$C3EL$: A Joint Model for Cross-Document Co-Reference Resolution and Entity Linking

Sourav Dutta
Max-Planck Institute for Informatics
Saarbrücken, Germany
sdutta@mpi-inf.mpg.de

Gerhard Weikum
Max-Planck Institute for Informatics
Saarbrücken, Germany
weikum@mpi-inf.mpg.de

Abstract

Cross-document co-reference resolution (CCR) computes equivalence classes over textual mentions denoting the same entity in a document corpus. Named-entity linking (NEL) disambiguates mentions onto entities present in a knowledge base (KB) or maps them to $null$ if not present in the KB. Traditionally, CCR and NEL have been addressed separately. However, such approaches miss out on the mutual synergies if CCR and NEL were performed jointly.

This paper proposes $C3EL$, an unsupervised framework combining CCR and NEL for jointly tackling both problems. $C3EL$ incorporates results from the CCR stage into NEL, and vice versa: additional global context obtained from CCR improves the feature space and performance of NEL, while NEL in turn provides distant KB features for already disambiguated mentions to improve CCR. The CCR and NEL steps are interleaved in an iterative algorithm that focuses on the highest-confidence still unresolved mentions in each iteration. Experimental results on two different corpora, news-centric and web-centric, demonstrate significant gains over state-of-the-art baselines for both CCR and NEL.

1 Introduction

With the advent of large knowledge bases (KB) like DBpedia, YAGO, Freebase, and others, entities (people, places, organizations, etc.) along with their attributes and relationships form the basis of smart applications like search, analytics, recommendations, question answering, and more. The major task that arises in both the KB construction process and the entity-centric applications involves precise *recognition*, *resolution*, and *linking* of named entities distributed across web pages, news articles, and social media.

Named Entity Recognition (NER) deals with the identification of entity *mentions* in a text and their classification into coarse-grained semantic types (person, location, etc.) (Finkel et al., 2005; Nadeau & Sekine, 2007; Ratinov & Roth, 2009). This involves segmentation of token sequences to obtain mention boundaries, and mapping relevant token spans to pre-defined entity categories. For example, NER on the text `Einstein won the Nobel Prize` identifies the mentions "Einstein" and "Nobel Prize" and marks them as *person* and *misc* type, respectively.

Named Entity Linking (NEL)[1] involves the *disambiguation* of textual mentions, based on context and semantic information, and their mapping to proper *entities in a KB* (Bunescu & Paşca, 2006; Cucerzan, 2007; Milne & Witten, 2008; Hoffart et al., 2011; Ratinov et al., 2011; Cornolti et al., 2013). For example, in the above text, the mention "Einstein" is linked to the physicist *Albert Einstein*.

Entity *Co-reference Resolution* (CR) (Haghighi & Klein, 2010; Ng, 2010; Lee et al., 2013) is essentially a clustering task to identify mentions (and anaphoras) within a document referring to the same entity, thus computing equivalence classes or *mention groups*. For example, mentions `Albert Einstein` and `Nobel laureate Einstein` both refer to the same entity *German physicist Albert Einstein*, but are different from the mention `Hans Albert Einstein`.

When CR is extended to an entire text corpus, in order to generate equivalence classes of co-referring mentions across documents, the task is known as *Cross-document Co-reference Resolution* (CCR) (Bagga & Baldwin, 1998; Culotta et al., 2007; Singh et al., 2011; Dutta & Weikum, 2015). Note that CCR is not the same as merely concatenating all documents in the corpus and utilizing existing CR methods. The linguistic diversity across documents and high computational cost for huge numbers of mentions in the corpus would typically make such a CR-based simulation perform poorly. Neither CR nor CCR links mention groups to corresponding KB entities. Thus, they represent both in-KB entities and out-of-KB entities (e.g., long-tail or emerging entities that do not have a Wikipedia article) in the same way.

[1] *Named Entity Disambiguation (NED)* and "Wikification" are often used to denote the same task. The latter may be more broadly used, though, to include the disambiguation of common nouns and phrases onto concepts, whereas NED restricts itself to noun phrases that denote individual entities.

Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 846–856,
Lisbon, Portugal, 17-21 September 2015.

State-of-the-Art and its Limitations: Established CR methods rely on rule-based methods or supervised learning techniques on syntactic paths between mentions, semantic compatibility, and other linguistic features (Haghighi & Klein, 2009), with additional use of distant features from KBs (Lee et al., 2013). Modern cluster-ranking (Rahman & Ng, 2011) and multi-sieve methods (Ratinov & Roth, 2012) involve incremental expansion of mention groups by considering semantic types and Wikipedia categories. CCR methods utilize transitivity-aware clustering techniques (Singh et al., 2011), by considering mention-mention similarities (Bagga & Baldwin, 1998) along with features extracted from external KBs (Dutta & Weikum, 2015).

NEL methods often harness the semantic similarity between mentions and entities and also among candidate entities for different mentions (in Wikipedia or other KBs) for contextualization and coherence disambiguation (Hoffart et al., 2011; Milne & Witten, 2008; Kulkarni et al., 2009; Ratinov et al., 2011). However, in the absence of CR mention groups, NEL has limited context and is bound to miss out on certain kinds of difficult cases.

Although NER, CR, CCR and NEL involve closely related tasks and their tighter integration has been shown to be promising (Chen & Roth, 2013; Zheng et al., 2013), they have mostly been explored in isolation. Recently, several *joint models* have been proposed for `CR-NER` (Haghighi & Klein, 2010; Singh et al., 2013), `CR-NEL` (Hajishirzi et al., 2013), and `NER-CR-NEL` (Durrett & Klein, 2014). However, to the best of our knowledge, no method exists for jointly handling CCR and NEL on large text corpora.

1.1 Approach and Contributions

This paper proposes the novel $C3EL$ *(Cross-doCument Co-reference resolution and Entity Linking)* framework for jointly modeling cross-document co-reference resolution (CCR) and linkage of mention groups to entities in a knowledge base (NEL).

Example: To illustrate the potential synergies between CCR and NEL, consider the 3 documents in Figure 1 containing 9 mentions (on the left) with candidate entities from a KB (on the right). CCR alone would likely miss the co-reference relation between `Logan` (Doc 1) and its alias `Wolverine` (Doc 2), leaving NEL with the difficult task of disambiguating "Logan" in a document with sparse and highly ambiguous context (Doc 1). On the other hand, NEL alone would likely map `Australia` (Doc 3) to the country (not the movie) and could easily choose the wrong link for mention "Hugh". Moreover, the presence of `Ava Eliot` as an out-of-KB mention complicates the task.

However, if we could more freely interleave

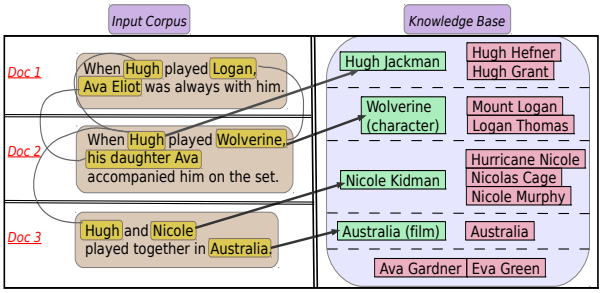

Figure 1: Joint CCR-NEL Example *(Green KB entries connected via arrows denote the correct entity linkage for the mention co-reference groups; while the red ones represent alternative incorrect candidates with similar surface forms)*

CCR and NEL and could iterate them several times, we would be in a much stronger position. An initial NEL step for the easiest mention, namely "Wolverine", maps it to the character of X-Men movies. This indicates that the three "Hugh" mentions could all be the same actor, and are thus easily merged into a co-reference group using CCR. We now have enough cues for NEL to choose the right entity for the "Hugh" mention group, which in turn enables the proper mapping of "Australia" to the movie. Finally, it becomes clear that mentions "Ava Eliot" and "his daughter Ava" should be merged into the same group and represented as an out-of-KB entity mapped to $null$.

The above example clearly demonstrates that interleaving CCR and NEL is highly beneficial. However, appropriate choices for the ordering of CCR and NEL steps are usually not obvious at all. The proposed $C3EL$ algorithm solves this problem: automatically determining an efficient interleaving of CCR and NEL.

Approach: $C3EL$ iteratively aggregates intermediate information obtained from alternating steps of CCR and NEL, thus forming a *feedback loop* for propagating mention features and entity knowledge. Intuitively, co-referring mentions obtained via CCR generate global context for improved NEL performance, while mentions linked to KB entities (by NEL) provide distant semantic features with additional cues for CCR. $C3EL$ couples several building blocks like unsupervised hierarchical clustering, *context summaries* for mentions and distant KB features for entities, drawing inspiration from the CCR-only method of (Dutta & Weikum, 2015). Mention linking to the KB (NEL) is performed using distant knowledge and co-occurring mentions.

In a nutshell, the major contributions of this paper are:

- the $C3EL$ framework for joint computation of cross-document co-reference resolution (CCR) and entity linking to a KB (NEL), based on propagating information across iterative CCR and NEL steps;
- techniques for considering co-occurring mentions in *context summaries* and for harnessing

context-based keywords for *link validation* in NEL, improving accuracy on out-of-KB entities;

- an experimental evaluation with two different corpora, one based on news articles and one based on web pages, demonstrating substantial gains for both CCR and NEL over state-of-the-art methods.

2 $C3EL$: Joint CCR-NEL Framework

Given an input corpus C of n documents, $C = \{D_1, D_2, \cdots, D_n\}$ with entity mentions $EM = \{m_{11}, m_{12}, \cdots, m_{21}, m_{22}, \cdots\}$ ($m_{ij} \in D_i$), $C3EL$ aims to *jointly compute*:

- CCR: an equivalence relation over EM with equivalence classes E_i, such that $E_i \cap_{i \neq j} E_j = \emptyset$ and $\cup_i E_i = EM$, and
- NEL: linking each of the classes E_i to entities present in a KB or map it to $null$ if there is no proper entity in the KB.

To this end, $C3EL$ consists of 3 algorithmic stages: **(i)** Pre-Processing, **(ii)** Interleaved NEL and CCR, and **(iii)** Finalization.

2.1 Pre-Processing Stage

HTML pages in the input corpus C are transformed into plain text using standard tools like `jsoup.org`. Recognition and markup of mentions are performed using the Stanford CoreNLP toolkit (`nlp.stanford.edu`), and a coarse-grained lexical type for each mention (e.g., person, location, organization, etc.) is obtained from the Stanford NER Tagger (Finkel et al., 2005). The multi-pass sieve algorithm for single-document CR (Raghunathan et al., 2010; Lee et al., 2011; Lee et al., 2013) then computes mention co-reference chains per document, and a *head mention* is chosen for each of the mention groups (chains). The head mention is typically represented by the most explicit denotation of the entity (e.g., person's full name with title, location name with country, etc.).

For each of the mention groups M_i, C3EL then constructs a *context summary* using:

- **Sentences** – all sentences in the document that contain mentions of group M_i; and
- **Co-occurrence** – all sentences for other mention groups that contain mentions co-occurring in any of the sentences of M_i (as obtained above).

Formally, for each mention group M_i, let $S(M_i) = \{sentence(m_j) \mid m_j \in M_i\}$ represent the set of extracted sentences, where $sentence(m_j)$ denotes the sentences in which mention m_j occurs. Also, let the co-occurring mention set of M_i be $Co(M_i) = \{m' \mid m' \in S(M_i) \wedge m' \notin M_i\}$. The context summary of M_i is defined as:

$$\mathbb{CS}(M_i) = S(M_i) \cup \left(\bigcup_{m' \in Co(M_i)} S(m') \right)$$

The context summaries intentionally do not include any distant KB features for mentions. The intuition is to minimize potential noise from overly speculative mappings to the KB at this initial stage.

2.2 Interleaved NEL & CCR Approach

After the preliminary CR step on each document and the construction of context summaries, $C3EL$ now performs an initial NEL step for each of the mention groups M_i, using the extracted sentences $S(M_i)$ as inputs to NEL. It obtains the best matching entity, the confidence of the match, and its corresponding Wikipedia page. Off-the-shelf NEL software (like WikipediaMiner or Illinois-Wikifier) is used for mention-entity mapping based on prior popularity of the named-entities (from the KB) and textual similarity between $S(M_i)$ (context of the mention group) and the entity descriptions in KB.

For each M_i, the entity link obtained (from NEL) is then "validated" using a similarity measure between features from the *context summary*, $\mathbb{CS}(M_i)$ (including co-occurring mentions) and distant KB labels – forming the *link validation* procedure of $C3EL$. This explicit use of co-occurring mentions' ($Co(M_i)$) contexts helps to better identify out-of-KB entities compared to direct full-fledged NEL using the entire input text (shown in Section 3). Also the use of NEL on $S(M_i)$ alone, makes $C3EL$ "light-weighted".

The mappings between the mention groups and KB entries are then classified, on the basis of the NEL confidence scores, into *Strong Evidence* (SE), *Weak Evidence* (WE), and *No Evidence* (NE) classes. For mention groups placed in SE, the KB features (obtained previously) are appended to their context summaries, while mentions strongly linked to same KB entities are considered to be co-referring and hence grouped together (performing implicit CCR).

Considering our example (Figure 1), we now outline the iterative steps of $C3EL$ interleaving NEL & CCR.

1. During **Iteration 1**, $C3EL$ performs:

- **NEL:** The initial NEL step maps the unambiguous mentions, `Wolverine` to the X-Men movie character and `Australia` to the country, with high confidence. However, link validation fails for "Australia" as there is very low similarity between the mention context features (e.g., Hugh, Wolverine, etc.) and the distant KB labels extracted from its Wikipedia page (e.g., Commonwealth, population, etc.); thus the link is dropped and the mention is added to NE. So only the mention "Wolverine" is added to the

SE class and enriched with KB features (e.g., alias Logan).

On the other hand, the 3 "Hugh" mentions exhibit low NEL confidence due to the high ambiguity of this first name and are therefore classified into WE. The remaining mentions have extremely low NEL confidence (due to sparse contextual information) and are added to NE.

- **CCR:** The WE and NE classes are fed separately to the CCR procedure. Based on the *context summary* similarities between mentions, $C3EL$ performs hierarchical clustering to group together the "Hugh" mentions (in the WE class) and creates a co-referring mention group with the individual mentions' context summaries concatenated. This merging of summaries grows and strengthens captured contexts, which propagates across documents. This concludes the first iteration of $C3EL$.

2. The above results are provided to the **second Iteration**:

- **NEL:** The *context summary* of the "Hugh" mention group in WE now provides definitive cues to correctly map it to the actor Hugh Jackman with high confidence, thus placing it in the SE class.
- **CCR:** The ensuing CCR step groups together "Ava Eliot" and "Ava" (in NE) using co-occurrence context of the co-referring Hugh mentions.

3. Subsequent NEL iterations (on WE and NE) identify "Ava" as an out-of-KB entity and correctly links "Australia" to the movie using CCR-generated mention-group contexts and link validation. CCR finally groups together "Logan" with "Wolverine" based on context similarity with distant KB features. This process of alternating CCR and NEL is repeated until all mention groups are strongly connected to KB entities (placed in SE), or no changes are made anymore.

The NEL and CCR procedures are performed separately on the different mention types (like PER, LOC, etc.), since different mention types rarely co-refer. We next present the internal working details of the NEL and CCR stages of $C3EL$.

2.2.1 Named-Entity Linking (NEL) Stage

In its NEL procedure, $C3EL$ disambiguates mentions to entities in the YAGO knowledge base (`yago-knowledge.org`). We perform NEL on the sentences ($S(M_i)$) of a mention group, using named-entity popularity statistics and context, to obtain the best matching entity, its confidence score, and the corresponding Wikipedia page (from $sameAs$ link in YAGO). Assume a mention group M_i to be mapped to an entity e_i with a confidence score of $\phi(M_i, e_i)$.

A. Link Validation: For each mention group (e.g., Hugh), we extract *distant KB labels* such as semantic types or categories (e.g., actor), title (e.g., Golden Globe winner), alias (e.g., Wolverine), location, and gender (for person) from the Wikipedia page infoboxes. The similarity of these features to keywords obtained from the context summary $\mathbb{CS}(M_i)$ is computed using IR-style term frequencies within a document (tf) and inverse document frequencies within the corpus (idf). We utilize the bag-of-words model based $tf \times idf$-weighted cosine similarity measure. If the similarity score is above a threshold, τ, the NEL result is accepted, otherwise it is discarded – thus avoiding noisy linkage of sparse mentions to prominent KB entries. This subtle introduction of *controlled distant supervision* within the $C3EL$ framework enables efficient detection of out-of-KB mentions.

B. Classification: To sift out well-known and long-tail entities from new ones, and prevent "noisy" interactions among the contexts of in-KB and out-of-KB mentions (with similar surface forms), mention groups M_i (linked to e_i with score $\phi(M_i, e_i)$) are classified into 3 classes by 2 threshold parameters, δ_s and δ_w, as:

- *Strong Evidence* (SE): For $\phi(M_i, e_i) \geq \delta_s$, mention group M_i exhibits high linkage confidence with e_i and is placed in SE. If two or more mentions in SE are independently mapped to the same KB entity, they co-refer transitively and are hence grouped together with their context summaries merged (implicit CCR). Distant KB features for mentions in SE are extracted and appended to $\mathbb{CS}(M_i)$, providing additional cues for later steps.
- *Weak Evidence* (WE): Mention groups with $\delta_w \leq \phi(M_i, e_i) < \delta_s$ are placed in this class. They mostly represent long-tail in-KB entities (sparsely represented in KB) with limited semantic information (for detection) but might also be new/emerging entities absent from KB.
- *No Evidence* (NE): $\phi(M_i, e_i) < \delta_w$ represents mentions groups that have been mapped to $null$ (or have near-zero match confidence) or have failed link validation during the NEL procedure. These entities are most likely to be out-of-KB and are allocated to this class.

2.2.2 Cross-Document CR (CCR) Stage

The CCR stage of $C3EL$ adopts the sampling-based hierarchical clustering approach of (Dutta & Weikum, 2015), to obtain co-referring mention clusters.

A. Similarity Measure: To infer whether two mention groups represent the same entity, the similarity between the context summaries are computed based on (i) *tf-idf*-weighted bag-of-words cosine distance, and (ii) partial-match scores of multi-word keyphrases in bounded text windows (Taneva et al., 2011). The context summaries (with stopwords removed) are re-interpreted as, (i) bag of words, and (ii) bag of keyphrases, to extract fea-

ture vectors for similarity computation. Finally, the mixture model of *bag-of-words* (BoW) and *keyphrases* (KP) of (Dutta & Weikum, 2015) is used to assign feature weights using *tf-idf* measure.

B. Hierarchical Clustering: s mention groups are uniformly randomly sampled and their similarities to the other groups (using context summary) are computed. A similarity-weighted graph with the mention groups as nodes and edge weights representing mention-mention similarities is constructed. Bisection-based hierarchical *balanced min-edge-cut graph partitioning* (Buluc et al., 2013) is performed, using the $METIS$ software (Karypis & Kumar, 1999)[2], to partition non-coreferent mentions groups. The *Bayesian Information Criterion* (BIC) (Schwarz, 1978; Hourdakis et al., 2010), a Bayesian variant of Minimum Description Length (Grünwald, 2007), is used as the cluster split stopping criterion, and the context summaries within each final cluster are merged.

CCR aims to process heterogeneous corpora that go beyond a single domain and style, such as Web collections.

2.3 Finalization Stage

For the remaining mention groups in WE, we finally perform threshold based disambiguation of mention clusters using the context summaries. For each mention group $M_i \in WE$, we compute (1) its context summary similarities (as in Section 2.2.2) to all other mention groups M_j in SE by also using distance features from the weakly linked KB entities, and (2) textual overlap between the mention group representatives. M_i is concatenated with the best matching entity M_k (in SE) if the similarity score is above a threshold θ; else M_i is marked as an out-of-KB entity (mapped to $null$) and is placed in the NE class. This helps in reducing propagated CR errors like erroneous mention boundary detection (in NER), omissions in co-reference chain, etc. (leading to "phantom" out-of-KB entities).

The obtained mention groups represent the final equivalence classes of co-referring mentions across documents – capturing both in-KB entities (with links to the KB) in the SE class and out-of-KB entities (mapped to $null$) in the NE class.

3 Experimental Evaluation

In this section, we empirically study the performance of $C3EL$ against various state-of-the-art methods. We analyze the individual gains in CCR and NEL due to the joint modeling.

Datasets: We use the following 2 publicly available corpora:

- **EventCorefBank (ECB) corpus**[3] (Bejan & Harabagiu, 2010)**:** contains 482 news and Web articles (classified into 43 topics) with a total of 5447 mentions corresponding to 1068 distinct named-entities. Entity co-reference annotations (across documents within each topic cluster) were provided by (Lee et al., 2012), and we performed manual examination of the annotations for KB linking of the entities to Wikipedia entries, if present; thus providing ground truth for both CCR and NEL.
- **ClueWeb2009 FACC1 dataset**[4] (Gabrilovich et al., 2013)**:** provides machine automated entity-linkage annotations of the *ClueWeb09* corpus (ca. 1 Billion crawled Web pages) with Freebase entries[5]. The corpus contains many topical domains and highly diverse documents from news, movie reviews, people home pages to blogs and other social media posts. We randomly select $500K$ documents containing 4.64 Million mentions associated with 1.29 Million distinct entities to form our corpus. For NEL ground-truth construction, we link the entities to their Wikipedia pages (using Freebase's "on the web" property). Since no explicit annotations of inter-document entity co-references exists, we consider two mentions (in different documents) to co-refer if they are linked with the same Freebase entity.

Evaluation: To assess the output quality of $C3EL$ we use the following established metrics:

- B^3 $F1$ **score** (Bagga & Baldwin, 1998): measures the F1 score as the harmonic mean of average *precision* and *recall* computed over all mention groups in the final equivalence classes. Precision (for a mention group) represents the ratio of the number of correctly reported co-references (or linking) to the actual number; while recall computes the fraction of the gold-standard annotations correctly identified.
- ϕ_3 $-$ $CEAF$ **score** (Luo, 2005): provides an alternate F1 score computed as in the B^3 measure; but calculates precision and recall of mention groups using the best 1-to-1 mapping (i.e., mapping with maximum mention overlap) between the resultant equivalence classes and those in the ground truth. Normalization with the number of mentions for each of the resultant classes yields the ϕ_4-CEAF score.

We consider only the 3 most notable mention types: person (PER), location (LOC), and organization (ORG) – accounting for 99.7% of entities present in the ECB corpus and 96.3% of our ClueWeb09 corpus. All experiments were conducted on a 4 core Intel i5 2.50 GHz processor with 8GB RAM running Ubuntu 12.04 LTS.

3.1 Parameter Tuning & Sensitivity Study

Validation of entity linkage to KB and their subsequent classification into *confidence classes* (as de-

[2] glaros.dtc.umn.edu/gkhome/metis/metis/overview

[3] faulty.washington.edu/bejan/data/ECB1.0.tar.gz

[4] lemurproject.org/clueweb09/FACC1

[5] Human analysis of a subset of the annotations generated revealed a precision of $80 - 85\%$ (Gabrilovich et al., 2013)

Approach	P	R	B^3	ϕ_3	ϕ_4
EECR	74.9	55.5	63.7	-	33.7
CROCS	73.11	75.28	74.18	67.35	-
C3EL	79.52	82.91	81.18	73.89	53.3

Table 2: CCR performance (%) comparison on ECB

scribed in Section 2) during the NEL step of $C3EL$ are based on 3 parameters: confidence thresholds (δ_s and δ_w) and validation threshold (τ); the values of which can be tuned based on *cross-validation* approach with *train* and *test* data subsets. Using the "gold annotations" of the train-set (30% of total data), parameter values providing the best precision score are individually learnt using *line search* with small step size.

In our experimental setup, we systematically vary the parameter values and observe its effects on $C3EL$ for the training data. With increase in δ_s, the number of mentions mapped to the *Strong Evidence* (SE) class decreases. This in turn limits the influx of external KB features, thus degrading CCR performance as observed in Table 1(a). While for low values of δ_s, even weak mention links are placed in SE, leading to a decrease in precision due to noisy KB feature inclusion. On the other hand, a high δ_w value increases the number of mentions in the NE class, while low values tends to accumulate mentions in the WE class. This adversely affects the detection of out-of-KB entities due to noise from other co-occurring similar KB mentions (refer Table 1(b)) during clustering in CCR step.

The effect of τ on $C3EL$ has be shown in Table 1(c). Similar to the behavior induced by δ_s, we observe that a high τ limits entity linking and possible KB feature inclusion, while an extremely low value (near to zero) allows for noisy feature incorporation – both situations leading to lowered CCR efficiency. However, since τ prevents gross mis-alignment of mentions to KB entities, a wide range of small value ($0.1 - 0.35$) is seen to provide comparable performance.

Hence, for our remaining experimental study we set $\delta_s = 0.11$ and $\delta_w = 0.06$ (as in (Hoffart et al., 2014)), while τ is set to 0.1, and threshold for the finalization stage $\theta = 2 \times \delta_s = 0.22$.

3.2 CCR Performance Results

We initially benchmark the performance improvement in cross-document co-reference resolution (CCR) procedure by $C3EL$ against two competing approaches:
(1) state-of-the art sampling based hierarchical clustering method, $CROCS$ (Dutta & Weikum, 2015); and
(2) iterative joint entity-event CCR, $EECR$ (Lee et al., 2012).

Table 2 tabulates the results obtained on the ECB dataset. We observe C3EL to decisively outperform both the existing methods, providing a B^3 F1 improvement of around 7% over $CROCS$ and 17% over EECR. We further attain around 6% $\phi_3 - CEAF$ score enhancement over CROCS, and a significant 20% improved $\phi_4 - CEAF$ score compared to EECR.

Approach	P (%)	R (%)	B^3 (%)	ϕ_3 (%)
$CROCS_G$	79.9	83.33	81.58	74.11
$C3EL_G$	84.74	89.9	87.24	80.5

Table 3: CCR results on ECB

Type	Approach	P (%)	R (%)	B^3 (%)
PER	$CROCS_G$	71.8	74.15	72.96
	$C3EL_G$	84.85	82.73	83.78
LOC	$CROCS_G$	78.23	85.41	81.66
	$C3EL_G$	81.41	94.31	87.29
ORG	$CROCS_G$	85.73	87.89	86.8
	$C3EL_G$	88.52	91.82	90.14

Table 4: CCR results on ECB for mention types

A. Gold Results: Errors introduced during the preprocessing stage of C3EL (e.g., mention omission, tag mis-classification, intra-document CR errors, etc., by the Stanford CoreNLP toolkit) propagate to subsequent computing stages and adversely impacts the overall system performance. To provide an unbiased viewpoint of the actual performance of $C3EL$, we manually provided "exact" mentions, mention tags, and intra-document CR mention chains for the ECB corpus; thereby obtaining *gold performance results*. From Table 3 we observe a 6% F1 points improvement (for both B^3 & CEAF-ϕ_3) in $C3EL$ compared to $CROCS$.

B. Mention Categorization: Person mention type (PER) provides the greatest challenge for CCR systems (compared to other types like LOC, ORG, etc.) due to associated nicknames, titles, and varied surface forms (abbreviations, spellings, etc.). We thus evaluate the CCR performance of $C3EL$ (and compare it with CROCS) on the ECB data, with "exact" input mentions, for the different mention categories. Table 4 validates that our joint modeling provides better global information cues, reporting a B^3 F1 score enhancement of around 11% over CROCS for PER mentions; along with improved results for the other mention types as well.

C. Large Data: To study the robustness of $C3EL$ and the effects of large datasets on CCR, we performed evaluations on the ClueWeb09-FACC1 dataset. Similar to the ECB dataset, $C3EL$ exhibits a B^3 F1 score improvement of nearly 10% and a ϕ_3-CEAF F1 improvement of 12% over $CROCS$ (refer Table 5).

The above experimental results showcase that a combined approach helps overcome challenges faced in CCR by entity linkage and corresponding distant KB feature extraction; improving the overall accuracy.

3.3 Named-Entity Linking (NEL) Results

We now benchmark the performance of named-entity linking (NEL) procedure for $C3EL$ against the state-of-the-art open-source AIDA software (`github.com/yago-naga/aida`). We separately inspect the precision of mention linking for *prominent entities* (in-KB) as well as *new/emerging* (out-of-KB) entities, and characterize the links as

Datasets	δ_s (B^3 F1)					δ_w (P)					τ (B^3 F1)				
	0.01	*0.05*	*0.10*	*0.15*	*0.20*	*0.01*	*0.02*	*0.04*	*0.06*	*0.08*	*0.03*	*0.10*	*0.20*	*0.35*	*0.50*
ECB	79.3	82.2	84.2	83.5	81.0	73.1	75.3	77.3	78.7	78.4	76.9	81.2	81.2	81.1	79.2
ClueWeb	70.1	77.2	81.5	81.0	78.7	78.2	81.1	83.6	85.1	85.1	70.3	79.1	78.2	78.8	76.4
	(a)					(b)					(c)				

Table 1: $C3EL$ performance (a) in CCR with δ_s, (b) in out-of-KB NEL with δ_w, and (c) in CCR with τ

Approach	P (%)	R (%)	B^3 (%)	ϕ_3 (%)
CROCS	68.66	70.96	69.79	62.85
C3EL	75.76	81.42	78.49	74.13

Table 5: CCR results on ClueWeb09-FACC1

Approach	Within-KB			Out-of-KB		Overall
	$\mathbb{C}$	$\mathbb{I}$	$\mathbb{U}$	$\mathbb{C}$	$\mathbb{I}$	**P** (%)
AIDA	86.5	13.5	0.0	63.9	36.1	83.4
C3EL	85.4	14.4	0.2	79.0	21.0	84.9

Table 6: NEL performance (%) comparison on ECB

Correct ($\mathbb{C}$), *Incorrect* ($\mathbb{I}$), or *Unlinked* ($\mathbb{U}$). The results on the ECB corpus are reported in Table 6. $C3EL$ attains comparable performance ($\sim$ 85% precision) to that of AIDA for well-known entity-mentions present in KB; albeit with a few mentions remaining unlinked due to our cautious *link validation* (using τ) approach. However, the use of τ reduces aggressive KB linking to provide a significant 15% improvement (over AIDA) in precise detection of new/emerging entities absent in KB. Overall, an 1.5% precision gain is observed by the joint formulation.

A. Large Data: The diverse nature of the web-scale ClueWeb09 dataset clearly portrays the performance gains in NEL procedure due to CCR generated information integration. For entities present in the KB, we observe an accuracy improvement of 0.5% over AIDA (refer Table 7). Similar to that of the ECB data, $C3EL$ attains a significant $\sim$ 14% improvement in the detection of new/emerging entities not represented in KB. For the 1 million mentions, $C3EL$ provides around 4% overall performance improvements.

Using a bootstrap re-sampling t-test (as in (Durrett & Klein, 2014)), we observed high statistical significance ($p < 0.01$) for Out-of-KB and Overall NEL, whereas the difference for Within-KB NEL is not statistically significant. Coping with Out-of-KB entities is essential for joint CCR+NEL, and an improved NEL performance using propagated information from CCR using semantics along with link validation enables highly efficient detection of new or emerging entities.

3.4 Comparison with Joint Models

Traditional CR methods fail to cope with the heterogeneity of mentions and contexts across multiple documents, and some form of clustering or joint reasoning over all mentions is thus mandatory. These methods have quadratic or cubic (sometimes even exponential) complexity, and hence running CR+NEL on a concatenated super-document works only for small corpora, and would be prohibitively expensive for large corpora, even in offline processing mode (Singh et al., 2011).

However, to study the behavior of existing CR-NEL joint models under "small" CCR environments, we compare $C3EL$ with:
(1) multi-sieve based $NECo$ (Hajishirzi et al., 2013)[6]; and
(2) conditional random field based BER (Durrett & Klein, 2014) [7].

Approach	Within-KB			Out-of-KB		Overall
	$\mathbb{C}$	$\mathbb{I}$	$\mathbb{U}$	$\mathbb{C}$	$\mathbb{I}$	**P** (%)
AIDA	88.5	10.6	1.0	69.6	30.4	84.6
C3EL	89.0	9.8	1.2	83.7	16.3	88.1

Table 7: NEL results (%) on ClueWeb09-FACC1 (statistical significance $p < 0.01$ for Out-of-KB entities)

Three topic clusters from the ECB corpus with 3, 4, and 5 articles respectively were selected, and the documents within each cluster were merged to form 3 "super-articles" (one per topic), forming a simulated CR setting. $NECo$ and BER were then used to perform CR and NEL on these 3 articles, and the results compared with that obtained by $C3EL$ on the original documents. We repeatedly sample 12 articles across 3 topic clusters, and execute the approaches to report the micro-averaged results across 5 independent runs.

From Table 8(a) we observe that the algorithms exhibit comparable co-reference resolution performance; thus validating propagation of global semantics in $C3EL$ due to the joint formulation. However, such CR methods using multi-sieves and CRF do not scale beyond few documents (upon concatenation), and require at least $4\times$ more runtime compared to $C3EL$. Hence, CCR cannot be efficiently tackled by simply employing CR methods on a "super-document".

However, harnessing of non-local mention features (via CCR) and efficient detection of new mentions using link validation enables $C3EL$ to achieve a gain of around 5% in NEL compared to others (see Table 8(b)). For both procedures, we observed statistically significant improvements of C3EL over BER and NECo with $p < 0.05$, using the bootstrap re-sampling t-test.

To further study the effect of larger corpus, we sampled 25 documents (with co-referring mentions) from the ClueWeb09 dataset and performed analysis among the algorithms. As previously, we observed significant computational complexity for traditional CR methods when applied to CCR setting making them far slower ($6-7\times$) than $C3EL$. Table 9 reports the CCR and NEL averaged results obtained across 5 independent runs. We attained comparable performance in CCR with around 3% improvement in NEL. All the algorithms are seen to achieve high NEL results due to the large presence of well-known (in-KB) entities.

[6] cs.washington.edu/research-projects/nlp/neco
[7] nlp.cs.berkeley.edu/projects/entity.shtml

Approach	P (%)	R (%)	B^3 (%)
NECo	87.77	82.09	84.84
BER	88.30	86.53	87.41
C3EL	87.54	88.11	87.82

(a)

Approach	$\mathbb{C}$ (%)	$\mathbb{I}$ (%)	$\mathbb{U}$ (%)
NECo	89.13	10.87	0.0
BER	89.89	10.11	0.0
C3EL	93.2	4.61	2.19

(b)

Table 8: Joint "Simulated" results on ECB subset for (a) CCR, and (b) NEL (statistical significance $p < 0.05$)

Approach	P (%)	R (%)	B^3 (%)
NECo	81.14	79.65	80.39
BER	84.36	83.01	83.68
C3EL	83.52	85.56	84.53

(a)

Approach	$\mathbb{C}$ (%)	$\mathbb{I}$ (%)	$\mathbb{U}$ (%)
NECo	94.71	5.29	0.0
BER	95.27	4.73	0.0
C3EL	98.23	1.5	0.27

(b)

Table 9: Joint "Simulated" results on ClueWeb09 subset for (a) CCR, and (b) NEL

3.5 Algorithmic Baseline Study

We explore the performance of variants of $C3EL$ (on both corpora) ablating various system components (see Table 10). Explicitly, we consider:

- **Co-occurring Mentions:** Removal of cooccurrence mentions context from the *context summaries* constructed, reduces semantic information and adversely affects both NEL and CCR procedures. We thus observe a sharp decrease in CCR performance and also a degradation in entity linking.
- **Link Validation:** Filtering of mention linking to KB entities using link validation step (with threshold τ) in $C3EL$ enables corroboration of mention context keywords with the linked entity features. This leads to enhanced detection of new or emerging entities by reducing induction of noise during the CCR phase. Removal of this process permits aggressive entity linking and introduces noise, affecting new/emerging entity detection. We observe (from Table 10) nearly 20% reduction of precision (on both datasets) in identification of out-of-KB entity-mentions compared to $C3EL$.
- **NEL Categorization:** The differentiation of mentions (into classes) confidently mapped to KB entity reduces the collusion of "strong" linked mentions with other "noisy" mention contexts. This reduces incorrect grouping of different mentions with similar surface forms, contexts, etc., thereby improving precision of the CCR process. Use of a single NEL classification approach is observed to degrade CCR results, which in turn increases spurious entity linkage, decreasing NEL efficiency (Table 10).
- **Distant KB features:** As observed in (Baker, 2012; Zheng et al., 2013), extracted external KB features provide global and enhanced information cues promoting CR. We similarly observe CCR to attain the lowest F1 scores (compared to other baselines) when KB features are ignored. This in turn affects the linking of (some) well-known entities due to reduced context, leading to incorrect or low confidence NEL. Since no feature inclusion is performed for out-of-KB mentions, no effect is observed.

We observe that a joint formulation encompassing multiple information sources (along with noise filtering) enables mutually enhanced CCR and NEL within the proposed iterative feedback based framework, $C3EL$.

4 Related Work

Co-reference Resolution (CR): Traditional intra-document CR methods involve syntactic and semantic feature combination for identifying the best antecedent (preceding name or phrase) for a mention. CR methods employ rules or supervised learning techniques based on linguistic features such as syntactic paths and mention distances to assess semantic compatibility (Haghighi & Klein, 2009; Raghunathan et al., 2010; Rahman & Ng, 2011), while syntactic features are derived by deep parsing of sentences and noun group parsing. Semantic features from background knowledge resources like encyclopedia were used in (Daumé & Marcu, 2005; Ponzetto & Strube, 2006; Ng, 2007). The use of Wikipedia and structured knowledge bases (such as YAGO) to obtain mention-type relation and fine-grained mention attributes was explored by (Haghighi & Klein, 2009; Rahman & Ng, 2011). An overview of CR methods is given in (Ng, 2010).

Recent methods involve the use of multi-phase sieve, applying a cascade of rules for narrowing down the antecedent candidates for a mention (Raghunathan et al., 2010). Cluster ranking functions have also been proposed (Rahman & Ng, 2011; Zheng et al., 2013) to extend this paradigm for incrementally expanding and merging mention groups with preceding candidate clusters using relatedness features (Ratinov & Roth, 2012) and distant knowledge inclusion (Durrett & Klein, 2013). Person name disambiguation, a specific variation of CR, dealing with only person names, titles, nicknames, and other surface form variations was introduced in (Chen & Martin, 2007).

Distant Knowledge Labels: For obtaining semantic features, additional knowledge resources such as Wikipedia, YAGO, and FrameNet have been considered (Rahman & Ng, 2011; Baker, 2012). CR methods with confidence-thresholds were proposed in (Ratinov & Roth, 2012; Lee et al., 2013), and (Zheng et al., 2013) generalized these tech-

Baseline	ECB Dataset								ClueWeb09-FACC1 Dataset							
	CCR result			NEL results					CCR result			NEL results				
				Within-KB			*Out-of-KB*					*Within-KB*			*Out-of-KB*	
	P	R	B^3	$\mathbb{C}$	$\mathbb{I}$	$\mathbb{U}$	$\mathbb{C}$	$\mathbb{I}$	P	R	B^3	$\mathbb{C}$	$\mathbb{I}$	$\mathbb{U}$	$\mathbb{C}$	$\mathbb{I}$
Ignored Mention Co-occurrence	72.5	74.4	73.4	80.2	19.6	0.2	74.4	25.6	69.3	72.2	70.7	83.8	14.6	1.6	80.6	19.4
Link Validation (τ) ignored	79.0	81.4	80.2	**85.5**	14.5	**0.0**	62.8	37.2	74.8	81.0	77.8	**88.9**	**10.1**	**1.0**	69.8	30.2
Removed NEL Classification	73.2	80.7	76.8	83.9	15.9	0.2	76.1	23.9	70.1	77.6	73.6	86.1	12.3	1.6	79.5	20.5
Distant KB feature dropped	68.9	73.1	70.9	82.8	17.0	0.2	**79.0**	**21.0**	66.4	72.9	69.5	85.4	13.0	1.6	**83.7**	**16.3**
C3EL (Complete)	**79.5**	**82.9**	**81.18**	85.4	**14.4**	0.2	**79.0**	**21.0**	**75.8**	**81.4**	**78.5**	88.3	**10.1**	1.6	**83.7**	**16.3**

Table 10: CCR and NEL results (%) of C3EL for different baseline variations

niques by ranking the matching entities for distant labeling. However, such prior methods utilize distance labels of the current mention and considers all matching mentions making the procedure expensive. On the other hand, we extract distant features for the strongly matching (best) candidate only, reducing the performance overhead.

Cross-Document CR (CCR): Early approaches towards CCR involved the use contextual information from input documents for IR-style similarity measures (e.g., tf×idf score, KL divergence, etc.) over textual features (Bagga & Baldwin, 1998; Gooi & Allan, 2004). Probabilistic graphical models jointly learning the mappings of mentions to equivalent classes (co-referring mentions) using features similar to local CR techniques were studied in (Culotta et al., 2007; Singh et al., 2010; Singh et al., 2011), A clustering approach coupled with statistical learning of parameters was studied in (Baron & Freedman, 2008). However, such methods fail to cope with large corpora, and hence a "light-weight" streaming variant of CCR was introduced by (Rao et al., 2010).

Co-occurring mentions context have been harnessed for disambiguating person names for CR in (Mann & Yarowsky, 2003; Niu et al., 2004; Chen & Martin, 2007; Baron & Freedman, 2008). However, these methods do not use KB and depend on information extraction (IE) methods, witnessing substantial noise due to IE quality variance. A CCR framework combining co-occurring mention context with distant KB features embedded in an active hierarchical clustering procedure (Dutta & Weikum, 2015) was recently shown to perform efficiently, and provides inspiration for parts of our proposed $C3EL$ approach.

Named Entity Linking (NEL): Named entity resolution and linking stems from SemTag (Dill et al., 2003), and similar frameworks like GLOW, WikipediaMiner, AIDA, and others (Milne & Witten, 2008; Ratinov et al., 2011). A collection of entity disambiguation models was presented in (Kulkarni et al., 2009). Other NEL approaches utilize the notion of semantic similarity of entities to corresponding Wikipedia pages (Milne & Witten, 2008), while co-referent mention graph construction modeling mention co-occurrences and context similarity from outgoing hyperlinks in Wikipedia was used by (Hoffart et al., 2011). An integer linear programming (ILP) formulation also based on Wikipedia page similarities was presented in (Ratinov et al., 2011). However, none of these methods involve the incorporation of CR results for NEL. The first study on the benefits of CR for NEL was by (Ratinov & Roth, 2012); but a joint model was not proposed, instead attributes from Wikipedia categories were used as features. An overview and evaluation of different NEL methods has been given by (Hachey et al., 2013).

Joint Models: Jointly solving CR for entities and events utilizing cluster construction based on feature semantic dependencies was devised in (Lee et al., 2012). The use of CR as a pre-processing step for subsequent NEL procedure using an ILP formulation was proposed by (Chen & Roth, 2013). Recently, (Hajishirzi et al., 2013) proposed a joint model for CR and NEL using the Stanford multi-pass cluster update CR system with automatic linking of mentions to Wikipedia. An integrated belief propagation-based framework for CR, NER, and relation extraction was developed in (Singh et al., 2013). Subsequently, the model was enhanced by the use of structured conditional random fields, to solve CR, NER, and NEL in combination (Durrett & Klein, 2014). Other works involving joint formulation of NER and NEL use uncertainty of mention boundaries along with segmentation information extracted from Wikipedia (Sil & Yates, 2013). However, to the best of our knowledge, this work provides the first approach to jointly tackle CCR and NEL across documents in an entire corpus.

5 Conclusions

This paper presented the novel $C3EL$ framework for joint computation of cross-document coreference resolution (CCR) and named-entity linking (NEL). Our approach utilizes: (1) *context summaries* including co-occurring mention groups allowing for global context and feature propagation, and (2) *link validation* for NEL using distant KB features. This is embedded in an interleaved CCR and NEL model allowing for global semantics and feature propagation. The iterative approach enables information feedback between CCR (provides corpus-wide cues) and NEL (providing distant KB features). Experimental results on news and web data demonstrate improved performance of both CCR and NEL compared to prior methods.

References

Amit Bagga and Breck Baldwin. Entity-Based Cross-Document Coreferencing Using the Vector Space Model. *In COLING-ACL* 1998, pages 79–85.

Collin F. Baker. FrameNet, Current Collaborations and Future Goals. *LREC* 2012, 46(2):269–286.

Alex Baron and Marjorie Freedman. Who is Who and What is What: Experiments in Cross-Document Co-Reference. *In EMNLP* 2008, pages 274–283.

Cosmin A. Bejan and Sanda Harabagiu. Unsupervised Event Coreference Resolution with Rich Linguistic Features. *In ACL* 2010, pages 1412–1422.

Aydin Buluc, Henning Meyerhenke, Ilya Safro, Peter Sanders, and Christian Schulz. Recent Advances in Graph Partitioning. Karlsruhe Institute of Technology, Technical Report 2013.

Razvan Bunescu and Marius Paşca. Using Encyclopedic Knowledge for Named Entity Disambiguation. *In EACL* 2006, pages 9–16.

Ying Chen and James Martin. Towards Robust Unsupervised Personal Name Disambiguation. *In EMNLP* 2007, pages 190–198.

Xiao Cheng and Dan Roth. Relational Inference for Wikification. *In EMNLP* 2013, pages 1787–1796.

Marco Cornolti, Paolo Ferragina, and Massimiliano Ciaramita. A Framework for Benchmarking Entity-Annotation Systems. *In WWW* 2013, pages 249–260.

Silviu Cucerzan. Large-Scale Named Entity Disambiguation Based on Wikipedia Data. *In EMNLP-CoNLL* 2007, pages 708–716.

Aron Culotta, Michael L. Wick, and Andrew McCallum. First-Order Probabilistic Models for Coreference Resolution. *In HLT-NAACL* 2007, pages 81–87.

Hal Daumé III, Daniel Marcu. A large-scale exploration of effective global features for a joint entity detection and tracking model. *In HLT-EMNLP* 2005, pages 97–104.

Stephan Dill, Nadav Eiron, David Gibson, Daniel Gruhl, Ramanathan V. Guha, Aanat Jhingran, Tapas Kanungo, Sridhar Rajagopalan, Andrew Tomkins, John A. Tomlin, and Jason Y. Zien. SemTag and Seeker: bootstrapping the semantic web via automated semantic annotation. *In WWW* 2003, pages 178–186.

Greg Durrett and Dan Klein. Easy victories and uphill battles in coreference resolution. *In EMNLP* 2013, pages 1971–1982.

Greg Durrett and Dan Klein. A Joint Model for Entity Analysis: Coreference, Typing, and Linking. *TACL* 2014, 2:477–490.

Sourav Dutta and Gerhard Weikum. Cross-Document Co-Reference Resolution using Sample-Based Clustering with Knowledge Enrichment. *TACL* 2015, 3:15–28.

Jenny R. Finkel, Trond Grenager, and Christopher D. Manning. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. *In ACL* 2005, pages 363–370.

Jenny R. Finkel and Christopher D. Manning. Joint Parsing and Named Entity Recognition. *In EMNLP-CoNLL* 2009, pages 326–334.

Evgeniy Gabrilovich, Michael Ringgaard, and Amarnag Subramanya. FACC1: Freebase annotation of ClueWeb corpora, Version 1 (Format version 1, Correction level 0). 2013.

Chung H. Gooi and James Allan. Cross-Document Coreference on a Large Scale Corpus. *In HLT-NAACL* 2004, pages 9–16.

Peter D. Grünwald. *The Minimum Description Length Principle*. MIT University Press, 2007.

Ben Hachey, Will Radford, Joel Nothman, Matthew Honnibal, and James R. Curran. Evaluating entity linking with Wikipedia. *Artificial Intelligence Journal* 2013, 194:130–150.

Aria Haghighi and Dan Klein. Simple Coreference Resolution with Rich Syntactic and Semantic Features. *In EMNLP* 2009, pages 1152–1161.

Aria Haghighi and Dan Klein. Coreference Resolution in a Modular, Entity-Centered Model. *In HLT-NAACL* 2010, pages 385–393.

Hannaneh Hajishirzi, Leila Zilles, Daniel S. Weld, and Luke Zettlemoyer. Joint Coreference Resolution and Named-Entity Linking with Multi-pass Sieves. *In EMNLP* 2013, pages 289–299.

Johannes Hoffart, Mohamed A. Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. Robust Disambiguation of Named Entities in Text. *In EMNLP* 2011, pages 782–792.

Johannes Hoffart, Yasemin Altun, and Gerhard Weikum. Discovering Emerging Entities with Ambiguous Names. *In WWW* 2014, pages 385–396.

Nikos Hourdakis, Michalis Argyriou, Euripides G. M. Petrakis, and Evangelos E. Milios. Hierarchical Clustering in Medical Document Collections: the BIC-Means Method. *Journal of Digital Information Management* 2010, 8(2):71–77.

George Karypis and Vipin Kumar. A Fast and Highly Quality Multilevel Scheme for Partitioning Irregular Graphs. *Journal on Scientific Computing* 1999, 20(1):359–392.

Akshay Krishnamurty, Sivaraman Balakrishnan, Min Xu, and Aarti Singh. Efficient Active Algorithms for Hierarchical Clustering. *In ICML* 2012, pages 887–894.

Sayali Kulkarni, Amit Singh, Ganesh Ramakrishnan, and Soumen Chakrabarti. Collective annotation of Wikipedia entities in Web text. *In KDD* 2009, pages 457–466.

Heeyoung Lee, Yves Peirsman, Angel Chang, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. Stanford's Multi-Pass Sieve Coreference Resolution System at the CoNLL-2011 Shared Task. *In CoNLL* 2011, pages 28–34.

Heeyoung Lee, Marta Recasens, Angel Chang, Mihai Surdeanu, and Dan Jurafsky. Joint Entity and Event Coreference Resolution across Documents. *In EMNLP* 2012, pages 489–500.

Heeyoung Lee, Angel Chang, Yves Peirsman, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. Deterministic Coreference Resolution based on Entity-centric, Precision-ranked Rules. *Computational Linguistics* 2013, 39(4): 885–916.

Xiaoqiang Luo. On Coreference Resolution Performance Metrics. *In EMNLP* 2005, pages 25–32.

Gideon S. Mann and David Yarowsky: Unsupervised Personal Name Disambiguation. *In CoNLL* 2003, pages 33–40.

David Milne and Ian H. Witten. Learning to Link with Wikipedia. *In CIKM* 2008, pages 509–518.

David Nadeau and Satoshi Sekine. A survey of named entity recognition and classification. *Lingvisticae Investigationes* 2007, 30(1):3–26.

Vincent Ng. Shallow semantics for coreference resolution. *In IJCAI* 2007, pages 1689–1694.

Vincent Ng. Supervised Noun Phrase Coreference Research: The First Fifteen Years. *In ACL* 2010, pages 1396–1411.

Cheng Niu, Wei Li, and Rohini K. Srihari. Weakly Supervised Learning for Cross-document Person Name Disambiguation Supported by Information Extraction. *In ACL* 2004, article 597.

Simone P. Ponzetto and Michael Strube. Exploiting semantic role labeling, WordNet and Wikipedia for coreference resolution. *In HLT-NAACL* 2006, pages 192–199.

Karthik Raghunathan, Heeyoung Lee, Sudarshan Rangarajan, Nathanael Chambers, Mihai Surdeanu, Dan Jurafsky, and Christopher Manning. A Multi-Pass Sieve for Coreference Resolution. *In EMNLP* 2010, pages 492–501.

Altaf Rahman and Vincent Ng. Coreference Resolution with World Knowledge. *In ACL* 2011, pages 814–824.

Altaf Rahman and Vincent Ng. Ensemble-Based Coreference Resolution. *In IJCAI* 2011, pages 1884–1889.

Delip Rao, Paul McNamee, and Mark Dredze. Streaming Cross Document Entity Coreference Resolution. *In COLING* 2010, pages 1050–1058.

Lev A. Ratinov and Dan Roth. Design Challenges and Misconceptions in Named Entity Recognition. *In CoNLL* 2009, pages 147–155.

Lev A. Ratinov, Dan Roth, Doug Downey, and Mike Anderson. Local and Global Algorithms for Disambiguation to Wikipedia. *In ACL* 2011, pages 1375–1384.

Lev A. Ratinov and Dan Roth. Learning-based Multi-Sieve Co-reference Resolution with Knowledge. *In EMNLP-CoNLL* 2012, pages 1234–1244.

Gideon E. Schwarz. Estimating the Dimension of a Model. *Annals of Statistics* 1978, 6(2):461–464.

Avirup Sil and Alexander Yates. Re-ranking for Joint Named-Entity Recognition and Linking. *In CIKM* 2013, pages 2369–2374.

Sameer Singh, Michael L. Wick, and Andrew McCallum. Distantly Labeling Data for Large Scale Cross-Document Coreference. CoRR abs/1005.4298, 2010.

Sameer Singh, Amarnag Subramanya, Fernando Pereira, and Andrew McCallum. Large-Scale Cross-Document Coreference Using Distributed Inference and Hierarchical Models. *In ACL* 2011, pages 793–803.

Sameer Singh, Sebastian Reidel, Brian Martin, Jiaping Zheng, and Andrew McCallum. Joint Inference of Entities, Relations, and Coreference. *In Workshop of AKBC* 2013, pages 1–6.

Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. YAGO: a Core of Semantic Knowledge. *In WWW* 2007, pages 697–706.

Bilyana Taneva, Mouna Kacimi, and Gerhard Weikum. Finding Images of Difficult Entities in the Long Tail. *In CIKM* 2011, pages 189–194.

Mohamed A. Yosef, Johannes Hoffart, Marc Spaniol, and Gerhard Weikum. AIDA: An Online Tool for Accurate Disambiguation of Named Entities in Text and Tables. *In VLDB* 2011, 4(12):1450–1453.

Jiaping Zheng, Luke Vilnis, Sameer Singh, Jinho D. Choi, and Andrew McCallum. Dynamic knowledge-base alignment for coreference resolution. *In CoNLL* 2013, pages 153–162.