Design Challenges for Entity Linking

Xiao Ling  Sameer Singh  Daniel S. Weld
University of Washington, Seattle WA
{xiaoling,sameer,weld}@cs.washington.edu

Abstract

Recent research on entity linking (EL) has introduced a plethora of promising techniques, ranging from deep neural networks to joint inference. But despite numerous papers there is surprisingly little understanding of the state of the art in EL. We attack this confusion by analyzing differences between several versions of the EL problem and presenting a simple yet effective, modular, unsupervised system, called VINCULUM, for entity linking. We conduct an extensive evaluation on nine data sets, comparing VINCULUM with two state-of-the-art systems, and elucidate key aspects of the system that include mention extraction, candidate generation, entity type prediction, entity coreference, and coherence.

1 Introduction

Entity Linking (EL) is a central task in information extraction — given a textual passage, identify entity mentions (substrings corresponding to world entities) and link them to the corresponding entry in a given Knowledge Base (KB, e.g. Wikipedia or Freebase). For example,

JetBlue begins direct service between Barnstable Airport and JFK International.

Here, “JetBlue” should be linked to the entity KB:JetBlue, “Barnstable Airport” to KB:Barnstable Municipal Airport, and “JFK International” to KB:John F. Kennedy International Airport.1 The links not only provide semantic annotations to human readers but also a machine-consumable representation of the most basic semantic knowledge in the text. Many other NLP applications can benefit from such links, such as distantly-supervised relation extraction (Craven and Kumlien, 1999; Riedel et al., 2010; Hoffmann et al., 2011; Koch et al., 2014) that uses EL to create training data, and some coreference systems that use EL for disambiguation (Hajishirzi et al., 2013; Zheng et al., 2013; Durrett and Klein, 2014). Unfortunately, in spite of numerous papers on the topic and several published data sets, there is surprisingly little understanding about state-of-the-art performance.

We argue that there are three reasons for this confusion. First, there is no standard definition of the problem. A few variants have been studied in the literature, such as Wikification (Milne and Witten, 2008; Ratinov et al., 2011; Cheng and Roth, 2013) which aims at linking noun phrases to Wikipedia entities and Named Entity Linking (aka Named Entity Disambiguation) (McNamee and Dang, 2009; Hoffart et al., 2011) which targets only named entities. Here we use the term Entity Linking as a unified name for both problems, and Named Entity Linking (NEL) for the subproblem of linking only named entities. But names are just one part of the problem. For many variants there are no annotation guidelines for scoring links. What types of entities are valid targets? When multiple entities are plausible for annotating a mention, which one should be chosen? Are nested mentions allowed? Without agreement on these issues, a fair comparison is elusive.

Secondly, it is almost impossible to assess approaches, because systems are rarely compared using the same data sets. For instance, Hoffart et al. (2011)
developed a new data set (AIDA) based on the CoNLL 2003 Named Entity Recognition data set but failed to evaluate their system on MSNBC previously created by (Cucerzan, 2007); Wikifier (Cheng and Roth, 2013) compared to the authors’ previous system (Ratinov et al., 2011) using the originally selected datasets but didn’t evaluate using AIDA data.

Finally, when two end-to-end systems are compared, it is rarely clear which aspect of a system makes one better than the other. This is especially problematic when authors introduce complex mechanisms or nondeterministic methods that involve learning-based reranking or joint inference.

To address these problems, we analyze several significant inconsistencies among the data sets. To have a better understanding of the importance of various techniques, we develop a simple and modular, unsupervised EL system, VINCULUM. We compare VINCULUM to the two leading sophisticated EL systems on a comprehensive set of nine datasets. While our system does not consistently outperform the best EL system, it does come remarkably close and serves as a simple and competitive baseline for future research. Furthermore, we carry out an extensive ablation analysis, whose results illustrate 1) even a near-trivial model using CrossWikis (Spitkovsky and Chang, 2012) performs surprisingly well, and 2) incorporating a fine-grained set of entity types raises that level even higher. In summary, we make the following contributions:

- We analyze the differences among several versions of the entity linking problem, compare existing data sets and discuss annotation inconsistencies between them. (Sections 2 & 3)
- We present a simple yet effective, modular, unsupervised system, VINCULUM, for entity linking. We make the implementation open source and publicly available for future research.² (Section 4)
- We compare VINCULUM to 2 state-of-the-art systems on an extensive evaluation of 9 data sets. We also investigate several key aspects of the system including mention extraction, candidate generation, entity type prediction, entity coreference, and coherence between entities. (Section 5)

## 2 No Standard Benchmark

In this section, we describe some of the key differences amongst evaluations reported in existing literature, and propose a candidate benchmark for EL.

### 2.1 Data Sets

Nine data sets are in common use for EL evaluation; we partition them into three groups. The UIUC group (ACE and MSNBC datasets) (Ratinov et al., 2011), AIDA group (with dev and test sets) (Hoffart et al., 2011), and TAC-KBP group (with datasets ranging from the 2009 through 2012 competitions) (McNamee and Dang, 2009). Their statistics are summarized in Table 1³.

Our set of nine is not exhaustive, but most other datasets, e.g. CSAW (Kulkarni et al., 2009) and AQUAINT (Milne and Witten, 2008), annotate common concepts in addition to named entities. As we argue in Sec. 3.1, it is extremely difficult to define annotation guidelines for common concepts, and therefore they aren’t suitable for evaluation. For clarity, this paper focuses on linking named entities. Similarly, we exclude datasets comprising Tweets and other short-length documents, since radically different techniques are needed for the specialized corpora.

Table 2 presents a list of recent EL publications showing the data sets that they use for evaluation. The sparsity of this table is striking — apparently no system has reported the performance data from all three of the major evaluation groups.

### 2.2 Knowledge Base

Existing benchmarks have also varied considerably in the knowledge base used for link targets. Wikipedia has been most commonly used (Milne and Witten, 2008; Ratinov et al., 2011; Cheng and Roth, 2013), however datasets were annotated using different snapshots and subsets. Other KBs include Yago (Hoffart et al., 2011), Freebase (Sil and Yates, 2013), DBpedia (Mendes et al., 2011) and a subset of Wikipedia (Mayfield et al., 2012). Given that almost all KBs are descendants of Wikipedia, we use Wikipedia as the base KB in this work.⁴

³An online appendix containing details of the datasets is available at https://github.com/xiaoling/vinculum/raw/master/appendix.pdf.

⁴Since the knowledge bases for all the data sets were around 2011, we use Wikipedia dump 20110513.
Table 1: Characteristics of the nine NEL data sets. Entity types: The AIDA data sets include named entities in four NER classes, Person (PER), Organization (ORG), Location (LOC) and Misc. In TAC KBP data sets, both Person (PER) and Organization entities (ORG) are defined differently from their NER counterparts and geo-political entities (GPE), different from LOC, exclude places like KB:Cental California. KB (Sec. 2.2): The knowledge base used when each data was being developed. Evaluation Metric (Sec. 2.3): Bag-of-Concept F1 is used as the evaluation metric in (Ratinov et al., 2011; Cheng and Roth, 2013). B3+F1 used in TAC KBP measures the accuracy in terms of entity clusters, grouped by the mentions linked to the same entity.

<table>
<thead>
<tr>
<th>Group</th>
<th>Data Set</th>
<th># of Mentions</th>
<th>Entity Types</th>
<th>KB</th>
<th># of NILs</th>
<th>Eval. Metric</th>
</tr>
</thead>
<tbody>
<tr>
<td>UIUC</td>
<td>ACE</td>
<td>244</td>
<td>Any Wikipedia Topic</td>
<td>Wikipedia</td>
<td>0</td>
<td>BOC F1</td>
</tr>
<tr>
<td></td>
<td>MSNBC</td>
<td>654</td>
<td>Any Wikipedia Topic</td>
<td>Wikipedia</td>
<td>0</td>
<td>BOC F1</td>
</tr>
<tr>
<td>AIDA</td>
<td>AIDA-dev</td>
<td>5917</td>
<td>PER,ORG,LOC,MISC</td>
<td>Yago</td>
<td>1126</td>
<td>Accuracy</td>
</tr>
<tr>
<td></td>
<td>AIDA-test</td>
<td>5616</td>
<td>PER,ORG,LOC,MISC</td>
<td>Yago</td>
<td>1131</td>
<td>Accuracy</td>
</tr>
<tr>
<td></td>
<td>TAC09</td>
<td>3904</td>
<td>PER^T,ORG^T,GPE</td>
<td>TAC⊂Wiki</td>
<td>2229</td>
<td>Accuracy</td>
</tr>
<tr>
<td></td>
<td>TAC10</td>
<td>2250</td>
<td>PER^T,ORG^T,GPE</td>
<td>TAC⊂Wiki</td>
<td>1230</td>
<td>Accuracy</td>
</tr>
<tr>
<td>TAC KBP</td>
<td>TAC10T</td>
<td>1500</td>
<td>PER^T,ORG^T,GPE</td>
<td>TAC⊂Wiki</td>
<td>426</td>
<td>Accuracy</td>
</tr>
<tr>
<td></td>
<td>TAC11</td>
<td>2250</td>
<td>PER^T,ORG^T,GPE</td>
<td>TAC⊂Wiki</td>
<td>1126</td>
<td>B3+F1</td>
</tr>
<tr>
<td></td>
<td>TAC12</td>
<td>2226</td>
<td>PER^T,ORG^T,GPE</td>
<td>TAC⊂Wiki</td>
<td>1049</td>
<td>B3+F1</td>
</tr>
</tbody>
</table>

Table 2: A sample of papers on entity linking with the data sets used in each paper (ordered chronologically). TAC-KBP proceedings comprise additional papers (McNamee and Dang, 2009; Ji et al., 2010; Ji et al., 2010; Mayfield et al., 2012). Our intention is not to exhaust related work but to illustrate how sparse evaluation impedes comparison.

<table>
<thead>
<tr>
<th>Data Set</th>
<th>ACE</th>
<th>MSNBC</th>
<th>AIDA-test</th>
<th>TAC09</th>
<th>TAC10</th>
<th>TAC11</th>
<th>TAC12</th>
<th>AQUAINT</th>
<th>CSAW</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cucerzan (2007)</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Milne and Witten (2008)</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kulkarni et al. (2009)</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ratinov et al. (2011)</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hoffart et al. (2011)</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Han and Sun (2012)</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>He et al. (2013a)</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>He et al. (2013b)</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cheng and Roth (2013)</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sil and Yates (2013)</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Li et al. (2013)</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cornolti et al. (2013)</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TAC-KBP participants</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

NIL entities: In spite of Wikipedia’s size, there are many real-world entities that are absent from the KB. When such a target is missing for a mention, it is said to link to a NIL entity (McNamee and Dang, 2009) (aka out-of-KB or unlinkable entity (Hoffart et al., 2014)). In the TAC KBP, in addition to determining if a mention has no entity in the KB to link, all the mentions that represent the same real world entities must be clustered together. Since our focus is not to create new entities for the KB, NIL clustering is beyond the scope of this paper. We only evaluate whether a mention with no suitable entity in the KB is predicted as NIL. The AIDA data sets similarly contain such NIL annotations whereas ACE and MSNBC omit these mentions altogether.

2.3 Evaluation Metrics

While a variety of metrics have been used for evaluation, there is little agreement on which one to use. However, this detail is quite important, since the choice of metric strongly biases the results. We describe the most common metrics below.

Bag-of-Concept F1 (ACE, MSNBC): For each document, a gold bag of Wikipedia entities is evaluated against a bag of system output entities requiring exact segmentation match. This metric may have its historical reason for comparison but is in fact flawed since it will obtain 100% F1 for an annotation in which every mention is linked to the wrong entity, but the bag of entities is the same as the gold bag.

Micro Accuracy (TAC09, TAC10, TAC10T): For a list of given mentions, the metric simply measures the percentage of correctly predicted links.
TAC-KBP $B^3+F_1$ (TAC11, TAC12): The mentions that are predicted as NIL entities are required to be clustered according to their identities (NIL clustering). The overall data set is evaluated using an entity cluster-based $B^3+F_1$.

NER-style $F_1$ (AIDA): Similar to official CoNLL NER $F_1$ evaluation, a link is considered correct only if the mention matches the gold boundary and the linked entity is also correct. A wrong link with the correct boundary penalizes both precision and recall.

We note that Bag-of-Concept $F_1$ is equivalent to the measure for Concept-to-Wikipedia task proposed in (Cornolti et al., 2013) and NER-style $F_1$ is the same as strong annotation match. In the experiments, we use the corresponding dataset-specific measures for comparison.

3 No Annotation Guidelines

Not only do we lack a common data set for evaluation, but most prior researchers fail to even define the problem under study, before developing algorithms. Often an overly general statement such as annotating the mentions to “referent Wikipedia pages” or “corresponding entities” is used to describe which entity link is appropriate. This section shows that failure to have a detailed annotation guideline causes a number of key inconsistencies between data sets. A few assumptions are subtly made in different papers, which makes direct comparisons unfair and hard to comprehend.

3.1 Entity Mentions: Common or Named?

Which entities deserve links? Some argue for restricting to named entities. Others argue that any phrase that can be linked to a Wikipedia entity adds value. Without a clear answer to this issue, any data set created will be problematic. It’s not fair to penalize a NEL system for skipping a common noun phrases; nor would it be fair to lower the precision of a system that “incorrectly” links a common concept. However, we note that including mentions of common concepts is actually quite problematic, since the choice is highly subjective.

Example 1 In December 2008, Hoke was hired as the head football coach at San Diego State University. (Wikipedia) At first glance, KB:American football seems the gold-standard link. However, there is another entity KB:College football, which is clearly also, if not more, appropriate. If one argues that KB:College football should be the right choice given the context, what if KB:College football does not exist in the KB? Should NIL be returned in this case? The question is unanswered.\(^5\)

For the rest of this paper, we focus on the (better defined) problem of solely linking named entities.\(^6\) AQUAINT and CSAW are therefore not used for evaluation due to an disproportionate number of common concept annotations.

3.2 How Specific Should Linked Entities Be?

It is important to resolve disagreement when more than one annotation is plausible. The TAC-KBP annotation guidelines (tac, 2012) specify that different iterations of the same organization (e.g. the KB:111th U.S. Congress and the KB:112th U.S. Congress) should not be considered as distinct entities. Unfortunately, this is not a common standard shared across the data sets, where often the most specific possible entity is preferred.

Example 2 Adams and Platt are both injured and will miss England’s opening World Cup qualifier against Moldova on Sunday. (AIDA)

Here the mention “World Cup” is labeled as KB:1998 FIFA World Cup, a specific occurrence of the event KB:FIFA World Cup.

It is indeed difficult to decide how specific the gold link should be. Given a static knowledge base, which is often incomplete, one cannot always find the most specific entity. For instance, there is no Wikipedia page for the KB:116th U.S. Congress because the Congress has not been elected yet. On the other hand, using general concepts can cause troubles for machine reading. Consider president-of relation extraction on the following sentence.

Example 3 Joe Biden is the Senate President in the 113th United States Congress. Failure to distinguish different Congress iterations would cause an information extraction system to falsely extracting the fact that KB:Joe Biden is...\(^5\)

Note that linking common noun phrases is closely related to Word Sense Disambiguation (Moro et al., 2014).

\(^6\)We define named entity mention extensionally: any name uniquely referring to one entity of a predefined class, e.g. a specific person or location.
3.3 Metonymy

Another situation in which more than one annotation is plausible is metonymy, which is a way of referring to an entity not by its own name but rather a name of some other entity it is associated with. A common example is to refer to a country’s government using its capital city.

Example 4  Moscow’s as yet undisclosed proposals on Chechnya’s political future have, meanwhile, been sent back to do the rounds of various government departments. (AIDA)

The mention here, “Moscow”, is labeled as KB:Government of Russia in AIDA. If this sentence were annotated in TAC-KBP, it would have been labeled as KB:Moscow (the city) instead. Even the country KB:Russia seems to be a valid label. However, neither the city nor the country can actually make a proposal. The real entity in play is KB:Government of Russia.

3.4 Named Entities, But of What Types?

Even in the data sets consisting of solely named entities, the types of the entities vary and therefore the data distribution differs. TAC-KBP has a clear definition of what types of entities require links, namely Person, Organization and Geo-political entities. AIDA, which adopted the NER data set from the CoNLL shared task, includes entities from 4 classes, Person, Organization, Location and Misc.\(^7\) Compared to the AIDA entity types, it is obvious that TAC-KBP is more restrictive, since it does not have Misc. entities (e.g. KB:FIFA World Cup). Moreover, TAC entities don’t include fictional characters or organizations, such as KB:Sherlock Holmes. TAC GPEs include some geographical regions, such as KB:France, but exclude those without governments, such as KB:Central California or locations such as KB:Murrayfield Stadium.\(^8\)

Figure 1 summarizes the substantial differences between the two type sets.

3.5 Can Mention Boundaries Overlap?

We often see one entity mention nested in another. For instance, a U.S. city is often followed by its state, such as “Portland, Oregon”. One can split the whole mention to individual ones, “Portland” for the city and “Oregon” for the city’s state. AIDA adopts this segmentation. However, annotations in an early TAC-KBP dataset (2009) select the whole span as the mention. We argue that all three mentions make sense. In fact, knowing the structure of the mention would facilitate the disambiguation (i.e. the state name provides enough context to uniquely identify the city entity). Besides the mention segmentation, the links for the nested entities may also be ambiguous.

Example 5  Dorothy Byrne, a state coordinator for the Florida Green Party, said she had been inundated with angry phone calls and e-mails from Democrats, but has yet to receive one regretful note from a Nader voter.

The gold annotation from ACE is KB:Green Party of Florida even though the mention doesn’t contain “Florida” and can arguably be linked to KB:US Green Party.

4 A Simple & Modular Linking Method

In this section, we present VINCULUM, a simple, unsupervised EL system that performs comparatively to the state of the art. As input, VINCULUM takes a plain-text document \(d\) and outputs a set of segmented mentions with their associated entities \(A_d = \{(m_i, l_i)\}\). VINCULUM begins with mention extraction. For each identified mention \(m\), candidate entities \(C_m = \{c_j\}\) are generated for linking. VINCULUM assigns each candidate a linking score \(s(c_j|m, d)\) based on the entity type compatibility, its coreference mentions, and other entity links around

\(^7\)http://www.cnts.ua.ac.be/conll2003/ner/annotation.txt

\(^8\)http://nlp.cs.rpi.edu/kbp/2014/elquery.pdf
all possible entities

Candidate Generation

Entity Type

Coreference

Evidencelapse

One most likely entity

Mention phrases

sentence

document

world knowledge

less context

more context

Figure 2: The process of finding the best entity for a mention. All possible entities are sifted through as VINCULUM proceeds at each stage with a widening range of context in consideration.

This mention. The candidate entity with the maximum score, i.e. \( l = \arg \max_{c \in C_m} s(c|m, d) \), is picked as the predicted link of \( m \).

Figure 2 illustrates the linking pipeline that follows mention extraction. For each mention, VINCULUM ranks the candidates at each stage based on an ever widening context. For example, candidate generation (Section 4.2) merely uses the mention string, entity typing (Section 4.3) uses the sentence, while coreference (Section 4.4) and coherence (Section 4.5) use the full document and Web respectively. Our pipeline mimics the sieve structure introduced in (Lee et al., 2013), but instead of merging coreference clusters, we adjust the probability of candidate entities at each stage. The modularity of VINCULUM enables us to study the relative impact of its subcomponents.

4.1 Mention Extraction

The first step of EL extracts potential mentions from the document. Since VINCULUM restricts attention to named entities, we use a Named Entity Recognition (NER) system (Finkel et al., 2005). Alternatively, an NP chunker may be used to identify the mentions.

4.2 Dictionary-based Candidate Generation

While in theory a mention could link to any entity in the KB, in practice one sacrifices little by restricting attention to a subset (dozens) precompiled using a dictionary. A common way to build such a dictionary \( D \) is by crawling Web pages and aggregating anchor links that point to Wikipedia pages. The frequency with which a mention (anchor text), \( m \), links to a particular entity (anchor link), \( c \), allows one to estimate the conditional probability \( p(c|m) \). We adopt the CrossWikis dictionary, which was computed from a Google crawl of the Web (Spitkovsky and Chang, 2012). The dictionary contains more than 175 million unique strings with the entities they may represent. In the literature, the dictionary is often built from the anchor links within the Wikipedia website (e.g., (Ratinov et al., 2011; Hoffart et al., 2011)).

In addition, we employ two small but precise dictionaries for U.S. state abbreviations and demonyms when the mention satisfies certain conditions. For U.S. state abbreviations, a comma before the mention is required. For demonyms, we ensure that the mention is either an adjective or a plural noun.

4.3 Incorporating Entity Types

For an ambiguous mention such as “Washington”, knowing that the mention denotes a person allows an EL system to promote KB:George Washington while lowering the rank of the capital city in the candidate list. We incorporate this intuition by combining it probabilistically with the CrossWikis prior:

\[
p(c|m, s) = \sum_{t \in T} p(c, t|m, s) = \sum_{t \in T} p(c|m, t, s)p(t|m, s),
\]

where \( s \) denotes the sentence containing this mention \( m \) and \( T \) represents the set of all possible types. We assume the candidate \( c \) and the sentential context \( s \) are conditionally independent if both the mention \( m \) and its type \( t \) are given. In other words, \( p(c|m, t, s) = p(c|m, t) \), the RHS of which can be estimated by renormalizing \( p(c|m) \) w.r.t. type \( t \):

\[
p(c|m) = \frac{p(c|m)}{\sum_{t \in T} p(c|m)},
\]

where \( c \leftrightarrow t \) indicates that \( t \) is one of \( c \)'s entity types.\(^9\) The other part of the equation, \( p(t|m, s) \), can be estimated by any off-the-shelf Named Entity Recognition system, e.g. Finkel et al. (2005) and Ling and Weld (2012).

4.4 Coreference

It is common for entities to be mentioned more than once in a document. Since some mentions are less ambiguous than others, it makes sense to use the most representative mention for linking. To this end, VINCULUM applies a coreference resolution system

\(^9\)We notice that an entity often has multiple appropriate types, e.g. a school can be either an organization or a location depending on the context. We use Freebase to provide the entity types and map them appropriately to the target type set.
(e.g. Lee et al. (2013)) to cluster coreferent mentions. The representative mention of a cluster is chosen for linking. While there are more sophisticated ways to integrate EL and coreference (Hajishirzi et al., 2013), VINCULUM’s pipeline is simple and modular.

4.5 Coherence

When KB:Barack Obama appears in a document, it is more likely that the mention “Washington” represents the capital KB:Washington, D.C. as the two entities are semantically related, and hence the joint assignment is coherent. A number of researchers found inclusion of some version of coherence is beneficial for EL (Cucerzan, 2007; Milne and Witten, 2008; Ratino et al., 2011; Hoffart et al., 2011; Cheng and Roth, 2013). For incorporating it in VINCULUM, we seek a document-wise assignment of entity links that maximizes the sum of the coherence scores between each pair of entity links predicted in the document, i.e. \[
\sum_{1 \leq i < j \leq |M_d|} \phi(l_{m_i}, l_{m_j})
\]
where \(\phi\) is a function that measures the coherence between two entities. \(M_d\) denotes the set of all the mentions detected in \(d\) and \(l_{m_i}\) (\(l_{m_j}\)) is one of the candidates of \(m_i (m_j)\). Instead of searching for the exact solution in a brute-force manner (\(O(|C| \cdot |M_d|)\)) where \(|C| = \max_{m \in M_d} |C_m|\), we isolate each mention and greedily look for the best candidate by fixing the predictions of other mentions, allowing linear time search (\(O(|C| \cdot |M_d|)\)). Specifically, for a mention \(m_i\) and each of its candidates, we compute a score, \(coh(c) = \frac{1}{|P_d| - 1} \sum_{p \in P_d \setminus \{p_m\}} \phi(p, c), c \in C_m\), where \(P_d\) is the union of all intermediate links \\{\(p_m\)\} in the document. Since both measures take values between 0 and 1, we denote the coherence score \(coh(c)\) as \(p_\phi(c | P_d)\), the conditional probability of an entity given other entities in the document. The final score of a candidate is the sum of coherence \(p_\phi(c | P_d)\) and type compatibility \(p(c|m, s)\).

Two coherence measures have been found to be useful: Normalized Google Distance (NGD) (Milne and Witten, 2008; Ratino et al., 2011) and relational score (Cheng and Roth, 2013). NGD between two entities \(c_i\) and \(c_j\) is defined based on the link structure between Wikipedia articles as follows:

\[
\phi_{NGD}(c_i, c_j) = 1 - \frac{\log(\max(|L_i|, |L_j|)) - \log(|L_i \cap L_j|)}{\log(|W|) - \log(\min(|L_i|, |L_j|))}
\]
where \(L_i\) and \(L_j\) are the incoming (or outgoing) links in the Wikipedia articles for \(c_i\) and \(c_j\) respectively and \(W\) is the total number of entities in Wikipedia. The relational score between two entities is a binary indicator whether a relation exists between them. We use Freebase \(^{11}\) as the source of the relation triples \(F = \{(sub, rel, obj)\}\). Relational coherence \(\phi_{REL}\) is thus defined as

\[
\phi_{REL}(e_i, e_j) = \begin{cases} 1 & \exists r, (e_i, r, e_j) \text{ or } (e_j, r, e_i) \in F \\ 0 & \text{otherwise.} \end{cases}
\]

5 Experiments

In this section, we present experiments to address the following questions:

- Is NER sufficient to identify mentions? (Sec. 5.1)
- How much does candidate generation affect final EL performance? (Sec. 5.2)
- How much does entity type prediction help EL? What type set is most appropriate? (Sec. 5.3)
- How much does coherence improve the EL results? (Sec. 5.4)
- How well does VINCULUM perform compared to the state-of-the-art? (Sec. 5.5)
- Finally, which of VINCULUM’s components contribute the most to its performance? (Sec. 5.6)

5.1 Mention Extraction

We start by using Stanford NER for mention extraction and measure its efficacy by the recall of correct mentions shown in Table 3. TAC data sets are not included because the mention strings are given in that competition. The results indicate that at least 10% of the gold-standard mentions are left out when NER alone, is used to detect mentions. Some of the missing mentions are noun phrases without capitalization, a well-known limitation of automated extractors. To

\(^{10}\)Note that the representative mention in coreference resolution is not always the best mention for linking. When the representative mention contains a relative clause, we use the submention without the clause, which is favorable for candidate generation. When the representative mention represents a location, a longer, non-conjunctive mention is preferred. We also apply some heuristic to find organization acronyms, etc.

\(^{11}\)The mapping between Freebase and Wikipedia is provided at https://developers.google.com/freebase.
In this section, we inspect the performance of candidate generation. We compare CrossWikis with an
intra-Wikipedia dictionary \(^\text{13}\) and Freebase Search API \(^\text{14}\). Each candidate generation component takes a mention string as input and returns an ordered list of candidate entities representing the mention. The candidates produced by Crosswikis and the intra-Wikipedia dictionary are ordered by their conditional probabilities given the mention string. Freebase API provides scores for the entities using a combination of text similarity and an in-house entity relevance score. We compute candidates for the union of all the non-NIL mentions from all 9 data sets and measure their efficacy by recall@k. From Figure 3, it is clear that CrossWikis outperforms both the intra-Wikipedia dictionary and Freebase Search API for almost all k. The intra-Wikipedia dictionary is on a par with CrossWikis at k = 1 but in general has a lower coverage of the gold candidates compared to CrossWikis \(^\text{15}\). Freebase API offers a better coverage than the intra-Wikipedia dictionary but is less efficient than CrossWikis. In other words, Freebase API needs a larger cut-off value to include the gold entity in the candidate set.

Using CrossWikis for candidate generation, we plot the recall@k curves per data set (Figure 4). To our surprise, in most data sets, CrossWikis alone can achieve more than 70% recall@1. The only exceptions are TAC11 and TAC12 because the organizers intentionally selected the mentions that are highly ambiguous such as “ABC” and/or incomplete such as “Brown”. For efficiency, we set a cut-off threshold at 30 (> 80% recall for all but one data set). Note that Crosswikis itself can be used a context-insensitive EL system by looking up the mention string and predicting the entity with the highest conditional probability. The second row in Table 4 presents the results using this simple baseline. Crosswikis alone, using only the mention string, has a fairly reasonable performance.

5.3 Incorporating Entity Types

Here we investigate the impact of the entity types on the linking performance. The most obvious

\(^{13}\)Adopted from AIDA (Hoffart et al., 2011)

\(^{14}\)https://www.googleapis.com/freebase/v1/search, restricted to no more than 220 candidates per query.

\(^{15}\)We also compared to another intra-Wikipedia dictionary (Table 3 in (Ratinov et al., 2011)). A recall of 86.85% and 88.67% is reported for ACE and MSNBC, respectively, at a cut-off level of 20. CrossWikis has a recall of 90.1% and 93.3% at the same cut-off.
A more appropriate choice is 112 fine-grained entity types introduced by Ling and Weld (2012) in FIGER, a publicly available package. These fine-grained types are not disjoint, i.e. each mention is allowed to have more than one type. For each mention, FIGER returns a set of types, each of which is accompanied by a score, \( t_{\text{FIGER}}(m) = \{ (t_j, g_j) : t_j \in T_{\text{FIGER}} \} \). A softmax function is used to probabilistically interpret the results as follows:

\[
p(t_j | m, s) = \begin{cases} \frac{1}{Z} \exp(g_j) & \text{if } (t_j, g_j) \in t_{\text{FIGER}}(m), \\ 0 & \text{otherwise} \end{cases}
\]

where \( Z = \sum_{(t_k, g_k) \in t_{\text{FIGER}}(m)} \exp(g_k) \).

We evaluate the utility of entity types in Table 4, which shows that using NER typically worsens the performance. This drop may be attributed to the rigid binary values for type incorporation; it is hard to output the probabilities of the entity types for a mention given the chain model adopted in Stanford NER. We also notice that FIGER types consistently improve the results across the data sets, indicating that a finer-grained type set may be more suitable for the entity linking task.

To confirm this assertion, we simulate the scenario where the gold types are provided for each mention (the oracle types of its gold entity). The performance is significantly boosted with the assistance from the gold types, which suggests that a better performing NER/FIGER system can further improve performance. Similarly, we notice that the results using FIGER types almost consistently outperform the ones using NER types. This observation endorses our previous recommendation of using fine-grained types for EL tasks.

### 5.4 Coherence

Two coherence measures suggested in Section 4.5 are tested in isolation to better understand their effects in terms of the linking performance (Table 5). In general, the link-based NGD works slightly better than the relational facts in 6 out of 9 data sets (comparing row “+NGD” with row “+REL”). We hypothesize that the inferior results of REL may be due to the incompleteness of Freebase triples, which makes it less robust than NGD. We also combine the two by taking the average score, which in most data set performs the best (“+BOTH”), indicating that two measures provide complementary source of information.

### 5.5 Overall Performance

To answer the last question of how well does \textsc{Vinculum} perform overall, we conduct an end-to-end comparison against two publicly available systems with leading performance:\footnote{We are also aware of other systems such as TagMe-2 (Ferragina and Scaiella, 2012), DBpedia Spotlight (Mendes et al., 2011) and WikipediaMiner (Milne and Witten, 2008). A trial test on the AIDA data set shows that both \textsc{Wikifier} and \textsc{AIDA} tops the performance of other systems reported in (Cornolti et al., 2013) and therefore it is sufficient to compare with these two systems in the evaluation.}

- **AIDA** (Hoffart et al., 2011): We use the recommended \textsc{GRAPH} variant of the \textsc{AIDA} package and are able to replicate their results when gold-standard mentions are given.
- **\textsc{Wikifier}** (Cheng and Roth, 2013): We are able to reproduce the reported results on ACE and MSNBC.
Table 5: Performance (%) after re-ranking candidates using coherence scores, comparing two coherence measures (NGD and REL). “no COH”: no coherence based re-ranking is used. “+BOTH”: an average of two scores is used for re-ranking. Coherence in general helps: a combination of both measures often achieves the best effect and NGD has a slight advantage over REL.

Table 6: End-to-end performance: We compare VINCULUM in different stages with two state-of-the-art systems, AIDA and WIKIFIER (%). The column “Overall” lists the average performance of nine data sets for each approach. CrossWikis appears to be a strong baseline. VINCULUM is 0.4% shy from WIKIFIER, each winning in four data sets; AIDA tops both VINCULUM and WIKIFIER on AIDA-test.

and obtain a close enough $B^3$+ F1 number on TAC11 (82.4% vs 83.7%). Since WIKIFIER overgenerates mentions and produce links for common concepts, we restrict its output on the AIDA data to the mentions that Stanford NER predicts.

Table 6 shows the performance of VINCULUM after each stage of candidate generation (CrossWikis), entity type prediction (+FIGER), coreference (+Coref) and coherence (+Coherence). The column “Overall” displays the average of the performance numbers for nine data sets for each approach. WIKIFIER achieves the highest in the overall performance. VINCULUM performs quite comparably, only 0.4% shy from WIKIFIER, despite its simplicity and unsupervised nature. Looking at the performance per data set, VINCULUM and WIKIFIER each is superior in 4 out of 9 data sets while AIDA tops the performance only on AIDA-test. The performance of all the systems on TAC12 is generally lower than on the other dataset, mainly because of a low recall in the candidate generation stage.

We notice that even using CrossWikis alone works pretty well, indicating a strong baseline for future comparisons. The entity type prediction provides the highest boost on performance, an absolute 1.4% increase, among other subcomponents. The coherence stage also gives a reasonable lift.

In terms of running time, VINCULUM runs reasonably fast. For a document with 20-40 entity mentions on average, VINCULUM takes only a few seconds to finish the linking process on one single thread.

5.6 System Analysis

We outline the differences between the three system architectures in Table 7. For identifying mentions to link, both VINCULUM and AIDA rely solely on NER detected mentions, while WIKIFIER additionally includes common noun phrases, and trains a classifier to determine whether a mention should be linked. For candidate generation, CrossWikis provides better coverage of entity mentions. For example, in Figure 3, we observe a recall of 93.2% at a cut-off of 30 by CrossWikis, outperforming 90.7% by AIDA’s dictionary. Further, Hoffart et al. (2011) report a precision of 65.84% using gold mentions on AIDA-test, while CrossWikis achieves a higher precision at 69.24%. Both AIDA and WIKIFIER use coarse NER types as features, while VINCULUM incorporates fine-grained types that lead to dramatically improved performance, as shown in Section 5.3. The differences in Coreference and Coherence are not crucial to performance, as they each provide relatively
small gains. Finally, VINCULUM is an unsupervised system whereas AIDA and WIKIFIER are trained on labeled data. Reliance on labeled data can often hurt performance in the form of overfitting and/or inconsistent annotation guidelines; AIDA’s lower performance on TAC datasets, for instance, may be caused by the different data/label distribution of its training data from other datasets (e.g. CoNLL-2003 contains many scoreboard reports without complete sentences, and the more specific entities as annotations for metonymic mentions).

We analyze the errors made by VINCULUM and categorize them into six classes (Table 8). “Metonymy” consists of the errors where the mention is metonymic but the prediction links to its literal name. The errors in “Wrong Entity Types” are mainly due to the failure to recognize the correct entity type of the mention. In Table 8’s example, the link would have been right if FIGER had correctly predicted the airport type. The mistakes by the coreference system often propagate and lead to the errors under the “Coreference” category. The “Context” category indicates a failure of the linking system to take into account general contextual information other than the fore-mentioned categories. “Specific Labels” refers to the errors where the gold label is a specific instance of a general entity, includes instances where the prediction is the parent company of the gold entity or where the gold label is the township whereas the prediction is the city that corresponds to the township. “Misc” accounts for the rest of the errors. In the example, usually the location name appearing in the byline of a news article is a city name; and VINCULUM, without knowledge of this convention, mistakenly links to a state with the same name.

The distribution of errors shown in Table 9 provides valuable insights into VINCULUM’s varying performance across the nine datasets. First, we observe a notably high percentage of metonymy-related errors. Since many of these errors are caused due to incorrect type prediction by FIGER, improvements in type prediction for metonymic mentions can provide substantial gains in future. The especially high percentage of metonymic mentions in the AIDA datasets thus explains VINCULUM’s lower performance there (see Table 6).

Second, we note that VINCULUM makes quite a number of “Context” errors on the TAC11 and TAC12 datasets. One possible reason is that when highly ambiguous mentions have been intentionally selected, link-based similarity and relational triples are insufficient for capturing the context. For example, in “... while returning from Freeport to Portland. (TAC)”, the mention “Freeport” is unbounded by the state, one needs to know that it’s more likely to have both “Freeport” and “Portland” in the same state (i.e. Maine) to make a correct prediction. Another reason may be TAC’s higher percentage of Web documents; since contextual information is more scattered in Web text than in newswire documents, this increases the difficulty of context modeling. We leave a more sophisticated context model for future work (Chisholm and Hachey, 2015; Singh et al., 2012).

Since “Specific Labels”, “Metonymy”, and “Wrong Entity Types” correspond to the annotation issues discussed in Sections 3.2, 3.3, and 3.4, the distribution of errors are also useful in studying annotation inconsistencies. The fact that the errors vary considerably across the datasets, for instance, VINCULUM makes many more “Specific Labels” mistakes in ACE and MSNBC, strongly suggests that annotation guidelines have a considerable impact on the final performance. We also observe that annotation inconsistencies also cause reasonable predictions to be treated as a mistake, for example, AIDA predicts KB:West Virginia...
Mountaineers football for "...Alabama offered the job to Rich Rodriguez, but he decided to stay at West Virginia. (MSNBC)" but the gold label is KB:West Virginia University.

6 Related Work

Most related work has been discussed in the earlier sections; see Shen et al. (2014) for an EL survey. Two other papers deserve comparison. Cornolti et al. (2013) present a variety of evaluation measures and experimental results on five systems compared head-to-head. In a similar spirit, Hachey et al. (2014) provide an easy-to-use evaluation toolkit on the AIDA data set. In contrast, our analysis focuses on the problem definition and annotations, revealing the lack of consistent evaluation and a clear annotation guideline. We also show an extensive set of experimental results conducted on nine data sets as well as a detailed ablation analysis to assess each subcomponent of a linking system.

7 Conclusion and Future Work

Despite recent progress in Entity Linking, the community has had little success in reaching an agreement on annotation guidelines or building a standard benchmark for evaluation. When complex EL systems are introduced, there are limited ablation studies for readers to interpret the results. In this paper, we examine 9 EL data sets and discuss the inconsistencies among them. To have a better understanding of an EL system, we implement a simple yet effective, unsupervised system, VINCULUM, and conduct extensive ablation tests to measure the relative impact of each component. From the experimental results, we show that a strong candidate generation component (CrossWikis) leads to a surprisingly good result; using fine-grained entity types helps filter out incorrect links; and finally, a simple unsupervised system like VINCULUM can achieve comparable performance with existing machine-learned linking systems and, therefore, is suitable as a strong baseline for future research.

There are several directions for future work. We hope to catalyze agreement on a more precise EL annotation guideline that resolves the issues discussed in Section 3. We would also like to use crowdsourcing (Bragg et al., 2014) to collect a large set of these annotations for subsequent evaluation. Finally, we hope to design a joint model that avoids cascading errors from the current pipeline (Wick et al., 2013; Durrett and Klein, 2014).
Acknowledgements The authors thank Luke Zettlemoyer, Tony Fader, Kenton Lee, Mark Yatskar for constructive suggestions on an early draft and all members of the LoudLab group and the LIL group for helpful discussions. We also thank the action editor and the anonymous reviewers for valuable comments. This work is supported in part by the Air Force Research Laboratory (AFRL) under prime contract no. FA8750-13-2-0019, an ONR grant N00014-12-1-0211, a WRF / TJ Cable Professorship, a gift from Google, and by TerraSwarm, one of six centers of STARnet, a Semiconductor Research Corporation program sponsored by MARCO and DARPA. Any opinions, findings, and conclusion or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the view of DARPA, AFRL, or the US government.

References

Xiao Cheng and Dan Roth. 2013. Relational inference for wikification. In EACL.


Ben Hachey, Joel Nothman, and Will Radford. 2014. Cheap and easy entity evaluation. In ACL.


Mitchell Koch, John Gilmer, Stephen Soderland, and Daniel S Weld. 2014. Type-aware distantly supervised relation extraction with linked arguments. In EMNLP.


Yang Li, Chi Wang, Fangqiu Han, Jiawei Han, Dan Roth, and Xifeng Yan. 2013. Mining evidences for named entity disambiguation. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1070–1078. ACM.


Wei Shen, Jianyong Wang, and Jiawei Han. 2014. Entity linking with a knowledge base: Issues, techniques, and solutions. *TKDE*.


Appendix A. Data Set Descriptions

UIUC: ACE, a newswire subset of the ACE coreference data set (Mitchell et al., 2005), was introduced in (Ratinov et al., 2011). The first nominal mention of each gold coreference chain is annotated by Amazon’s Mechanical Turk workers. MSNBC was developed by (Cucerzan, 2007), which consists of 20 MSNBC news articles on different topics.

AIDA: Based on CoNLL 2003 Named Entity Recognition data, Hoffart et al. (2011) hand-annotated all these proper nouns with corresponding entities in YAGO2. Both the dev set (AIDA-dev) and the test set (AIDA-test) are included in the benchmark.

TAC-KBP: From the annual TAC-KBP competitions\(^\text{20}\), the evaluation sets from 2009 to 2012 are included (as well as a training set from 2010, TAC10T). Each data set consists of a series of linking queries for named entities. A query provides the surface form of the mention and the source document id. The source documents mainly come from newswire and web documents.

---

Appendix B. VINCULUM Algorithm

Subroutines: A mention extractor \(E\), a candidate generator \(D\), an entity type predictor \(TP\), a coreference system \(R\) and a coherence function \(\phi\).

Input: Document \(d\).

Output: Entity Link Annotations \(\{(m, l_m)\}\)

Extract mentions \(M = E(d)\).

Run coreference resolution \(R(d)\) and obtain coreference clusters of mentions. Denote the cluster containing a mention \(m\) as \(r(m)\) and the representative mention of a cluster \(r\) as \(rep(r)\).

\[
\text{for } m \in M \text{ do } \\
\quad \text{if } m = rep(r(m)) \text{ then } \\
\quad \quad \text{Generate candidates } C_m = D(m) \text{ (Sec. 4.2); } \\
\quad \quad \text{use } TP \text{ to predict the entity types (Sec. 4.3); } \\
\quad \quad \text{for } c \in C_m \text{ do } \\
\quad \quad \quad \text{Compute the probability of each candidate } p(c|m, s_m) \text{ based on the predicted types. } \\
\quad \quad \text{end } \\
\quad \text{else } \\
\quad \quad \text{use the representative mention } \\
\quad \quad \text{rep}(r(m)) \text{ for linking (Sec. 4.4). } \\
\quad \text{end } \\
\text{Set } p_m = \arg \max_{c \in C_m} p(c|m, s_m); \\
\text{end } \\
\text{Let } P_d = \bigcup_{m_i \in M} \{p_{m_i}\} \text{ (Sec. 4.5); } \\
\text{for } m \in M \text{ do } \\
\quad \text{for } c \in C_m \text{ do } \\
\quad \quad \text{Compute } p_\phi(c|P_d) \text{ using the given coherence function } \phi \text{ and the final score } \\
\quad \quad \text{s}(c|m, d) = p(c|m, s_m) + p_\phi(c|P_d); \\
\quad \text{end } \\
\text{Set the final link } l_m = \arg \max_{c \in C_m} s(c|m, d) \\
\text{end } \\
\text{return } \{(m, l_m) : m \in M\}
\]

\(^{20}\text{http://www.nist.gov/tac/2014/KBP/}\)