Entity Linking to One Thousand Knowledge Bases

Ning Gao[1] and Silviu Cucerzan[2]

[1]University of Maryland, College Park
ninggao@umd.edu
[2]Microsoft Research
silviu@microsoft.com

Abstract. We address the task of entity linking to multiple knowledge bases (KB). In particular, we investigate the use of over one thousand domain-specific KBs derived from Wikia.com collections in conjunction with the Wikipedia collection as a background-knowledge repository. Our system employs a two-step approach: for each document, a supervised model with a large set of features detects whether there exists a Wikia collection whose domain matches the document; when such a collection is available, the system extracts and resolves the entity mentions in the document to the KB obtained by merging the Wikipedia KB and the KB corresponding to the matched Wikia collection. Otherwise, the system employs only the background KB for analysis, in a standard entity-detection-and-linking framework. On a Web news articles dataset, our system achieves 90% precision in detecting domain-accurate Wikia collections while providing also high linking accuracy (93%) to the KB of the matched Wikia collection.

1 Introduction

Entity detection and linking (EDL), also known as entity recognition and disambiguation (ERD), is the task of identifying mentions of entities in text (*detection/recognition*) and assigning the detected mentions to entities in a large knowledge base (*linking/disambiguation*). The establishment of large encyclopedic collections such as Wikipedia and Freebase has drawn considerable attention towards this task [2, 3, 6, 10, 14, 16]. However, focusing exclusively on Wikipedia or similar general knowledge repositories is not sufficient in many real-world scenarios, such as entity-based indexing of corporate data and entity-based indexing of news because numerous domain-specific entities and concepts are absent from general-use encyclopedic KBs (e.g., finance, food, fiction).

Our study finds that more than 80% of the entities in the domain-specific collections used in this paper are likely not to have entries in Wikipedia. In particular, while there exists a comprehensive Wikia collection for the Pokémon world with over 14,000 entity pages, Wikipedia contains only a few tens of pages for Pokémon entities. Similarly, there is a Wikia collection for the "Survivor" television franchise with over 3,000 entity pages, of which only very few have pages in Wikipedia.

Most of the previous work on entity linking employs only one single KB as linking target. To the best of our knowledge, there has been very little done to address the problem of entity linking to a large number of KBs automatically. We propose a framework in which one large encyclopedic KB is employed as a comprehensive repository of general knowledge in conjunction with over one thousand independent domain-specific KBs. For the former, we employ Wikipedia; for the latter, we employ a large set of Wikia collections from `http://www.wikia.com/Wikia`.

Rather than attempting to merge all KBs into one single knowledge base, we employ a paradigm in which the domain-specific KBs are kept separate, and they *get activated on the fly* and used in conjunction with the general KB when they are needed for analyzing a document pertaining to those respective domains. This paradigm is interesting because of several aspects: First, merging a large collection of diverse ontologies/KBs is very challenging. Second, in an enterprise context, many domain-specific KBs are designed to be protected with authority access and are not allowed to be merged into a general, externally-maintained repository. For example, aeronautical companies such as Boeing and Airbus deal with domain/company-specific terminologies as well as numerous technologies that need be kept private (from the ouside world or from each other) in addition to many entities and concepts that are part of common knowledge. Conflating such terminologies and knowledge repositories into one KB or with each other would be both daunting and undesirable. Third, facts and relations from some domain-specific collections, such as those that target fictional work, may be valid only with respect to that particular domain. For example, the city of London has entries in *Harry Potter* Wikia, *Baker Street* Wikia and hundreds of other Wikia collections. However, the *British Ministry of Magic* being located in *London* is a “fact” only in the *Harry Potter* domain. Merging directly the knowledge for entities in different domains into one canonical entity entry can lead to inaccuracies, conflicting information, and noise.

2 Related Work

The tasks of linking to multiple KBs and merging KBs for entity linking have been tackled only to a small degree previously. Most entity linking work, starting with the works of Bunescu and Paşca [1], Cucerzan [3], and Mihalcea and Csomai [10], has employed Wikipedia for deriving a reference KB for the task. The Text Analysis Conference track on Knowledge Base Population (TAC-KBP) established the evaluation framework for entity linking [7, 8], in which a target KB with over 800,000 entities was derived from the Wikipedia collection as of October 2008, and thousands of documents from a large corpus of news and Web text were annotated with entity mentions. Interestingly, 57% of the evaluated entities were not in the targeted KB [9].

Ruiz-Casado et al. [13] and Niemann et al. [11] studied the task of automatically assigning Wikipedia entries to WordNet synsets, which can be considered as simple one-direction merging of *two* KBs. However, extending those approaches to bi-directional merging over thousands of KBs seems extremely difficult. Sil et al. [15] proposed an open-database named entity disambiguation system that is able to resolve entity mentions detected in text to an arbitrary KB provided in Boyce-Codd normal form. However, this work focuses on distant supervision and domain-adaption, and relies on *manually* identifying a KB that matches the analyzed documents, without addressing the tasks of detecting domain-specific KBs or maintaining a multi-KB structure automatically. Demartini et al. [5] used probabilistic reasoning and crowdsourcing techniques for the task of entity linking over *four* KBs (DBpedia, Freebase, Geonames and New

		Wikipedia	**Wikia set**	
			macro	**micro**
number of pages		4,591,935	3,059,412	
page length	*mean*	4,284	2,573	2,454
	stdev	7,901	2,992	8,801
inner links	*mean*	31.6	8.4	11.8
	stdev	77.8	6.9	36.7
out links	*mean*	0	0.14	0.09
	stdev	0.09	0.65	2.18
cross links	*mean*		0.03	0.03
	stdev		0.1	4.19

Table 1. Page and linkage statistics for Wikipedia and Wikia.

York Times). However, the KBs are simply “merged”, and then the candidate entities are triaged by TF-IDF methods. Pereira [12] proposed an idea of resolving the task of entity linking to multiple KBs by using different textual and KB features, and ontology modularization to select entities in the same semantic context, although the detailed structure is not discussed in the paper.

These approaches have dealt with a small number of KBs, for either the purpose of entity linking or similar tasks, none of which has the ability to deal with the complexity caused by a very large number of KBs. Employing such a large repository of KBs and automatically mapping documents to domain-specific KBs from this repository to perform entity linking has not been reported until now.

3 Datasets

We employed the English Wikipedia collection from August 11, 2014, and we crawled all 1,163 available Wikia collections in English within the top 5,000 Wamranked list[1] as of June 13, 2014. Table 1 compares page and linkage statistics between Wikipedia and the employed Wikia collections. *Number of pages* denotes the total number of entity pages in Wikipedia and all Wikia collections. *Page length* shows the mean and standard deviation (SD) for the length (in characters) of Wikipedia and Wikia pages. *Micro* shows the average over all the Wikia pages, while *Macro* shows the average over all the Wikia collections. While the Wikipedia pages are on average 1.7 times longer than Wikia pages, the difference between the average page length of the two sources is not significant given the high standard deviations inside these collections.

Table 1 also shows statistics for the existing linkage between collections as created by the wiki contributors. *Inner links* is the average number of links on a page from one collection (whether Wikipedia or a Wikia collection) to other pages in the same collection. *Out links* for Wikipedia is the average number of links from a Wikipedia page to pages in any Wikia collection. *Out links* for Wikias is the average number of links from a Wikia page to Wikipedia pages. *Cross links* is the average number of links

[1] Wamrank is the official ranking from the Wikia website, which evaluates the health and vitality of collections.

	Editorial Links	**Automatic Links**
Wikia collections with links to Wikipedia	920	1,163
Wikia-to-entity linkage	58	616
entity-to-entity linkage	4,456	518,335
mention-to-entity linkage	196,928	1,801,203

Table 2. Statistics for *editorial links* (contributor-created) and *automatic links* (as generated by NEMO) from Wikia to Wikipedia.

from pages in one Wikia to pages in other Wikias. Compared to the rich *inner links*, both *out links* and *cross links* for Wikipedia and Wikia are sparse, suggesting that the Wikipedia and the domain-specific Wikias are relatively isolated from each other.

4 Linking Wikia Collections to Wikipedia

Because editorial links between Wikipedia and the Wikias are quite rare, we attempt to connect more strongly the Wikia collections in our study to Wikipedia by employing NEMO, a state-of-the-art Wikipedia-based EDL system [4]. The text of each Wikia page is analyzed with NEMO to identify entity mentions and automatically link them to Wikipedia when possible. Table 2 shows more statistics for both editorial links and automatically generated links from Wikia collections to Wikipedia. When accounting only for editorial links, 243 out of the 1,163 Wikias are completely isolated from Wikipedia (i.e., they do not contain any links to Wikipedia). As expected, the automatic linking process is able to connect all Wikia collections to Wikipedia. We split both types of links from Wikia to Wikipedia into three mutually exclusive sets, as follows:

Wikia-to-entity are the links for which the entity mention appears on the homepage of the Wikia collection and is string-wise identical with both the name of the Wikia collection and the mapped Wikipedia entity. For example, in the text of the *Harry Potter* Wikia's homepage, there is a mention of "Harry Potter" that gets identified and linked to the Wikipedia page *Harry Potter*. These types of links are likely to capture cases in which there exists a whole Wikia collection dedicated to one entity in Wikipedia.

Entity-to-entity are the links for which the entity mention in the text of a Wikia page is identical to the title of the Wikia page. For example, on the page *J.K. Rowling* in the *Harry Potter* Wikia, there is a mention "J.K. Rowling", which gets linked to the Wikipedia entity *J.K. Rowling*. These links are likely to indicate duplicate coverage of an entity in both Wikipedia and the analyzed Wikia collection.

Mention-to-entity are the links that are not included in the former two categories. For example, on the page *Zubeida Khan* in the *Harry Potter* Wikia, there is a mention "Pakistan" linked to the Wikipedia entity *Pakistan*. These links are likely to indicate the case when a Wikipedia entity being mentioned in the analyzed Wikia collection.

Using all editorial links as ground-truth, we evaluate the recall and accuracy for the automatic entity detection (ED) and entity linking (EL) processes. The results are shown in Table 3. Because of the large number of Wikia contributors, who are not trained for NLP style annotations, the linkage is inconsistent (for example, some links include determiners, possessive particles, or titles/occupations, while others do not),

ED recall for overlapping mentions	0.74
ED recall for exact boundaries	0.63
EL Wikia-to-entity accuracy	1.00
EL entity-to-entity accuracy	0.86
EL mention-to-entity accuracy	0.84

Table 3. Performance of entity detection (ED) and linking (EL) from Wikia to Wikipedia.

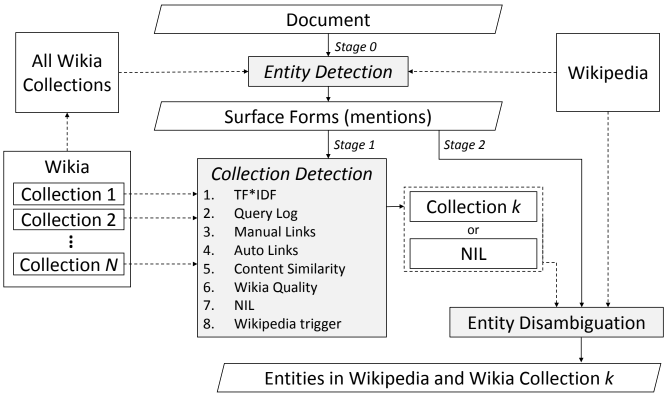

Fig. 1. System architecture.

which we refer to as boundary inconsistency. In our error analysis of the missed links, we found out that 4.2% of those are numbers and 33% are lowercase words (e.g., blurb, film, novel). For the detected links with identical boundaries, we further evaluate the disambiguation accuracy of the employed linking system for each of the three types of links. We obtain 100% accuracy for Wikia-to-entity links, and close to 85% for the other two types of links. We also measured the linking accuracy for the detected overlapping mentions with different boundaries and obtained that over half of those were resolved to the same Wikipedia entities as those chosen by the Wikia contributors.

5 System Architecture

We propose an approach in which we use Wikipedia as a background encyclopedic KB and Wikia collections as domain-specific KBs. As shown in Figure 1, for each input document, our multi-KB entity linking system first employs in *Stage 0* the information from all Wikia collections and Wikipedia to detect candidate entity mentions, referred to as *surface forms* henceforth. It then detects in *Stage 1* the best matching Wikia collection (or returns NIL if no matching Wikia collection is found). In *Stage 2*, it links the surface forms in the document to either Wikipedia or the selected Wikia collection. This final stage employs a modified version of a state-of-the-art EDL system [4], which targets Wikipedia and the selected Wikia collection as one single KB, but with two strongly connected subcomponents. When a mention can be resolved to both Wikipedia and Wikia pages, we give preference to the disambiguation in Wikia, based on the intuition that pages from the domain-specific collections are more relevant in the context of a document that belongs to the respective domain. For example, in an article about the “Sherlock and Holmes” movie, linking the mention “London” to the entity *London* in the *Baker Street* Wikia should be preferable to linking it to *London* in Wikipedia.

The detection of an appropriate Wikia collection is vital in this framework because incorrect collection choice can lead to wrong linked entities. We frame the collection detection task as a supervised ranking problem, in which the Wikia collections are scored and ranked based on how well they match the input document. We employ a boosted-tree learning-to-rank framework in which we investigate eight groups of features that attempt to measure the topical matching between a document and each Wikia collection, as described further. Several feature groups are novel by task design (i.e., Wikia to Wikipedia Linkage, Wikia Collection Quality, Wikipedia Triggers). Other feature groups (e.g., TFIDF, Content Matching) have been widely used in other tasks such as document classification and document retrieval.

TF*IDF. We employ a group of 13 features that are variants of similarities between a document and a Wikia collection, as inspired by the TF*IDF framework. Formally, for an input document d with m detected surface forms $S_d = \{s_d^1, \cdots, s_d^m\}$ and for a candidate Wikia collection c, let tf_i denote the frequency of s_d^i in the input document, $|c|$ the number of pages in c, w_i the term frequency of s_d^i in collection c, and idf_i the inverse document frequency of s_d^i in all the Wikia collections, computed as $idf_i = \log_2(1 + \frac{N}{N(s_d^i)})$, where N=1,163 is the number of Wikia collections in our repository and $N(s_d^i)$ is the number of those collections that contain s_d^i as a linked mention (either editorial or automatic). The formulas for the features employed are as following (all sums are over all mentions from s_d^1 to s_d^m):

$$\begin{array}{lll}
\sum_{i=1}^m tf_i * idf_i & \sum_{i=1}^m \sqrt{tf_i} * idf_i & \sum_{i=1}^m \frac{tf_i * idf_i * w_i}{\sqrt{|c|}} \\
\sum_{i=1}^m \frac{tf_i * idf_i}{\sqrt{|c|}} & \sum_{i=1}^m \frac{\log_2(tf_i) * idf_i}{\sqrt{|c|}} & \sum_{i=1}^m \frac{\sqrt{tf_i * w_i * idf_i}}{\sqrt{|c|}} \\
\sum_{i=1}^m tf_i * idf_i * w_i & \sum_{i=1}^m \sqrt{tf_i * w_i} * idf_i & \sum_{i=1}^m \log_2(tf_i) * idf_i \\
\sum_{i=1}^m \frac{\log_2(tf_i * w_i) * idf_i}{\sqrt{|c|}} & \sum_{i=1}^m \frac{\sqrt{tf_i} * idf_i}{\sqrt{|c|}} & |c| \\
\sum_{i=1}^m \log_2(tf_i * w_i) * idf_i & &
\end{array}$$

Web Search Logs. We attempt to capture the relatedness between Wikipedia pages and Wikia collections by mining the query logs of a major Web search engine to identify queries for which a user visited both a Wikipedia page and a Wikia page for more than 30 seconds each after retrieving them as search results. In such cases, we create a connection between the respective Wikipedia page and Wikia collection. For example, if we detect that a user submitted the query “jeff moss”, then visited the Wikipedia *Jeff Moss* page and also the *Jeff Moss* page in the *Muppet* Wikia, which were returned by the search engine in the top ranked results, we create a connection between the Wikipedia entity *Jeff Moss* and the Wikia collection *Muppet*. In this way, we associate to each Wikia collection c a set of Wikipedia entities $E_c = \{e_c^1, \cdots, e_c^{m_c}\}$. We were able to extract in total 72,030 such connections. Given an input document, we analyze it first with the Wikipedia-based EDL system and obtain a set $E_d = \{e_d^1, \cdots, e_d^m\}$ of extracted Wikipedia entities. We compute the relatedness of the input document and a Wikia collection c as the cardinality of the intersection $|E_d \cap E_c|$.

Wikia to Wikipedia Linkage. The editorial and automatic links from Wikia collections to Wikipedia pages are also used to calculate the similarity between an input document and each Wikia collection. Let e_d^i be a referenced Wikipedia entity in the analyzed document, and l_c^i the number of links from Wikia collection c to the Wikipedia entity e_d^i. Then $\sum_{i=1}^{m} l_c^i$ is the total number of links from the Wikia collection c to the entities $E_d = \{e_d^1, \cdots, e_d^m\}$ referenced in the input document. This number can be employed as a similarity score between the document and the collection c. Since we have both editorial links and automatic links from the Wikia collections to Wikipedia, further organized into three categories (Wikia-to-entity, entity-to-entity, and mention-to-entity), we can compute in this manner a total of six features.

Content Matching. We devise a group of 5 features to measure the importance of the surface forms detected in a document d with respect to a candidate Wikia collection c. As previously, let S_d denote the set of surface forms in the input document and $w_c(s)$ denote the frequency of a surface form s as a linked mention (either editorially or automatically) in collection c. We denote with L_c the set of all surface forms employed in linked entity mentions in collection c. We calculate as features the number of surface forms from the document that are in the collection $|S_d \cap L_c|$, the total frequency of the matched surface forms in the collection $\sum_{s \in S_d \cap L_c} w_c(s)$, the coverage of the matched surface forms in the Wikia collection $\frac{|S_d \cap L_c|}{|L_c|}$, the frequency-based coverage $\frac{\sum_{s \in S_d \cap L_c} w_c(s)}{\sum_{l \in L_c} w_c(l)}$, as well as a binary indicator whether the title/first line of the document contains the Wikia collection name as a substring.

Wikia Collection Quality. We employ 5 features to measure the quality of Wikia collections, as well as the novelty provided by each Wikia collection with respect to Wikipedia. We employ the Wamrank of Wikia collections as a measure of quality. For novelty, we use the number of surface forms in the collection that are novel with respect to Wikipedia, and the percentage of novel surface forms in the target Wikia. Additionally, we compute the number and the percentage of surface forms from the input document that are in the candidate Wikia but not in Wikipedia.

NIL. We employ a binary feature as a NIL indicator, which is set to 1 for NIL and 0 for any candidate collection c.

Wikipedia Triggers. After we train Stage 1 by using only the previously discussed 31 features on the training set employed in the experiments on news articles (as discussed in the next section), we employ it to pre-analyze all Wikipedia pages and to detect for each page the best matching Wikia collection. We obtain that 31% of the Wikipedia pages *trigger* a Wikia collection, while 69% get assigned the NIL class. For example, the Wikipedia page *Albus Dumbledore* triggers the *Harry Potter* Wikia, while the page *Piotr Kuncewicz* does not trigger any Wikia collection. We can devise now an extra binary feature for the matching of a Wikia collection c to an input document d, by using the triggers of the entities extracted by the Wikipedia-based EDL system from the d. We assign to this feature the value 1 if there exists a Wikipedia entity extracted from d that triggers c, and 0 otherwise.

average number of candidates	206
recall	0.95
precision (accuracy/recall)	0.84
accuracy	0.79
MRR	0.85

Table 4. Wikia collection detection for Wikia pages.

6 Experiments

As noted in Section 2, there are no existing systems or data collections designed for the task of entity linking to a large number of KBs. To evaluate our work, we employ two sets of documents, consisting of Wikia pages and news stories.[2]

6.1 Wikia Pages

From all Wikia pages with more than three linked entity mentions in the 1,163 collections employed in our study, we randomly select a quarter for training the boosted tree ranking system of Stage 1, a quarter for dev-test, and the remaining half (1,568,325 pages) for final testing. Each Wikia page is employed as an input document to our system. We use the collection to which a page belongs as ground-truth for Wikia detection (Stage 1), and the editorial links in the text of each Wikia page to other Wikia pages or Wikipedia pages as ground-truth for linking (Stage 2). Note that because any Wikia page in these sets belongs to one of the candidate Wikias, we do not have NIL triggers in this setting.

Wikia Collection Detection. Table 4 shows the performance of Wikia detection for the Wikia pages in the test set. The *average number of trigger candidates* to Wikia pages is 206. The *recall* and *precision* @1 are 0.95 and 0.84 respectively. The *accuracy* for detecting the ground-truth Wikia collection is 0.79, and the Mean Reciprocal Rank (MRR) is 0.85. In our error analysis, we found that there are 316 Wikia collection pairs that have more than 100 pages wrongly detected to each other as the target Wikia collection, which account for 43.4% of the detection errors. Table 5 shows the top five such pairs. For example, there are 5,302 pages in the *starwars* Wikia and the *swfanon* Wikia wrongly detected as being from the other collection. The former contains information about the “Star Wars” universe, including movies, characters, video games, etc., while the latter contains information about “Star Wars” gathered or written by fans. We manually browsed the home pages of these detected Wikia pairs and judged which pairs are *comparable* (cover the same domain). We found that 89.6% of them are comparable collections, and thus, linking entities to one another is actually informative.

Feature Study.

To further analyze the performance of each feature group, we randomly select 20 documents with at least three linked entity mentions from each Wikia collection. We split this into two groups of 11,630 Wikia pages for training and testing. Figure 2 shows

[2] The annotations can be downloaded at `http://www.umiacs.umd.edu/~ninggao/publications`.

Wikia pairs	**confusion count**
swfanon, starwars	5,302
stexpanded, memory-beta	2,427
pokemonfanon, pokemon	2,387
classiccars, automobile	1,737
doctor-who-collectors, tardis	1,728

Table 5. Top five Wikia pairs that get confused to one other

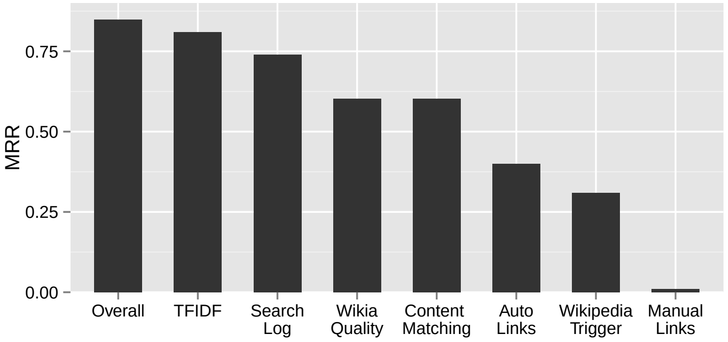

Fig. 2. Performance of single feature groups

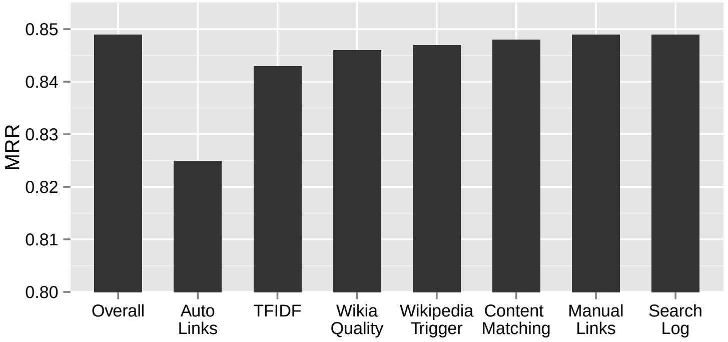

Fig. 3. Feature group ablation study

the MRR obtained overall (0.85) and by using individual feature groups. The TF*IDF group performs the best for Wikia detection (0.82), followed by the Web search logs group (0.74), Wikia quality (0.60), and the content matching features (0.60). A rather surprising finding is the extremely poor performance of the editorial (manual) linkage subgroup (0.01) in comparison to the automatic linkage subgroup (0.40). While the features based on automatic links do not achieve very high MRR by themselves in our

Total number of links to Wikia	8,305,241	*Total number of links to Wikipedia*	271,744
ED recall for overlapping mentions	0.77	*ED recall for overlapping mentions*	0.66
ED recall for exact boundaries	0.67	*ED recall for exact boundaries*	0.53
EL Wikia accuracy	0.93	*EL Wikipedia accuracy*	0.74
Wikipedia predicted links / Wikia ground-truth	0.01	*Wikia predicted links / Wikipedia ground-truth*	0.12

Table 6. Entity detection and linking results

feature ablation study shown in Figure 3, we note that that the overall accuracy gets the largest drop when the auto links features are suppressed (from 0.85 to 0.82).

Entity Linking Accuracy.

For the Wikia pages with correctly triggered Wikia in Stage 1, we evaluate the linking of the detected surface forms to Wikipedia and the triggered Wikia. As previously, we use the links created by the Wikia contributors as the linkage ground truth. The results are shown in Table 6. The EL evaluation is done only for exact boundary matching. As expected, the performance of *ED recall* and *EL accuracy* is substantially better for Wikia than for Wikipedia, partly because the input documents belong to the domain of the triggered Wikia, and partly because the ambiguity of surface forms in each Wikia is much lower than the ambiguity in Wikipedia.

Table 6 also shows that 1% of the surface forms linked by Wikia contributors to other pages in Wikia get linked by our system to Wikipedia. Conversely, 12% of surface forms that are linked by contributors to Wikipedia are resolved by the employed EDL system to Wikia. While we did not perform an exhaustive analysis of those errors, we noticed that a large number of them are outer links to Wikipedia created for the purpose of interlinking the collections, as in the sentence “Reference Wikipedia Harry Potter page”, in which “Harry Potter” is linked to *Harry Potter* in Wikipedia, while our systems links it to the page with the same name in the triggered *Harry Potter* Wikia.

6.2 News Articles

Wikia pages provide a large pre-annotated dataset for training and evaluating. However, there are two shortcomings with the strategy of employing them for training and testing. First, there are no NIL examples in these data. Second, the Wikia pages in the test are likely to be more similar content-wise to the pages on which the training is done. Therefore, we also evaluate the performance of our multi-KB entity linking system on a set of news articles crawled from the Web.

Wikia Detection for News Articles. For training, we use: (1) 1,000 local news articles from a local news station in a major city (names not disclosed to preserve anonymity of the review process). We automatically assign NIL as triggers by adding “NIL” as a trigger candidate; (2) 3,000 randomly selected Wikia pages, each annotated with the collection to which it belongs as trigger; (3) 246 news articles from the same news source, with the property that their title contains one of the names of the Wikia collections employed in our study. For example “Batman star Christian Bale visits shooting victims” contains the Wikia name *Batman*, which we assign to this article as its Wikia trigger. (4) 300 randomly selected Wikipedia pages with manually assigned Wikia triggers by human annotators as ground truth. For testing, we use 260 news articles selected from 15 popular news sites (e.g. CNN), selected through a process that

Non-NIL (65%)	**correct**	**correct + interesting**
precision@1	0.90	0.95

NIL(35%)	**correct**
precision@1	0.77

Table 7. News article Wikia triggering.

	Wikia	**Wikipedia**
ED accuracy	1	0.942
EL accuracy	0.928	0.805

Table 8. Evaluation for news article entity detection and linking.

uses the query logs of a large Web search engine as follows: after submitting a query, the user must open a Wikipedia page, a Wikia page, and a news article from the selected websites in any order, and spend at least 30 seconds on each of those pages.

Table 7 shows the trigger Wikia detection results for the tested news articles. To obtain the ground-truth, two judges read the news article and judged post-hoc the top 3 returned trigger Wikia collections (possibly including NIL), by employing three labels: *correct*, *interesting* or *wrong*. For example, for a news article with title “Game Of Thrones author George RR Martin talks season 4”, the Wikia collection of *Game of Thrones* was the correct answer. While the Wikia collection of *Ice and Fire* is not correct, since this Wikia focuses on the book series rather than the TV series, it might still be labeled as an interesting collection for linking. Note that for a news article, there could be more than one correct trigger Wikia collection. For example, both collections *marvel-movies* and *marvel* are judged as correct for a news article titled “Agent Coulsons’ secret is out”. The annotation inter-agreement is 91.5%. Table 7 shows the results of this evaluation, with a relatively high precision @1 of 0.90 for Wikia triggers and 0.77 for NIL when employing the strict definition of correctness. Moreover, for the news articles with Non-NIL trigger Wikia collections, 83% of the top 2 Non-NIL Wikia triggers are correct, with an additional 6% judged as interesting.

Entity Detection and Linking Accuracy. We further evaluate the effectiveness of ED and EL by using the triggered *Wikia* (the first prediction) or *Wikipedia* as linking targets, shown in table 8. By employing all the information from Wikia collections and Wikipedia, 2882 surface forms from news articles are detected, in which 500 are randomly selected as evaluation mentions. In table 8, *Wikia* employs the information from the triggered Wikia and Wikipedia in the step of entity detection, and linking the detected surface forms to the triggered Wikia. On the other hand, *Wikipedia* employs only the information from Wikipedia for entity detection, and linking the mentions to Wikipedia. Two independent annotators evaluate the correctness of the ED and EL steps. As shown in table 8, by including all the information from the triggered Wikia, the system is able to find additional 6.2% of mentions comparing with using only Wikipedia. More importantly, the system achieves a 15.3% relative improvement (63.1% error reduction) on linking accuracy by using the triggered Wikia as linking target rather than Wikipedia.

7 Conclusion

We investigated a multi-KB entity linking framework that employs one general knowledge KB (Wikipedia) and a large set of domain-specific KBs (Wikia collections) as linking targets. We developed a supervised model with a large and diverse set of features to detect when a domain-specific KB matches a document targeted for entity analysis. The system obtained high performance for both Wikia detection and entity linking to Wikipedia and Wikia. The performance of both entity detection and entity disambiguation improved by targeting in conjunction the triggered Wikia and Wikipedia as opposed to only the Wikipedia collection.

References

1. Bunescu, R. and Paşca, M.: Using encyclopedic knowledge for named entity disambiguation. In: NACL. pp. 9–16 (2006)
2. Cassidy, T., Ji, H., Ratinov, L.A., Zubiaga, A., Huang, H.: Analysis and enhancement of Wikification for microblogs with context expansion. In: COLING. pp. 441–456 (2012)
3. Cucerzan, S.: Large-scale named entity disambiguation based on Wikipedia data. In: EMNLP-CoNLL. pp. 708–716 (2007)
4. Cucerzan, S.: Named Entities Made Obvious: the participation in the ERD 2014 evaluation. In: ERD@SIGIR. pp. 95–100, http://dblp.uni-trier.de/db/conf/sigir/erd2014 (2014)
5. Demartini, G., Difallah, D.E., Cudré-Mauroux, P.: Zencrowd: Leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking. In: WWW. pp. 469–478 (2012)
6. Hoffart, J., Yosef, M.A., Bordino, I., Fürstenau, H., Pinkal, M., Spaniol, M., Taneva, B., Thater, S., Weikum, G.: Robust disambiguation of named entities in text. In: EMNLP. pp. 782–792 (2011)
7. Ji, H., Grishman, R., Dang, H.T., Griffitt, K., Ellis, J.: Overview of the TAC 2010 knowledge base population track. In: TAC (2010)
8. McNamee, P., Dang, H.T.: Overview of the TAC 2009 knowledge base population track. In: TAC. vol. 17, pp. 111–113 (2009)
9. Mcnamee, P., Dredze, M., Gerber, A., Garera, N., Finin, T., Mayfield, J., Piatko, C., Rao, D., Yarowsky, D., Dreyer, M.: HLTCOE approaches to knowledge base population at TAC 2009. In: TAC (2009)
10. Mihalcea, R., Csomai, A.: Wikify!: Linking documents to encyclopedic knowledge. In: CIKM. pp. 233–242 (2007)
11. Niemann, E., Gurevych, I.: The people's web meets linguistic knowledge: automatic sense alignment of wikipedia and wordnet. In: IWCS. pp. 205–214 (2011)
12. Pereira, B.: Entity linking with multiple knowledge bases: An ontology modularization approach. In: The Semantic Web–ISWC 2014, pp. 513–520 (2014)
13. Ruiz-Casado, M., Alfonseca, E., Castells, P.: Automatic assignment of wikipedia encyclopedic entries to wordnet synsets. In: Advances in Web Intelligence, pp. 380–386 (2005)
14. Shen, W., Wang, J., Luo, P., Wang, M.: Linking named entities in Tweets with knowledge base via user interest modeling. In: SIGKDD. pp. 68–76 (2013)
15. Sil, A., Cronin, E., Nie, P., Yang, Y., Popescu, A.M., Yates, A.: Linking named entities to any database. In: EMNLP-CoNLL. pp. 116–127 (2012)
16. Zheng, Z., Si, X., Li, F., Chang, E.Y., Zhu, X.: Entity disambiguation with freebase. In: IEEE/WIC/ACM. pp. 82–89 (2012)