Stanford-UBC Entity Linking at TAC-KBP, Again

Angel X. Chang,[†] Valentin I. Spitkovsky,[†]
Eneko Agirre,[‡] Christopher D. Manning[†]
[†]Computer Science Department, Stanford University, Stanford, CA, USA
[‡]IXA NLP Group, University of the Basque Country, Donostia, Basque Country

Abstract

This paper describes the joint Stanford-UBC knowledge base population system for the entity linking tasks. We participated in both the English and the cross-lingual tasks, using a dictionary from strings to possible Wikipedia titles, taken from our 2009 submission. This dictionary is based on frequencies of Wikipedia back-links, and it provides a strong context-independent baseline. For the English track, we improved on the results given by the dictionary by disambiguating entities using a distantly supervised classifier, trained on context extracted from Wikipedia. Since we did not use any text from the Wikipedia pages associated with the knowledge base nodes for the dictionary, we submitted that run to the *no wiki text* track, and the one using the distantly supervised classifier to the *wiki text* track. Our work focused on disambiguating among articles, allowing for very simple NIL strategies: the system returned NIL whenever selected Wikipedia articles were not present in the KB; moreover, NILs were then clustered only according to the target string. These simple approaches were sufficient for our runs to score above the median entry in each of their respective tracks for the English task; for the cross-lingual task, there was only one track, and our submissions (using the English-specific, context-independent dictionaries) fell below the median.

1 Introduction

The 2011 Text Analysis Conference (TAC) Knowledge Base Population (KBP) track includes two tasks: entity linking and slot filling. We participated in the first task with a system similar to the one we prepared for the 2009 and 2010 TAC-KBP evaluations [ACJ+09, CSY+10], consisting of several techniques inspired by the word sense disambiguation (WSD) literature [AE06].

In the TAC-KBP entity linking task, the input is a set of queries, each consisting of a target string and a document in which the target string can be found. Using a knowledge base (KB, also given) derived from Wikipedia, the task is to establish connections between queries and the KB entities to which they refer. If the query does not match an entity in the KB, then the system should return NIL, with an ID, by clustering non-matching entities and assigning a unique ID to each cluster. In 2009 and 2010 [MD09, JGD+10], evaluation was based on simple micro-averaging accuracy (without NIL-clustering), which becomes problematic because of different percentages of NILs across datasets. In 2011, evaluation was changed to take into account the clustering of NILs by using a modified B-Cubed metric (B-Cubed+) that requires non-NIL responses to be correctly linked to the appropriate entity ID in the KB.

Since the KB is based on (a subset of the English) Wikipedia, our strategy was to first map a query to a Wikipedia title. We performed this initial mapping using a context-independent dictionary from strings to ordered lists of possible Wikipedia titles; for each Wikipedia title, the dictionary provides a score, indicative of the conditional probability of the Wikipedia article given the string. This dictionary is built based primarily on the anchor-texts of web-links: both internal inter-Wikipedia links and external links from the web into the English Wikipedia (the dictionary and its variants are described in §2). After determining the list of potential Wikipedia titles, we use a disambiguation component to rerank this list, based on the context from the associated document (we used a distantly supervised approach, described in §3).

From the list of potential Wikipedia titles and their scores, we take the top Wikipedia title as our answer. If the Wikipedia title corresponds to a KB entity, we return that entity's ID; and if the Wikipedia title does not correspond to a KB entity, we use the Wikipedia title as the NIL ID. If no Wikipedia titles match, we explore two baseline strategies for assigning NIL IDs:

1. Strategy 1 (N1): use a NIL ID that is unique to the target string;
2. Strategy 2 (N2): always assign a new NIL ID, unique to the query.

2 The Dictionary

Dictionaries are rudimentary static mappings from strings to relevant Wikipedia articles. We constructed two such context-free components using several modules from our previous submissions: the remapper [CSY+10, §2.1]; the core EXCT, LNRM and FUZZ dictionaries [CSY+10, §2.2]; and the GOOG dictionary [CSY+10, §2.4]. The first, a simple *cascade of dictionaries* combines different components into a single dictionary, to be used down-stream, as in the previous years; the second, a *heuristic combination*, incorporates more, noisier data, and is novel this year.

2.1 A Cascade of Dictionaries

Traditionally, we have combined several dictionaries in a simple cascade (`exct`→`lnrm`) to get overall, context-independent scores. For each target mention, we provided a ranked list of potential matches by using the following rules:

- if the EXCT dictionary has a match, use scores from the EXCT dictionary;
- else if the LRNM dictionary has a match, use scores from the LNRM dictionary;
- otherwise, do not suggest anything.

We did not include the FUZZ dictionary in the cascade since such fall-throughs resulted in lower micro-averaged accuracies, in our earlier experiments.

2.2 A Heuristic Combination

This year, we made a new effort to filter out some of the noisier mappings proposed by the original dictionary flavors. The heuristic combination is specifically designed to retain some of the suggestions provided by the FUZZ dictionary, to increase the recall over EXCT and LNRM, while discarding obviously bad choices. We used the following filtering strategy:

1. discard disambiguation pages;
2. discard date pages;
3. discard list pages;
4. discard pages only suggested by FUZZ, unless one of the following is true:
   - it is an acronym;
   - the string is similar to the title;
   - the string is a substring of the title name.
5. discard low in-degree pages, unless one of the following is true:
   - the page is a potential disambiguation for the string;
   - the string is the title of the page.

The resulting heuristic combination dictionary (`dictheur`) is characterized by high precision but a lower recall. For example, for the string *ABC*, the EXCT dictionary offers 191 options, LNRM has 62, and FUZZ has 3,274. Using the filters described above, we were able to trim potential Wikipedia titles down to a final list of just 110 options. The reduced set of candidates made it easier for disambiguation components to rerank the remaining options based on document context. Of course, in some cases, our heuristics discard correct mappings while keeping inferior choices. For instance, a feasible mapping *Angela Dorothea Kasner* → *Angela_Merkel* is discarded, but the undesirable mapping *Angela Merkel* → *German_federal_election,_2005* is kept.

3 Distantly Supervised Disambiguation

In addition to plain lookups, we applied machine learning techniques to do supervised disambiguation of entities. For each target string, we trained a multi-class classifier to distinguish the more relevant of the possible articles. Using the text surrounding the target string in the document, this classifier scored each of the entities provided by the combined dictionary, `dictheur`.

Inter-Wikipedia links and the surrounding snippets of text comprised our training data. When one article linked to another, we assumed that the anchor text was synonymous with the linked Wikipedia entity (as collapsed by the remapper, mentioned above). Since the dictionary could return entities having no incoming Wikipedia links for the target string, we restricted the list of possible entities to those with actual links for the target string. Our snippets were spans of text, consisting of 100 tokens to the left and 100 to the right of a link, extracted from the October 8th, 2008 English Wikipedia dump.

Using binary features extracted from these spans [CSY$^+$10, §4], we trained maximum entropy log-linear models, producing one classifier for each target string that had more than one possible entity. For unambiguous target strings, we did not train a classifier and instead returned the original top title suggested by the simple cascade of dictionaries, `exct`$\rightarrow$`lnrm`.

4 Results

We participated in both the English and the cross-lingual entity-linking tasks. The English task was split into two tracks: the primary track (*wiki text*), in which text from Wikipedia pages associated with KB nodes could be used; and an optional track (*no wiki text*), in which using such text was not allowed. We submitted two runs: a context-independent approach for *no wiki text*, relying mainly on the dictionary we developed in 2009; and a context-sensitive approach for *wiki text*, using distantly supervised classification. For the cross-lingual task, we submitted only baselines, again using our dictionaries from 2009.

We evaluated our entity linking system using the scorer provided for the 2011 TAC KBP entity linking task, reporting both the micro-average accuracy (average over all queries) and the B-Cubed+ F_1 that takes into account NIL-clustering.

4.1 English Evaluation

	test2009 (News)			train2010 (Web)			test2010 (News + Web)		
	KB MicroAve	$B^3 F_1$ (N1)	$B^3 F_1$ (N2)	KB MicroAve	$B^3 F_1$ (N1)	$B^3 F_1$ (N2)	KB MicroAve	$B^3 F_1$ (N1)	$B^3 F_1$ (N2)
`nil`	0.571	0.408	0.166	0.284	0.269	0.192	0.547	0.511	0.302
`exct`$\rightarrow$`lnrm`	0.749	0.623	0.616	0.873	0.858	0.846	0.814	0.781	0.742
`supervised`	0.812	0.669	0.661	0.886	0.867	0.855	0.844	0.805	0.764

Table 1: Stanford-UBC results for English entity linking on datasets from 2009 and 2010.

Using data from prior TAC-KBP entity-linking competitions, we evaluated (i) a NIL baseline; (ii) a context-independent strategy using the dictionary; and (iii) a context-sensitive approach based on distant supervision:

1. NIL *baseline* (`nil`) — a baseline that always returns an empty set of Wikipedia titles;
2. *cascade of dictionaries* (`exct`$\rightarrow$`lnrm`) — an overall context-independent score for each entity, which defaults to the LNRM (lower-cased normalized) dictionary in case of an EXCT (exact match) dictionary miss;
3. *distant supervision* (`supervised`) — a context-dependent score, obtained by running the distantly supervised classifier.

For each approach, we evaluated both proposed NIL strategies. All results, using the three approaches, are shown in Table 1.[1] Note that NIL strategy N1 is consistently better than N2.

Since our dictionary did not use text from Wikipedia pages associated with the KB nodes, we submitted `exct`$\rightarrow$`lnrm` to the *no wiki text* track and `supervised` to the primary track. Unfortunately, due to an operator error, we submitted with NIL strategy N2 instead of N1, which cost us about 1%. Of the systems that did not access the web during the evaluation period, ours still scored above the median in their respective tracks. The context-independent entry (`exct`$\rightarrow$`lnrm`) was close in B-Cubed+ F_1 to the highest scoring system in the *no wiki text* track, and the distantly supervised context-sensitive approach (`supervised`) led to an additional significant improvement over the static dictionary (see Table 2).

[1] Since the 2010 reference data did not disambiguate NILs, we constructed one set of plausible groupings ourselves, manually.

		test2011 (News + Web)				
		KB MicroAve	P	R	$B^3 F_1$	
	nil (N2)	0.500	0.500	0.359	0.418	
	nil (N1)	0.500	0.455	0.497	0.475	
no wiki text	median				0.521	
	exct→lnrm (N2)	0.722	0.680	0.687	0.684	Stanford_UBC-1
	exct→lnrm (N1)	0.722	0.680	0.718	0.696	
	highest				0.714	
wiki text	median				0.716	
	supervised (N2)	0.790	0.757	0.743	0.750	Stanford_UBC-1
	supervised (N1)	0.790	0.753	0.774	0.763	
	highest				0.846	

Table 2: Stanford-UBC results for the 2011 English entity-linking task.

4.2 Cross-Lingual Evaluation

For the cross-lingual entity linking task, we submitted three runs that began to investigate the suitability of the dictionaries that we developed back in 2009 for non-English queries. Below, we compare the results of our exploratory submissions, along with a NIL baseline:

1. NIL *baseline* (`nil`) — a baseline that always returns an empty set of Wikipedia titles, as in the previous section;
2. *cascade of dictionaries* (`exct→lnrm`) — an overall context-independent score for each entity, also as before;
3. *Google baseline 1* (`goog, 2009 remapper`) — a baseline using the GOOG dictionary, which queries the Google search engine with the string and maps English Wikipedia results to KB titles, using the original remapper, from 2009;
4. *Google baseline 2* (`goog, 2011 remapper`) — another GOOG baseline, using an updated version of the remapper.

	train2011				test2011				
	KB MicroAve	P	R	$B^3 F_1$	KB MicroAve	P	R	$B^3 F_1$	
nil (N1)	0.537	0.427	0.516	0.467	0.501	0.398	0.497	0.442	
goog, 2011 remapper (N1)	0.677	0.585	0.658	0.620	0.650	0.564	0.636	0.598	Stanford_UBC-3
exct→lnrm (N1)	0.686	0.585	0.661	0.621	0.665	0.569	0.647	0.605	Stanford_UBC-1
goog, 2009 remapper (N1)	0.698	0.602	0.678	0.638	0.660	0.572	0.647	0.607	Stanford_UBC-2
median								0.675	
highest								0.788	

Table 3: Stanford-UBC results for the 2011 cross-lingual entity linking task.

All of our results using the old dictionaries scored below the median (see Table 3). Nevertheless, they provide interesting baselines for an approach that is both context-independent and English-specific. We can conclude that dictionaries that focus exclusively on strings associated with the English Wikipedia articles are not well-suited to cross-lingual information retrieval.[2]

5 Conclusions

We described the joint Stanford-UBC knowledge base population systems for the entity linking tasks. In 2009, we developed several (mostly context-independent) approaches for monolingual entity linking. These were based on frequencies of back-links, training on contexts of anchors, overlaps of contexts with an entity's Wikipedia text, and their various heuristic and supervised combinations. In 2010, we focused on a simple context-sensitive approach for entity disambiguation, which performed surprisingly well on newswire data from the 2009 evaluation corpus, but did less well on the 2010 web data.

This year, for English entity linking, we submitted a simplified system using a baseline dictionary and a distantly supervised disambiguation component. Our approach focuses on the disambiguation among Wikipedia articles, and thus the treatment of NILs is trivial: we return NIL only when an entity not in the KB is selected, and NILs are clustered according to the target string. The resulting above-median scores suggest that we constructed a strong algorithm for disambiguating entities, which could be further improved using a less naïve treatment of NILs, in particular using more sophisticated NIL-clustering techniques.

For cross-lingual entity linking, our monolingual components did less well, but a related approach presented by the Stanford team fared better [SC11]. Their cross-lingual dictionaries and associated components are in the process of being released [SC12].

[2]However, several conceptually-simple extensions to how our monolingual dictionaries were constructed can restore much of the lost performance [SC11].

Acknowledgments

This material is based upon work supported by the Air Force Research Laboratory (AFRL) under prime contract no. FA8750-09-C-0181. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the view of the Air Force Research Laboratory.

Angel X. Chang is supported by a Stanford Graduate Fellowship; Eneko Agirre partially supported by KNOW2 (TIN2009-14715-C04-01).

We thank the task organizers for their effort.

References

[ACJ+09] Eneko Agirre, Angel X. Chang, Daniel S. Jurafsky, Christopher D. Manning, Valentin I. Spitkovsky, and Eric Yeh. Stanford-UBC at TAC-KBP. In *TAC*, 2009.

[AE06] Eneko Agirre and Philip Edmonds, editors. *Word Sense Disambiguation: Algorithms and Applications*. Springer, 2006.

[CSY+10] Angel X. Chang, Valentin I. Spitkovsky, Eric Yeh, Eneko Agirre, and Christopher D. Manning. Stanford-UBC entity linking at TAC-KBP. In *TAC*, 2010.

[JGD+10] Heng Ji, Ralph Grishman, Hoa T. Dang, Kira Griffitt, and Joe Ellis. Overview of the TAC 2010 Knowledge Base Population track. In *TAC*, 2010.

[MD09] Paul McNamee and Hoa Dang. Overview of the TAC 2009 Knowledge Base Population track. In *TAC*, 2009.

[SC11] Valentin I. Spitkovsky and Angel X. Chang. Strong baselines for cross-lingual entity linking. In *TAC*, 2011.

[SC12] Valentin I. Spitkovsky and Angel X. Chang. A cross-lingual dictionary for English Wikipedia concepts. In *LREC*, 2012.