1-3% abs. improvement in CoNLL F_1), while a RELIC-based encoder with GRINCH performs better on the last dataset (+9% abs. improvement in CoNLL F_1). In cases where existing approaches perform well, we also find that better performance can be obtained by using a combination of neural and feature-based mention encoders. Lastly, we observe that by using relatively simple memory management policies, e.g. removing old and redundant mentions from the mention cache, bounded memory models can achieve performance near on-par with unbounded models while storing only a fraction of the mentions (in one case we observe a 2% abs. drop in CoNLL F_1 caching only 10% of the mentions).

2 Streaming Cross-Document Entity Coreference (CDC)

2.1 Task Overview

The key goal of cross-document entity coreference (CDC) is to identify mentions that refer to the same entity. Formally, let $\mathcal{M} = \{m_1, \ldots, m_{|\mathcal{M}|}\}$ denote a corpus of mentions, where each mention consists of a surface text $m.\text{surface}$ (e.g., the colored text in Figure 1a), as well as its surrounding context $m.\text{context}$ (e.g., the text in black). Provided $\mathcal{M}$ as an input, a CDC system produces a disjoint clustering over the mentions $C = \{\mathcal{C}_1, \ldots, \mathcal{C}_{|C|}\}$, $|C| \leq |\mathcal{M}|$, as the output, where each cluster $\mathcal{C}_e = \{m \in \mathcal{M} \mid m.\text{entity} = e\}$ is the set of mentions that refer to the same entity.

In streaming CDC, there are two additional requirements: 1) mentions arrive in a fixed order

($\mathcal{M}$ is a list) and are clustered incrementally, and 2) memory is constrained so that only a fixed number of mentions can be stored. This can be formulated in terms of the above notation by adding a time index t, so that $\mathcal{M}_T = \{m_t \in \mathcal{M} \mid t \leq T\}$ is the set of all mentions observed at or before time T, $\widetilde{\mathcal{M}}_T \subseteq \mathcal{M}_T$ is a subset of "active" mentions whose size does not exceed a fixed memory bound k, e.g., $|\widetilde{\mathcal{M}}_T| \leq k$, and C_T is comprised of clusters that only contain mentions in $\widetilde{\mathcal{M}}_T$. Due to the streaming nature, $\widetilde{\mathcal{M}}_T - \{m_T\} \subset \widetilde{\mathcal{M}}_{T-1}$, i.e., a mention cannot be added back to $\widetilde{\mathcal{M}}_T$ if it was previously removed. When the memory bound is reached, mention are removed from $\widetilde{\mathcal{M}}$ according to a memory management policy Φ.

An illustrative example is provided in Figure 1. Mentions arrive in left-to-right order (Figure 1a), with the clustering process depicted in Figure 1b (memory bound $k = 3$). At time $T = 4$, the mention m_1 is removed from $\widetilde{\mathcal{M}}_4$. Note that, even though m_1 is removed, it is still possible to disambiguate mentions of all previously observed entities, whereas this would not be possible had m_3 or m_4 been removed. This illustrates the effect the memory management policy can have on performance.

2.2 Background and Motivation

Cross Document Entity Coreference As we show later, we employ a two-stage CDC pipeline where mentions are first encoded as vectors, and subsequently clustered. This approach is used in most existing work on CDC (Bagga and Baldwin, 1998; Mann and Yarowsky, 2003; Gooi and Allan, 2004). In the past decade, research on CDC has mainly focused in improving scalability (Singh et al., 2011), and jointly learning to perform CDC with other tasks such as entity linking (Dutta and Weikum, 2015) and event coreference (discussed in the next paragraph). This work similarly investigates whether entity linking is beneficial for CDC, however we use entity linkers that are pretrained separately and kept fixed during inference.

Recently, there has been a renewed interest in performing CDC jointly with cross-document event coreference (Barhom et al., 2019; Meged et al., 2020; Cattan et al., 2020; Caciularu et al., 2021) on the ECB+ dataset (Cybulska and Vossen, 2014). Although we do not evaluate methods from this line of research in this work, we hope that the benchmark we compile will be useful for future evaluation of these systems.

Streaming Cross Document Coreference The methods mentioned in the previous paragraphs disambiguate mentions all at once, and are thus unsuitable for applications where a large number of mentions appear over time. Rao et al. (2010) propose to address this issue using an incremental clustering approach where each new mention is either placed into one of a number of candidate clusters, or a new cluster if similarity does not exceed a given threshold (Allaway et al. (2021) use a similar approach for joint entity and event coreference). Shrimpton et al. (2015) note that the this incremental clustering does not process mentions in constant time/memory, and thus is not "truly streaming". They present the only truly streaming approach for CDC by introducing a number of memory management policies that limit the size of $\widetilde{\mathcal{M}}$, which we describe in more detail in Section 3.3.

One of the key problems inhibiting further research on streaming CDC is a lack of suitable evaluation datasets for measuring system performance. The datasets used in Rao et al. (2010) are either small in size (few hundreds of mentions), contain few annotated entities, or are expensive to procure. Additionally, they do not include any canonical ordering of the mentions, which precludes consistent evaluation of streaming systems. Meanwhile, Tweets annotated by Shrimpton et al. (2015) only cover two surface texts (Roger and Jessica) and are no longer accessible via the Twitter API.[2] To address this we collect a new evaluation benchmark, comprised of 3 existing publicly available datasets, covering a diverse collection of topics (News, Biomedical Articles, Wikias) with natural orderings (e.g., chronological, categorical). This benchmark is described in detail in Section 4.1.

Entity Linking CDC is similar to the task of entity linking (EL, Mihalcea and Csomai (2007)), which also addresses the problem of named entity disambiguation, with the key distinction that EL is formulated as a supervised classification problem (list of entities is known at training and test time), while CDC is an unsupervised clustering problem. In particular, CDC is similar to *time-aware* EL (Agarwal et al., 2018)—where temporal context is used to help disambiguate mentions—and *zero-shot* EL (Zeshel, Logeswaran et al. (2019))—where the set of entities linked to during evaluation does not overlap with the set of entities observed during training. Streaming CDC can also be con-

[2] At this time, only 56 of the first 100 tweets were available.

sidered a method for time/order-aware zero-shot named entity disambiguation, however, it is strictly more challenging as it does not assume access to a curated list of entities at prediction time, or any supervised training data.

Although CDC is formulated as a strictly unsupervised clustering task, this does not preclude the usage of labeled data for transfer learning. One of the primary goals in this work is to investigate whether the mention encoders learned by entity linking systems provide useful representations in the first step of the CDC pipeline. Specifically, we consider mention encoders for two state-of-the-art entity linking architectures: RELIC (Ling et al., 2020) and the BLINK bi-encoder (Wu et al., 2020).

Emerging Entity Detection Streaming CDC is also related to the task of *emerging entity detection* (EED, Färber et al. (2016)), which, given a mention that cannot be linked, seeks to predict whether it should produce a new KB entry. Although both tasks share similar motivations, they adopt different approaches (EED is formulated as a binary classification task), and CDC does not require deciding which entities should and should not be added to a knowledge base. However, in many practical applications, it may make sense to apply streaming CDC only to emerging entities.

3 Building Streaming CDC Systems

Following previous work, we adopt a two-step approach to performing streaming cross-document coreference. In the first step, an encoder is used to produce a vector representation of the incoming mention $\boldsymbol{m}_t = \text{Enc}(m_t)$. In the second step, these vectors are input into an incremental clustering algorithm to update the predicted clustering $C_t = \text{Clust}(C_{t-1}, \boldsymbol{m}_t)$. In the following sections we describe in detail the mention encoders and clustering algorithms used in this work.

3.1 Mention Encoders

The primary goal of mention encoders $\text{Enc}(m_t)$ is to produce a compact representation of the mention, including both the surface and the context text.

Feature-Based Encoders Existing models for streaming cross-document coreference exclusively make use of feature-based mention encoders. While there are many feature engineering options explored in the literature, in this work we consider the mention encoding approach proposed by Shrimpton et al. (2015), which uses character skip bigram indicator vectors to encode the surface text, and tf-idf vectors to represent contexts. When using this encoding scheme, similarity scores are computed independently for the surface and context embeddings, and a weighted average is taken to produce the final similarity score. We use the same setup and parameters as Shrimpton et al. (2015).

Masked Language Model Encoders We also consider mention encodings produced by masked language models, particularly BERT (Devlin et al., 2019). We encode the mention by feeding the contiguous text of the mention (containing both the surrounding and surface text) into BERT and concatenating the contextualized vectors associated with the first and last word-piece of the surface text. That is, let $s, e \in \mathbb{N}$ denote the start and end of the mention surface text within the complete mention, and let $\boldsymbol{M} = \text{BERT}(m)$ denote the contextualized word vectors output by BERT. Then the mention encoding is given by: $\text{Enc}_{\text{MLM}}(m) = [\boldsymbol{M}[s]; \boldsymbol{M}[e]]$.

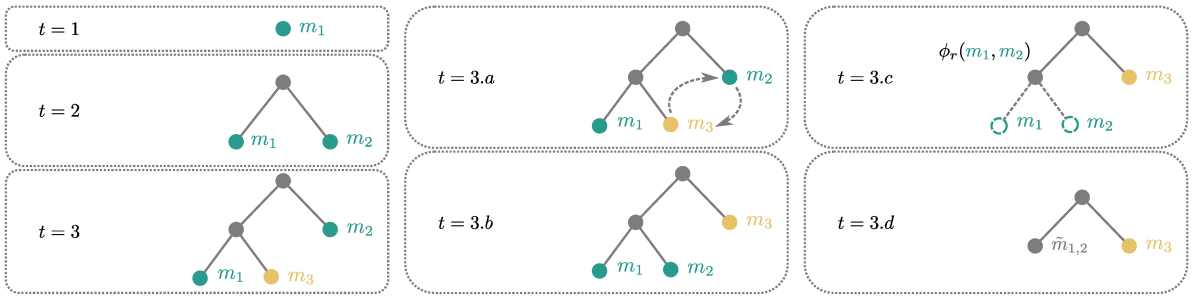
Entity Linker-Based Encoders We consider producing mention encodings using the bi-encoder-based neural entity linkers: RELIC (Ling et al., 2020) and BLINK (Wu et al., 2020). The bi-encoder architecture is comprised of two components—a mention encoder Enc_m, and an entity encoder Enc_e—and is trained to maximize a similarity score (e.g., dot-product) between the mention encoding and the encoding of its underlying entity, while simultaneously minimizing the score for other entities. We use Enc_m from pre-trained entity linkers to encode mentions for CDC.

Hybrid Encoder We also consider a *hybrid* encoder which combines feature-based and neural mention encoders. We retain the feature-based character skip bigram surface text encoder, but use one of the neural encoders from entity linkers in place of tf-idf context representation. Similarity scores are computed by averaging the two without any weights, unlike by Shrimpton et al. (2015).

3.2 Clustering Algorithms

Here we describe incremental clustering approaches, $\text{Clust}(C_{t-1}, \boldsymbol{m}_t)$, that compute a new clustering when m_t is added to the mentions under consideration ($\widetilde{\mathcal{M}}$).

Greedy Nearest Neighbors Clustering Shrimpton et al. (2015) and Rao et al. (2010) both evaluate



(a) Greedy Agglomerative Clustering (b) GRINCH (c) Bounded GRINCH

Figure 2: **Bounded GRINCH**. (a) Greedy agglomerative clustering produces a sub-optimal tree structure due to the order points are received. (b) The GRINCH (Monath et al., 2019) algorithm recovers from this mistake by reconfiguring the tree structure (in this case using a *rotation* operation). (c) To ensure the memory used by GRINCH remains bounded, we add a new operation—*remove*—that prunes two leaf nodes when the number of leaves exceeds a given size. Nodes are selected for removal using a scoring function ϕ_r. In this case nodes m_1 and m_2 are selected, and their parent $\tilde{m}_{1,2}$ becomes a new leaf.

CDC using a single linkage incremental clustering approach that clusters each new mention m to its nearest neighbor $m' = \arg\min_{m' \in \widetilde{\mathcal{M}}} \text{sim}(m, m')$, if the similarity exceeds some threshold τ. We use a similar approach here, however we cluster m with *all* $m' \in \widetilde{\mathcal{M}}$ such that $\text{sim}(m, m') > \tau$ thus allowing previously separate clusters to be merged if m is similar to both of them.

GRINCH Gooi and Allan (2004) find that average link hierarchical agglomerative clustering can outperform greedy single link approaches. However, agglomerative approaches are typically not used for streaming CDC because running the algorithm at each time step is too expensive, and incremental variants of the approach are not able to recover from incorrect choices made early on (Figure 2a). The recently introduced GRINCH clustering algorithm (Monath et al., 2019) uses `rotate` and `graft` operations that reconfigure the tree, thereby avoiding these issues (Figure 2b). We defer to the original paper for details, however note that, for our application, each interior node of the cluster tree is computed as a weighted average of its children's representations (where the weights are proportional to the number of leaves). Thus at each interior node, it is possible to compute the similarity score between that node's children. This allows us to produce a *flat* clustering from the cluster tree by thresholding the similarity score, just as in the greedy clustering case.

3.3 Memory Management Policies

As described in Section 2.1, memory management policies decide which mentions to remove from $\widetilde{\mathcal{M}}$ to prevent its size from exceeding the memory bound, providing scalable, *memory-bound* variants of the clustering algorithms.

Bounded Memory Greedy NN Clustering For bounded memory greedy nearest neighbors clustering, we consider the following memory management policies of Shrimpton et al. (2015):

- *Window*: Remove the oldest mention in $\widetilde{\mathcal{M}}$.
- *Cache*: Remove the oldest mention in the least recently updated cluster $\mathcal{C}_{\text{LRU}}$.
- *Diversity*: Remove the most similar mention to mention just added, i.e. $\arg\max_m \text{sim}(m, m_t)$
- *Diversity-Cache*: A combination of the diversity and cache strategies, where the diversity strategy is used if the similarity score exceeds a given threshold $\text{sim}(m, m_t) > \alpha$, and the cache strategy is used otherwise.

Bounded Memory GRINCH Memory management for GRINCH is more complicated than for greedy clustering, as instead of maintaining a flat clustering of mentions, GRINCH instead maintains a cluster hierarchy in the form of a binary cluster tree. Every time a mention is inserted into the tree, two new nodes are created: one node for the mention itself, and a new parent node linking the mention to its sibling (Figure 2a). Accordingly, when the memory bound is reached, the memory management policy for GRINCH must remove two nodes from the tree. Furthermore, in order to preserve the tree's binary structure, the removed nodes must be leaf nodes as well as siblings. Because the original GRINCH algorithm only includes routines for inserting nodes into the tree, and reconfiguring the tree's structure, we modify GRINCH to

	$\lvert\mathcal{M}\rvert$	$\lvert\mathcal{E}\rvert$	% Seen	MAE
AIDA				
Train	18.5K	4.1K	100%	1.1K
Dev	4.8K	1.6K	23%	290
Test	4.5K	1.6K	16%	263
MedMentions				
Train	121K	18K	100%	4.7K
Dev	42K	8.8K	27%	1.8K
Test	39K	8.3K	26%	1.7K
Zeshel				
Train	81K	32K	100%	9.3K
Dev	18K	7.5K	0%	2.9K
Test	17K	7.2K	0%	3.3K

Table 1: **Dataset Statistics**. $\lvert\mathcal{M}\rvert$: #mentions, $\lvert\mathcal{E}\rvert$: #unique entities, % Seen: fraction of entities observed during training, MAE: maximum active entities, e.g., the number of mentions an ideal streaming CDC system would need to store to perfectly cluster the data.

include a new *remove* operation that prunes two nodes satisfying the these criteria. The parent of these nodes then becomes a leaf node, whose vector representation is produced by combining the vector representations of its former children using a weighted average (this is conceptually similar to the *collapse* operation described in Kobren et al. (2017)). We consider the following policies here:

- *Window*: Remove the nodes whose parent was least recently added to the tree.
- *Diversity*: Remove the pair of nodes that are most similar to each other.

4 Benchmarking Streaming CDC

In this section, we describe our proposed benchmark for evaluating streaming CDC systems.

4.1 Datasets

Current research on CDC is inhibited by a lack of large, publicly accessible datasets. We address this by compiling datasets for streaming CDC by adapting existing entity linking datasets: AIDA CoNLL-YAGO, MedMentions, and Zeshel.

AIDA AIDA CoNLL-YAGO (Hoffart et al., 2011) contains news articles from the Reuters Corpus written between August and December 1996 with annotations linking mentions to YAGO and Wikipedia. We create a canonical ordering for this dataset by ordering articles by date. As the original train, dev, and test splits respect this ordering, we use the original splits in our benchmark.

MedMentions The MedMentions (Mohan and Li, 2019) corpus contains abstracts for biomedical articles published to PubMed in 2016, and annotated with links to the UMLS medical ontology. We order abstracts by publication date[3] to create a canonical ordering. Since the original dataset is not ordered by date, we create new train, dev, and test splits of comparable size that respect this ordering.

Zeshel The Zeshel (Logeswaran et al., 2019) dataset consists of Wikia articles for different FANDOMs. In addition to the original set of annotated mentions, we use the provided entity descriptions as an additional source of mentions. We impose an ordering that groups all mentions belonging to the same Wikia together, and otherwise retains their original order in the Zeshel data. This is an interesting scenario for streaming CDC as no clusters need be retained when transitioning to a new Wikia.

Analysis Statistics for the benchmark data are provided in Table 1, which list the number of mentions and unique entities for each dataset. We also list the percentage overlap between entities in the training set, and entities in the dev and test sets (% Seen), as well as the *maximum active entities* (MAE). MAE is a quantity introduced by Toshniwal et al. (2020), which measures the maximum number of "active entities" (e.g., entities that have been previously mentioned, and will be mentioned in the future) for a given dataset, which can alternatively be interpreted as the smallest possible memory bound that can be used in order to ensure that a CDC system can cluster each mention with at least one other mention of the same entity. Importantly, this number is a small fraction of the total number of mentions in each dataset, indicating that these datasets are appropriate for the streaming setting and to compare memory management policies.

4.2 Evaluation Metrics

We evaluate CDC performance using the standard evaluation metrics: MUC (Vilain et al., 1995), B^3 (Bagga and Baldwin, 1998), CEAFe (Luo, 2005), and CoNLL F_1 which is an average of the previous three. In order to perform evaluation when memory is bounded, we perform the following bookkeeping to track nodes which have been removed by the memory management policy. For *bounded memory greedy NN clustering*, we keep track of the removed node's predicted cluster (e.g., if the node was removed from cluster $\mathcal{C}$, then it is considered an element of $\mathcal{C}$ during evaluation).

[3] 6 abstracts were omitted due to missing metadata

This is similar to the evaluation used by Toshniwal et al. (2020). For *bounded memory GRINCH*, we keep track of the removed node's place within the tree structure, and produce a flat clustering using the thresholding approach described in Section 3.2 as if the node were never removed. Because leaf nodes (and accordingly removed nodes) are never updated by insertion or removal operations, nodes belonging to the same cluster before they are pruned they will always remain in the same cluster during evaluation, which is the same assumption used for the greedy NN evaluation.

4.3 Hyperparameters

Vocabulary and inverse document frequency (idf) weights are estimated using each dataset's train set. For masked language model encoders, we use an unmodified BERT-base architecture, with model weights provided by the HuggingFace transformers library (Wolf et al., 2020). For BLINK, we use the released BERT-large bi-encoder weights.[4] Our bounded memory variant of GRINCH is based on the official implementation.[5] Note that GRINCH does not currently support sparse inputs, so we do not include results for feature-based mention encoders. RELIC model weights are initialized from BERT-base, and then finetuned to perform entity linking in the following settings:

- *RELIC (wiki)*: Trained on the same Wikipedia data used to train the BLINK bi-encoder.
- *RELIC (in-domain)*: Trained on respective benchmark's training dataset; a separate model is trained for each benchmark.

Training is performed using hyperparameters suggested by Ling et al. (2020).[6] For each benchmark, the hybrid mention encoder uses the best performing RELIC variant on that benchmark. Cluster thresholds τ are chosen so that the number of predicted clusters on the dev dataset approximately matches the number of unique entities.

5 Results

In this section, we provide a comprehensive evaluation of the design choices that define the existing and proposed approaches for streaming CDC.

[4] https://github.com/facebookresearch/BLINK

[5] https://github.com/iesl/grinch

[6] Trained on a server w/ 754 GB RAM, Intel Xeon Gold 5218 CPU and 4x NVIDIA Quadro RTX 8000 GPUs.

Choice of Encoder We include the results for CDC systems with unbounded memory on the benchmark datasets in Table 2, as well as results for two baselines: 1) a system that clusters together all mentions with the same surface forms (**exact match**), and 2) a system that only considers gold within-document clusters and does not merge clusters across documents (**oracle within-doc**). We observe that, in general, neural mention encoders are not sufficient to obtain good CDC performance. With the exception of the RELIC (In-Domain) on MedMentions, no neural mention encoders are able to outperform the feature-based greedy NN approach, and furthermore, the MLM and BLINK mention encoders do not even surpass the exact match baseline. However, note that for AIDA and Zeshel, best results are obtained using a hybrid mention encoder. Thus, in these domains, we can conclude that while neural mention encoders are useful for encoding contexts, CDC systems require an additional system to model surface texts to achieve good performance. The results on MedMentions provide an interesting contrast to this conclusion. Here the RELIC (In-Domain) mention encoder outperforms both the feature-based and hybrid mention encoders. In the error analysis below, we find that this is due mainly to improved performance clustering mentions of entities seen when training the mention encoder.

Choice of Clustering Algorithm Comparing greedy nearest neighbors clustering to GRINCH, we do not observe a consistent trend across mention encoders or datasets. While the best performance on AIDA and Zeshel is achieved using greedy nearest neighbor clustering, the best performance on MedMentions is achieved using GRINCH. These results highlight the importance of benchmarking CDC systems on a number of different datasets; patterns observed on a single dataset do not extrapolate well to other settings. It is also interesting to observe that a much simpler approach often works better than the more complex one.

Error Analysis We characterize the errors of these models by investigating: a) the entities whose mentions are *conflated* (e.g., are wrongly clustered together) and *split* (e.g., wrongly grouped into separate clusters) using the approach of Kummerfeld and Klein (2013), and b) differences in performance on entities that are *seen* vs. *unseen* during training for models that use in-domain data. A sub-

	AIDA				MedMentions				Zeshel			
	MUC	B^3	CEAF	Avg.	MUC	B^3	CEAF	Avg.	MUC	B^3	CEAF	Avg.
Exact Match	90.2	84.1	81.0	85.1	78.8	66.0	54.0	66.3	28.3	64.3	46.4	46.3
Oracle Within-Doc	15.2	46.8	47.1	36.4	16.5	32.8	34.8	28.0	-	-	-	-
Greedy NN												
Feature-Based	94.2	89.0	87.3	90.2	83.6	67.0	58.0	69.5	39.6	60.9	52.6	51.0
MLM	75.9	71.1	58.1	68.4	70.8	52.1	42.0	55.0	16.2	53.8	44.8	38.3
BLINK (Wiki)	58.2	56.3	56.6	57.0	59.4	39.2	43.2	47.3	36.6	36.9	40.9	41.6
RELIC (Wiki)	92.4	89.4	83.6	88.5	73.2	56.1	42.4	57.2	36.2	58.1	48.7	47.7
RELIC (In-Domain)	93.2	80.7	84.5	86.1	**86.8**	69.5	**62.4**	72.9	28.2	61.4	42.5	44.0
Hybrid	**94.7**	**90.1**	**88.5**	**91.1**	85.6	70.5	59.9	72.0	44.0	**64.5**	**53.3**	**54.0**
GRINCH												
MLM	37.8	59.2	41.5	46.2	70.8	52.1	42.0	55.0	49.0	38.0	33.1	40.0
BLINK (Wiki)	64.3	26.9	23.2	38.1	83.2	17.1	11.9	37.4	45.6	24.8	21.7	30.7
RELIC (Wiki)	91.6	88.3	82.5	87.5	73.9	57.9	42.2	58.0	**72.6**	4.2	4.3	27.0
RELIC (In-Domain)	82.8	84.0	69.5	78.8	85.4	**73.3**	61.8	**73.5**	27.3	57.5	40.1	41.6

Table 2: **Unbounded Memory Results**. CoNLL F_1 scores for each valid combination of clustering algorithm and mention encoder. Similiarity threshold clustering + hybrid mention encoder works best on AIDA and Zeshel, whereas GRINCH clustering + in-domain RELIC mention encoder works best for MedMentions.

Feature-Based + Greedy NN	
FIFA World Cup	. . . ' ' Japan , co-hosts of the *World Cup* in 2002 and ranked 20th in the world by . . .
1995 Rugby World Cup	. . . team . Cuttitta announced his retirement after the *1995 World Cup* , where he took issue with being dropped from . . .
Rugby World Cup	. . . Australia to defeat with a last-ditch drop-goal in the *World Cup* quarter-final in Cape Town . . .
RELIC (Wiki) + Greedy NN	
FC Volendam	. . . leaders PSV Eindhoven romped to a 6-0 win over *Volendam* on Saturday . Their other marksmen were Brazilian . . .
Feyenoord	. . . game . They boast a nine-point lead over *Feyenoord* , who have two games in hand , and . . .
RKC Waalwijk	. . . division soccer match played on Friday : *RKC Waalwijk* 1 (Starbuck 76) Willem II . . .

Table 3: **Most Conflated Entities on AIDA.** Left: Unique entity ID. Right: Mention with entity surface form in italics. Results for remaining models are provided in the Appendix.

set of our results is provided in Table 3, with full results available in Tables 4–11 in the Appendix.

In aggregate, these error metrics closely track the results in Table 2, where better models make fewer errors of all types. We do, however, observe that in-domain training improves RELIC's performance considerably on MedMentions (+15 CoNLL F_1 on seen entities, and +18 on unseen entities), and is the primary reason underlying the improved performance over feature-based encoders (72.6 vs. 60.7 CoNLL F_1 on seen entities, while performance on unseen entities is comparable).

Comparing mentions of the most conflated entities provides a qualitative sense of the failure modes of each method. We note that the feature-based method tends to fail at distinguishing entities with the same surface form, e.g., world cups of different sports, while neural entity linkers tend to conflate entities with similar contexts, particularly when surface forms are split into multiple word pieces in the model's vocabulary (each surface form in the bottom of Table 3 gets broken into 3+ word pieces).

Effect of Bounded Memory Results for the bounded memory setting are illustrated in Figure 3. In these experiments we take the best neural mention encoder for each benchmark dataset (RELIC (Wiki) for AIDA and Zeshel, and RELIC (In-Domain) for MedMentions), and plot the CoNLL F_1 score for each of the memory management policies described in Section 3.3. We measure performance for memory bounds at the maximum number of active entities (MAE) and total unique entities ($|\mathcal{E}|$) for each dataset (as well as 1/2x, and 2x multiples of these numbers). In sum, these results provide strong evidence that CDC systems can reliably cluster mentions in a truly streaming setting, even when memory is bounded to a small fraction of the number of entities encountered by the system. Most impressively, using the diversity-

cache memory management policy, a greedy nearest neighbors bounded memory model achieves a CoNLL F_1 score within 2% of the best performing unbounded memory model, while only storing approximately 10% (i.e., $\mathcal{E}/2$) of the mentions.

We notice a few fairly consistent trends across datasets. The first is that increasing the memory bound has diminishing returns; while there is a large benefit incurred by increasing the bound from MAE/2 to MAE, the difference in performance attained from increasing the bound from $\mathcal{E}$ to $2\mathcal{E}$ is often negligible. We also find that naïve memory management policies that store recent mentions (i.e., window, W, and cache, C) tend to perform better than the policies that attempt to remove redundant mentions (i.e., diversity, D). This effect is particularly pronounced for small memory bounds. While this is somewhat surprising—storing mentions of the same entity is particularly harmful when memory is limited, so encouraging diversity should be a good thing—one possible explanation is that the diversity policy is actually removing mentions of entities that appear within the same context, as we saw earlier that neural mention encoders appear to focus more on mention context than surface text. Lastly, regarding the comparison of greedy nearest neighbors clustering to GRINCH we again see that inconsistency in performance across datasets; GRINCH appears to perform better at larger cache sizes for AIDA and MedMentions, while greedy nearest neighbors clustering has much better performance than GRINCH on Zeshel.

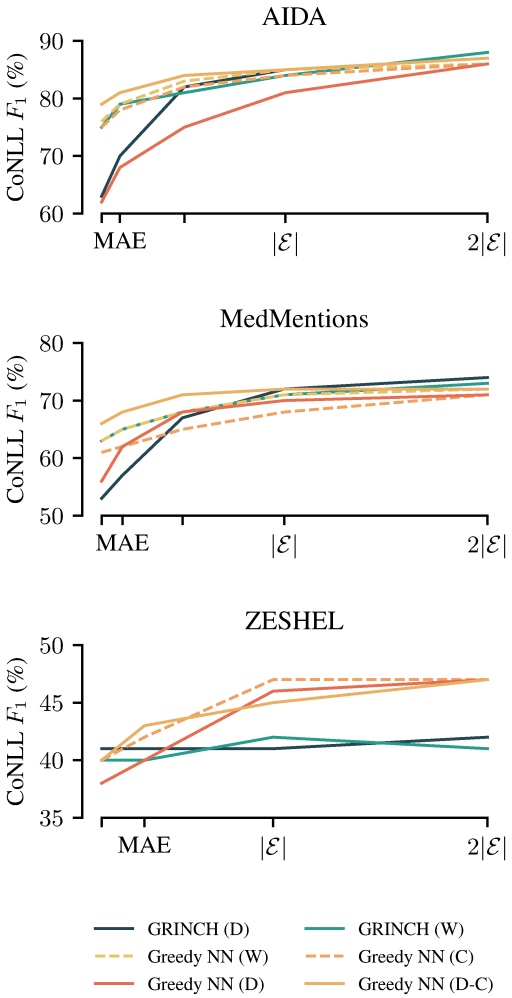


Figure 3: **Effect of Bounded Memory**. CoNLL F_1 scores as we vary the bound. MAE: Max. active entities (defined in Section 4.1). $|\mathcal{E}|$: #unique entities.

6 Conclusion and Future Work

Streaming cross document coreference has a number of compelling applications, especially concerning processing streams of data such as research publications, social media feeds, and news articles where new entities are frequently introduced. Despite being well-motivated, this task has received little attention from the NLP community. In order to foster a more welcoming environment for research on this task, we compile a diverse benchmark dataset for evaluating CDC, comprised of existing datasets that are free and publicly available. We additionally evaluate the performance of a collection of existing approaches for CDC, as well as introduce new approaches that leverage modern neural architectures. Our results highlight a number of challenges for future CDC research, such as how to better incorporate surface level features into neural mention encoders, as well as alternative policies for memory management that improve upon the naïve baselines studied in this work. Benchmark data and materials needed to reproduce our results are provided at: https://github.com/rloganiv/streaming-cdc.

Acknowledgements

The authors would like to thank Sanjay Subramanian, Nitish Gupta, Keith Hall, Ryan McDonald, Livio Baldini Soares, Nicholas FitzGerald, and Tom Kwiatkowski for their technical guidance and helpful comments while conducting this work. We would also like to thank thank the anonymous ACL reviewers for their valuable feedback. This project is supported in part by NSF award no. 1817183, and the DARPA MCS program under Contract No. N660011924033.

Broader Impact Statement

This paper focuses on systems that perform entity disambiguation without reliance on an external knowledge base. The potential benefit of such systems is an improved ability to track mentions of rare and emergent entities (e.g., natural disasters, novel disease variants, etc.); however, this is also relevant in digital surveillance settings, and may result in reduced privacy.

References

Prabal Agarwal, Jannik Strötgen, Luciano del Corro, Johannes Hoffart, and Gerhard Weikum. 2018. diaNED: Time-aware named entity disambiguation for diachronic corpora. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 686–693, Melbourne, Australia. Association for Computational Linguistics.

Emily Allaway, Shuai Wang, and Miguel Ballesteros. 2021. Sequential cross-document coreference resolution. *arXiv preprint arXiv:2104.08413*.

Amit Bagga and Breck Baldwin. 1998. Entity-based cross-document coreferencing using the vector space model. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 79–85, Montreal, Quebec, Canada. Association for Computational Linguistics.

Shany Barhom, Vered Shwartz, Alon Eirew, Michael Bugert, Nils Reimers, and Ido Dagan. 2019. Revisiting joint modeling of cross-document entity and event coreference resolution. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4179–4189, Florence, Italy. Association for Computational Linguistics.

Avi Caciularu, Arman Cohan, Iz Beltagy, Matthew E Peters, Arie Cattan, and Ido Dagan. 2021. Cross-document language modeling. *arXiv preprint arXiv:2101.00406*.

Arie Cattan, Alon Eirew, Gabriel Stanovsky, Mandar Joshi, and Ido Dagan. 2020. Streamlining cross-document coreference resolution: Evaluation and modeling. *arXiv preprint arXiv:2009.11032*.

Agata Cybulska and Piek Vossen. 2014. Using a sledgehammer to crack a nut? lexical diversity and event coreference resolution. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4545–4552, Reykjavik, Iceland. European Language Resources Association (ELRA).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Sourav Dutta and Gerhard Weikum. 2015. Cross-document co-reference resolution using sample-based clustering with knowledge enrichment. *Transactions of the Association for Computational Linguistics*, 3:15–28.

Michael Färber, Achim Rettinger, and Boulos El Asmar. 2016. On emerging entity detection. In *Knowledge Engineering and Knowledge Management*, pages 223–238, Cham. Springer International Publishing.

Chung Heong Gooi and James Allan. 2004. Cross-document coreference on a large scale corpus. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 9–16, Boston, Massachusetts, USA. Association for Computational Linguistics.

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*, pages 782–792, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Ari Kobren, Nicholas Monath, Akshay Krishnamurthy, and Andrew McCallum. 2017. A hierarchical algorithm for extreme clustering. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 255–264.

Nikolaos Kolitsas, Octavian-Eugen Ganea, and Thomas Hofmann. 2018. End-to-end neural entity linking. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 519–529, Brussels, Belgium. Association for Computational Linguistics.

Jonathan K. Kummerfeld and Dan Klein. 2013. Error-driven analysis of challenges in coreference resolution. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 265–277, Seattle, Washington, USA. Association for Computational Linguistics.

Jeffrey Ling, Nicholas FitzGerald, Zifei Shan, Livio Baldini Soares, Thibault Févry, David Weiss, and Tom Kwiatkowski. 2020. Learning cross-context entity representations from text. *arXiv preprint arXiv:2001.03765*.

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*, pages 3449–3460, Florence, Italy. Association for Computational Linguistics.

Xiaoqiang Luo. 2005. On coreference resolution performance metrics. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 25–32, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Gideon Mann and David Yarowsky. 2003. Unsupervised personal name disambiguation. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 33–40.

Yehudit Meged, Avi Caciularu, Vered Shwartz, and Ido Dagan. 2020. Paraphrasing vs coreferring: Two sides of the same coin. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4897–4907, Online. Association for Computational Linguistics.

Rada Mihalcea and Andras Csomai. 2007. Wikify! linking documents to encyclopedic knowledge. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 233–242.

Sunil Mohan and Donghui Li. 2019. Medmentions: A large biomedical corpus annotated with {umls} concepts. In *Automated Knowledge Base Construction (AKBC)*.

Nicholas Monath, Ari Kobren, Akshay Krishnamurthy, Michael R. Glass, and Andrew McCallum. 2019. Scalable hierarchical clustering with tree grafting. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, page 1438–1448, New York, NY, USA. Association for Computing Machinery.

Delip Rao, Paul McNamee, and Mark Dredze. 2010. Streaming cross document entity coreference resolution. In *Coling 2010: Posters*, pages 1050–1058, Beijing, China. Coling 2010 Organizing Committee.

Luke Shrimpton, Victor Lavrenko, and Miles Osborne. 2015. Sampling techniques for streaming cross document coreference resolution. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1391–1396, Denver, Colorado. Association for Computational Linguistics.

Sameer Singh, Amarnag Subramanya, Fernando Pereira, and Andrew McCallum. 2011. Large-scale cross-document coreference using distributed inference and hierarchical models. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 793–803, Portland, Oregon, USA. Association for Computational Linguistics.

Shubham Toshniwal, Sam Wiseman, Allyson Ettinger, Karen Livescu, and Kevin Gimpel. 2020. Learning to Ignore: Long Document Coreference with Bounded Memory Neural Networks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8519–8526, Online. Association for Computational Linguistics.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. In *Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

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 (EMNLP)*, pages 6397–6407, Online. Association for Computational Linguistics.

A Error Analysis

A.1 Seen vs. Unseen Performance

We evaluate CoNLL F_1 scores for mentions of entities that are seen vs. unseen in the AIDA and MedMentions training datasets (Zeshel is excluded since no test entities are seen in the training data). Results are provided in Table 4, with performance of models that are trained using the in-domain training datasets reported in bold.

	AIDA		MedMentions	
	Seen	Unseen	Seen	Unseen
Greedy NN				
Feature-Based	91.2	92.2	60.7	80.8
MLM	67.2	71.7	54.5	61.5
BLINK (Wiki)	39.0	39.2	45.1	59.0
RELIC (Wiki)	89.3	89.9	57.6	62.2
RELIC (In-Dom.)	**87.1**	**86.4**	**72.6**	**80.3**
Hybrid	92.8	92.2	**71.8**	**80.5**
GRINCH				
MLM	46.2	43.0	30.6	24.9
BLINK (Wiki)	39.0	39.2	38.1	33.7
RELIC (Wiki)	88.2	89.1	58.4	62.8
RELIC (In-Dom.)	**84.2**	**70.3**	**73.1**	**78.1**

Table 4: **CoNLL F_1 Scores** on mentions of entities that are seen vs. unseen in the AIDA and MedMentions training datasets. Zeshel is excluded since all entities in the test data are unseen. Bolded numbers indicate that the mention encoder is trained on seen mentions.

A.2 Clustering Mistakes

Kummerfeld and Klein (2013) define a system for categorizing coreference errors into a number of underlying error types. Because gold mention boundaries are provided in our task setup, the main error types of relevance are *divided entities*, i.e., mentions of the same entity that occur in different clusters, and *conflated entities*, i.e., mentions of different entities that are grouped into the same clusters. We can quantify these error types by counting the number of times clusters need to be merged together vs. split, respectively. The overall error counts are provided in Table 5.

In addition to providing the overall error counts, we also render a sample of mentions from predicted clusters containing the most conflated entities in Tables 6–11.

	AIDA		MedMent.		Zeshel	
	Confl.	Div.	Confl.	Div.	Confl.	Div.
Greedy NN						
Feature-Based	173	156	5.0K	5.2K	5.4K	6.2K
MLM	764	676	8.9K	9.1K	6.0K	8.6K
BLINK (Wiki)	1.6K	749	13.1K	12.3K	7.4K	5.9K
RELIC (Wiki)	243	209	8.1K	8.4K	5.8K	6.5K
RELIC (In-Dom.)	178	219	4.1K	4.1K	3.5K	7.8K
Hybrid	155	156	4.4K	4.5K	4.5K	5.9K
GRINCH						
MLM	1.2K	829	8.4K	0	6.9K	5.2K
BLINK (Wiki)	1.7K	749	7.9K	3.2K	8.0K	2.8K
RELIC (Wiki)	284	214	7.8K	8.2K	6.8K	6.6K
RELIC (In-Dom.)	101	794	3.3K	5.4K	2.9K	8.7K

Table 5: **Clustering Mistakes**. *Conflated*: Number of times mentions of different entities are grouped into the same cluster. *Divided*: Number of times mentions of the same entity are grouped into different clusters.

	Feature-Based
FIS Ski Jumping World Cup	. . .) 228.1 (129.4 / 98.7) Leading *World Cup* standings (after three events) : 1. . . .
FIFA World Cup	. . . his squad to face Macedonia next week in a *World Cup* qualifier . Midfielder Valentin Stefan and striker Viorel . . .
1966 FIFA World Cup	. . . 35 caps and was a key member of the *1966 World Cup* winning team with his younger brother , Bobby . . . .
	MLM
Sheffield Wednesday FŎ02eCŎ02e	. . . . 0-1 . 19,306 Liverpool 0 *Sheffield Wednesday* 1 (Whittingham 22) . 0-1 . . . .
Aston Villa FŎ02eCŎ02e	. . . 0 Leeds 0 . 30,018 Southampton 0 *Aston Villa* 1 (Townsend 34) . 0-1 . . . .
Newcastle United FŎ02eCŎ02e	. . . Villa 17 9 3 5 22 15 30 *Newcastle* 15 9 2 4 26 17 29 Manchester . . .
	BLINK (Wiki)
Japan national football team	. . . SOCCER - *JAPAN* GET LUCKY WIN , CHINA IN SURPRISE DEFEAT . . . .
China PR national football team	. . . SOCCER - JAPAN GET LUCKY WIN , *CHINA* IN SURPRISE DEFEAT . Nadim Ladki AL-AIN . . .
Al Ain	. . . CHINA IN SURPRISE DEFEAT . Nadim Ladki *AL-AIN* , United Arab Emirates 1996-12-06 Japan began the . . .
	RELIC (Wiki)
RKC Waalwijk	. . . division soccer match played on Friday : *RKC Waalwijk* 1 (Starbuck 76) Willem II Tilburg 2 . . .
Willem II (football club)	. . . : RKC Waalwijk 1 (Starbuck 76) *Willem II Tilburg* 2 (Konterman 45 , Van der Vegt . . .
PSV Eindhoven	. . . SOCCER - *PSV* HIT VOLENDAM FOR SIX . AMSTERDAM 1996-12-07 . . .
	RELIC (In-Domain)
Japan national football team	. . . SOCCER - *JAPAN* GET LUCKY WIN , CHINA IN SURPRISE DEFEAT . . . .
China PR national football team	. . . SOCCER - JAPAN GET LUCKY WIN , *CHINA* IN SURPRISE DEFEAT . Nadim Ladki AL-AIN . . .
United Nations	. . . 1975 and annexed the following year . The *United Nations* has never recognised Jakarta 's move . Alatas . . .
	Hybrid
FIFA World Cup	. . . ' ' Japan , co-hosts of the *World Cup* in 2002 and ranked 20th in the world by . . .
1966 FIFA World Cup	. . . 35 caps and was a key member of the *1966 World Cup* winning team with his younger brother , Bobby . . . .
Rugby World Cup	. . . Australia to defeat with a last-ditch drop-goal in the *World Cup* quarter-final in Cape Town . " Campo has . . .

Table 6: **Most Conflated Entities on AIDA using Greedy NN Clustering.** Left: Unique entity ID. Right: Mention with entity surface form in italics.

	MLM
Japan national football team	. . . SOCCER - *JAPAN* GET LUCKY WIN , CHINA IN SURPRISE DEFEAT . . . .
China PR national football team	. . . SOCCER - JAPAN GET LUCKY WIN , *CHINA* IN SURPRISE DEFEAT . Nadim Ladki AL-AIN . . .
Japan national football team	. . . Ladki AL-AIN , United Arab Emirates 1996-12-06 *Japan* began the defence of their Asian Cup title with . . .
	BLINK (Wiki)
Japan national football team	. . . SOCCER - *JAPAN* GET LUCKY WIN , CHINA IN SURPRISE DEFEAT . . . .
China PR national football team	. . . SOCCER - JAPAN GET LUCKY WIN , *CHINA* IN SURPRISE DEFEAT . Nadim Ladki AL-AIN . . .
Al Ain	. . . CHINA IN SURPRISE DEFEAT . Nadim Ladki *AL-AIN* , United Arab Emirates 1996-12-06 Japan began the . . .
	RELIC (Wiki)
RKC Waalwijk	. . . division soccer match played on Friday : *RKC Waalwijk* 1 (Starbuck 76) Willem II Tilburg 2 . . .
Willem II (football club)	. . . : RKC Waalwijk 1 (Starbuck 76) *Willem II Tilburg* 2 (Konterman 45 , Van der Vegt . . .
PSV Eindhoven	. . . SOCCER - *PSV* HIT VOLENDAM FOR SIX . AMSTERDAM 1996-12-07 . . .
	RELIC (In-Domain)
English people	. . . Mahindra International . The Australian brushed aside unseeded *Englishman* Mark Cairns 15-7 15-6 15-8 . Top-seeded Eyles . . .
England	. . . 16-year-old who attended Sale Grammar School in the northern *England* city of Manchester died less than a day after . . .
England	. . . a last-minute goal to salvage a 2-2 draw for *English* premier league leaders Arsenal at home to Derby on . . .

Table 7: **Most Conflated Entities on AIDA using GRINCH Clustering.** Left: Unique entity ID. Right: Mention with entity surface form in italics.

	Feature-Based
Transcription, Genetic	. . . of eight tissues. The Mef2c promoter had the higher *transcriptional activity* in differentiated C2C12 cells than that in proliferating C2C12 . . .
Transcriptional Activation	. . . in proliferating C2C12 cells, which was accompanied by the *up-regulation* of mRNA expression of Mef2c gene. Function deletion and . . .
Regulation of Biological Process	. . . (GPCRs) is a key event for cell signaling and *regulation* of receptor function. Previously, using tandem mass spectrometry, we . . .
	MLM
Left Ventricular Function	. . . computed tomography angiography, we assessed 3 primary outcome measures: *left ventricular (LV) systolic function* (left ventricular ejection fraction), LV diastolic function (early relaxation . . .
Left Ventricular Ejection Fraction	. . . 3 primary outcome measures: left ventricular (LV) systolic function (*left ventricular ejection fraction*), LV diastolic function (early relaxation velocity), and coronary atherosclerosis . . .
Endoscopy (Procedure)	. . . Comprehensive Cancer Network (NCCN) guidelines, which are based on *rigid endoscopic* measurements. The medical records of patients scheduled to receive . . .
	BLINK (Wiki)
Medical Records	. . . guidelines, which are based on rigid endoscopic measurements. The *medical records* of patients scheduled to receive curative surgery for histologically . . .
Lower - Spatial Qualifier	. . . with rectal cancer located in the upper (Ra) or *lower* (Rb) division using double-contrast barium enema. The median values . . .
Asymptomatic (Finding)	. . . two bladder urothelial cancer metastatic to the penis with *no relevant clinical symptoms* . Namely, a 69 years-old man with a warthy lesions . . .
	RELIC (Wiki)
Lysome-associated . . .	. . . Herein, we demonstrated that Zn(2+) could induce deglycosylation of *lysosome-associated membrane protein 1* and 2 (LAMP-1 and LAMP-2), which primarily locate in . . .
Sialic Acid . . .	. . . In this study, we set out to define how *CD169(+)* phagocytes contribute to neuroinflammation in MS. CD169 - diphtheria . . .
SMAD3 gene	. . . Epigenome-wide analysis links *SMAD3* methylation at birth to asthma in children of asthmatic . . .
	RELIC (In-Domain)
Individual	. . . Cardiovascular Toxicity of Illicit Anabolic-Androgenic Steroid Use Millions of *individuals* have used illicit anabolic-androgenic steroids (AAS), but the long-term . . .
Pharmicologic Substance	. . . steroids (AAS), but the long-term cardiovascular associations of these *drugs* remain incompletely understood. Using a cross-sectional cohort design, we . . .
Finding	. . . complications occurred in the studied neonates. Based on these *findings* , IC - ECG -guided tip placement appears to be . . .
	Hybrid
Study	. . . marker for activated phagocytes in inflammatory disorders. In this *study* , we set out to define how CD169(+) phagocytes contribute . . .
Evaluation	. . . to provide holistic end-of-life care and assisted in the *overall assessment* of palliative care patients, identifying areas that might not . . .
Research Study	. . . which is hardly visible in clinically applied CT-imaging. This *experimental study* investigates ten different PSI designs and their effect to . . .

Table 8: **Most Conflated Entities on MedMentions using Greedy NN Clustering.** Left: Unique entity ID. Right: Mention with entity surface form in italics.

	MLM
Cardiovascular	. . . *Cardiovascular* Toxicity of Illicit Anabolic-Androgenic Steroid Use Millions of individuals . . .
Toxic Effect	. . . Cardiovascular *Toxicity* of Illicit Anabolic-Androgenic Steroid Use Millions of individuals have . . .
Steroids	. . . Cardiovascular Toxicity of Illicit *Anabolic-Androgenic Steroid* Use Millions of individuals have used illicit anabolic-androgenic steroids . . .
	BLINK (Wiki)
Cardiovascular	. . . *Cardiovascular* Toxicity of Illicit Anabolic-Androgenic Steroid Use Millions of individuals . . .
Toxic Effect	. . . Cardiovascular *Toxicity* of Illicit Anabolic-Androgenic Steroid Use Millions of individuals have . . .
Steroids	. . . Cardiovascular Toxicity of Illicit *Anabolic-Androgenic Steroid* Use Millions of individuals have used illicit anabolic-androgenic steroids . . .
	RELIC (Wiki)
Sialic Acid . . .	. . . In this study, we set out to define how *CD169(+)* phagocytes contribute to neuroinflammation in MS. CD169 - diphtheria . . .
USP17L2 Protein	. . . *DUB3* and USP7 de-ubiquitinating enzymes control replication inhibitor Geminin: molecular . . .
USP7 Protein	. . . DUB3 and *USP7* de-ubiquitinating enzymes control replication inhibitor Geminin: molecular characterization and . . .
	RELIC (In-Domain)
Protein Expression	. . . by vascular endothelial growth factor (VEGF) signaling. We describe *spatiotemporal expression* of vegf and vegfr and experimental manipulations targeting VEGF . . .
Genes, Homeobox	. . . cell adhesion, and newly identified processes, including transcription and *homeobox genes* . We identified mutations in protein binding sites correlating with . . .
Gene Expression	. . . identified mutations in protein binding sites correlating with differential *expression* of proximal genes and experimentally validated effects of mutations . . .

Table 9: **Most Conflated Entities on MedMentions using GRINCH Clustering.** Left: Unique entity ID. Right: Mention with entity surface form in italics.

Feature-Based	
Yu - Gi - Oh ! - Episode 004	. . . *Yu - Gi - Oh ! - Episode 004* " Into the Hornet ' s Nest " , known . . .
Yu - Gi - Oh ! ZEXAL - Episode 082	. . . *Yu - Gi - Oh ! ZEXAL - Episode 082* " Sphere Cube Calamity : Part 1 " , known . . .
Yu - Gi - Oh ! Duelist - Duel 168	. . . *Yu - Gi - Oh ! Duelist - Duel 168* " The Waiting Grave " , also known as " . . .
MLM	
Yami Yugi and Rafael ' s first Duel	. . . *Yami Yugi and Rafael ' s first Duel* Yami Yugi goes to duel against Rafael as the message . . .
Yu - Gi - Oh ! - Episode 004	. . . *Yu - Gi - Oh ! - Episode 004* " Into the Hornet ' s Nest " , known . . .
Yu - Gi - Oh ! ZEXAL - Episode 082	. . . *Yu - Gi - Oh ! ZEXAL - Episode 082* " Sphere Cube Calamity : Part 1 " , known . . .
BLINK (Wiki)	
Robin (Friends)	. . . , Sarah and Maya . Emma has a horse called *Robin* , dog called Lady and a cat called Jewel . . . .
41003 Olivia ' s Newborn Foal	. . . a pet bird , Goldie . Olivia also has a *new pet foal* , which she takes care of frequently . She seems . . .
41007 Heartlake Pet Salon	. . . its neck . Background . Joanna brings her poodle to *the pet salon* , where Emma pampers her up . ¡ br ¿ . . .
RELIC (Wiki)	
Ro Gale	. . . , as was Maquis leader Macias . Ro recalled that *her father* made the strongest " hasperat " she ' d ever . . .
Unnamed shuttlepods (22nd century)	. . . . " ( ) The Federation starship carried at least *one shuttlepod* until the time of its disappearance in the mid - . . .
Founders ' homeworld (2372)	. . . As she is reluctant to reveal the location of the *Founders ' new homeworld* , but respects Sisko ' s loyalty to Odo when . . .
RELIC (In-Domain)	
Yu - Gi - Oh ! - Episode 004	. . . *Yu - Gi - Oh ! - Episode 004* " Into the Hornet ' s Nest " , known . . .
Yu - Gi - Oh ! ZEXAL - Episode 082	. . . *Yu - Gi - Oh ! ZEXAL - Episode 082* " Sphere Cube Calamity : Part 1 " , known . . .
Yu - Gi - Oh ! Duelist - Duel 168	. . . *Yu - Gi - Oh ! Duelist - Duel 168* " The Waiting Grave " , also known as " . . .
Hybrid	
Yu - Gi - Oh ! - Episode 004	. . . *Yu - Gi - Oh ! - Episode 004* " Into the Hornet ' s Nest " , known . . .
Yu - Gi - Oh ! ZEXAL - Episode 082	. . . *Yu - Gi - Oh ! ZEXAL - Episode 082* " Sphere Cube Calamity : Part 1 " , known . . .
Yu - Gi - Oh ! Duelist - Duel 168	. . . *Yu - Gi - Oh ! Duelist - Duel 168* " The Waiting Grave " , also known as " . . .

Table 10: **Most Conflated Entities on Zeshel using Greedy NN Clustering.** Left: Unique entity ID. Right: Mention with entity surface form in italics.

MLM	
Moondeep Sea	. . . Larynda Telenna was the high priestess of Kiaransalee in the *Vault of Gnashing Teeth* beneath Vaasa . She was also the leader of Kiaransalee . . .
Tabaxi (tribe)	. . . the Chultan Peninsula , consisting primarily of members of the *Tabaxi tribe* . Description . Chultans were tall and had dark , . . .
New Velar	. . . from the Moonsea Ride , as it would have connected *Harrowdale Town* with this major road . To avoid ambushes by the . . .
BLINK (Wiki)	
Astral projection	. . . also possible to escape with " teleportation " spells or *astral travel* , though the force blocked ethereal travel . A captive . . .
Krakentua (Shinkintin)	. . . and force newly hatched krakentua spawn to fight . A *krakentua* related these events via dreams to adventurers in the Fochu . . .
Generic temple guard	. . . two to attempt a crossing were Father Sambar and a *temple guard* . Sambar died horrifically , but the guard survived as . . .
RELIC (Wiki)	
Moondeep Sea	. . . Larynda Telenna was the high priestess of Kiaransalee in the *Vault of Gnashing Teeth* beneath Vaasa . She was also the leader of Kiaransalee . . .
Tabaxi (tribe)	. . . the Chultan Peninsula , consisting primarily of members of the *Tabaxi tribe* . Description . Chultans were tall and had dark , . . .
Kaedlaw Burdun	. . . the Silver Wyrm in 1369 DR , Queen Brianna , *her newborn child* , Avner , Tavis Burdun , and Basil retreated to . . .
RELIC (In-Domain)	
Vetrix Family	. . . *Vetrix Family* The Vetrix Family , known as the Tron Family in . . .
Yu - Gi - Oh ! - Episode 004	. . . *Yu - Gi - Oh ! - Episode 004* " Into the Hornet ' s Nest " , known . . .
Yu - Gi - Oh ! ZEXAL - Episode 082	. . . *Yu - Gi - Oh ! ZEXAL - Episode 082* " Sphere Cube Calamity : Part 1 " , known . . .

Table 11: **Most Conflated Entities on Zeshel using GRINCH Clustering.** Left: Unique entity ID. Right: Mention with entity surface form in italics.