Entity Linking with a Unified Semantic Representation

Zhaochen Guo
Department of Computing Science
University of Alberta
zhaochen@ualberta.ca

Denilson Barbosa
Department of Computing Science
University of Alberta
denilson@ualberta.ca

ABSTRACT

Entity Linking (EL) consists in linking mentions in a document to their referent entities in a Knowledge Base. Current approaches fall into two main categories: *local approaches*, in which mentions are linked independently of each other, and *global approaches*, in which all mentions are linked collectively. Local approaches often ignore the semantic relatedness of entities, and while global approaches incorporate the semantic relatedness, they tend to focus only on directly connected entities, ignoring indirect connections which might be useful. We present a global EL approach that unifies the representation of the semantics of entities and documents–the probability distribution of entities being visited during a random walk on an entity graph–that accounts for direct and indirect connections. An experimental evaluation shows that our method outperforms five state-of-the-art EL systems and two very strong baselines.

1. INTRODUCTION

Entity linking is the task of assigning identifiers of *entities* in a Knowledge Base (KB) to *mentions* of named entities in a text document. EL is key for Information Extraction (IE) but has many other applications. For instance, it enables expanding or correcting a KB with facts extracted from documents–this task is called Knowledge Base Population [11]. Another application is Semantic Search, an emerging paradigm of Web search that combines traditional Information Retrieval approaches over document corpora with KB-style query answering and reasoning to offer more accurate and concise answers to Web searches.

EL is challenging due to the inherent ambiguity of natural language. Most entities can be mentioned in many synonymous ways. For instance, retired basketball player *Michael Jordan* is commonly mentioned as *Air Jordan, Michael J. Jordan, MJ23*, depending on the context. Another problem is that the same mention may refer to many different entities (*polysemy*), again depending on context. For example, Wikipedia also has entries for a *mycologist*, a *researcher*, and a *football player* called Michael Jordan. Having a complete and accurate synonym dictionary for all entities is impractical and would not account for other sources of ambiguity, such as misspellings in the text. For this reason, EL systems often operate in two steps (1) selecting a broad list of candidate entities (2) disambiguating them by context.

The current crop of EL systems differ, primarily, in the disambiguation phase. *Local* EL approaches (e.g., [2, 3]) treat mentions independently and use local features such as contextual words or named entities for mention disambiguation. Typically, they rank candidates by similarity of their feature vectors with that of the article of the mention. One drawback of local approaches is that they ignore the semantic relatedness [15] between features and mentions, which can help to solve the feature sparsity issue. For example, *New York City* and *Big Apple*, which are semantically related, will be less likely to be linked because their spellings are not similar.

Global EL approaches (e.g., [5, 9, 12, 18]), on the other hand, take the semantic relatedness between mentions into consideration and perform EL collectively on all mentions. Leveraging semantic relatedness can provide additional context information: for example, linking *NBA* to *National Basketball Association* will make it easier to disambiguate the mention to *Michael Jordan*. These methods seek to find an assignment for mentions such that entities in the assignment not only are compatible with the mentions, but also have maximum internal coherence among all possible assignments. As finding such an assignment with maximum coherence is NP-Hard [12], all global approaches turn to approximate algorithms or heuristics.

Our approach. One limitation of current global EL methods is that the way they compute semantic relatedness, by considering only directly connected entities as the semantic representation of entities. Doing so, however, ignores entities that are indirectly connected but semantically related to the target entity. We improve on this as follows.

We perform a random walk with restart [19], which, as demonstrated in the personalized PageRank algorithm [8], can propagate information along a graph and provide a relatedness measure between indirectly connected entities. We use the resulting probability distribution as a notion of relatedness between all nodes in the graph and the target node. We refer to the distribution as a *semantic signature*, and use the semantic signature for the relatedness measure in the EL task. The semantic signature represents the semantics of entities in a more fined-grained manner than the 0-1

WWW'14 Companion, April 7–11, 2014, Seoul, Korea.
ACM 978-1-4503-2745-9/14/04.
http://dx.doi.org/10.1145/2567948.2579705.

weighting in local approaches. Also, by applying the random walk with restart on a set of target entities, we can compute the semantic signature of sentences and documents (represented by the set of target entities). In other words, the semantic signature is a unified representation that can capture the semantics of entities, sentences, or documents, and can greatly facilitate mention disambiguation–the key part of global EL approaches.

The main contributions of this work are:

- We use the semantic signature to represent the semantics of entities and documents, and help with the entity linking task.
- We experimentally evaluate our EL system on several benchmark datasets, and make comparisons with five state-of-the-art EL systems and two strong baselines.

2. RELATED WORK

Earlier work [2, 3] on Entity Linking exploited local features, often neighboring words, to select referent entities based on their contextual similarity with mentions. Recent EL systems take into account the semantic relations between mentions and entities, employing various measures of semantic relatedness computed over an entity graph. Semantic relatedness has been widely used in recent EL systems [6, 7, 9, 12, 16, 18]. Cucerzan [5] used related named entities and entity categories for this purpose, while Milne and Witten [15] exploited links within Wikipedia. However, they consider only those entities directly linked to the target entities in their measures.

The approach closest to ours is that of Han et. al [6], which uses a random walk with restart to obtain a rank vector for all candidates of mentions, and considers the rank value in the vector to be the relatedness between a mention and its candidate. Not unlike the other semantic relatedness measures, their measure can only compute the relatedness between two entities. Instead, we use a unified semantic representation for both documents and entities. As a result, we can find the similarity of multiple entities and mentions at once.

The idea of random walk with restart has been applied on graphs constructed from the WordNet [14], with the stationary distribution to represent the semantics of words. It has been shown to be effective in the word similarity measurement [1, 10], and word sense disambiguation [17]. However, we are not aware of any previous work applying random walk with restart for the EL task.

3. PROBLEM FORMALIZATION

Definition 1 (Entity Linking). *Given a set of mentions $M = \{m_1, \ldots, m_n\}$ in a document D, and a knowledge base KB whose entity set is E, the problem of entity linking is to find an assignment $\Gamma : M \rightarrow E \cup \{NIL\}$.*

As customary, NIL is used to indicate mentions which cannot be linked to any entity in KB. Also, as introduced above, global approaches aim to find an assignment Γ that is contextually compatible with mentions, and has maximum coherence. Formally, the solution to the EL problem is an assignment Γ^* maximizing the following objective function:

$$\Gamma^* = \arg\max_{\Gamma} \left(\sum_{1}^{N} \phi(m_i, e_i) + \Psi(\Gamma) \right), \qquad (1)$$

in which $\phi(m_i, e_i)$ measures the context similarity of m_i and e_i, and $\Psi(\Gamma)$ measures the coherence of Γ.

3.1 Disambiguation with Semantic Signatures

Our approach is substantially different from the scheme in Equation 1. First, for each mention $m \in D$, we find $CS(m)$: the top-k candidates[1] from an entity alias list, ranked by their prior probability $P(e_i|m)$. These candidates (and their neighboring entities) yield a subset of the KB, which we call KB_D. We compute the semantic signature of D from this KB_D, and also the signatures for each candidate entity separately. (The details of computing semantic signatures are given in the next section.)

Let $SS(e_i)$ be the semantic signature of an entity in $CS(m)$, and $SS(D)$ be the semantic signature of the document D. We use the cosine similarity to compute the semantic coherence between e_i and D as follows:

$$\Psi(e_i, D) = \frac{SS(e_i) \cdot SS(D)}{\| SS(e_i) \| \| SS(D) \|} \qquad (2)$$

Our solution to the EL problem is thus: given a mention m and its candidate entities $CS(m)$ as above, we assign m to the entity maximizing:

$$\Gamma(m) = \arg\max_{e_i \in CS(m)} \Psi(e_i, D) \qquad (3)$$

Linking to NIL. There are two situations in which our method will assign NIL to a mention. The first is when the mention does not have any good candidates–i.e., it is dissimilar to all entities in the alias list. The second is when the semantic coherence of the entity maximizing Equation 3 and the document is not high enough. In both cases, we have application-specific thresholds. In future work, we will study these thresholds in a variety of settings.

4. SEMANTIC SIGNATURES

We start by describing the construction of an entity graph from a knowledge base, and then explain the random walk model and how to compute the semantic signature using the random walk model on the entity graph.

4.1 Entity Graph Construction

As usual, our KB is derived from Wikipedia. A directed graph $G = (V, E)$ is constructed from Wikipedia in which V is the set of entities, and E is the set of links linking entities. However, since there are over 4 million entities in Wikipedia, efficiency will be an issue when computing the probability distribution necessary for ranking entities. Thus instead of using the whole graph, we construct a subgraph to improve the efficiency without sacrificing the performance.

Candidate Generation. The graph construction starts with the candidate entities of mentions. Given a mention m_i, the first step is to generate the candidates of m_i from the knowledge base. We use an alias dictionary, which maps aliases to their referred entities, for the candidate generation. The alias dictionary is built from the Wikipedia page title, redirect page, disambiguation page, and anchor text of

[1]We set $k = 10$ in the tests reported here.

links. The candidates of mention m_i, are then generated by matching the mention name against the alias dictionary.

Graph Construction. The entity graph is then initialized with the candidate entities. To construct a graph that can better measure the semantics of entities, we expand this initial graph with entities directly connected to the candidate entities. When expanding the graph, we prune entities with low connectivity to limit the size of the graph so as to improve the efficiency of the random walk. In our experiment, we prune (expanded) entities with in-degree below a threshold (40 in the experiments reported here). However, we always keep *candidates* in the entity graph even if their in-degree is below the threshold, so that uncommon mentions are still linked. Finally, the expanded graph becomes the entity graph on which we will run the random walk model to compute semantic signatures.

4.2 Random Walk with Restart

Given the entity graph and initial values for entities in the graph [2], the random walk algorithm traverses the graph and propagates the value of entities to their neighbours in a proportional way. This process continues until the value of entities converges to a stationary distribution.

Following links in the entity graph, we visit an entity e_i from entities linking to e_i, and compute its value as follows:

$$r_i^{t+1} = \sum_{e_j \in IN(e_i)} r_j^t * P(e_i|e_j) \tag{4}$$

in which $IN(e_i)$ is the set of entities linking to e_i, and $P(e_i|e_j)$ is the probability to move from entity e_j to entity e_i.

As customary, we incorporate a random restart probability in the preference vector to avoid the issues caused by sinks and guarantee convergence. Formally, the random walk model can be modelled as:

$$r^{t+1} = \alpha \times r^t \times M + (1-\alpha) \times \vec{v}, \tag{5}$$

where r^t and r^{t+1} are the value vector of entities at iteration t and $t+1$, and M is the transition matrix of the graph with $M_{ij} = P(e_i|e_j)$. The preference vector $\vec{v}$ defines the probability that a surfer randomly jumps to an entity, and $\sum v_i = 1$.

When a random walk process converges to a stationary state, the value vector of entities is called *stationary distribution* R ($\sum r_i = 1$), which is what we use as our semantic signature.

4.3 Semantic Signature Computation

We describe how to compute the semantic signature of entities and documents separately.

4.3.1 Semantic Signature of an Entity

To compute the semantic signature of an entity e_i, we need the stationary distribution to be biased towards e_i. In other words, entities that are more semantically related to e_i should be given higher values. A random walk with restart from e_i always restarts from e_i with probability $1-\alpha$, and automatically assigns high value to entities close to e_i. So when we set $v_i = 1$ and $v_{j(j\neq i)} = 0$, a random walk with restart will generate the semantic signature of entity e_i.

[2]As the initial value does not affect the stationary distribution, any values summed up to 1 are acceptable, e.g. $\frac{1}{N}$

4.3.2 Semantic Signature of a Document

Theoretically, computing the semantic signature of a document is the same as that of an entity. Suppose we represent a document with entities in an assignment Γ, and we set the preference weight of each entity $e_j \in \Gamma$ to be $\frac{1}{|\Gamma|}$ in the preference vector $\vec{v}$, then the semantic signature of the document is the stationary distribution computed through the random walk with restart from entities in Γ.

However, there are two issues here. First, the true assignment Γ is not available and finding Γ is the task of an EL system. This is a chicken-and-egg problem. Second, a uniform weight $\frac{1}{|\Gamma|}$ in the preference vector for each entity may not reflect the importance of these entities in the document. To solve these two issues, we adopt the following strategies.

For the first issue, we replace the referent entity of mention m with m's candidate entities. Given that the prior probability $P(e_i|m)$ is a strong baseline for entity linking (see the experiment), we propose to approximate the referent entity using the candidate entities of m, each of which is weighted by their prior probability. In terms of the preference vector, we assign the weight of each candidate entity in $\vec{v}$ to be $P(e_i|m)$. Formally:

$$\vec{v}_i = P(e_i|m) \tag{6}$$

and the prior probability (aka. *commonness* [13]) can be approximated using the alias dictionary as follows:

$$P(e_i|m) = \frac{count(e_i, m)}{count(m)} \tag{7}$$

in which $count(e_i, m)$ is the number of times m refers to e_i in Wikipedia, and $count(m)$ is the number of times m appears as a link in Wikipedia.

As we will show in the experiments, entities with the highest prior probability tend to be the referent entities with very high probability. The advantage of this approximation is that, on one hand, higher weight is given to entities with higher prior probability so that the candidate entity set can capture the semantics of the document, and on the other hand, referent entities with low prior probability are still assigned with weights in the preference vector and can contribute to the semantic representation of the document.

For the second issue, we weight each mention by its importance in the document, for example, its tf-idf. Integrating the importance weighting and the approximate entity set, we define the preference vector as follows:

$$\vec{v}_i = importance(m) * P(e_i|m) \tag{8}$$

in which $importance(m)$ measures the importance of mention m in the document (tf-idf in our experiment),

With the preference vector, the semantic signature of a document can be computed from a random walk process with $\vec{v}$.

4.3.3 Computational Efficiency

The efficiency of computing semantic signature is a concern in our EL system. Suppose there are total K candidates

	Datasets								
	MSNBC			AQUAINT			ACE2004		
Systems	**Accuracy**	**F1@MI**	**F1@MA**	**Accuracy**	**F1@MI**	**F1@MA**	**Accuracy**	**F1@MI**	**F1@MA**
PriorProb	85.98	86.50	87.15	84.87	87.27	87.16	84.82	85.49	87.13
Local	77.43	77.91	72.30	66.44	68.32	68.09	61.48	61.96	56.95
Cucerzan	87.80	88.34	87.76	76.62	78.67	78.22	78.99	79.30	78.22
M&W	68.45	78.43	80.37	79.92	85.13	84.84	75.54	81.29	84.25
Han'11	87.65	88.46	87.93	77.16	79.46	78.80	72.76	73.48	66.80
GLOW	65.55	75.37	77.33	75.65	83.14	82.97	75.49	81.91	83.18
RI	88.57	90.22	90.87	85.01	87.72	87.74	82.35	86.60	87.13
SemSig	**91.62**	**92.18**	**92.10**	**85.83**	**88.14**	**88.02**	**87.16**	**87.50**	**88.43**

Table 1: Performance of various EL systems on the MSNBC, AQUAINT, and ACE2004 datasets.

for all mentions in M, we need to compute $K+1$ semantic signatures (K for candidate entities, and 1 for the document). To improve the efficiency, we adopt the following methods. First, we rank candidates by their prior probability and keep only the top candidates for each mention (10 in our experiments). Second, we limit the size of the entity graph by pruning uncommon entities as described above. Third, we parallelize the computation of the semantic signatures for the candidates.

5. EXPERIMENTS

We now report on an experimental comparison of our EL method against the state-of-the-art in EL and two baselines.

Datasets. We used 3 benchmarks: (1) MSNBC [5], with 20 news articles from 10 different topics (2 articles per topic) and 739 mentions in total; (2) AQUAINT, compiled by Milne and Witten [16], with 50 documents and 727 mentions from a news corpus from the Xinhua News Service, the New York Times, and the Associated Press; and (3) the ACE2004 dataset [18], a subset of the ACE2004 Coreference documents with 57 articles and 306 mentions, annotated through *crowdsourcing*.

We used a Wikipedia dump from June 2013 to test all systems, except GLOW and RI, which were published with a dump from May 2013 (we assumed the differences were minor, and decided that keeping the original data used by their authors would be preferable). We aligned the annotations in the benchmarks to the June 2013 dump by replacing old annotations with their redirected entities and removing annotations that do not exist any more. The only change was for the MSNBC dataset, for which only 739 (of the original 755) could be linked[3].

Evaluation Measures. We use the standard *accuracy, precision, recall*, and *F1*:

$$\begin{aligned} Accuracy &= \frac{|TP|+|TN|}{|TP|+|TN|+|FP|+|FN|} \\ Precision &= \frac{|TP|}{|TP|+|FP|} \\ Recall &= \frac{|TP|}{|TP|+|FP|+|FN|} \\ F1 &= \frac{2*Precision*Recall}{Precision+Recall} \end{aligned}$$

[3]The datasets available at `http://www.cs.ualberta.ca/~denilson/data/deos14_ualberta_experiments.tgz`.

where TP (true positives) is the set of mentions correctly linked to an entity; TN (true negatives) is the set of mentions correctly linked to *NIL*; FP (false positives) is the set of mentions incorrectly linked to entities; and FN (false negatives) is the set of mentions incorrectly linked to *NIL*.

5.1 Results

The two baselines are *PriorProb*, which links mentions to the entities with the highest prior probability $P(e_i|m)$, and *Local*, which chooses the candidate that maximizes the local compatibility $\phi(e_i, m)$ with the mention. The five competitor systems are *Cucerzan* [5]–the first collective EL system, M&W [16]–a machine learning system for the EL task, Han'11 [6]–a collective EL system exploiting an entity graph to compute the relatedness and jointly link mentions, GLOW [18]–a system combining local and global features for entity linking, and RI [4]–the start-of-the art EL system using relational inference for mention disambiguation. Table 1 lists the results of these EL systems on the 3 datasets in terms of overall accuracy and F1, both across mentions (micro-averaged, indicated as **F1@MI**) and across documents (macro-averaged, **F1@MA**).

The performance of the *Local* baseline indicates that text features alone cannot solve EL. Combining local compatibility with the semantic relatedness, as in *Han'11*, provides substantial gains. The strongest baseline *PriorProb* outperforms many EL systems, which also points to limitations in the benchmarks–they are biased towards popular entities. The 5 previous EL systems all build on the assumption that the true assignment should have the maximum coherence. Our system, *SemSig*, which does not make that assumption, outperforms them, which suggests that the coherence between entities and the document as a whole could be a better semantic relatedness measure. One possible explanation is that entities that are close in a document are more likely semantically related, while entities that are far from each other may not be semantically related. Thus the assignment with the maximum coherence may not be the best assignment.

5.2 TAC Entity Linking 2011

We also evaluated our system on the TAC 2011[4] Entity Linking task. One difference between the TAC dataset and the above datasets is that it contains many more abbreviations and acronyms, making the mentions more ambiguous. We perform a query expansion on the abbreviation mentions

[4]`http://www.nist.gov/tac/2011/`

System	Accuracy
LCC	86.1
MS-MLI	86.8
RI	86.1
NUSchime	86.3
SemSig	86.4

Table 2: Accuracy of various EL systems on the TAC 2011 Entity Linking task.

by extracting definitions of abbreviations using patterns *definition (Abbrev)* and *Abbrev (definition)*. Table 2 shows the results of our *SemSig* system, *RI* and a few top EL systems in the submissions, in terms of linking accuracy. *SemSig* is very competitive overall, virtually tying with RI and the top submissions to the TAC contest. It is worth noting that the MS-MLI system exploits external web search logs for candidate generation and additional training datasets.

6. CONCLUSION

Collectively linking entities in a document exploits the assumption that all mentions to named entities in a document are coherent with one another. Therefore, the choice of how to measure coherence among entities is crucial for achieving good results. In this work, we propose to use a probability distribution computed from a random walk with restart over an entity graph to measure the semantics of entities and documents in a unified way. Our encouraging preliminary results indicate that this semantic representation can greatly improve the entity linking results with better performance than the state-of-the-art EL systems.

Future work. The method described here uses prior probabilities to choose the top-k candidates for each mention, and is thus biased towards popular entities. This approach is fine for the current benchmarks, as indicated in our results, and for applications that process news text, which often deal with popular entities as well. However, these methods will not work as well for entities in the "long tail". Designing more challenging benchmarks that cannot be solved with simple baselines like *PriorProb* would help advance the field.

Our system performs multiple PageRank computations, making it time consuming if implemented naively. Therefore, designing proper system infrastructure with the appropriate indexes and/or parallel computing frameworks to speed up these computations would be interesting. Moreover, other state-of-the-art systems perform other expensive operations as well, such as accessing the web. Designing objective and fair benchmarks for comparing these different approaches in a more holistic way would be of great value.

Finally, our approach, like most other systems, has many application-specific parameters (recall Section 3) and depends on specific similarity measures (e.g., to filter candidate entities). Further studies are needed to understand how the choice of similarity measures and configuration of parameters affect the accuracy of our approach.

Acknowledgements

This work was supported in part by grants from the Natural Sciences and Engineering Council of Canada, through its Business Intelligence Network, and Alberta Innovates Technology Futures.

7. REFERENCES

[1] E. Agirre, E. Alfonseca, K. Hall, J. Kravalova, M. Pasca, and A. Soroa. A study on similarity and relatedness using distributional and wordnet-based approaches. In *HLT-NAACL*, pages 19–27, 2009.

[2] A. Bagga and B. Baldwin. Entity-based cross-document coreferencing using the vector space model. In *COLING-ACL*, pages 79–85, 1998.

[3] R. C. Bunescu and M. Pasca. Using encyclopedic knowledge for named entity disambiguation. In *EACL*, pages 9–16, 2006.

[4] X. Cheng and D. Roth. Relational inference for wikification. In *EMNLP*, pages 1787–1796, 2013.

[5] S. Cucerzan. Large-scale named entity disambiguation based on wikipedia data. In *EMNLP-CoNLL*, pages 708–716, 2007.

[6] X. Han, L. Sun, and J. Zhao. Collective entity linking in web text: a graph-based method. In *SIGIR*, pages 765–774, 2011.

[7] X. Han and J. Zhao. Named entity disambiguation by leveraging wikipedia semantic knowledge. In *CIKM*, pages 215–224, 2009.

[8] T. H. Haveliwala. Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search. *IEEE Trans. Knowl. Data Eng.*, 15(4):784–796, 2003.

[9] J. Hoffart, M. A. Yosef, I. Bordino, H. Fürstenau, M. Pinkal, M. Spaniol, B. Taneva, S. Thater, and G. Weikum. Robust disambiguation of named entities in text. In *EMNLP*, pages 782–792, 2011.

[10] T. Hughes and D. Ramage. Lexical semantic relatedness with random graph walks. In *EMNLP-CoNLL*, pages 581–589, 2007.

[11] H. Ji and R. Grishman. Knowledge base population: Successful approaches and challenges. In *ACL*, pages 1148–1158, 2011.

[12] S. Kulkarni, A. Singh, G. Ramakrishnan, and S. Chakrabarti. Collective annotation of wikipedia entities in web text. In *KDD*, pages 457–466, 2009.

[13] O. Medelyan, I. H. Witten, and D. Milne. Topic indexing with wikipedia. In *WIKIAI*, pages 19–24, 2008.

[14] G. A. Miller. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41, 1995.

[15] D. Milne and I. H. Witten. An effective, low-cost measure of semantic relatedness obtained from wikipedia links. In *WIKIAI*, pages 25–30, 2008.

[16] D. Milne and I. H. Witten. Learning to link with wikipedia. In *CIKM*, pages 509–518, 2008.

[17] M. T. Pilehvar, D. Jurgens, and R. Navigli. Align, disambiguate and walk: A unified approach for measuring semantic similarity. In *ACL*, pages 1341–1351, 2013.

[18] L.-A. Ratinov, D. Roth, D. Downey, and M. Anderson. Local and global algorithms for disambiguation to wikipedia. In *ACL*, pages 1375–1384, 2011.

[19] H. Tong, C. Faloutsos, and J.-Y. Pan. Fast random walk with restart and its applications. In *ICDM*, pages 613–622, 2006.