Joint Coreference Resolution and Named-Entity Linking with Multi-pass Sieves

Hannaneh Hajishirzi **Leila Zilles** **Daniel S. Weld** **Luke Zettlemoyer**
Department of Computer Science and Electrical Engineering
University of Washington
{hannaneh,lzilles,lsz,weld}@cs.washington.edu

Abstract

Many errors in coreference resolution come from semantic mismatches due to inadequate world knowledge. Errors in named-entity linking (NEL), on the other hand, are often caused by superficial modeling of entity context. This paper demonstrates that these two tasks are complementary. We introduce NECo, a new model for named entity linking and coreference resolution, which solves both problems jointly, reducing the errors made on each. NECo extends the Stanford deterministic coreference system by automatically linking mentions to Wikipedia and introducing new NEL-informed mention-merging sieves. Linking improves mention-detection and enables new semantic attributes to be incorporated from Freebase, while coreference provides better context modeling by propagating named-entity links within mention clusters. Experiments show consistent improvements across a number of datasets and experimental conditions, including over 11% reduction in MUC coreference error and nearly 21% reduction in F1 NEL error on ACE 2004 newswire data.

1 Introduction

Coreference resolution and named-entity linking are closely related problems, but have been largely studied in isolation. This paper demonstrates that they are complementary by introducing a simple joint model that improves performance on both tasks.

Coreference resolution is the task of determining when two textual mentions name the same individual. The biggest challenge in coreference resolution — accounting for 42% of errors in the state-of-the-art Stanford system — is the inability to reason effectively about background semantic knowledge (Lee et al., 2013). For example, consider the sentence in Figure 1. "President" refers to "Donald Tsang" and "the park" refers to "Hong Kong Disneyland," but automated algorithms typically lack the background knowledge to draw such inferences. Incorporating knowledge is challenging, and many efforts to do so have actually hurt performance, e.g. (Lee et al., 2011; Durrett and Klein, 2013).

[Michael Eisner]$_1$ and [Donald Tsang]$_2$ announced the grand opening of [[Hong Kong]$_3$ Disneyland]$_4$ yesterday. [Eisner]$_1$ thanked [the President]$_2$ and welcomed [fans]$_5$ to [the park]$_4$.

Figure 1: A text passage illustrating interactions between coreference resolution and NEL.

Named-entity linking (NEL) is the task of matching textual mentions to corresponding entities in a knowledge base, such as Wikipedia or Freebase. Such links provide rich sources of semantic knowledge about entity attributes — Freebase includes *president* as Tsang's title and `Disneyland` as having the attribute *park*. But NEL is itself a challenging problem, and finding the correct link requires disambiguating based on the mention string and often non-local contextual features. For example, "Michael Eisner" is relatively unambiguous but the isolated mention "Eisner" is more challenging. However, these mentions could be clustered with a coreference model, allowing for improved NEL through link propagation from the easier mentions.

Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 289–299,
Seattle, Washington, USA, 18-21 October 2013.

We present NECO, a new algorithm for jointly solving named entity linking and coreference resolution. Our work is related to that of Ratinov and Roth (2012), which also uses knowledge derived from an NEL system to improve coreference. However, NECO is the first joint model we know of, is purely deterministic with no learning phase, does automatic mention detection, and improves performance on both tasks.

NECO extends the Stanford's sieve-based model, in which a high recall mention detection phase is followed by a sequence of cluster merging operations ordered by decreasing precision (Raghunathan et al., 2010; Lee et al., 2013). At each step, it merges two clusters only if all available information about their respective entities is consistent. We use NEL to increase recall during the mention detection phase and introduce two new cluster-merging sieves, which compare the Freebase attributes of entities. NECO also improves NEL by initially favoring high precision linking results and then propagating links and attributes as clusters are formed.

In summary we make the following contributions:

- We introduce NECO, a novel, joint approach to solving coreference and NEL, demonstrating that these tasks are complementary by achieving joint error reduction.
- We present experiments showing improved performance at coreference resolution, given both gold and automatic mention detection: e.g., 6.2 point improvement in MUC recall on ACE 2004 newswire text and 3.1 point improvement in MUC precision the CoNLL 2011 test set.
- NECO also leads to better performance at named-entity linking, given both gold and automatic linking, improving F1 from 61.7% to 69.2% on a newly labeled test set.[1]

2 Background

We make use of existing models for coreference resolution and named entity linking.

[1]Our corpus and the source code for NECO can be downloaded from `https://www.cs.washington.edu/research-projects/nlp/neco`.

2.1 Coreference Resolution

Coreference resolution is the the task of identifying all text spans (called *mentions*) that refer to the same entity, forming mention clusters.

Stanford's Sieve Model is a state-of-the-art coreference resolver comprising a pipeline of "sieves" that merge coreferent mentions according to deterministic rules. Mentions are automatically predicted by selecting all noun phrases (NP), pronouns, and named entities. Each sieve either merges a cluster to its single best antecedent from a list of previous clusters, or declines to merge.

Higher precision sieves are applied earlier in the pipeline according to the following order, looking at different aspects of the text, including: (1) speaker identification, (2-3) exact and relaxed string matches between mentions, (4) precise constructs, including appositives, acronyms and demonyms, (5-9) different notions of strict and relaxed head matches between mentions, and finally (10) a number of syntactic and distance cues for pronoun resolution.

2.2 Named Entity Linking

Named-entity linking (NEL) is the task of identifying mentions in a text and linking them to the entity they name in a knowledge base, usually Wikipedia. NECO uses two existing NEL systems: GLOW (Ratinov et al., 2011) and WikipediaMiner (Milne and Witten, 2008).

WikipediaMiner links mentions based on a notion of semantic similarity to Wikipedia pages, considering all substrings up to a fixed length. Since there are often many possible links, it disambiguates by choosing the entity whose Wikipedia page is most semantically related to the nearby context of the mention. The semantic scoring function includes n-gram statistics and also counts shared links to other unambiguous mentions in the text.

GLOW finds mentions by selecting all the NPs and named entities in the text. Linking is framed as an integer linear programming optimization problem that takes into account using similar local constraints but also includes global constraints such as entity link co-occurrence.

Both systems return confidence values. To maintain high precision, NECO uses an ensemble of

- Let $Exemplar(c)$ be a representative mention of the cluster c, computed as defined below
- Let c_j be an antecedent cluster of c_i if c_j has a mention which is before the first mention of c_i
- Let $l(m)$ be a Wikipedia page linked to mention m or $\varnothing$ if there is no link
- Let $l(c)$ be a Wikipedia page linked to mention $Exemplar(c)$ or $\varnothing$ if there is no link

1. **Initialize Linked Mentions:**
   (a) Let $M_{NEL} = \{m_i \mid i = 1 \dots p\}$ be the NEL output mentions, m_i, each with a link $l(m_i)$
   (b) Let $M_{CR} = \{m_i \mid i = 1 \dots q\}$ be the mentions m_i from coreference mention detection
   (c) Let $M \leftarrow M_{CR} \cup M_{NEL}$ (Sec. 3.1)
   (d) Update entity links for all $m \in M$ and prune M (Sec. 3.2)
   (e) Extract attributes from Wikipedia and Freebase for all $m \in M$ (Sec. 3.3)
   (f) Let $C \leftarrow M$ be singleton mention clusters where $Exemplar(c_i) = m_i$, $l(c_i) = l(m_i)$
2. **Merge Clusters:** For every sieve S (including NEL sieves, Sec. 3.6) and cluster $c_i \in C$
   (a) For every cluster $c_j, j = [i-1 \dots 1]$ (traverse the preceding clusters in reverse order)
      i. **NEL constraints**: Prevent merge if $l(c_i) \neq l(c_j)$ (Sec. 3.4)
      ii. If all rules of sieve S are satisfied for clusters c_i and c_j
         A. $c_k \leftarrow Merge(c_i, c_j)$, including entity link and attribute updates (Sec. 3.5)
         B. $C \leftarrow C \cup \{c_k\} \setminus \{c_i, c_j\}$
3. **Output:** Coreference clusters C and linked Wikipedia pages $l(c_i) \forall c_i \in C$

Figure 2: NECO: A joint algorithm for named-entity linking and coreference resolution.

GLOW and WikipediaMiner, selecting only high confidence links.

3 Joint Coreference and Linking

We introduce a joint model for coreference resolution and NEL. Building on the Stanford sieve architecture, our algorithm incrementally constructs clusters of mentions using deterministic coreference rules under NEL constraints.

Figure 2 presents the complete algorithm. The input to NECO is a document and the output is a set C of coreference clusters, with links $l(c)$ to Wikipedia pages for a subset of the clusters $c \in C$. Step 1 detects mentions, merging the outputs of the base systems (Sec. 3.1). Step 2 repeatedly merges coreference clusters, while ensuring that NEL constraints (Sec. 3.4) are satisfied. It uses the original Stanford sieves and also two new NEL-informed sieves (Sec. 3.6). NEL links are propagated to new clusters as they are formed (Sec. 3.5).

3.1 Mention Detection

In Steps 1(a-c) in Fig. 2, NECO combines mentions from the base coreference and NEL systems.

Let M_{CR} be the set of mentions returned by using Stanford's rule-based mention detection algorithm (Lee et al., 2013). Let M_{NEL} be the set of mentions output by the two NEL systems. NECO creates an initial set of mentions, M, by taking the union of all the mentions in M_{NEL} and M_{CR}. In practice, taking the union increases diversity in the mention pool. For example, it is often the case that M_{NEL} will include sub-phrases such as "Suharto" when they are part of a larger mention "ex-dictator Suharto" that is detected in M_{CR}.

3.2 Mention Entity Links and Pruning

Step 1(d) in Fig. 2 assigns Wikipedia links to a subset of the detected mentions.

For mentions m output by the base NEL systems, we assign an *exact* link $l(m)$ if the entire mention span is linked. Mentions m' that differ from an exact linked mention m by only a pre- or post-fix stop word are similarly assigned exact links $l(m') = l(m)$. For example, the mention "the president" will be assigned the same link as "president" but "The governor of Alaska Sarah Palin" would not be assigned an exact link to `Sarah Palin`.

For mentions m' that do not receive an exact link, we assign a *head* link $h(m')$ if the head word[2] m has been linked, by setting $h(m') = l(m)$. For instance, the head link for the mention "President Clinton" (with "Clinton" as head word) will be the Wikipedia title of `Bill Clinton`. We use head links for the Relaxed NEL sieve (Sec. 3.6).

Next, we define $L(m)$ to be the set con-

[2] A head word is assigned to every mention with the Stanford parser head finding rules (Klein and Manning, 2003).

country	*president*	*city*	*area*
company	*state*	*region*	*location*
place	*agency*	*power*	*unit*
body	*market*	*park*	*province*
manager	*organization*	*owner*	*trial*
site	*prosecutor*	*attorney*	*county*
senator	*stadium*	*network*	*building*
attraction	*government*	*department*	*person*
origin	*plant*	*airport*	*kingdom*
capital	*operation*	*author*	*period*
nominee	*candidate*	*film*	*venue*

Figure 3: The most commonly used fine-grained attributes from Freebase and Wikipedia (out of over 500 total attributes).

taining $l(m)$ and $l(m')$ for all sub-phrases m' of m. We add the sub-phrase links only if their confidence is higher than the confidence for $l(m)$. For instance, assuming appropriate confidence values, $L(m)$ would include the pages for {`List of governors of Alaska`, `Alaska`, `Sarah Palin`} given the mention "The governor of Alaska Sarah Palin." We will use $L(m)$ for NEL constraints and filtering (Sec. 3.4).

After updating the entity links for all mentions, NECo prunes spurious mentions that begin or end with a stop word where the remaining subexpression of the mention exists in M. It also removes time expressions and numbers from M if they are not included in M_{NEL}.

3.3 Mention Attributes

Step 1(e) in Fig. 2 also assigns attributes for a mention m linked to Wikipedia page $l(m)$, at both *coarse* and *fine-grained* levels, based on information from the Freebase entry corresponding to exact link $l(m)$ or head link $h(m)$.

The coarse attributes include gender, type, and NER classes such as PERSON, LOCATION, and ORGANIZATION. These attributes are part of the original Stanford coreference system and are used to avoid merging conflicting clusters. We use the Freebase values for these attributes when available. For instance, if the linked entity contains the Freebase type *location* or *organization*, we include the coarse type to LOCATION or ORGANIZATION respectively. In order to account for both links to specific people (Barack Obama) and generic links to positions held by people (President), we include the type PERSON if the linked entity has any of the Freebase types *person*, *job_title*, or *government_office_or_title*. If no coarse Freebase types are available for an attribute, we default to predicted NER classes.

We add fine-grained attributes from Freebase and Wikipedia by importing additional type information. We use all of the Freebase *notable types*, a set of hundreds of commonly used Freebase types, ranging from *us_president* to *tropical_cyclone* and *synthpop_album*. We also include all of the Wikipedia categories, on average six per entity. For example, the mention "Indonesia" is assigned fine-grained attributes such as *book_subject*, *military_power*, and *olympic_participating_country*. Since many of these fine-grained attributes are extremely specific, we use the last word of each attribute to define an additional fine-grained attribute (see Fig. 3). These fine-grained attributes are used in the Relaxed NEL sieve (Sec. 3.6).

3.4 NEL Constraints

While applying sieves to merge clusters in Figure 2 Step 2(a), NECo uses NEL constraints to eliminate some otherwise acceptable merges.

We avoid merging inconsistent clusters that link to different entities. Clusters c_i and c_j are inconsistent if both are linked (i.e., both clusters have non-null entity assignments) and $l(c_i) \neq l(c_j)$ or $h(c_i) \neq h(c_j)$. Also, in order to consider an antecedent cluster c as a merge candidate, we require a pair of entities in the set of linked entities $L(c)$ to be related to one another in Freebase. Two entities are related in Freebase if they both appear in a relation; for example, `Bill Clinton` and `Arkansas` are related because Bill Clinton has a "governor-of" relation with Arkansas.

3.5 Merging Clusters and Update Entity Links

When two clusters c_i and c_j are merged to form a new cluster c_k, the entity link information $L(c_k)$, $l(c_k)$, and $h(c_k)$ must be updated (Step 2 of Fig. 2). We set $L(c_k)$ to the union of the linked entities found in $l(c_i)$ and $l(c_j)$ and merge coarse attributes at this point.

In order to set the exact and head entity links $l(c_k)$ and $h(c_k)$, we use the exemplar mention

$Exemplar(c_k)$ that denotes the most representative mention of the cluster. $Exemplar(c)$ is selected according to a set of rules in the Stanford system, based on textual position and mention type (proper noun vs. common). We augment this function by considering information from exact and head entity links as well. Mentions appearing earlier in text, proper mentions, and mentions that have exact or head named-entity links are preferred to those which do not. Given exemplars, we set $l(c_k) = l(Exemplar(c_k))$ and $h(c_k) = h(Exemplar(c_k))$.

3.6 NEL Sieves

Finally, we introduce two new sieves that use NEL information at the beginning and end of the Stanford sieves pipeline in the merging stage (Step 2 of Fig. 2).

Exact NEL sieve The Exact NEL sieve merges two clusters c_i and c_j if both are linked and their links match, $l(c_i) = l(c_j)$. For example, all mentions that have been linked to `Barack Obama` will become members of the same coreference cluster. Because the Exact NEL sieve has high precision, we place it at the very beginning of the pipeline.

Relaxed NEL sieve The Relaxed NEL sieve uses fine-grained attributes of the linked mentions to merge proper nouns with common nouns when they share attributes. For example, this sieve is able to merge the proper mention "Disneyland" with the "the mysterious park", because *park* is one of the fine-grained attributes assigned to `Disneyland`.

More formally, let $m_i = Exemplar(c_i)$ and $m_j = Exemplar(c_j)$. For every common noun mention m_i, we merge c_i with an antecedent cluster c_j if (1) m_j is a linked proper noun, (2) if m_i or the title of its linked Wikipedia page is in the list of fine-grained attributes of m_j, or (3) if $h(m_j)$ is related to the head link $h(m_i)$ according to Freebase as defined above.

Because this sieve has low precision, we only allow merges between mentions that have a maximum distance of three sentences between one another. We add the Relaxed NEL sieve near the end of the pipeline, just before pronoun resolution.

4 Experimental Setup

Core Components and Baselines The Stanford sieve-based coreference system (Lee et al., 2013), the GLOW NEL system (Ratinov et al., 2011), and WikipediaMiner (Milne and Witten, 2008) provide core functionality for our joint model, and are also the state-of-the-art baselines against which we measure performance.

Parameter Settings Based on performance on the development set, we set the GLOW's confidence parameter to 1.0 and WikipediaMiner's to 0.4 to assure high-precision NEL. We also optimized for the set of fine-grained attributes to import from Wikipedia and Freebase, and the best way to incorporate the NEL constraints into the sieve architecture.

Datasets We report results on the following three datasets: ACE2004-NWIRE, CONLL2011, and ACE2004-NWIRE-NEL. ACE2004-NWIRE, the newswire subset of the ACE 2004 corpus (NIST, 2004), includes 128 documents. The CONLL2011 coreference dataset includes text from five different domains: broadcast conversation (BC), broadcast news (BN), magazine (MZ), newswire (NW), and web data (WB) (Pradhan et al., 2011). The broadcast conversation and broadcast news domains consist of transcripts, whereas magazine and newswire contain more standard written text. The development data includes 303 documents and the test data includes 322 documents.

We created ACE2004-NWIRE-NEL by taking a subset of ACE2004-NWIRE and annotating with gold-standard entity links. We segment and link all the expressions in text that refer to Wikipedia pages, allowing for nested linking. For instance, both the phrase "Hong Kong Disneyland," and the sub-phrase "Hong Kong" are linked. This dataset includes 12 documents and 350 linked entities.

Metrics We evaluate our system using MUC (Vilain et al., 1995), B^3 (Bagga and Baldwin, 1998), and pairwise scores. MUC is a link-based metric which measures how many clusters need to be merged to cover the gold clusters and favors larger clusters; B^3 computes the proportion of intersection between predicted and gold clusters for every mention and favors singletons (Recasens and Hovy, 2010). We computed the scores using the Stanford

Method	MUC			B^3			Pairwise		
	P	R	F1	P	R	F1	P	R	F1
Stanford Sieves	39.9	46.2	42.8	67.9	71.8	69.8	44.2	29.7	35.6
NECo	**46.8**	**52.5**	**49.5**	70.4	**72.6**	**71.5**	**51.5**	**34.6**	**41.4**
No NEL Mentions	46.1	48.3	47.2	71.4	70.0	70.9	49.7	30.9	38.1
No Mention Pruning	43.6	45.6	44.6	70.5	69.9	70.2	46.2	29.4	35.9
No Attributes	45.9	47.4	46.6	**71.8**	69.7	70.7	48.6	27.0	34.7
No Constraints	42.3	49.3	45.5	68.3	72.3	70.2	44.2	28.6	34.7

Table 1: Coreference results on ACE2004-NWIRE with predicted mentions and automatic linking.

coreference software for ACE2004 and using the CoNLL scorer for the CoNLL 2011 dataset.

5 Experimental Results

We first look at NECo's performance at coreference resolution and then evaluate its ability at NEL.

5.1 Coref. Results with Predicted Mentions

Overall System Performance on ACE Data Table 1 shows NECo's performance at coreference resolution on ACE-2004 compared to the Stanford sieve implementation (Lee et al., 2013). The table shows that NECo has both significantly improved precision and recall compared to the Stanford baseline, across all metrics. We generally observe larger gains in MUC due to better mention detection and the Relaxed NEL Sieve.

Contribution of System Components Table 1 also details the performance of four variants of our system that ablate various components and features. Specifically, we consider the following cases:

- **No NEL Mentions:** We discard additional mentions, M_{NEL}, provided by NEL (Sec. 3.1). This increases B^3 precision at the expense of recall. Inspection shows that some of the errors introduced by M_{NEL} are actually due to correctly linked entities that were not annotated as mentions in the dataset, but also some improperly linked mentions.
- **No Mention Pruning:** We disable the initial step of updating mention boundaries and removing spurious mentions (Sec. 3.2). As expected, removing this step drops precision and recall significantly, even compared to the No NEL Mentions variant.
- **No Attributes:** Ablating coarse and fine-grained attributes (Sec. 3.3) drops F1 and recall measures across all metrics. To understand this effect, note that NECo uses attributes in two different settings. Updating coarse attributes tends to increase precision because it prevents dangerous merges, such as merging "Staples" with the mention "it" in a situation when "Staples" refers to the person entity `Todd Staples`. Fine-grained attributes also help with recall, when merging a specific name of an entity with a mention that uses a more general term; for instance, "Hong Kong Disneyland" can be merged with "the mysterious park" because "park" is a fine-grained attribute for `Disneyland`. However, when fine-grained attributes are used, precision sometimes drops (e.g., when "president" might merge with "Bush" when it should really merge with "Clinton").
- **No NEL Constraints:** Removing these constraints (Sec. 3.4) drops precision dramatically leading to drop in F1. In the case of incorrect linking, however, NEL constraints can affect recall. For instance, NEL constraints might prevent merging "Staples" with "Todd Staples" if the former were linked to the company and the latter to the politician.

Overall System Performance on CoNLL Data We also compare our full system (with added NEL sieves, constraints, and mention pruning[3]) with the Stanford sieve coreference system on CoNLL data

[3]Due to CoNLL annotation guidelines, a named entity is added to the mention list if it is *not* inside a larger mention with an exact named entity link.

Category: Method	MUC			B^3		
	P	R	F1	P	R	F1
BC: NECo	**62.1**	64.7	**63.4**	**69.8**	57.8	**63.2**
BC: Stanford Sieves	60.9	**65.0**	62.9	69.2	**58.0**	63.1
BN: NECo	**69.3**	**59.4**	**64.0**	78.8	**60.8**	**68.6**
BN: Stanford Sieves	68.0	58.9	63.1	**79.0**	60.2	68.3
MZ: NECo	**67.6**	62.9	**65.2**	**78.4**	61.1	68.7
MZ: Stanford Sieves	66.0	**63.4**	64.9	77.9	**61.5**	68.7
NW: NECo	**62.0**	**54.5**	**58.0**	74.9	**57.4**	**65.0**
NW: Stanford Sieves	60.0	54.2	56.9	**75.3**	57.0	64.9

Table 3: Coreference results on the individual categories of CoNLL 2011 development data. (BC=broadcast conversation, BN=broadcast news, MZ=magazine, NW=newswire)

Method	MUC			B^3		
	P	R	F1	P	R	F1
Development Data						
NECo	**64.1**$^+$	**59.4**	**61.7**$^+$	74.7	**58.7**	**65.7**
Stanford	62.7	59.0	60.8	**74.8**	58.3	65.6
NECo*	**56.4**$^+$	50.0	**53.0**$^+$	**72.6**	**51.6**	**60.3**
Stanford*	53.5	50.0	51.6	71.8	51.3	59.9
Test Data						
NECo	**61.2**$^+$	58.4	**59.8**$^+$	**72.2**	**56.4**	**63.3**
Stanford	59.2	**58.8**	59.0	71.3	56.1	62.8
NECo*	**55.1**$^+$	51.7	**53.3**$^+$	**70.0**	50.8	**58.8**
Stanford*	52.0	**52.3**$^+$	52.1	68.9	50.8	58.5

Table 2: Coreference results on CoNLL 2011 development and test data, using predicted mentions. Rows denoted with * indicate runs using the fully automated Stanford CoreNLP pipeline rather than the predicted annotations provided with the CoNLL data. Given the relatively close results, we ran the Mann-Whitney U test for this table; values with the $^+$ superscript are significant with $p < 0.05$.

(Table 2). We ran NECo and the baseline in two settings: in the first, we use the standard predicted annotations (for POS, parses, NER, and speaker tags) provided with the CoNLL data, and in the second, we use the automated Stanford CoreNLP pipeline to predict this information. On both the development and test sets, we gain about 1 point in MUC F1 as well as a smaller improvement in B^3. Closer inspection indicates that our system increases precision primarily due to mention pruning and NEL constraints. Due to the differences in mention annotation guidelines between ACE and CoNLL, performance on ACE benefits more from improved mention detection from NEL. Moreover, the ACE corpus is all newswire text, which contains more entities that can benefit from linking. CoNLL, on the other hand, contains a wider variety of texts, some of which do not mention many named entities in Wikipedia.

To examine the performance of our system on the different domains covered by the CoNLL data, we also test our system on each domain separately (Table 3). We found NEL provided the biggest improvement for the news domains, broadcast news (BN) and newswire (NW). These domains especially benefit from the improved mention detection and pruning provided by NEL, and strong linking benefitted both precision and recall in these domains. We found that the magazine (MZ) section of the corpus benefited the least from NEL, as there were relatively few entities that our NEL systems were able to connect to Wikipedia.

5.2 Coreference Results with Gold Linking

Some of the errors introduced in our system are due to incorrect or incomplete links discovered by the automatic linking system. To assess the effect of NEL performance on NECo, we tested on a portion of ACE2004-NWIRE dataset for which we hand-labeled correct links for the gold and predicted mentions. "NECo + Gold NEL" denotes a version of our system which uses gold links instead of those predicted by NEL. As shown in Table 4, gold linking significantly improves the performance of our system across all measures. This suggests that further work to improve automatic NEL may have substantial reward.

Gold linking improves precision for two main rea-

Method	MUC			B^3			Pairwise		
	P	R	F1	P	R	F1	P	R	F1
Gold Mentions									
NECo + Gold NEL	**85.8**	**75.5**	**80.3**	**91.4**	**81.2**	**86.0**	89.1	**68.0**	**77.1**
NECo	84.6	74.0	78.9	90.5	80.4	85.2	83.9	66.0	73.9
Stanford Sieves	84.5	72.2	77.8	89.9	77.7	83.4	**89.9**	57.3	68.1
Predicted Mentions									
NECo + Gold NEL	**56.4**	**58.8**	**57.5**	**78.2**	**78.3**	**78.3**	**68.0**	**54.3**	**60.4**
NECo	51.3	53.5	52.4	76.5	76.4	76.5	61.2	45.6	52.2
Stanford Sieves	43.9	46.4	45.1	74.4	74.2	74.3	51.3	36.1	42.4

Table 4: Coreference results on ACE2004-NWIRE-NEL with gold and predicted mentions and gold or automatic linking.

Method	MUC			B^3			Pairwise		
	P	R	F1	P	R	F1	P	R	F1
NECo	**85.0**	**76.6**	**80.6**	**87.6**	**76.4**	**81.6**	79.3	**56.1**	**65.8**
Stanford Sieves	84.6	75.1	79.6	87.3	74.1	80.2	**79.4**	50.1	61.4
Haghighi and Klein (2009)	77.0	75.9	76.5	79.4	74.5	76.9	66.9	49.2	56.7
Poon and Domingos (2008)	71.3	70.5	70.9	-	-	-	62.6	38.9	48.0
Finkel and Manning (2008)	78.7	58.5	67.1	86.8	65.2	74.5	76.1	44.2	55.9

Table 5: Coreference results on ACE2004-NWIRE with gold mentions and automatic linking.

sons. First, it reduces the coreference errors caused by incorrect NEL links. For instance, gold linking replaces the erroneous link generated by our NEL systems for "Nasser al-Kidwa" to the correct Wikipedia entity. As another example, two mentions of "Rutgers" will not be merged if one links to the university and the other links to their football team. Second, gold linking leads to better mention detection and better linked mentions. For instance, under gold linking, the whole mention, "The governor of Alaska, Sarah Palin," is linked to the politician, while automatic linking systems only link the substring containing her name, "Sarah Palin." Still, gold NEL cannot compensate for all coreference errors in cases of generic or unlinked entities.

5.3 Coreference Results with Gold Mentions

Many of the previous papers evaluate coreference resolution assuming gold mentions so we also run under that condition (Table 5) using ACE2004-NWIRE data. As the table shows, with gold mentions our system outperforms Haghighi and Klein (2009), Poon and Domingos (2008), Finkel and Manning (2008) and the Stanford sieve algorithm across all metrics. Our method shows a relatively smaller gain in precision, because this condition adds no benefit to our technique of using NEL information for pruning mentions.

5.4 Improving Named Entity Linking

While our previous experiments show that named-entity linking can improve coreference resolution, we now address the question of whether coreference techniques can help NEL. We compare NECo with a baseline ensemble[4] composed of GLOW (Ratinov et al., 2011) and WikipediaMiner (Milne and Witten, 2008) on our ACE2004-NWIRE-NEL dataset (Table 6). Our system gains about 8% in absolute recall and 5% in absolute precision. For instance, our system correctly adds links from "Bullock" to the entity `Sandra Bullock` because coreference resolution merges two mentions. In another example, it correctly links "company" to `Nokia`. Overall, there is a 21% relative reduction in F1 error.

[4]We take the union of all the links returned by GLOW and WikipediaMiner, but if they link a mention to two different entities, we use only the output of WikipediaMiner.

Method	F1	Precision	Recall
NECo	**70.6**	**72.0**	**69.2**
Baseline NEL	64.4	67.4	61.7

Table 6: NEL performance of our system and the ensemble baseline linker on ACE2004-NWIRE-NEL.

5.5 Error Analysis

We analyzed 90 precision and recall errors and present our findings in Table 7. Spurious mentions accounted for the majority of non-semantic errors. Despite the improvements that come from NEL, a large portion of coreference errors can still be attributed to incomplete semantic information, including precision errors caused by incorrect linking. For instance, the mention "Disney" sometimes refers to the company, and other times refers to the amusement park; however, the NEL systems we used had difficulty disambiguating these cases, and NECo often incorrectly merges such mentions. Overly general fine-grained attributes caused precision errors in cases where many proper noun mentions were potential antecedents for a common noun. Although attributes such as *country* are useful for resolving a generic "country" mention, this information is insufficient when two distinct mentions such as "China" and "Russia" both have the *country* attribute.

However, many recall errors are also caused by the lack of fine-grained attributes. Finding the ideal set of fine-grained attributes remains an open problem.

6 Related Work

Coreference resolution has a fifty year history which defies brief summarization; see Ng (2010) for a recent survey. Section 2.1 described the Stanford multi-pass sieve algorithm, which is the foundation for NECo.

Earlier coreference resolution systems used shallow semantics and pioneered knowledge extraction from online encyclopedias (Ponzetto and Strube, 2006; Daumé III and Marcu, 2005; Ng, 2007). Some recent work shows improvement in coreference resolution by incorporating semantic information from Web-scale structured knowledge bases. Haghighi and Klein (2009) use a rule-based system to extract fine-grained attributes for mentions by analyzing precise constructs (e.g., appositives) in Wikipedia articles. Subsequently, Haghighi and Klein (2010) used a generative approach to learn entity types from an initial list of unambiguous mention types. Bansal and Klein (2012) use statistical analysis of Web n-gram features including lexical relations.

Rahman and Ng (2011) use YAGO to extract type relations for all mentions. These methods incorporate knowledge about all possible meanings of a mention. If a mention has multiple meanings, extraneous information might be associated with it. Zheng et al. (2013) use a ranked list of candidate entities for each mention and maintain the ranked list when mentions are merged. Unlike previous work, our method relies on NEL systems to disambiguate possible meanings of a mention and capture high-precision semantic knowledge from Wikipedia categories and Freebase notable types.

Ratinov and Roth (2012) investigated using NEL to improve coreference resolution, but did not consider a joint approach. They extracted attributes from Wikipedia categories and used them as features in a learned mention-pair model, but did not do mention detection. Unfortunately, it is difficult to compare directly to the results of both systems, since they reported results on portions of ACE and CoNLL datasets using gold mentions. However, our approach provides independent evidence for the benefit of NEL, and joint modeling in particular, since it outperforms the state-of-the-art Stanford sieve system (winner of the CoNLL 2011 shared task (Pradhan et al., 2011)) and other recent comparable approaches on benchmark datasets.

Our work also builds on a long trajectory of work in named entity resolution stemming from SemTag (Dill et al., 2003). Section 2.2 discussed GLOW and WikipediaMiner (Ratinov et al., 2011; Milne and Witten, 2008). Kulkarni et al. (2009) present an elegant collective disambiguation model, but do not exploit the syntactic nuances gleaned by within-document coreference resolution. Hachey et al. (2013) provide an insightful summary and evaluation of different approaches to NEL.

7 Conclusions

Observing that existing coreference resolution and named-entity linking have complementary strengths

Error Type	Percentage	Example
Extra mentions	31.1	**The other thing Paula really important** is that they talk a lot about the fact ...
Pronoun	27.7	However , [*all 3 women gymnasts , taking part in the internationals for the first time*], performed well , because they had strong events and their movements had difficulty .
Contextual semantic	16.6	[*The Chinese side*] hopes that each party concerned continues to make constructive efforts to ...Considering the requirements of the Korean side , ... the Chinese government decided to ...
NEL semantic	13.3	The most important thing about **Disney** is that it is a global brand. ... The subway to Disney has already been constructed.
Attributes	11.1	The **Hong Kong** government turned over to Disney Corporation [*200 hectares of land ...*]. ... this area has become a prohibited zone in Hong Kong.

Table 7: Examples of different error categories and the relative frequency of each. For every example, the mention to be resolved is underlined, and the correct antecedent is *italicized*. For precision errors, the wrongly merged mention is **bolded**. For recall errors, the missed mention is surrounded by [brackets].

and weaknesses, we present a joint approach. We introduce NECO, a novel algorithm which solves the problems *jointly*, demonstrating improved performance on both tasks.

We envision several ways to improve the joint model. While the current implementation of NECO only introduces NEL once, we could also integrate predictions with different levels of confidence into different sieves. It would be interesting to more tightly integrate the NEL system so it operates on clusters rather than individual mentions — after each sieve merges an unlinked cluster, the algorithm would retry NEL with the new context information. NECO uses a relatively modest number of Freebase attributes. While using more semantic knowledge holds the promise of increased recall, the challenge is maintaining precision. Finally, we would also like to explore the extent to which a joint probabilistic model (e.g., (Durrett and Klein, 2013)) might be used to learn how to best make this tradeoff.

8 Acknowledgements

The research was supported in part by grants from DARPA under the DEFT program through the AFRL (FA8750-13-2-0019) and the CSSG (N11AP20020), the ONR (N00014-12-1-0211), and the NSF (IIS-1115966). Support was also provided by a gift from Google, an NSF Graduate Research Fellowship, and the WRF / TJ Cable Professorship. The authors thank Greg Durrett, Heeyoung Lee, Mitchell Koch, Xiao Ling, Mark Yatskar, Kenton Lee, Eunsol Choi, Gabriel Schubiner, Nicholas FitzGerald, Tom Kwiatkowski, and the anonymous reviewers for helpful comments and feedback on the work.

References

Amit Bagga and Breck Baldwin. 1998. Algorithms for scoring coreference chains. In *International Conference on Language Resources and Evaluation Workshop on Linguistics Coreference*.

Mohit Bansal and Dan Klein. 2012. Coreference semantics from web features. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*.

Hal Daumé III and Daniel Marcu. 2005. A large-scale exploration of effective global features for a joint entity detection and tracking model. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*.

Stephen Dill, Nadav Eiron, David Gibson, Daniel Gruhl, R. Guha, Anant Jhingran, Tapas Kanungo, Sridhar Rajagopalan, Andrew Tomkins, John A. Tomlin, and Jason Y. Zien. 2003. SemTag and Seeker: bootstrapping the semantic web via automated semantic annotation. In *Proceedings of the 12th International Conference on World Wide Web*.

Greg Durrett and Dan Klein. 2013. Easy victories and uphill battles in coreference resolution. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Jenny Rose Finkel and Christopher D. Manning. 2008. Enforcing transitivity in coreference resolution. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies*.

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

Aria Haghighi and Dan Klein. 2009. Simple coreference resolution with rich syntactic and semantic features. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Aria Haghighi and Dan Klein. 2010. Coreference resolution in a modular, entity-centered model. In *Human Language Technologies: Annual Conference of the North American Chapter of the Association for Computational Linguistics*.

Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*.

Sayali Kulkarni, Amit Singh, Ganesh Ramakrishnan, and Soumen Chakrabarti. 2009. Collective annotation of Wikipedia entities in Web text. In *Proceedings of the 2009 Conference on Knowledge Discovery and Data Mining*.

Heeyoung Lee, Yves Peirsman, Angel Chang, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2011. Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 shared task. In *Proceedings of the Conference on Computational Natural Language Learning*.

Heeyoung Lee, Angel Chang, Yves Peirsman, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2013. Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Computational Linguistics*, 39(4).

Dan Milne and Ian H. Witten. 2008. Learning to link with Wikipedia. In *Proceedings of the ACM Conference on Information and Knowledge Management*.

Vincent Ng. 2007. Shallow semantics for coreference resolution. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*.

Vincent Ng. 2010. Supervised noun phrase coreference research: The first fifteen years. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*.

NIST. 2004. The ACE 2004 evaluation planXPToolkit architecture.

Simone Paolo Ponzetto and Michael Strube. 2006. Exploiting semantic role labeling, Wordnet and Wikipedia for coreference resolution. In *Proceedings of the North American Association for Natural Language Processing on Human Language Technologies*.

Hoifung Poon and Pedro Domingos. 2008. Joint unsupervised coreference resolution with Markov logic. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Sameer Pradhan, Lance Ramshaw, Mitchell Marcus, Martha Palmer, Ralph Weischedel, and Nianwen Xue. 2011. CoNLL-2011 shared task: modeling unrestricted coreference in OntoNotes. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*.

Karthik Raghunathan, Heeyoung Lee, Sudarshan Rangarajan, Nathanael Chambers, Mihai Surdeanu, Dan Jurafsky, and Christopher Manning. 2010. A multi-pass sieve for coreference resolution. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Altaf Rahman and Vincent Ng. 2011. Coreference resolution with world knowledge. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*.

Lev Ratinov and Dan Roth. 2012. Learning-based multi-sieve co-reference resolution with knowledge. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Lev Ratinov, Dan Roth, Doug Downey, and Mike Anderson. 2011. Local and global algorithms for disambiguation to Wikipedia. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*.

Marta Recasens and Eduard Hovy. 2010. Coreference resolution across corpora: languages, coding schemes, and preprocessing information. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. In *Proceedings of the 6th conference on Message Understanding*.

Jiaping Zheng, Luke Vilnis, Sameer Singh, Jinho D. Choi, and Andrew McCallum. 2013. Dynamic knowledge-base alignment for coreference resolution. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*.