Unsupervised Entity Linking with Abstract Meaning Representation

Xiaoman Pan[1], Taylor Cassidy[2], Ulf Hermjakob[3], Heng Ji[1], Kevin Knight[3]
[1]Computer Science Department, Rensselaer Polytechnic Institute
{panx2,jih}@rpi.edu
[2]IBM Research & Army Research Laboratory
taylor.cassidy.civ@mail.mil
[3]Information Sciences Institute, University of Southern California
{ulf,knight}@isi.edu

Abstract

Most successful Entity Linking (EL) methods aim to link mentions to their referent entities in a structured Knowledge Base (KB) by comparing their respective contexts, often using similarity measures. While the KB structure is given, current methods have suffered from impoverished information representations on the mention side. In this paper, we demonstrate the effectiveness of Abstract Meaning Representation (AMR) (Banarescu et al., 2013) to select high quality sets of entity "collaborators" to feed a simple similarity measure (Jaccard) to link entity mentions. Experimental results show that AMR captures contextual properties discriminative enough to make linking decisions, without the need for EL training data, and that system with AMR parsing output outperforms hand labeled traditional semantic roles as context representation for EL. Finally, we show promising preliminary results for using AMR to select sets of "coherent" entity mentions for collective entity linking [1].

1 Introduction

The Entity Linking (EL) task (Ji et al., 2010; Ji et al., 2011; Ji et al., 2014) aims at automatically linking each named entity mention appearing in a *source* text document to its unique referent in a *target* knowledge base (KB). For example, consider the following sentence posted to a discussion forum during the 2012 U.S. presidential election: "*Where would McCain be without Sarah?*". An Entity Linker should link the entity mentions "*McCain*" and "*Sarah*" to the entities `John McCain` and `Sarah Palin`, respectively, which serve as unique identifiers for the real people.

A typical EL system works as follows. Given a mention m (a string in a source document), the top N most likely entity referents from the KB are enumerated based on prior knowledge about which entities are most likely referred to using m. The candidate entities are re-ranked to ultimately link each mention to the top entity in its candidate list. Reranking consists of two key elements: *context representation* and *context comparison*. For a given mention, candidate entities are re-ranked based on a comparison of information obtained from the context of m with known structured and/or unstructured information associated with the top N KB entities, which can be considered the "context" of the KB entity[2]. The basic intuition is that the entity referents of m and related mentions should be similarly connected in the KB.

However, there might be many entity mentions in the context of a target entity mention that could potentially be leveraged for disambiguation. In this paper, we show that a deeper semantic knowledge representation - including the Abstract Meaning Representation (AMR) (Banarescu et al., 2013) - can capture contextual properties that are discriminative enough to disambiguate entity mentions that current state-of-the-art systems cannot handle, without the need for EL training data. Specifically, for a given

[1]The web service of this EL system is at: blender02.cs.rpi.edu:3300 and some related AMR tools are at: github.com/panx27/amr-reader

[2]Most work uses Wikipedia and related resources to derive the KB, prior link likelihood, and entity information (e.g., Wikipedia article text and infoboxes).

Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL, pages 1130–1139,
Denver, Colorado, May 31 – June 5, 2015.

entity mention, using AMR provides a rich context representation, facilitating the selection of an optimal set of *collaborator* entity mentions, i.e., those co-occurring mentions most useful for disambiguation. In previous approaches, collaborator sets have tended to be too narrow or too broad, introducing noise. We then use unsupervised graph inference for context comparison, achieving results comparable with state-of-the-art supervised methods and substantially outperforming context representation based on traditional Semantic Role Labeling.

In addition, most state-of-the-art EL approaches now rely on *collective inference*, where a set of *coherent mentions* are linked simultaneously by choosing an "optimal" or maximally "coherent" set of named entity targets - one target entity for each mention in the coherent set. We show preliminary results suggesting that AMR is effective for the partitioning of all mentions in a document into coherent sets for collective linking.

We evaluate our approach using both human and automatic AMR annotation, limiting target named entity types to person (PER), organization (ORG), and geo-political entities (GPE) [3].

2 Related Work

In most recent collective inference methods for EL (e.g., (Kulkarni et al., 2009; Pennacchiotti and Pantel, 2009; Fernandez et al., 2010; Radford et al., 2010; Cucerzan, 2011; Guo et al., 2011; Han and Sun, 2011; Ratinov et al., 2011; Chen and Ji, 2011; Kozareva et al., 2011; Dalton and Dietz, 2013)), the target entity mention's "collaborators" may simply include all mentions which co-occur in the same discourse (sentence, paragraph or document) (Ratinov et al., 2011; Nguyen et al., 2012). But this approach usually introduces many irrelevant mentions, and it's very difficult to automatically determine the scope of discourse. In contrast, some recent work exploited more restricted measures by only choosing those mentions which are topically related (Cassidy et al., 2012; Xu et al., 2012), bear a relation from a fixed set (Cheng and Roth, 2013), coreferential (Nguyen et al., 2012; Huang et al., 2014), socially related (Cassidy et al., 2012; Huang et al., 2014), dependent (Ling et al., 2014), or a combination of these through meta-paths (Huang et al., 2014). These measures can collect more precise collaborators but suffer from low coverage of predefined information templates and the unsatisfying quality of state-of-the-art coreference resolution, relation and event extraction.

In this paper, we demonstrate that AMR is an appropriate and elegant way to acquire, select, represent and organize deeper knowledge in text. Together with our novel utilization of the rich structures in merged KBs, the whole framework carries rich enough evidence for effective EL, without the need for any labeled data, collective inference, or sophisticated similarity.

3 Knowledge Network Construction from Source

Abstract Meaning Representation (AMR) (Banarescu et al., 2013) is a sembanking language that captures whole sentence meanings in a rooted, directed, labeled, and (predominantly) acyclic graph structure. AMR utilizes multi-layer linguistic analysis such as PropBank frames, non-core semantic roles, coreference, named entity annotation, modality and negation to represent the semantic structure of a sentence. AMR strives for a more logical, less syntactic representation. Compared to traditional dependency parsing and semantic role labeling, the nodes in AMR are entities instead of words, and the edge types are much more fine-grained[4]. AMR thus captures deeper meaning compared with other representations more commonly used to represent mention context in EL.

We use AMR to represent semantic information about entity mentions expressed in their textual context. Specifically, given an entity mention m, we use a rule based method to construct a Knowledge Network, which is a star-shaped graph with m at the hub, with leaf nodes obtained from entity mentions reachable by AMR graph traversal from m, as well as AMR node attributes such as entity type. A subset of the leaf nodes are selected as m's collaborators using rules presented in the following subsec-

[3]The mapping from AMR entity types to these three main types is at: amr.isi.edu/lib/ne-type-sc.txt

[4]AMR distinguishes between entities and concepts, the former being instances of the latter. We consider AMR concepts as entity mentions, and use AMR entity annotation for coreference resolution.

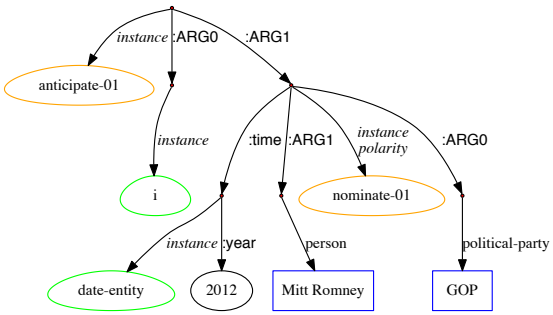
tions. Note that while we only evaluate linking of PER, ORG, and GPE entities, collaborators may be of any type. We also outline preliminary efforts to use AMR to create sets of coherent entity mentions.

In each of the following subsections we describe elements of AMR useful for context representation in EL. For each element we explain how our current system makes use of it (primarily, by using it to add entity mentions to a particular entity mention's set of collaborators). In doing so, we mainly refer to several examples from political discussion forums about "*Mitt Romney*", "*Ron Paul*" and "*Gary Johnson*". Their AMR graphs are depicted in Figure 1.

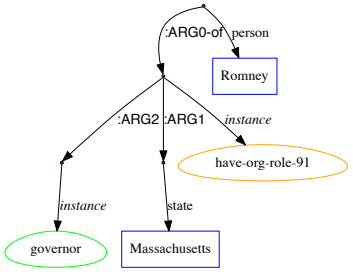
3.1 Entity Nodes

Each AMR node represents an entity mention, and contains its canonical name as inferred from sentential context. This property is called *name expansion*. Consider the following sentence: "*Indonesia lies in a zone where the **Eurasian**, **Philippine** and **Pacific** plates meet and occasionally shift, causing earthquakes and sometimes generating tsunamis.*". Here, the nodes representing the three plates will be labeled as "*Eurasian Plate*", "*Philippine Plate*" and "*Pacific Plate*" respectively, even though these strings do not occur in the sentence. Note that these labels may be recovered primarily by appealing to syntactic reasoning, without consulting a KB. In our implementation we consider these expanded names as mentions (these strings supersede raw mentions as input to the salience based candidate enumeration (Section 5.2)). Because the initial enumeration of entity candidates depends heavily on the mention's surface form, independent of context, name expansion will help us link "*Philippine*" to "`Philippine Sea Plate`" as opposed to the country.

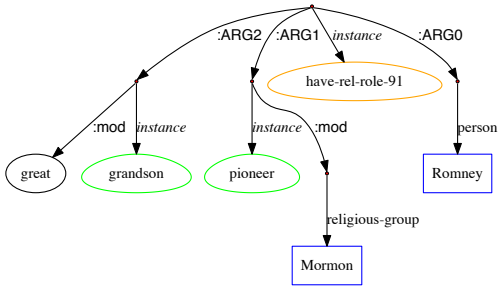
An AMR node also contains an entity type. AMR defines 8 main entity types (Person, Organization, Location, Facility, Event, Product, Publication, Natural object, Other) and over one hundred fine-grained subtypes. For example, company, government organization, military, criminal organization, political party, school, university, research institute, team and league are subtypes of organization. The fine-grained entity types defined in AMR help us restrict KB entity candidates for a given mention by encouraging entity type matching. For example, in "*The **Yuri dolgoruky** is the first in a series of new nuclear submarines to be commissioned this*



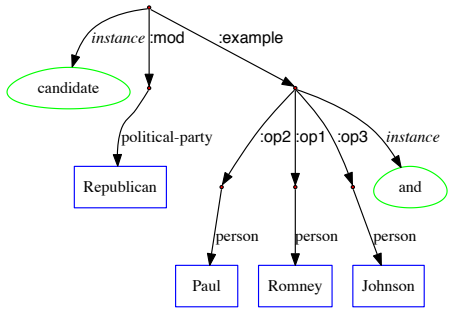
(a) *I am cautiously anticipating the GOP nominee in 2012 not to be **Mitt Romney**.*



(b) ***Romney** was the Governor of Massachusetts...*



(c) ***Romney** is the great-great-grandson of a Mormon pioneer...*



(d) *Republican candidates like **Romney**, **Paul**, and **Johnson**...*

Figure 1: AMR for the Walk-through Example

year but the bulava nuclear-armed missile developed to equip the submarine has failed tests and the deployment prospects are uncertain.", AMR labels "*Yuri dolgoruky*" as a product instead of a person. We manually mapped AMR entity types to equivalent DBpedia types to inform type matching restrictions [5]. However, to make our context comparison algorithm less dependent on the quality of this mapping, and on automatic AMR name type assignment, we add a mention's type to its collaborators [6]. In future work we plan to investigate the effects of different type matching techniques, varying degrees of strictness.

3.2 Semantic Roles

AMR defines core roles based on the OntoNotes (Hovy et al., 2006) semantic role layer. Each predicate is associated with a sense and frame description. If a target entity mention *m* and a context entity mention *n* are both playing core roles for the same predicate, we consider *n* as a collaborator of *m*. Consider the following post: "*Did **Palin** apologize to **Giffords**? He needs to conduct a beer summit between **Palin** and **NBC**.*". We add "*Giffords*" and "*NBC*" as collaborators of "*Palin*", because they play core roles in both the "*apologize-01*" and "*meet-03*" events.

AMR defines new core semantic roles which did not exist in PropBank (Palmer et al., 2005), NomBank (Meyers et al., 2004), or Ontonotes (Hovy et al., 2006). Intuitively, the following special roles should provide discriminative collaborators:

- The ARG2 role of the *have-org-role-91* frame indicates the title held by an entity (ARG0), such as President and Governor, within a particular organization (ARG1).
- ARG2 and ARG3 of *have-rel-role-91* are used to describe two related entities of the same type, such as family members.

AMR defines a rich set of general semantic relations through non-core semantic roles. We choose the following subset of non-core roles to provide collaborators for entity mentions: *domain*, *mod*, *cause*, *concession*, *condition*, *consist-of*, *extent*, *part*, *purpose*, *degree*, *manner*, *medium*, *instrument*, *ord*, *poss*, *quant*, *subevent*, *subset*, *topic*.

3.3 Background Time and Location

AMR provides rich temporal and spatial information about entities and events. Types instantiated in AMR include time, year, month, day, source, destination, path and location. We exploit time and location entities as collaborators for entity mentions when they each play a role in the same predicate. For example, in the following post, the time role of the "*die-01*" event is "*2008*": "*I just Read of Clark's death in 2008*". We can link "*Clark*" to `Arthur C Clark` in the KB, which contains the triple: <`Arthur C Clark`, *date-of-death*, 2008-03-19> (see Section 4). Similarly, it's very challenging to link the abbreviation "*BMKG*", in the following sentence, to the correct target entity `Indonesian Agency for Meteorology, Climatology and Geophysics`, whose headquarters are listed as `Jakarta` in the KB: "*It keeps on shaking. **Jakarta** BMKG spokesman Mujuhidin said*". Here, "*Jakarta*" is added as a collaborator of "*BMKG*" since AMR labels it as the location of the organization, which facilitates the correct link because in DBpedia `Jakarta` is listed as its headquarter.

Authors often assume that readers will infer implicit temporal information about events. In fact, half of the events extracted by information extraction (IE) systems lack time arguments (Ji et al., 2009). Therefore if an AMR parse includes no time information, we use the document creation time as an additional collaborator for mention in question. For example, knowing the document creation time "*2005-06-05*" can help us link "*Hsiung Feng*" in the following sentence "*The BBC reported that Taiwan has successfully test fired the **Hsiung Feng**, its first cruise missile.*" to `Hsiung Feng IIE`, which was deployed in 2005. Similarly, we include document creation location as a global collaborator.

3.4 Coreference

For linking purposes, we treat a coreferential chain of mentions as a single "mention". In doing so, the collaborator set for the entire chain is computed as the union over all of the chain's mentions' collabo-

[5] The mapping from three main types and AMR entity types to Dbpedia types is at: nlp.cs.rpi.edu/amrel/dbtype.txt

[6] A more strict approach might disallow type mismatches between entity mentions and their target KB entities outright.

rator sets. From here on we refer to a coreferential chain of mentions as simply a "mention".

AMR currently only represents sentence-level coreference resolution. In order to construct a knowledge network across sentences, we use the following heuristic rules. If two names have a substring match (on a token-wise basis with stop words removed), or one name consists of the initials of another in all capital letters, then we mark them as coreferential. We replace all names in a coreferential chain with their canonical name, which may have been derived via name expansion (Section 3.1): full names for people and abbreviations for organizations.

3.5 Knowledge Networks for Coherent Mentions

AMR defines a rich set of conjunction relations: "*and*", "*or*", "*contrast-01*", "*either*", "*compared to*", "*prep along with*", "*neither*", "*slash*", "*between*" and "*both*". These relations are often expressed between entities that have other relations in common. We therefore group mentions connected by conjunction relations into sets of coherent mentions. This representation is used only in preliminary experiments on collective entity linking.

Figure 2 shows the expanded knowledge network that includes results from individual networks for each of the coherent mentions from the walk-through example. For each coherent set, we merge the knowledge networks of all of its mentions [7].

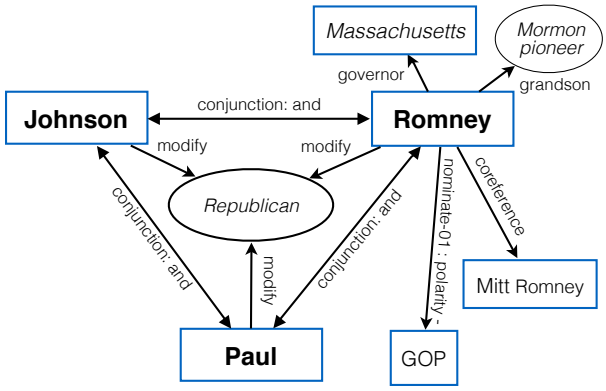


Figure 2: Knowledge Network for Mentions in Source

[7] recall that by mention, we mean a coreferential chain of mentions that may extend across sentences

4 Knowledge Network Construction from KB

We combine Wikipedia with derivative resources to create the KB. The KB is a single knowledge network in which nodes are entities (Wikipedia articles) or constant values (e.g. a dollar amount or date), and the edges represent relations. We use this structure for context representation for entities, which together with context representation for mentions (Section 3) feeds re-ranking based on context comparison.

The KB is formally represented by triples:

$$< Entity, EdgeLabel, Node >$$

where *Entity* is the entity's unique identifier, *EdgeLabel* is relation type, and *Node* is the corresponding relation value - either another Entity or a constant. These triples are derived from typed relations expressed within Wikipedia infoboxes, Templates, and Categories, untyped hyperlinks within Wikipedia article text, typed relations within DBpedia (dbpedia.org) and Freebase (www.freebase.com), and Google's "people also searched for" list [8]. Figure 3 shows a portion of the KB pertaining to the example in Figure 1.

In order to merge nodes from multiple KBs, we use the Wikipedia title as a primary key, and then use DBpedia *wikiPageID* and Freebase *Key* relations.

5 Linking Knowledge Networks

5.1 Overview

In this section we present our detailed algorithm to link each mention to a KB entity using a simple similarity measure over knowledge networks. Recall that a rule-based method has already been employed to construct star-shaped knowledge networks for individual mentions and entities (see sections 3 and 4; A KB knowledge network is the subnetwork of the entire KB centered at a candidate entity).

For each mention to be linked, an initial list of candidate entities are enumerated based on entity salience with respect to the mention, independent of mention context (Section 5.2)[9]. *Context collaborator* re-ranking proceeds in an unsupervised fashion

[8] In response to a query entity Google provides a list of entities that "people also search for" - we add them to the entity's network.

[9] Here, "mention" means coreferential chain of mentions.

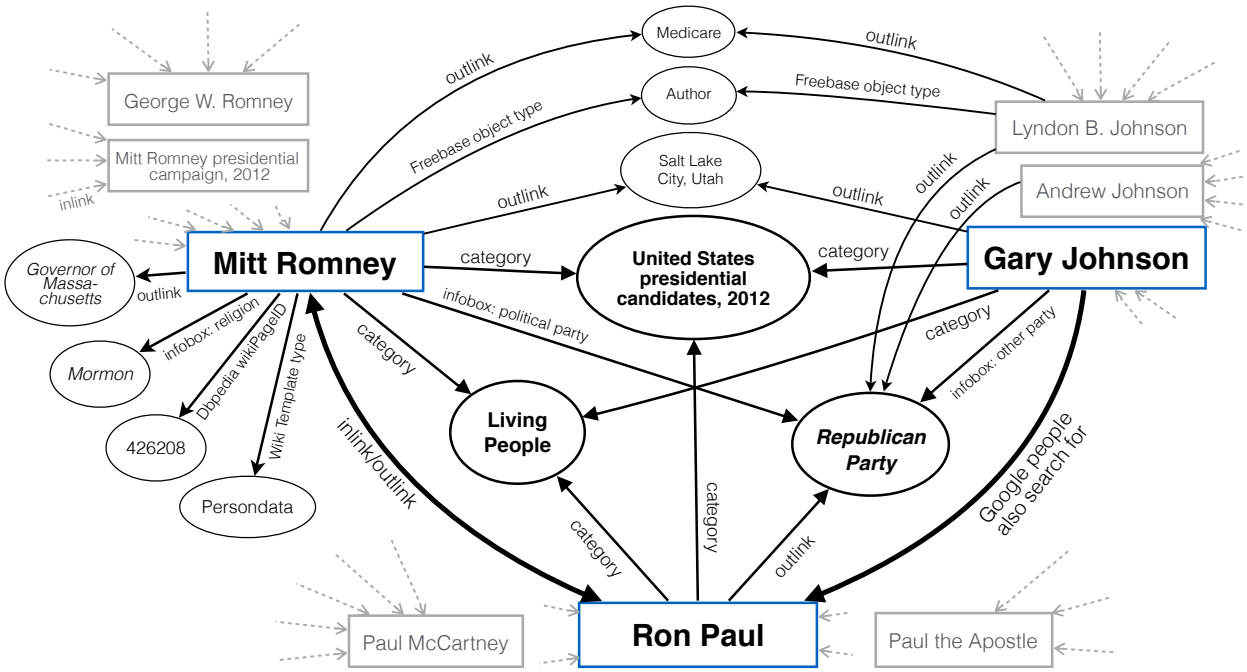


Figure 3: Knowledge Network for Entities in KB

agnostic to knowledge network edge labels using the Jaccard similarity measure computed between the mention and each entity, by taking their collaborator sets as inputs (Section 5.3).

We also describe *Context Coherence* re-ranking in terms of KB knowledge networks only, which constitutes preliminary steps toward unsupervised collective entity linking in section 5.4 based on the notion of coherence described in section 3.5. We leave a combination of the two re-ranking approaches to future work.

5.2 Salience

We use commonness (Medelyan and Legg, 2008) as a measure of context independent salience for each mention m, to generate an initial ranked list of candidate entities $E = (e_1, ..., e_N)$ where N is the cutoff for number of candidates. In all experiments, we used N = 15 which can give us an oracle accuracy score 97.58%.

$$Commonness(m, e) = \frac{count(m, e)}{\sum_{e'} count(m, e')}$$

Here, $count(m, e)$ is the number of hyperlinks with anchor text m and entity e within all of Wikipedia. As illustrated in Figure 3, using this salience measure "*Romney*" is successfully linked to `Mitt Romney`. For the mention "*Paul*", the politician `Ron Paul` is ranked at top 2 (less popular than the musician `Paul McCartney`). For the mention "*Johnson*", the correct entity `Gary Johnson` is ranked at top 9, after more popular entities such as `Lyndon B. Johnson` and `Andrew Johnson`.

5.3 Context Collaborator Based Re-ranking

Context collaborator based re-ranking is driven by the similarity between mention and entity knowledge networks. We construct knowledge network $g(m)$ for each mention m, and knowledge network $g(e_i)$ for each entity candidate e_i in m's entity candidate list E. We re-rank E according to Jaccard Similarity, which computes the similarity between $g(m)$ and $g(e_i)$:

$$J(g(m), g(e_i)) = \frac{|g(m) \cap g(e_i)|}{|g(m) \cup g(e_i)|}$$

Note that the edge labels (e.g., *nominate-01* for a mention, or *infobox: religion* for an entity) are ignored, as the similarity metric operates over sets of collaborators (leaf nodes in the knowledge networks). For set intersection and union computation, elements are treated as lists of lower-cased tokens with stop words removed, and two elements are considered equal if and only if they have one or

more token in common. Due to the support from their neighbor `Republican` in the KB (Figure 3) which matches the neighbor "*Republican*" of mentions "*Paul*" and "*Johnson*" (Figure 2), `Ron Paul` and `Gary Johnson` are promoted to top 1 and top 3 respectively. `Gary Johnson` is still behind two former U.S. presidents `Andrew Johnson` and `Lyndon B. Johnson` who also shares the neighbor `Republican` in the KB.

5.4 Context Coherence Based Re-ranking

Context coherence based re-ranking is driven by the similarity among KB entities. Let R_m be a set of coherent entity mentions, and R_E be the set of corresponding entity candidate lists, which are generated according to salience. Given R_E, we generate every combination of possible top candidate lists for the mentions in R_m, and denote the set of these combinations C_m. Formally, C_m is the Cartesian product of all candidate lists $E \in R_E$. In the walk-through example, R_m contains ["*Romney*", "*Paul*", "*Johnson*"], and C_m contains [`Mitt Romney, Ron Paul, Gary Johnson`], [`Mitt Romney, Paul McCartney, Lyndon Johnson`], etc. We compute coherence for each combination $c \in C_m$ as Jaccard Similarity, by applying a form of Equation 5.3 generalized to take any number of arguments to the set of knowledge networks for all entities in c, i.e., $\{g(e)|e \in c\}$. The highest similarity combination is selected, yielding a top candidate for each $m \in R_m$. For example, compared to `Andrew Johnson` and `Lyndon Johnson`, `Gary Johnson` is more coherently connected with `Mitt Romney` and `Ron Paul`, therefore it is promoted to top 1 with the coherence measure.

6 Experiments

6.1 Data And Scoring Metric

For our experiments we use a publicly available AMR R3 corpus (LDC2013E117) that includes manual EL annotations for all entity mentions (LDC2014E15) [10].

For evaluation we used all the discussion forum posts (DF), and news documents (News) that were sorted according to alphabetic order of document IDs and taken as a tenth. The detailed data statistics are presented in Table 1 [11].

	PER	ORG	GPE	All
News	159	187	679	1,025
DF	235	129	224	588
All	394	316	903	1,613

Table 1: Total # of Entity Mentions in Test Set

For each mention, we check whether the KB entity returned by an approach is correct or not. We compute accuracy for an approach as the proportion of mentions correctly linked.

6.2 Experiment Results

We focus primarily on context collaborator based reranking results. We compare our results with several baseline and state-of-the-art approaches in Table 2. In Table 3 we present preliminary results for collective linking.

Our Unsupervised Context Collaborator Approach substantially outperforms the popularity based methods. More importantly, we see that AMR provides the best context representation for collaborator selection. Even system AMR outperformed not only baseline co-occurrence based collaborator selection methods, but also outperforms the collaborator selection method based on human annotated core semantic roles. Figure 4 depicts accuracy increases as more AMR annotation is used in selecting collaborators. From the commonness baseline, additional knowledge about individual names leads to substantial gains followed by additional gains after incorporating links denoting semantic roles. Note that coreference here includes cross-sentence co-reference not based on AMR (Section 3.4). Furthermore, the results using human annotated AMR outperform the state-of-the-art supervised methods trained from a large scale EL training corpus, which rely on collective inference[12]. These results all verify the importance of incorporating a wider range of deep knowledge. Finally, Table 2 presents results in which our

[10] EL annotations are available to KBP shared task registrants (nlp.cs.rpi.edu/kbp/2014) via Linguistic Data Consortium (www.ldc.upenn.edu).

[11] The list of document IDs in the test set is at: nlp.cs.rpi.edu/amrel/testdoc.txt

[12] Note that the ground-truth EL annotation for the test set was created by correcting the output from supervised methods, so it may even favor these methods.

Approach		Definition	News	DF	Total
Popularity	Commonness	based on the popularity measure as described in section 5.2.	89.76	68.99	82.20
	Google Search	use the top Wikipedia page returned by Google search using the mention as a key word.	88.10	77.17	84.12
Supervised	State-of-the-art	supervised re-ranking using multi-level linguistic features for collaborators and collective inference, trained from 20,000 entity mentions from TAC-KBP2009-2014. We combined two systems (Chen and Ji, 2011; Cheng and Roth, 2013) using rules to highlight their strengths.	93.07	87.41	**91.01**
Unsupervised Context Collaborator Approach	Sen. Level Cooccurrence	sentence-level co-occurrence based collaborator selection	93.17	73.25	85.92
		(collaborators limited to human AMR-labeled named entities)	90.77	70.31	83.31
	Doc. Level Cooccurrence	document-level co-occurrence based collaborator selection	90.05	69.86	82.69
		(collaborators limited to human AMR-labeled named entities)	87.51	69.37	80.90
	Human AMR	using human annotated AMR nodes and edges.	93.56	86.88	**91.13**
	System AMR	using AMR nodes and edges automatically generated by an AMR parser (Flanigan et al., 2014).	90.15	85.69	**88.52**
	Human SRL	using human annotated core semantic roles defined in PropBank (Palmer et al., 2005) and NomBank (Meyers et al., 2004): ARG0, ARG1, ARG2, ARG4 and ARG5.	93.27	71.21	85.24
Unsupervised Combined Approach	Human AMR	coherence approach used where possible (215 mentions), collaborator approach elsewhere (remaining 1398 mentions), using human annotated AMR nodes and edges.	94.34	88.25	**92.12**

Table 2: Accuracy (%) on Test Set (1613 mentions)

context coherence method is used where possible (i.e., those 215 mentions that are members of coherent sets according to our criteria as described in Section 3.5), and the context collaborator approach based on human AMR annotation is applied elsewhere.

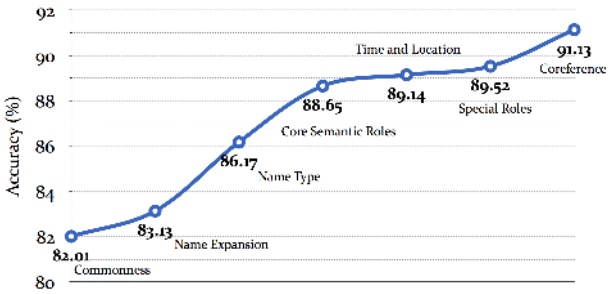


Figure 4: AMR Annotation Layers Effects on Accuracy

Table 3 focuses on the 215 mentions that met our narrow criteria for forming a coherent set of mentions. We applied the context coherence based reranking method (Section 5.4) to collectively link those mentions. This approach substantially outperforms the co-occurrence baseline, and even outperforms the context collaborator approach applied to those 215 mentions, especially for discussion forum data.

Approach Description	News	DF	All
Coherence: coherence set built from within-sentence collaborators limited to human AMR-labeled Named Entities.	72.64	76.85	75.47
Coherence: coherence set built from human AMR conjunctions (Sec. 3.5)	96.73	95.16	**96.28**
Collaborator: used coherent set based on human AMR as collaborators.	91.50	82.26	88.84

Table 3: Context Coherence Accuracy (%) on 215 Mentions which Can Form Coherent Sets

6.3 Remaining Error Analysis and Discussion

A challenging source of errors pertains to the *knowledge gap* between the source text and KB. News and social media are source text genres that tend to focus on new information, trending topics, breaking events, or even mundane details about the entity. In contrast, the KB usually provides a snapshot summarizing only the entity's most representative and important facts. A source-KB similarity driven approach alone will not suffice when a mention's context differs substantially from anything on the KB side. AMR annotation's synthesis of words and phrases from the surface texts into concepts only provides a first step toward bridging the knowledge gap. Successful linking may require (1) reasoning using general knowledge, or (2) retrieval of other sources that contain additional useful linking information. Table 4 illustrates two relevant examples

Type	Source	Knowledge Base
General Knowledge	**[Christies]$_{\mathbf{m}}$** denial of ***marriage*** privledges to ***gays*** will alienate independents and his "I wanted to have the people ***vote*** on it" will ring hollow.	**[Chris Christie]$_{\mathbf{e}}$** has said that he favoured New Jersey's law allowing ***same-sex*** couples to form civil unions, but would veto any bill legalizing ***same-sex marriage*** in New Jersey.
External Knowledge	Translation out of hype-speak: some kook made ***threatening*** noises at **[Brownback]$_{\mathbf{m}}$** and go ***arrested***.	**[Samuel Dale "Sam" Brownback]$_{\mathbf{e}}$** (born September 12, 1956) is an American politician, the 46th and current ***Governor*** of Kansas.

Table 4: Examples of Knowledge Gap

that our system does not correctly link. In the first example, if we don't already know that Christie is the topic of discussion, as humans we might use our general knowledge that "governors veto bills" to pick the correct entity. Using this type of knowledge presents interesting challenges (e.g., governors don't always veto bills, nor are they the only ones who can do so). In the second example, the rumor about this politician is not important enough to be reported in his Wikipedia page. We might first figure out, using cross-document coreference techniques, that a news article with the headline "*Man Accused Of Making Threatening Phone Call To Kansas Gov.* ***Sam Brownback*** *May Face Felony Charge...*" is talking about the same rumor. Then we might use biographical facts (e.g., Brownback is the governor of Kansas) from the article to enrich Brownback's knowledge network on the source side.

Sometimes helpful neighbor concepts are omitted because the current collaborator selection criteria are too restricted. For example, "*armed*" and "*conflicts*" are informative words for linking "*The Stockholm Institute*" to `Stockholm International Peace Research Institute` in the following sentence "***The Stockholm Institute*** *stated that 23 of 25 major armed conflicts in the world in 2000 occurred in impoverished nations.*", but they were not selected as context collaborators. In addition, our cross-sentence coreference resolution is currently limited to proper names. Expanding it to include nominals could further enrich context collaborators to overcome some remaining errors. For example, in the sentence, "*The first woman to serve on* ***SCOTUS***", if we know "*The first woman*" is coreferential with "*Sandra Day O'Connor*" in the previous sentence, we can link "*SCOTUS*" to `Supreme Court of the United States` instead of `Scotus College`.

7 Conclusions and Future Work

EL requires a representation of the relations among entities in text. We showed that the Abstract Meaning Representation (AMR) can better capture and represent the contexts of entity mentions for EL than previous approaches. We plan to improve AMR representation as well as automatic annotation. We showed that AMR enables EL performance comparable to the supervised state of the art using an unsupervised, non-collective approach. We plan to combine collaborator and coherence methods into a unified approach, and to use edge labels in knowledge networks for context comparison (note that the last of these is quite challenging due to normalization, polysemy, and semantic distance issues). We have only applied a subset of AMR representations to the EL task, but we aim to explore how more AMR knowledge can be used for other more challenging Information Extraction and Knowledge Base Population tasks.

Acknowledgments

This work was supported by the U.S. DARPA DEFT Program No. FA8750-13-2-0041 and FA8750-13-2-0045, DARPA BOLT Program No. HR0011-12-C-0014, ARL NS-CTA No. W911NF-09-2-0053, NSF Awards IIS-0953149 and IIS-1523198, AFRL DREAM project, DHS CCICADA, gift awards from IBM, Google, Disney and Bosch. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

References

L. Banarescu, C. Bonial, S. Cai, M. Georgescu, K. Griffitt, U. Hermjakob, K. Knight, P. Koehn, M. Palmer, and N. Schneider. 2013. Abstract meaning representation for sembanking. In *Proc. ACL 2013 Workshop on Linguistic Annotation and Interoperability with Discourse*.

T. Cassidy, H. Ji, L. Ratinov, A. Zubiaga, and H. Huang. 2012. Analysis and enhancement of wikification for microblogs with context expansion. In *Proc. International Conference on Computational Linguistics (COLING 2012)*.

Z. Chen and H. Ji. 2011. Collaborative ranking: A case study on entity linking. In *Proc. Empirical Methods in Natural Language Processing (EMNLP 2011)*.

X. Cheng and D. Roth. 2013. Relational inference for wikification. In *Proc. Empirical Methods in Natural Language Processing (EMNLP 2013)*.

S. Cucerzan. 2011. Tac entity linking by performing full-document entity extraction and disambiguation. In *Proc. Text Analysis Conference (TAC 2011)*.

J. Dalton and L. Dietz. 2013. A neighborhood relevance model for entity linking. In *Proc. Open Research Areas in Information Retrieval (OAIR 2013)*.

N. Fernandez, J. A. Fisteus, L. Sanchez, and E. Martin. 2010. Webtlab: A cooccurence-based approach to kbp 2010 entity-linking task. In *Proc. Text Analysis Conference (TAC 2010)*.

J. Flanigan, S. Thomson, J. Carbonell, C. Dyer, and N. A. Smith. 2014. A discriminative graph-based parser for the abstract meaning representation. In *Proc. Association for Computational Linguistics (ACL 2014)*.

Y. Guo, W. Che, T. Liu, and S. Li. 2011. A graph-based method for entity linking. In *Proc. International Joint Conference on Natural Language Processing (IJCNLP 2011)*.

X. Han and L. Sun. 2011. A generative entity-mention model for linking entities with knowledge base. In *Proc. Association for Computational Linguistics: Human Language Technologies (ACL-HLT 2011)*.

E.d Hovy, M. Marcus, M. Palmer, L. Ramshaw, and R. Weischedel. 2006. Ontonotes: The 90% solution. In *Proc. Human Language Technology conference - North American chapter of the Association for Computational Linguistics (HLT-NAACL 2006)*.

H. Huang, Y. Cao, X. Huang, H. Ji, and C. Lin. 2014. Collective tweet wikification based on semi-supervised graph regularization. In *Proc. Association for Computational Linguistics (ACL 2014)*.

H. Ji, R. Grishman, Z. Chen, and P. Gupta. 2009. Cross-document event extraction and tracking: Task, evaluation, techniques and challenges. In *Proc. Recent Advances in Natural Language Processing (RANLP 2009)*.

H. Ji, R. Grishman, H. T. Dang, K. Griffitt, and J. Ellis. 2010. Overview of the tac 2010 knowledge base population track. In *Proc. Text Analysis Conference (TAC 2010)*.

H. Ji, R. Grishman, and H. T. Dang. 2011. Overview of the tac 2011 knowledge base population track. In *Proc. Text Analysis Conference (TAC 2011)*.

H. Ji, J. Nothman, and H. Ben. 2014. Overview of tac-kbp2014 entity discovery and linking tasks. In *Proc. Text Analysis Conference (TAC 2014)*.

Z. Kozareva, K. Voevodski, and S. Teng. 2011. Class label enhancement via related instances. In *Proc. Empirical Methods in Natural Language Processing (EMNLP 2011)*.

S. Kulkarni, A. Singh, G. Ramakrishnan, and S. Chakrabarti. 2009. Collective annotation of wikipedia entities in web text. In *Proc. Knowledge Discovery and Data Mining (KDD 2009)*.

X. Ling, S. Singh, and D. S. Weld. 2014. Context representation for named entity linking. In *Proc. Pacific Northwest Regional NLP Workshop (NW-NLP 2014)*.

O. Medelyan and C. Legg. 2008. Integrating cyc and wikipedia: Folksonomy meets rigorously defined common-sense. In *Proc. AAAI 2008 Workshop on Wikipedia and Artificial Intelligence*.

A. Meyers, R. Reeves, C. Macleod, R. Szekely, V. Zielinska, B. Young, and R. Grishman. 2004. The nombank project: An interim report. In *Proc. HLT-NAACL 2004 Workshop on Frontiers in Corpus Annotation*.

H. Nguyen, H. Minha, T. Cao, and T. Nguyenb. 2012. Jvn-tdt entity linking systems at tac-kbp2012. In *Proc. Text Analysis Conference (TAC 2012)*.

M. Palmer, D. Gildea, and P. Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

M. Pennacchiotti and P. Pantel. 2009. Entity extraction via ensemble semantics. In *Proc. Empirical Methods in Natural Language Processing (EMNLP 2009)*.

W. Radford, B. Hachey, J. Nothman, M. Honnibal, and J. R. Curran. 2010. Cmcrc at tac10: Document-level entity linking with graph-based re-ranking. In *Proc. Text Analysis Conference (TAC 2010)*.

L. Ratinov, D. Roth, D. Downey, and M. Anderson. 2011. Local and global algorithms for disambiguation to wikipedia. In *Proc. Association for Computational Linguistics: Human Language Technologies (ACL-HLT 2011)*.

J. Xu, Q. Lu, J. Liu, and R. Xu. 2012. Nlpcomp in tac 2012 entity linking and slot-filling. In *Proc. Text Analysis Conference (TAC 2012)*.