Language and Domain Independent Entity Linking with Quantified Collective Validation

Han Wang[*,1], **Jin Guang Zheng**[*,2], **Xiaogang Ma**[1], **Peter Fox**[1], **and Heng Ji**[2]
{Tetherless World Constellation[1], Computer Science Department[2]}
Rensselaer Polytechnic Institute
Troy, NY, USA
{wangh17, zhengj6, max7, pfox, jih}@rpi.edu

Abstract

Linking named mentions detected in a source document to an existing knowledge base provides disambiguated entity referents for the mentions. This allows better document analysis, knowledge extraction and knowledge base population. Most of the previous research extensively exploited the linguistic features of the source documents in a supervised or semi-supervised way. These systems therefore cannot be easily applied to a new language or domain. In this paper, we present a novel unsupervised algorithm named Quantified Collective Validation that avoids excessive linguistic analysis on the source documents and fully leverages the knowledge base structure for the entity linking task. We show our approach achieves state-of-the-art English entity linking performance and demonstrate successful deployment in a new language (Chinese) and two new domains (Biomedical and Earth Science). Experiment datasets and system demonstration are available at http://tw.rpi.edu/web/doc/hanwang_emnlp_2015 for research purpose.

1 Introduction and Motivation

The entity linking (EL) task aims at analyzing each named entity mention in a source document and linking it to its referent in a knowledge base (KB). Consider the following example: "*One day after released by the **Patriots**, **Florida** born **Caldwell** visited the **Jets**. ..... **The New York Jets** have six receivers on the roster: **Cotchery**, **Coles**, ...*". Here "*Caldwell*" is an ambiguous mention because not only are there thousands of people with different professions named "*Caldwell*", but even if as an American football player, as most people would recognize it from the context, there are several "*Caldwell*"s who are/were associated with either "*the Patriots*" or "*the Jets*". An EL system should be able to disambiguate the mention by carefully examining the context and then identify the correct KB referent, which is `Reche Caldwell` in this case.

Although EL has attracted a lot of community attention in the recent years, most research efforts have been focused on developing systems only effective for generic English corpora. When these systems are migrated to a new language or domain, their performance will usually suffer from a noticeable decline due to the following reasons:

1) State-of-the-art EL systems have developed comprehensive linguistic features from the source documents to generate advanced representations of the mentions and their context. While this methodology has been proved rewarding for a resource-rich language such as English, it prevents the systems from being adopted to a new language, especially to one with limited linguistic resources. One can imagine that it would be very difficult, if not impossible, for an English EL system that benefits from the part-of-speech tagging, dependency parsing, and named entity recognition to be deployed to a new language such as Chinese that has quite different linguistic characteristics.

2) The current EL approaches mostly target at people, organizations, and geo-political entities which are widely present in a general KB such as Wikipedia. However, domain-specific EL tends to pay more attention to entities beyond the above three types. For instance, in the biomedical science domain, protein is a major class of entities that greatly interest scientists. Conventional EL systems are very likely to fail in linking protein mentions in the text due to the lack of labeled

[*]These authors contributed equally to this work.

training data. Moreover, their reliance on general reference KBs seems insufficient for a specific domain. Take "*A20*", a type of protein as an example. Wikipedia has more than a few items listed under the name of "*A20*" and their types range from aircrafts to roads. This diversified information inevitably introduces noise for a biomedical EL application.

One potential solution to tackle these limitations is, instead of concentrating on the source documents, to conduct more deliberate study on the KB. Structured KBs such as DBpedia[1] typically offer detailed descriptions about entities, a large collection of named relations between entities, and a growing number of multi-lingual entity surface forms. By embracing these ready-for-use information and linked structures, we will be able to obtain sufficient contextual information for disambiguation without generating a full list of linguistic features from the source documents, and therefore eliminate the language dependency. Moreover, currently there exist numerous publicly available domain ontology repositories such as BioPortal[2] and OBO Foundry[3] which provide significantly more domain knowledge than general KBs for EL to leverage. By incorporating these domain ontologies, we can easily increase the entity coverage and reduce noise for deploying EL in various new domains.

In order to make the most of the KB structure, the mention context should be matched against the KB such that the relevant KB information can be extracted. A collective way of aligning co-occurred mentions to the KB graph has been proved to be a successful strategy to better represent the source context (Pennacchiotti and Pantel, 2009; Fernandez et al., 2010; Cucerzan, 2011; Han et al., 2011; Ratinov et al., 2011; Dalton and Dietz, 2013; Zheng et al., 2014; Pan et al., 2015). We take a further step to consider quantitatively differentiating entity relations in the KB in order to evaluate entity candidates more precisely. Meanwhile, we jointly validate these candidates by aligning them back to the source context and integrating multiple ranking results. This novel EL framework deeply exploits the KB structure with a light weight representation of the source context, and thus enables a smooth migration to new languages and domains.

The main novel contributions of this paper are summarized as follows: 1) We design an unsupervised EL algorithm, namely, Quantified Collective Validation (QCV) that builds KB entity candidate graphs with quantified relations for the purpose of collective disambiguation and inference. 2) We develop a procedure of building language and domain independent EL systems by incorporating various ontologies into the QCV component. 3) We demonstrate that our system is able to achieve state-of-the-art performance in English EL, and it can also produce promising results for Chinese EL as well as EL in Biomedical Science and Earth Science.

2 Baseline Collective EL

As a baseline, we adopt a competitive unsupervised collective EL system (Zheng et al., 2014) utilizing structured KBs. It defines entropy based weights for the KB relations, and embeds them in a two-step candidate ranking process to produce the EL results.

Structured KB Terminologies: In a structured KB, a fact is usually expressed in the form of a triple: (e_h, r, e_t) where e_h, e_t are called the head entity and the tail entity, respectively, and r is the relation between e_h and e_t.

Entropy Based KB Relation Weights: The goal is to leverage various levels of granularity of KB relations. The calculation of the relation weight $H(r)$ is given in Equation (1):

$$H(r) = -\sum_{e_t \in E_t(r)} P(e_t) \log(P(e_t)) \qquad (1)$$

where $E_t(r)$ is the tail entity set for r in the KB, and $P(e_t)$ is the probability of e_t appearing as the tail entity for r in the KB.

Salience Ranking: As the first ranking step, we examine the candidates without the context and prefers those with higher importance in the KB. Equation (2) computes the salience score $S_a(c)$ for a candidate c:

$$S_a(c) = \sum_{r \in R(c), e_t \in E_t(r)} H(r) \frac{S_a(e_t)}{L(e_t)} \qquad (2)$$

where $R(c)$ is the relation set for c in the KB; $H(r)$ is given by Equation (1); $E_t(r)$ is the tail entity set with c being the head entity and r being the connecting relation in the KB; $L(e_t)$ denotes the

[1] http://wiki.dbpedia.org
[2] http://bioportal.bioontology.org
[3] http://www.obofoundry.org

cardinality of the tail entity set with e_t being the head entity in the KB. $S_a(c)$ is recursively computed until convergence.

Collective Ranking: The similarity $Sim^F(m, c)$ between a candidate c and its mention m is defined using Equation (3) as the final ranking score:

$$Sim^F(m,c) = \alpha \cdot JS(m,c) \cdot S_a(c) + \beta \cdot \sum_{r \in R(c)} H(r) \cdot \sum_{n \in E_t(r) \cap C(m)} S_a(n) \quad (3)$$

where $JS(m, c)$ is the Jaccard similarity between the string surface forms of m and c; $S_a(c)$ and $S_a(n)$ are both evaluated by Equation (2); $C(m)$ denotes the candidate set for mention m; α and β are hyperparameters.

3 Quantified Collective Validation

Incorporating the KB relation weighing mechanism of the baseline system, our QCV algorithm constructs a number of candidate graphs for a given set of collaborative mentions, and then performs a two-level ranking followed by a collective validation on those candidate graphs to acquire the linking results. Because this procedure minimally relies on linguistic analysis of the source documents while mainly uses the KB structure which by nature keeps detached from any specific language or domain, we claim that QCV comes with language and domain independence.

3.1 Candidate Graph Construction

The KB entity candidate graphs are constructed based on a mention context graph and a KB graph. We will introduce them in order as follows.

Mention Context Graph: To avoid abusing linguistic knowledge from the source documents, we construct a mention context graph G_m simply involving mention co-occurrence. Figure 1 depicts a constructed G_m for the *Caldwell* example at the beginning of Section 1. In this figure, mentions "*New York Jets*", "*Cotchery*" and "*Coles*" are brought into G_m through the coreference between "*Jets*" and "*New York Jets*" since the three of them are outside the context window of "*Caldwell*", "*Florida*", "*Patriots*", and "*Jets*". G_m contains a set of vertices representing the mentions extracted from the source document and a set of undirected edges. There will be an edge between two mention vertices if both of them fall into a context window with width w_m in the source document. Ideally, w_m should cover a single discourse according to the one sense per discourse assumption (Gale et al., 1992), but for simplicity we heuristically set w_m to be 7-sentence wide as a hyperparameter. Two mention vertices will be connected via a dashed edge if they are coreferential but are not located in the same context window. Here we determine the coreference by performing substring matching and abbreviation expansion. The dashed edge indicates the out-of-context coreferential mention together with its neighbors will be indirectly included in G_m as extended context to later facilitate the candidate graph collective validation. Note that all of these loose settings comply with our intention of generating a light-weight source context representation born with domain and language independence.

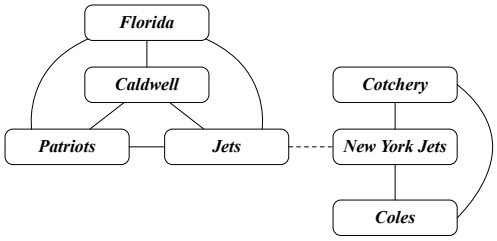

Figure 1: Mention context graph for the *Caldwell* example.

KB Graph: A structured KB such as DBpedia can be represented as a weighted graph G_k that consists of a set of vertices representing the entities and a set of directed edges labeled with relations between entities. The weights of relations are computed using Equation (1). In order to further enrich the KB relations, we add a type of relation named "wiki link" between two entities if one of them appears in the Wikipedia article of the other. Figure 2 presents a subgraph of the DBpedia KB graph containing the relevant entities in the *Caldwell* example.

Candidate Graph: The candidate graph is a set of graphs G_c^i $(i = 1, 2, ...)$ used for computing ranking scores for the KB entity candidates. For each of the mentions extracted from the source context, we first select a list of entity candidates from G_k with heuristic rules such as fuzzy string matching, synonyms, Wikipedia redirect, etc. Then we pick one candidate from each of the mentions to constitute the vertices of a G_c^i. In each G_c^i, we add an edge between two vertices if they are connected in G_k by some relation r and their mentions are connected in G_m. The edge label r from G_k is transferred to G_c^i. Upon comple-

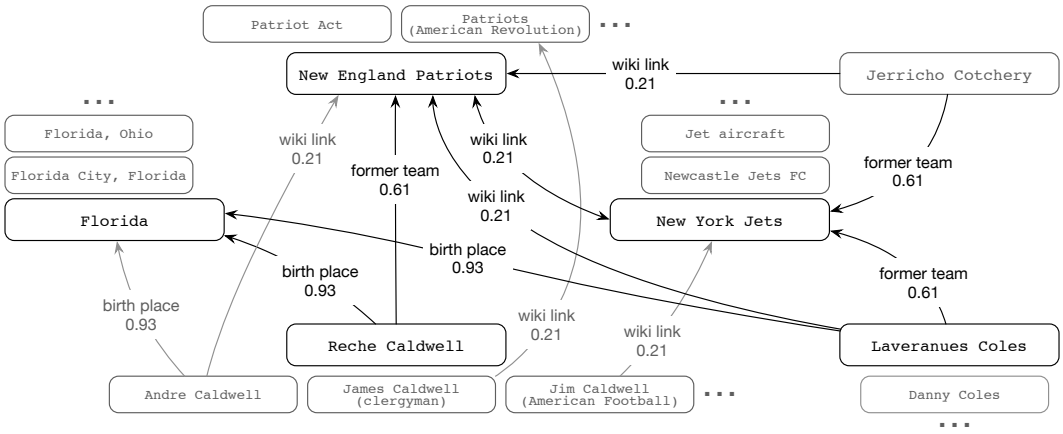

Figure 2: KB graph for the *Caldwell* example.

tion, every G_c^i represents a collective linking solution to the given mention set. Figure 3 shows three of the constructed candidate graphs for the *Caldwell* example. One can see that the first two graphs are very likely to be good solutions since they inherit many of the relation edges from G_K, while the third one is probably a poor collection as the candidates barely connect to one another. In the next section, we will more formally reveal how to rank these candidate graphs to obtain the optimal linking results.

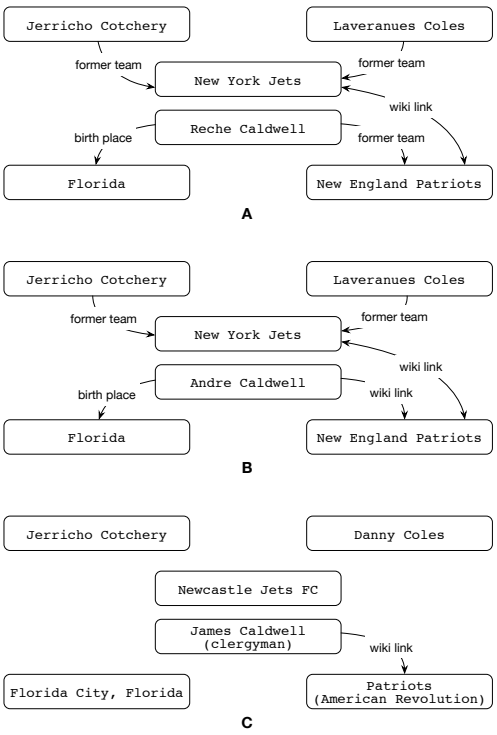

Figure 3: Candidate graphs for the *Caldwell* example.

3.2 Candidate Ranking

With the constructed candidate graphs, QCV performs two levels of ranking. First, it uses Equation (2) to compute the candidates' salience scores as a priori ranking. Then it compares each candidate graph with the mention context graph, and evaluates their vertex set similarity for context similarity ranking. Finally, by considering the relation weights in the candidate graphs as well as previous ranking scores, QCV collectively validates all the candidates and assembles the linking results. Below we will focus on introducing the context similarity ranking and the collective validation since the salience ranking resembles that of our baseline system.

Context Similarity Ranking: As shown in Figure 3, among the constructed candidate graphs, some of them contain many connected vertices while some are otherwise quite disconnected. Intuitively we would like to measure this structure difference by comparing each candidate graph G_c^i with its mention context graph G_m. Granted, we can only assert co-occurrence between two connected mentions in G_m, but it should be of great probability that two co-occurring mentions have their entity referents connected by some relation in the KB. In other words, the more a G_c^i is structurally similar to its G_m, the better the candidates in this G_c^i represent their mentions in G_m. Therefore, we define the context similarity $S_m(m_c, c)$ between a candidate c and its mention m_c using Jaccard similarity in Equation (4):

$$S_m(m_c, c) = \frac{|\Theta^{G_m}(m_c) \cap \Theta^{G_c^i}(c)|}{|\Theta^{G_m}(m_c) \cup \Theta^{G_c^i}(c)|} \quad (4)$$

where $\Theta^{G_m}(m_c)$ and $\Theta^{G_c^i}(c)$ denote m_c's neighbor set in G_m and c's neighbor set in G_c^i, respectively. The intersection takes the candidates of those mentions in $\Theta^{G_m}(m_c)$ that appear in $\Theta^{G_c^i}(c)$, and the union is equivalent to $\Theta^{G_m}(m_c)$ due to the way we construct G_c^i. We rank G_c^i using the summation of the context similarity of every c in G_c^i. Note that our baseline system uses Jaccard similarity to achieve approximate string match between the surface forms of a mention and a candidate, while we alternatively use it to capture the graph's structural similarity. After ranking with the context similarity, those G_c^i with more connected vertices such as Figure 3A and Figure 3B will get closer to the top of the ranked candidate graph list.

Candidate Graph Collective Validation: Besides the salience, the context similarity provides another ranking score for each candidate c in G_c^i, and it promotes those candidates remaining connected in G_c^i. However, it fails to differentiate how two candidates are connected. In Figure 3A, `Reche Caldwell` is a former player of `New England Patriots`, and in Figure 3B, `Andre Caldwell`'s Wikipedia article includes a hyperlink pointing to `New England Patriots`. The former seems a "tighter" relation than the latter. Although these two distinct relations imply that these two candidate pairs are related with different relation types, the context similarity rankings for these two candidate graphs are identical. Based on this observation, assuming that a "tighter" relation between two candidates is more likely to be an appropriate representation of the relation between their co-occurring mentions in the source context, we propose a novel validation step that not only considers the two previous ranking scores of each candidate but also quantitatively examines the relations between candidates. We transfer the calculated relation weights from G_k to G_c^i as positive indicators of how tightly two candidates are related, and then define the composite graph weight $W(G_c^i)$ for each G_c^i in Equation (5) as the final ranking metric:

$$W(G_c^i) = \sum_{c \in V(G_c^i)} S_a(c) S_m(m_c, c) + \sum_{r \in E(G_c^i)} H(r) \quad (5)$$

where $V(G_c^i)$ and $E(G_c^i)$ are the vertex set and the edge set of G_c^i; $S_a(c)$, $S_m(m_c, c)$, and $H(r)$ are given by Equation (2), Equation (4), and Equation (1), respectively. With this composite graph weight, since the relation "*former team*" has a greater weight than "*wiki link*", the candidate graph in Figure 3A outweighs that in Figure 3B, and therefore is ranked to the top.

4 Experiments

In this section, we first show QCV's performance on generic English corpora and compare it with our baseline together with other state-of-the-art EL systems. Then we move to a new language (Chinese) and two new domains (Biomedical Science and Earth Science) to demonstrate the language and domain independent nature of our algorithm.

4.1 EL on Generic English Corpora

For this evaluation, we used the TAC-KBP2013 EL dataset[1], which contains 2,190 mentions extracted from English newswire, web blogs, and discussion forums. We selected a subset of 1,090 linkable mentions that have entity referents in the KB for our experiment. DBpedia 3.9, which was generated from the Wikipedia dump in early 2013 and includes more than 4 million entities and more than 470 million facts[2], was used as our KB. We followed the KBP EL track using `B-Cubed+` (Ji et al., 2011) as the evaluation metric. Table 1 presents the results of QCV, our baseline system, as well as the top 3 supervised participant systems[3]and the top 3 unsupervised participant systems[3] of the TAC-KBP2013 EL track.

System	B^3+ F_1
Supervised 1st	0.724[4]
Supervised 2nd	0.721[4]
Supervised 3rd	0.718[4]
Unsupervised 1st	0.632[4]
Unsupervised 2nd	0.576[4]
Unsupervised 3rd	0.573[4]
Baseline (unsupervised)	0.697
QCV (unsupervised)	**0.749**

Table 1: Performance on the TAC-KBP2013 EL Dataset (1,090 linkable mentions).

[1]http://www.nist.gov/tac/2013/KBP/data.html

[2]http://wiki.dbpedia.org/services-resources/datasets/dataset-39

[3]Due to NIST policy, the names of the TAC-KBP2013 participant systems are not revealed.

[4]http://www.nist.gov/tac/publications/2013/papers.html

As shown in Table 1, QCV not only substantially outperforms the best unsupervised systems but also beats the best supervised systems from the KBP participants. In order to understand this notable advancement, we broke down our system into components and evaluated them accumulatively using the same dataset as above. The experiment results are summarized in Table 2.

Components	B^3+ P	B^3+ R	B^3+ F_1
SR	0.680	0.598	0.636
SR + CS	0.699	0.624	0.659
SR + CS + CV	**0.789**	**0.712**	**0.749**

Table 2: QCV Performance by Component.

In Table 2, *SR*, *CS*, and *CV* correspond to the Salience Ranking, the Context Similarity Ranking, and the Collective Validation in our QCV algorithm, respectively. It can be seen that *SR* already outperforms the best KBP unsupervised systems from Table 1. This is mainly attributed to the engagement of the entropy based relation weights which injects the impact of different relations into the entity salience. Notwithstanding being somewhat effective, *SR* solely depends on the KB and plays its role without the source context. It should be straightforward that the system performance gets improved after enabling *CS* since the source context has been incorporated. However, it was a little puzzling that the performance boost by enabling *CS* turned out to be relatively small. We took a careful look at the intermediate experiment results and discovered that although *CS* did not produce a lot more correct linking results than *SR* did, it did promote a great number of good candidates to the top of the ranking list. For example, in the *Caldwell* case, *CS* successfully raised the rankings of the context-related candidates such as `Reche Caldwell`, `Andre Caldwell`, and `Jim Caldwell`, despite the fact that it delivered `Andre Caldwell` instead of `Reche Caldwell` as the final linking result. This convincingly implies that *CS* is able to well capture the context of the target mentions, but meanwhile it is deficient in recognizing the subtle contextual difference among similar candidates. In Table 2 there is a significant performance gain after enabling *CV*. As described in Section 3.2, *CV* collectively validates the candidates of the target mention "*Caldwell*" and the mentions in its context such as "*Florida*", "*Patriots*", and "*Jets*" by integrating their *SR* and *CS* scores as well as the weights of the KB relations between them. Therefore this improvement is reasonably substantial.

By investigating the remaining errors, we identified several potential causes: 1) Our system occasionally could not capture enough context for the target mention. This happened more frequently for web blogs and discussion forums, where the language was informal and casual. Without any linguistic analysis on the source documents, it was difficult for us to extract additional context words. 2) Our simple coreference rules sometimes failed to work correctly and introduced false candidates, which, without clear context to disambiguate, could lead to linking errors. 3) Our KB had limited knowledge about some entities in a way that certain relations were missing. This kept us from creating necessary links in the candidate graphs and further effectively validating the graphs.

4.2 EL on Generic Chinese Corpora

Using Chinese as a case study, we evaluate the language portability of our approach. We used the TAC-KBP2012 Chinese EL dataset[1], and selected a subset of 1,240 linkable mentions out of the total 2,122 mentions extracted from Chinese newswire, web blogs, and discussion forums. For KB, we still used DBpedia because it contains multilingual surface forms for its entities. For instance, the entity `Barack Obama` has surface forms in over 30 languages including the Chinese one: "贝拉克·奥巴马". This cross-lingual surface form mapping naturally provides us with a convenient translation tool. Table 3 shows the linking performance comparison among QCV, our baseline system, and the top 3 participant systems of the KBP Chinese EL track. Again, we employed the `B-Cubed+` metric.

System	B^3+ F_1
Clarke et al. (2012) (supervised)	0.493
Monahan and Carpenter (2012) (supervised)	0.660
Fahrni et al. (2012) (supervised)	**0.736**
Basline (unsupervised)	0.648
QCV (unsupervised)	0.671

Table 3: Performance on the TAC-KBP2012 Chinese EL Dataset (1240 linkable mentions).

As shown in Table 3, the best performance is

[1]http://www.nist.gov/tac/2012/KBP/data.html

achieved by Fahrni et al. (2012), a supervised system using over 20 fine-tuned features and many linguistic resources. In contrast, our QCV is an unsupervised approach without using any labeled data or linguistic resources. During the error analysis, we found that in this dataset multiple mentions are often the variants of the surface form of a single KB entity. For example, "奥巴马" and "欧巴马", being just different Chinese transliterations, both refer to "*Obama*". This fact tends to result in a low recall for our system because one or more of the mention variants may not exist in the KB. We decided to heuristically apply a substring matching in addition to the Wikipedia redirection mapping to boost the recall. However, as one can imagine, this simple strategy will impair the system precision due to the introduced noise. Take "奥巴马" again for example. If we only match its second and third characters, "欧巴马" will be correctly picked, but "巴马镇" (a small town in China) will also be falsely included. Fortunately, our QCV algorithm was able to select and rank candidates complying with the source context. Consequently most of this kind of noise got filtered out, and we thus could produce balanced precision and recall.

We acknowledge that, without performing deeper linguistic analysis on the source documents, the cross-language surface form mapping of the KB plays a crucial role in our approach. One can replace it with any machine translation product which, however, is not always available especially for a low-resource language. We should take advantage of the existing KBs where such cross-lingual mapping has already been widely created. The latest DBpedia provides localized versions in 125 languages[1], for instance.

4.3 EL in Biomedical Science

To demonstrate the domain portability of our approach, we first take the biomedical science domain as a case study. We conducted our experiment using the evaluation dataset created by Zheng et al. (2014) which contains 208 linkable mentions extracted from several biomedical publications. We built our KB with over 300 domain ontologies downloaded from BioPortal. Table 4 compares the linking accuracy of QCV and our baseline system.

As shown in Table 4, our approach achieves similar performance to our baseline system which is the state-of-the-art to our knowledge. However, we were curious why QCV did not improve the baseline system in the biomedical domain as much as it did in the general domain. After some in-depth analysis of the experiment results, we discovered that in this dataset the candidates of the related mentions (*i.e.* those mentions within the same context window) mostly have similar relations in the KB. In other words, for each mention, the candidate entity types are not as diverse as those in the general domain. As a consequence, the collective validation step in QCV does not take much effect since the weights of the involved relations are quite close to one another. On such a dataset, the context similarity ranking will play a major part for the disambiguation, and QCV will not be able to function at its full power. Nonetheless, from the results we can see that our approach can be efficiently and effectively adapted to this new domain.

System	Correct	Total	Accuracy
Baseline	173	208	83.17%
QCV	177	208	**85.10%**

Table 4: Biomedical Science EL Performance.

4.4 EL in Earth Science

Now we move to another new domain, Earth Science. As far as we know, we are the first to study EL in this domain. In order to create an evaluation dataset, our domain expert selected three scientific papers about Early Triassic discovery, Global Stratotype Section, and Triassic crisis, which are three different aspects of Earth Science related discovery, and then identified 296 mentions that can be linked to DBpedia entities. Table 5 presents the linking accuracy comparison between QCV and our baseline system. We can see that QCV provided significant gains.

System	Correct	Total	Accuracy
Baseline	221	296	74.66%
QCV	236	296	**79.73%**

Table 5: Earth Science EL Performance.

The linking errors were mainly caused by the following reasons: 1) As a general KB, DBpedia has introduced certain noise for our domain-

[1] http://wiki.dbpedia.org/about

specific EL. For example, in Geology, the term "*Beds*" mostly refers to "*Geology Bed*", which is a division of a geologic formation. But in general, "*Beds*" usually means the beds people sleep on. Much more common in the KB, the latter had such a significantly higher salience score than the former that the final ranking score of our system got biased. 2) Some relations between Earth Science related entities are not clearly defined in DBpedia. For instance, in geology time scale, the period "*Chattian*" is immediately preceded by the period "*Rupelian*". An explicit relation such as "*preceded by*" should be inserted between these two period entities. Instead, only a vague "*wiki link*" relation is present in our KB. This directly diminishes the differentiating power of our system on the KB relations.

It is worth mentioning that there exists a large number of well established ontologies for different sub-domains of Earth Science. SWEET ontologies[1], for example, widely capture Earth and Environmental terminologies. By adopting these ontologies, we will be able to considerably improve our domain EL performance, and the benefits of EL in the domain will further get revealed.

4.5 System Complexity

We indexed our KB and ontologies in the format of triples using Apache Lucene[2] such that retrieving entity candidates of a mention is $O(1)$. We precomputed all the entropy-based relation weights and entity salience scores with complexities of $O(n_r \cdot n_e)$ and $O(n_e \cdot k)$, respectively, where n_r is the number of KB relations, n_e is the number of KB entities, and k is the number of iterations it took for the salience score to get converged. For the final QCV score computation, the upper bound of the computing time to link all the mentions in a document is $O(n_m \cdot n_c \cdot n_{n_c} \cdot n_{n_m})$, where n_m is the number of linkable mentions in the document, n_c is the number of candidates for each mention, and n_{n_c} is the number of neighbor nodes of a candidate, and n_{n_m} is the number of neighbors of a mention.

5 Related Work

In recent years, collective inference methods for EL have become increasingly popular. Many efforts have been devoted to encoding linguistic features from the source documents in order to precisely select collaborator mentions for collective inference. These features include topic modeling (Xu et al., 2012; Cassidy et al., 2012), relation constraint (Cheng and Roth, 2013), coreferential chaining (Nguyen et al., 2012; Huang et al., 2014), and dependency restriction (Ling et al., 2014). Some recent work utilized multi-layer linguistic analysis integration to capture contextual properties for better mention collection (Pan et al., 2015). While many of these approaches have been proved to be effective, the dependency on deep linguistic knowledge makes it difficult to migrate them to a new language or domain. In contrast to these methods, we establish a very loose setting for the mention selection, and rely on the quantified information computed from the structured KB to collectively evaluate and validate the entity candidates. Since the KB is relatively universal to languages and domains, our approach inherently is language and domain independent.

Recent cross-lingual EL approaches can be divided into two types. The first type (McNamee et al., 2011; Cassidy et al., 2011; McNamee et al., 2012; Guo et al., 2012; Miao et al., 2013) translated entity mentions and source documents from the new language into English and then ran English mono-lingual EL to link to English KB. The second type (Monahan et al., 2011; Fahrni and Strube, 2011; Fahrni et al., 2012; Monahan and Carpenter, 2012; Clarke et al., 2012; Fahrni et al., 2013) developed EL systems on the new language and used cross-lingual KB links to map the link results back to English KB. While the bottleneck of the former method usually is on translation errors, the latter approach heavily relies on the linguistic resources and the KB of the new language. In comparison, our system mainly uses the English KB and a mention surface form mapping that can either come from translation or cross-lingual KB links, and requires minimal linguistic resources from the new language.

There is a limited amount of research work in the literature that focused solely on domain-specific EL (Zheng et al., 2014). In the biomedical domain, a few studies have been found on EL-related tasks such as scientific name discovery (Akella et al., 2012), gene name normalization (Hirschman et al., 2005; Fang et al., 2006; Dai et al., 2010), biomedical named entity recognition (Usami et al., 2011; Van Landeghem et al.,

[1] http://sweet.jpl.nasa.gov

[2] https://lucene.apache.org/

2012) and concept mention extraction (Tsai et al., 2013). The baseline system (Zheng et al., 2014) in this paper is the work most similar to ours in a sense of collectively aligning mentions to structured KBs. However, our system differs by integrating a context similarity ranking and a candidate validation to conduct a two-way collective inference with better performance.

6 Conclusions and Future Work

Language and domain independence is a new requirement to EL systems and this capability is particularly welcome by low-resource language related applications and domain scientists. In this paper we demonstrated a high-performance EL approach that can be easily migrated to new languages and domains due to the minimal reliance on linguistic analysis and the deep utilization of structured KBs. In the future, we plan to improve the source document processing such that the system can better extract the mention context without involving extensive linguistic knowledge. We are also experimenting with our collective validation algorithm to incorporate the impact of more distant KB entities other than just the neighbors.

7 Acknowledgement

This work was supported by the U.S. DARPA DEFT Program No. FA8750-13-2-0041, ARL NS-CTA No. W911NF-09-2-0053, NSF CAREER Award IIS-1523198, DARPA LORELEI, AFRL DREAM project, 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. M. Akella, C. N. Norton, and H. Miller. 2012. NetiNeti: Discovery of Scientific Names from Text Using Machine Learning Methods. *BMC Bioinformatics*, 13:211.

T. Cassidy, Z. Chen, J. Artiles, H. Ji, H. Deng, L. Ratinov, J. Zheng, J. Han, and D. Roth. 2011. CUNY-UIUC-SRI TAC-KBP2011 Entity Linking System Description. In *Proceedings of Text Analysis Conference 2011*.

T. Cassidy, H. Ji, L. Ratinov, A. Zubiaga, and H. Huang. 2012. Analysis and Enhancement of Wikification for Microblogs with Context Expansion. In *Proceedings of the 25th International Conference on Computational Linguistics*.

X. Cheng and D. Roth. 2013. Relational Inference for Wikification. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*.

J. Clarke, Y. Merhav, G. Suleiman, S. Zheng, and D. Murgatroyd. 2012. Basis Technology at TAC 2012 Entity Linking. In *Proceedings of Text Analysis Conference 2012*.

S. Cucerzan. 2011. TAC Entity Linking by Performing Full-Document Entity Extraction and Disambiguation. In *Proceedings of Text Analysis Conference 2011*.

H. Dai, P. Lai, and R. T. Tsai. 2010. Multistage Gene Normalization and SVM-Based Ranking for Protein Interactor Extraction in Full-Text Articles. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 7(3):412–420.

J. Dalton and L. Dietz. 2013. A Neighborhood Relevance Model for Entity Linking. In *Proceedings of the 10th Conference on Open Research Areas in Information Retrieval*.

A. Fahrni and M. Strube. 2011. HITS' Cross-lingual Entity Linking System at TAC2011: One Model for All Languages. In *Proceedings of Text Analysis Conference 2011*.

A. Fahrni, T. Göckel, and M. Strube. 2012. HITS' Monolingual and Cross-lingual Entity Linking System at TAC 2012: A Joint Approach. In *Proceedings of the Text Analysis Conference 2012*.

A. Fahrni, B. Heinzerling, T. Gockel, and M. Strube. 2013. HITS' Monolingual and Cross-lingual Entity Linking System at TAC 2013. In *Proceedings of Text Analysis Conference 2013*.

H. Fang, K. Murphy, Y. Jin, J. S. Kim, and P. S. White. 2006. Human Gene Name Normalization Using Text Matching with Automatically Extracted Synonym Dictionaries. In *Proceedings of the Workshop on Linking Natural Language Processing and Biology: Towards Deeper Biological Literature Analysis*, pages 41–48.

N. Fernandez, J. A. Fisteus, L. Sanchez, and E. Martin. 2010. WebTLab: A Cooccurence-Based Approach to KBP 2010 Entity-Linking Task. In *Proceedings of Text Analysis Conference 2010*.

W. A. Gale, K. W. Church, and D. Yarowsky. 1992. One Sense per Discourse. In *Proceedings of the Fifth DARPA Speech and Natural Language Workshop*.

Z. Guo, Y. Xu, F. Mesquita, D. Barbosa, and G. Kondrak. 2012. ualberta at TAC-KBP 2012: English and Cross-Lingual Entity Linking. In *Proceedings of Text Analysis Conference 2012*.

X. Han, L. Sun, and J. Zhao. 2011. Collective Entity Linking in Web Text: A Graph-Based Method. In *Proceedings of the 34th Annual ACM SIGIR Conference*.

L. Hirschman, M. Colosimo, A. Morgan, and A. Yeh. 2005. Overview of BioCreAtivE task 1B: Normalized gene lists. *BMC Bioinformatics*, 6.

H. Huang, Y. Cao, X. Huang, H. Ji, and C. Lin. 2014. Collective Tweet Wikification based on Semi-supervised Graph Regularization. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*.

H. Ji, R. Grishman, and H.T. Dang. 2011. Overview of the TAC 2011 Knowledge Base Population Track. In *Proceedings of Text Analysis Conference 2011*.

X. Ling, S. Singh, and D. S. Weld. 2014. Context Representation for Named Entity Linking. In *Proceedings of the 3rd Pacific Northwest Regional NLP Workshop*.

P. McNamee, J. Mayfield, D. Lawrie, D. W. Oard, and D. Doermann. 2011. Cross-Language Entity Linking. In *Proceedings of the 5th International Joint Conference on Natural Language Processing*.

P. McNamee, V. Stoyanov, J. Mayfield, T. Finin, T. Oates, T. Xu, D. W. Oard, and D. Lawrie. 2012. HLTCOE Participation at TAC 2012: Entity Linking and Cold Start Knowledge Base Construction. In *Proceedings of Text Analysis Conference 2012*.

Q. Miao, R. Fang, Y. Meng, and S. Zhang. 2013. FRDC's Cross-lingual Entity Linking System at TAC 2013. In *Proceedings of Text Analysis Conference 2013*.

S. Monahan and D. Carpenter. 2012. Lorify: A Knowledge Base from Scratch. In *Proceedings of Text Analysis Conference 2012*.

S. Monahan, J. Lehmann, T. Nyberg, J. Plymale, and A. Jung. 2011. Cross-Lingual Cross-Document Coreference with Entity Linking. In *Proceedings of Text Analysis Conference 2011*.

H. Nguyen, H. Minha, T. Cao, and T. Nguyenb. 2012. JVN-TDT Entity Linking Systems at TAC-KBP2012. In *Proceedings of Text Analysis Conference 2012*.

X. Pan, T. Cassidy, U. Hermjakob, H. Ji, and K. Knight. 2015. Unsupervised Entity Linking with Abstract Meaning Representation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics – Human Language Technologies*.

M. Pennacchiotti and P. Pantel. 2009. Entity Extraction via Ensemble Semantics. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing EMNLP2009*.

L. Ratinov, D. Roth, D. Downey, and M. Anderson. 2011. Local and Global Algorithms for Disambiguation to Wikipedia. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*.

C. Tsai, G. Kundu, and D. Roth. 2013. Concept-based analysis of scientific literature. In *CIKM*.

Y. Usami, H. Cho, N. Okazaki, and J. Tsujii. 2011. Automatic Acquisition of Huge Training Data for Bio-medical Named Entity Recognition. In *Proceedings of BioNLP 2011 Workshop*.

S. Van Landeghem, J. Björne, T. Abeel, B. De Baets, T. Salakoski, and Y. Van de Peer. 2012. Semantically Linking Molecular Entities in Literature through Entity Relationships. *BMC Bioinformatics*, 13.

J. Xu, Q. Lu, J. Liu, and R. Xu. 2012. NLPComp in tac 2012 entity linking and slot-filling. In *Proceedings of Text Analysis Conference 2012*.

J. Zheng, D. Howsmon, B. Zhang, J. Hahn, D. McGuinness, J. Hendler, and H. Ji. 2014. Entity Linking for Biomedical Literature. In *Proceedings of the ACM 8th International Workshop on Data and Text Mining in Bioinformatics*, pages 3–4, New York, NY, USA. ACM.