UMass CIIR at TAC KBP 2013 Entity Linking: Query Expansion using Urban Dictionary

Jeffrey Dalton
University of Massachusetts, Amherst
jdalton@cs.umass.edu

Laura Dietz
University of Massachusetts, Amherst
dietz@cs.umass.edu

Abstract

This paper describes the system submitted to the TAC 2013 entity linking task of the Knowledge Base Population track. The core of the approach is probabilistic information retrieval over a search index of the knowledge base, including the text of Wikipedia. The retrieval results are further reranked using a supervised learning-to-rank model. The submission this year builds on the neighborhood and context extraction methods for query expansion introduced in 2012. In 2013, two new models are added. The first is query expansion using Urban Dictionary. For mentions in forum documents, we search Urban Dictionary for the mention string and perform query expansion. The second method is a multi-pass linking model. Instead of linking only the query mention, all entity mentions in the document are linked with features that encourage coherence among the linked entities. The results show that the method incorporating Urban Dictionary expansion run performed the best.

1 Introduction

A typical TAC KBP entity linking system has five steps: 1) query expansion, 2) candidate generation, 3) candidate ranking, 4) NIL detection, and 5) NIL clustering. The goal of the first two steps is to achieve a high-recall set of KB entities. The third step performs ranking and lastly these are filtered. The remaining 'out-of-kb' mentions are further clustered. The first three stages are typical of an information retrieval system.

The runs submitted by UMass CIIR to the 2013 evaluation are based on the open source Knowledge Base (KB) Bridge[1] system (Dalton and Dietz, 2013a). The KB Bridge system evolved from two systems developed for the UMass submissions to the TAC KBP entity linking 2012 (Dietz and Dalton, 2012) and the TREC Knowledge Base Acceleration (KBA) Cumulative Citation Recommendation 2012 task (Dalton and Dietz, 2013b). KB Bridge includes models for both starting from a text document and linking entity mentions to an existing KB, and starting from an entity in a KB and retrieving relevant documents. At the core of both there are models for extracting context from an entity or mention and transforming the context into a search query.

There are several notable changes to the system from TAC 2012 beyond a myriad of small improvements and fixes. First, in TAC 2013 a significant fraction of the mentions in the evaluation set are from forum data. Unlike previous newswire and web data, the mentions in this data appear to contain a high fraction of creative slang and pop culture terms. Examples of such slang query mentions include: [McSame], [MCCane], [Biebs], [Obamessiah], [Nobama], [Turd Blossom], and [uz-becky-becky-becky-stan-stan]. The existing sources of aliases from anchor text and structured metadata are unlikely to contain these informal references. To address the vocabulary mismatch, we use Urban Dictionary. Urban Dictionary is a crowd-sourced online web dictionary with more than seven million definitions, focused on slang and pop culture phrases not found in standard dictionaries. For example, [McSame] has the definition: "John McCain. He considers himself a straight talking maverick, when in reality he is merely running on the promise of four more years of George W. Bush." We leverage the entries as a source for query expansion to include in the retrieval context.

The second major change in the 2013 system is the use of fully disambiguated document representations. In this model, entity extraction and disambiguation is performed on all entities in the document in a first pass. Then, a second step that leverages the features from disambiguated mentions re-ranks the possible links with a model that includes entity-to-entity compatibility features. This is

[1] http://ciir.cs.umass.edu/~jdalton/kbbridge

a joint assignment model similar to the techniques employed by other leading systems (Cucerzan, 2011; Monahan et al., 2011; Ratinov et al., 2011). Although this model proved promising on training data, we found that the supervised model did not generalize well to the 2013 data distribution. We hypothesize that it did not perform as well as expected because of limited training data.

2 KB Bridge System Description

The submitted system is an evolution of the KB Bridge system described in (Dalton and Dietz, 2013a). It employs the Galago[2] (Cartright et al., 2012) search engine for probabilistic retrieval over both the KB and TAC Corpora. In the first stage of analysis, mention detection is performed on the document. Queries for the mentions are generated and run over a search index of the Knowledge Base. An optional further reranking step employs a supervised learning-to-rank model using RankLib[3]. NIL detection is performed using a score threshold tuned on training data. NIL clustering is performed using a sieve approach: first linking to a more recent dump of Wikipedia, then by string matching.

2.1 Probabilistic Retrieval

To efficiently identify relevant Wikipedia and TAC source documents, we build upon the Markov Random Field model for Information Retrieval (Metzler and Croft, 2005). The query model scores the documents in the corpus using a log-linear weighted combination of language model probabilities of multi-word concepts. The query model allows for arbitrary composition of unigram and dependence models.

We include three types a of concepts with corresponding weights λ_A in the query: the mention text t, a set of name variants $\vec{v}$, and a set of neighboring NER spans $\vec{e}$. For each document d in the collection, the score $f(d)$ is given by the proportionality in Equation 1, with type-based weights λ_T, λ_V, and λ_E, concept-based weights $\vec{\phi}$, and ψ which is a real-valued log-score of the concept under the document's language model.

$$f(d) \propto \exp\left\{ \sum_{a \in \{t,v,e\}} \lambda_A \frac{1}{|\vec{a}|} \sum_i \phi_i^A \psi(d, a_i) \right\} \quad (1)$$

Concept-based weights $\vec{\phi}$ which are assumed to be uniform if omitted, and are re-normalized to form a multinomial distribution.

In this work, we use sequential dependence language models (Metzler and Croft, 2005) for ψ, which incorporate word, phrase, and proximity from adjacent concept words. The model from Equation 1 can be expressed using the Galago query language as specified in Figure 1.

```
#combine:0=λ_T:1=λ_V:2=λ_E(
  #sdm(t)
  #combine(#sdm(v_0) ... #sdm(v_V))
  #combine:0 = φ_0^E : ... : k = φ_k^E(
    #sdm(e_0), ..., #sdm(e_k)
  )
)
```

Figure 1: Query for retrieving relevant stream documents in Galago query syntax.

2.2 Knowledge Base Representation

Our system addresses text-driven knowledge bases in which each entity is associated with free text, with relationships between entities from hyperlinks or other sources. In order to efficiently support the queries above, we create an extended index of Wikipedia with Galago. The index is based on a Freebase Wikipedia Extraction (WEX) dump of English Wikipedia from January 2012 which provides the Wikipedia page in machine-readable XML format and relational data in tabular format. The Freebase dump contains 5,841,791 entries. We filter out non-article entries, such as category pages. The resulting index contains 3,811,076 articles and over 60 billion words.

The index contains fields for anchor text (within Wikipedia as well as from the web), Wikipedia categories, Freebase names, Freebase types, redirects, article titles, and full-text for each article. Most of this information is contained in the WEX dump. We incorporate external web anchor text using the Google Cross-Wiki dictionary (Spitkovsky and Chang, 2012), which contains 3 billion links and 297 million associations from 175 million unique anchor text strings. The search index allows us to both efficiently retrieve articles as well as compute features (e.g. link probability).

2.3 Document Analysis

The first step in linking is to identify the entity query span q in the document and to find disambiguating contextual information for the query model. This includes name variations v, and other neighboring mentions m. In the single pass model we link only the target entity mention. In the two pass model, every mention in all query documents are linked to the KB.

In the TAC KBP challenge, the entities of type person, organization, or location are the main focus of the link-

[2] http://www.lemurproject.org/galago.php
[3] http://cs.umass.edu/~vdang/ranklib.html

Feature Set	Type	Description
Character Similarity	q, v	Lower-cased normalized string similarity: Exact match, prefix match, Dice, Jaccard, Levenstein, Jaro-Winkler
Token Similarity	q,v	Lower-cased normalized token similarity: Exact match, Dice, Jaccard
Acronym match	q	Tests if query is an acronym, if first letters match, and if KB entry name is a possible acroynm expansion
Field matches	q, v	Field counts and query likelihood probabilities for title, anchor text, wiki redirects, and freebase names
Link Probability	q, v	p (anchor | KB entry) - the fraction of internal and external total anchor strings targeting the entity
Inlink count	document prior	Log of the number of internal and external links to the target KB entry
Text Similarity	document	Normalized text similarity of document and KB entity: Cosine with TF-IDF, KL, JS, Jaccard token overlap
Neighborhood text similarity	document	Normalized neighborhood similarity: KL Divergence, Number of matches, match probability
Neighborhood link similarity	document	Neighborhood similarity with in/out links: KL divergence, Jensen-Shannon Divergence, Dice overlap, Jaccard
Rank features	retrieval	Raw retrieval log likelihood, Normalized posterior probability, 1/retrieval_rank
Context Rank Features	retrieval	Retrieval scores for each contextual components

Table 1: Features of the mention-to-entity similarity.

Feature Set	Description
Category IDs	Intersection, Misses, Dice, Jaccard, Cosine
Category Words	Jaccard, Jensen-Shannon Divergence, Cosine with TF-IDF, Unweighted cosine
Article Text	Jaccard, Jensen-Shannon Divergence, Cosine similarity
Text Mentions	Contains entity name, Both articles contain name
Inlinks	Pointwise mutual information, ProxPMI (wikifier), Intersection, Jaccard, Dice, Google Norm. Distance
Outlinks	Pointwise mutual information, ProxPMI (wikifier), Intersection, Jaccard, Dice, Google Norm. Distance
Inlinks + Outlinks	Pointwise mutual information, ProxPMI (wikifier), Intersection, Jaccard, Dice, Google Norm. Distance
Shared Links	Linked, mutal link

Table 2: Features of the entity-to-entity similarity.

ing effort and so the system detects entities using standard named entity recognition tools, namely UMass's factorie[4] NER toolkit. These provide the mention spans to derive query mentions q, name variations v, and neighboring entities m.

Given a target entity mention, q, the system needs to identify name variations, v, in the document, such as "Steve" to "Steve Jobs" or "IOC" to "International Olympic Committee". The goal of this step is to identify alternative names that are less ambiguous than the query mention. We use the set of all mentions in the documents as candidates and use string matching heuristics (prefix and suffix token matches and acronym matching) to extract name variations v. For weighting neighboring entities m, we use the 'local' document model described by Gottipati and Jiang (Gottipati and Jiang, 2011). This models weights the neighboring entities by their maximum likelihood probability in the document.

2.4 KB Entity Ranking

The query model from document analysis is executed against the search index of KB entries and the top 250 entity results are retrieved. The initial ranking may be reranked using supervised machine learning in a learning-to-rank (LTR) model. The refinement employs more extensive feature comparisons which would be expensive to compute over the entire collection or are not directly supported in the Galago query language. For these experiments we use a Multiple Additive Regression Tree (MART) model that is state-of-the-art and captures non-linear dependencies in the data. The model includes dozens of features. A description of the features used in the ranking model is found in Table 1.

[4]http://factorie.cs.umass.edu/

2.5 NIL Handling

After the entities are ranked, the last step is to determine if the top-ranked entity for a mention is correct and should be linked to the KB entry or instead refers to an entity not in the knowledge base, in which case NIL should be returned. For these experiments, we use a simple NIL handling strategy. We return NIL, if the supervised score of the top ranked entity is below a threshold τ. The NIL threshold τ is tuned on the training data. For the special case of the TAC KBP challenge, the reference knowledge base is a subset of Wikipedia. We exploit this fact by returning NIL when the linked entity is in Wikipedia, but is not contained in the reference knowledge base. The remaining NIL mentions not linkedable to Wikipedia are clustered by normalized string equality.

3 Novel enhancements for 2013

3.1 Urban Dictionary Expansion

To address the vocabulary mismatch between the language used in forum data and the knowledge base, we leverage Urban Dictionary. We take the mention string an issue it as a query against the urban dictionary web service to retrieve definitions. We retrieve the full-text of the definition, as well as the tags. For simplicity, we focused on the article tags, which often include the name of the entity being described. These tags were added to the neighboring entities m and consequently as part of the retrieval query. The main goal of this step is to improve the recall of our retrieved entities from the knowledge base.

3.2 Entity-to-Entity Compatibility Model

A recent trend in entity linking has been joint or 'collective' assignment of mentions in a document. The HLTCOE introduced the Context Aware Linker of Entities (CALE) using local context entities (Stoyanov et al., 2012). Language Computer Corporation (LCC) uses features from a subset of the closest unambiguous mentions (Monahan et al., 2011). The Microsoft system for TAC builds a context vector from the union of candidates for all entities (Cucerzan, 2011). UIUIC's GLOW system uses 'global' similarity features from a first pass linking model (Ratinov et al., 2011).

We implemented an extension to our supervised ranking model that incorporates features similar to Wikifier. We first perform a first pass ranking, taking all mentions that would be predicted as non-NIL as context links. For documents with large numbers of entities, we limited the context to the 50 links with the highest compatibility score. We use the features described by GLOW as well as those from MSR. A full list of the features are given in the Table 2.

3.3 Name Variant NIL Handling

In previous years, the entities that were not matched to Wikipedia were matched based on matching mention strings. This year, we added a small evolution to this approach that uses the set of name variants $\vec{v}$ discovered for the entity in the document. We observe that these aliases provide an improved canonical name that is helpful for clustering people and acronyms.

4 Experiments

4.1 Parameters

For our retrieval model, we need to tune the Dirichlet smoothing parameters μ and the parameters for the sequential dependence model weights. These are trained on 50 queries of the 2010 data set, yielding $\mu = 96400$, $\phi^{\mathrm{t}} = 0.29$, $\phi^{\mathrm{o}} = 0.21$, $\phi^{\mathrm{u}} = 0.5$. For each of the compared methods, we train separate λ parameters on the training data using a coordinate ascent learning algorithm. For the QVM_local model the estimated parameters are: $\lambda^{\mathrm{Q}} = 0.31$, $\lambda^{\mathrm{V}} = 0.38$, and $\lambda^{\mathrm{M}} = 0.31$. For training the mention-to-entity RankLib model we use the TAC data from 2009-2012, omitting the 2010 training data. We perform a random 80-20 training-validation split. Unfortunately, due to time constraints we were only able to train the entity-to-entity model on the 2012 data. We believe that the limited size of the training set (and bias in the 2012 data distribution) resulted in a model that did not generalize well to the 2013 data, which differs significantly. One change from last year is that we use a larger retrieval pool for re-ranking from 250, up from 100 because of increased ambiguity for the forum queries.

The optimal NIL score threshold across all years on the training data is 0.5. The optimal threshold for 2012 is 3.0, except for the e2e model which is 0.5. The runs submitted use the threshold tuned for 2012, which proved to be sub-optimal because the 2013 data differs significantly from 2012. The global NIL threshold would provide better effectiveness.

4.2 Submitted Runs

We submitted five runs to the TAC KBP English monolingual Entity Linking Task testing the unsupervised and supervised models.

- UMass_CIIR1 - This is a retrieval only run. This run performs query expansion using the local neighborhood weighting. The query is expanded to include name variations and neighboring entities from the document.

- UMass_CIIR2 - This uses retrieval run from UMass_CIIR1 and performs additional mention-to-entity re-ranking using a supervised model. The results are re-ranked using a MART learning-to-rank method using only mention-to-entity features.

- UMass_CIIR3 - This model is similar to UMass_CIIR2, but does not perform query expansion using entities from the neighborhood. Query expansion is performed using only name variations. The results are re-ranked using the same MART learning-to-rank method using only mention-to-entity features. NILs are clustered with additional clustering based on extracted canonical name variations.

- UMass_CIIR4 - This run is the same as UMass_CIIR2 with the addition of query expansion using Urban Dictionary. The query is expanded to include name variations and neighboring entities from the document. Additional query

Approach short description	Run ID	Accuracy	B^3+ Precision	B^3+ Recall	B^3+F1
qvm_local-retrieval	UMass_CIIR1	0.577	0.573	0.317	0.408
qvm_local-m2e	UMass_CIIR2	0.729	0.716	0.462	0.561
qv-m2e-nvNil	UMass_CIIR3	0.802	0.781	0.571	0.660
qvm_local-urbdict-m2e	UMass_CIIR4	0.806	0.785	0.584	0.670
qvm_local-e2e	UMass_CIIR5	0.746	0.730	0.503	0.595
Median		0.746	0.718	0.496	0.574
Best		0.833	0.826	0.689	0.746

Table 3: Overall effectiveness on the Entity Linking task.

Approach short description	Run ID	News	Web	Forum
qvm_local-retrieval	UMass_CIIR1	0.493	0.528	0.202
qvm_local-m2e	UMass_CIIR2	0.637	0.609	0.414
qv-m2e-nvNil	UMass_CIIR3	0.743	0.615	0.547
qvm_local-urbdict-m2e	UMass_CIIR4	0.745	0.620	0.572
qvm_local-e2e	UMass_CIIR5	0.667	0.638	0.457
Median		0.645	0.525	0.488
Best		0.829	0.678	0.662

Table 4: B^3+ F1 by document type.

	UMass_CIIR1	UMass_CIIR2	UMass_CIIR3	UMass_CIIR4	UMass_CIIR5	Median	Best
PER	0.576	0.671	0.694	0.709	0.722	0.627	0.778
ORG	0.590	0.638	0.626	0.639	0.662	0.542	0.737
GPE	0.091	0.399	0.657	0.660	0.424	0.552	0.746

Table 5: B^3+ F1 by entity class.

	All		NIL		In-KB	
	Accuracy	B^3+ F1	Accuracy	B^3+ F1	Accuracy	B^3+ F1
qvm_local-urbdict-m2e-globalNil	0.804	0.691	0.860	0.687	0.756	0.692
qvm_local-urbdict-m2e	0.806	0.670	0.905	0.678	0.722	0.654
Median	0.746	0.574	0.880	0.566	0.626	0.554
Best	0.833	0.746	1.000	0.777	0.788	0.722

Table 6: NIL vs. Non-NIL effectiveness for best runs.

expansion is performed by searching for the query mention on Urban Dictionary. Tags from the top returned urban dictionary entries are addded as neighbors in the query.

- UMass_CIIR5 - This run disambiguates all entities mentions in a two-pass supervised model. It uses UMass_CIIR3 to link all mentions in the document in a first pass. In the second linking pass, the disambiguated neighboring entities are used as features for a model that includes both mention-to-entity and entity-to-entity similarity.

4.3 Result analysis

The overall results of our runs are shown in Table 3. The results by document type are in Table 4 and by entity class in Table 5. Unlike the results in 2012, the unsupervised retrieval model, UMass_CIIR1, performed significantly below the median, especially on the forum data with a B^3+ F1 value of only 0.202. It also struggled with GPE entities with a B^3+ F1 of only 0.091. However, it performs above the median on ORGs. It is clear that more effective context models are needed for both GPEs and forum data.

The second model, UMass_CIIR2 applies supervised re-ranking to the entity results from UMass_CIIR1. The results improve dramatically over UMass_CIIR1. It is only slightly below the median overall. The B^3+ F1 score on forum data nearly doubles to 0.414. The GPE effectiveness improves over 300% to 0.399, but is still below the median of 0.552. It is above the median for the other entity classes.

Our second best performing run overall is UMass_CIIR3, which is more conservative in its retrieval strategy and does not use the entity neighborhood context expansion. It performs competitively, significantly above the median in B^3+ F1. Compared with UMass_CIIR2 this run improves the effectiveness on newswire and forum documents. The per-entity class results show improvements across PER and GPE entity types, but a decrease in effectiveness on ORGs.

The best submitted run is UMass_CIIR4. It performs well, significantly above the median overall. This run is similar to UMass_CIIR2, but also uses Urban Dictionary expansion. This improves the effectiveness on forum data. Examining the entity class results shows that it improves PER and greatly improves GPE effectiveness over UMass_CIIR2.

We analyze UMass_CIIR4 run further in Table 6. The table shows a breakdown of the effectiveness for NIL and Non-NIL mentions. We want to study this because the focus of our system is mainly on In-KB entities, with simple NIL clustering heuristics. The tables includes a modification of our best run with a different NIL threshold, referred to as globalNil, which is a run (not-submitted) where the NIL threshold is tuned across all TAC years instead of only on the 2012 data. The globalNil threshold is lower, predicting more (correct) non-NIL entity links. This would have improved overall effectiveness as well as effectiveness on In-KB mentions, where this threshold leads correct predictions for an additional 3.8% of In-KB queries.

We now analyze UMass_CIIR5, which was our best performing run on validation data, but did not perform as well as expected. It builds upon the results of the UMass_CIIR3 run and adds a second pass model with additional global entity to entity features. However, the overall result in Table 3 shows that it performs worse than UMass_CIIR3. We hypothesize that one cause of this behavior is that the second pass entity-to-entity model was trained only on 2012 data. The model appears to be biased towards the distribution of entitiy types for that year. We examine this further in Table 5. It shows that the entity-to-entity model improves effectiveness for people and organizations, but has a dramatic decrease in effectiveness on GPEs. It does not perform well on the forum data, with a B^3+ F1 of 0.457 compared with 0.547 for UMass_CIIR3. We plan to investigate this further by re-training the model across all years as well as incorporating the Urban Dictionary expansion which we believe may improve recall for forum data.

5 Conclusion

We find that the baseline unsupervised retrieval model with the neighborhood query expansion performs poorly on forum data and GPEs. This finding motivates work on improving context modeling techniques for these classes of mentions. To improve forum data, we experiment using Urban Dictionary for query expansion, which effectively improved recall. We also introduce a new linking model for 2013 that performs full-document entity extraction and joint disambiguation, incorporating global entity compatibility features. Although the results on 2013 are lower than expected, we believe this is due to significant differences between the training and evaluation data distributions.

Overall, the KB Bridge linking system performs competitively, significantly above the median for several runs. In comparison with other systems, there is room for further improvement in improved handling of NIL entities.

Acknowledgements

This work was supported in part by the Center for Intelligent Information Retrieval, in part by IBM subcontract #4913003298 under DARPA prime contract #HR001-12-C-0015, and in part by NSF grant #IIS-0910884. Any opinions, findings and conclusions or recommendations

expressed in this material are those of the authors and do not necessarily reflect those of the sponsor.

References

Marc-Allen Cartright, Samuel Huston, and H Field. 2012. Galago: A modular distributed processing and retrieval system. In *SIGIR 2012 Workshop on Open Source Information Retrieval*, pages 25–31.

S. Cucerzan. 2011. Tac entity linking by performing full-document entity extraction and disambiguation. *Proceedings of the Text Analysis Conference*.

Jeffrey Dalton and Laura Dietz. 2013a. A Neighborhood Relevance Model for Entity Linking. In *Proceedings of the 10th International Conference in the RIAO series (OAIR), 2013*.

Jeffrey Dalton and Laura Dietz. 2013b. Bi-directional linkability from wikipedia to documents and back again: Umass at trec 2012 knowledge base acceleration track. *TREC'12*.

Laura Dietz and Jeffrey Dalton. 2012. Across-document neighborhood expansion: Umass at tac kbp 2012 entity linking. *Proceedings of the Text Analysis Conference (TAC KBP)*.

Swapna Gottipati and Jing Jiang. 2011. Linking entities to a knowledge base with query expansion. In *EMNLP*.

Donald Metzler and W. Bruce Croft. 2005. A markov random field model for term dependencies. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '05, pages 472–479.

S. Monahan, J. Lehmann, T. Nyberg, J. Plymale, and A. Jung. 2011. Cross-lingual cross-document coreference with entity linking. *Proceedings of the Text Analysis Conference*.

L. Ratinov, D. Roth, D. Downey, and M. Anderson. 2011. Local and global algorithms for disambiguation to wikipedia. In *ACL*.

V.I. Spitkovsky and A.X. Chang. 2012. A cross-lingual dictionary for english wikipedia concepts. In *Eighth International Conference on Language Resources and Evaluation (LREC 2012) Open access*.

Veselin Stoyanov, James Mayfield, Tan Xu, Douglas W Oard, Dawn Lawrie, Tim Oates, and Tim Finin. 2012. A context-aware approach to entity linking. In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction*, pages 62–67. Association for Computational Linguistics.