Strong Baselines for Cross-Lingual Entity Linking

Valentin I. Spitkovsky, Angel X. Chang

Computer Science Department, Stanford University, Stanford, CA, 94305
Google Research, Google Inc., Mountain View, CA, 94043
{valentin, angelx}@{cs.stanford.edu, google.com}

Abstract

We describe several context-independent baselines for tackling the cross-lingual entity linking task. Our methods are quite basic, reducing to efficient look-ups in static, pre-computed tables. Despite their simplicity, however, such approaches scored well in a recent knowledge-base population competition. Moreover, these language-independent techniques still perform strongly on English entity linking tasks.

Introduction

The entity linking task — as defined in Knowledge-Base Population (KBP) tracks at the Text Analysis Conference (TAC) — is a challenge to associate string mentions in documents with articles in a knowledge base (KB). In the two earliest TAC-KBPs, the KB was a subset of the English Wikipedia, and the documents were also in English (McNamee and Dang, 2009; Ji et al., 2010). In 2011, the conference's organizers created a new, cross-lingual track, in which mentions and documents could be in English or in Chinese, although the KB still remained English-centric.

Somewhat surprisingly, context-independent techniques developed by Stanford-UBC — which ignore the documents and focus on just the mention of each query — have managed to score above the median entries in all previous English entity linking evaluations (Agirre et al., 2009; Chang et al., 2010; Chang et al., 2011). At the core of that approach were several static, English-specific dictionaries for mapping short strings of natural language text to canonical article titles from the English Wikipedia. Since the dictionaries were English-specific, unmodified look-up methods led to below-median performance on the cross-lingual entity linking task (Chang et al., 2011). We will show how such dictionaries can be improved — using conceptually-simple modifications — to again score above the median.

New and Improved Components

We will now describe several key dictionary components and our improvements over the original Stanford-UBC dictionary from 2009 (which was reused in 2010 and 2011).

Remapper

The remapper attempts to group various English Wikipedia titles that, in fact, refer to the same article, by mapping them to a canonical URL (Agirre et al., 2009, §2.1). This year, we improved the original remapper in several ways, the most important of which was disallowing merging two clusters if both of them contain an entry from the KB. Other improvements had to do with better handling of KB entries whose Wikipedia pages are now redirects and preferential treatment of URLs that start with upper-case characters, among non-KB pages. This component remained English-specific.

Cross-Mapper

The cross-mapper is a new, multi-lingual component which groups together all Wikipedia articles corresponding to the same English counterpart, by mapping them to the canonical English Wikipedia URL. It is also English-centric — as is the KB — since it ignores clusters of parallel Wikipedia articles that aren't available in English.

GOOG Dictionary

The GOOG "dictionary" disambiguates a string by querying the Google search engine in English (`hl=en`), with the `site:en.wikipedia.org` directive, scoring any returned URLs beginning with `http://en.wikipedia.org/wiki/` using the inverses of their ranks (Agirre et al., 2009, §2.4).

Our first language-independent baseline is a simple modification of GOOG, which drops `hl=en`, relaxes the restriction to just `site:wikipedia.org` and keeps not only the English but now also any non-English Wikipedia pages covered by the cross-mapper; scores for canonical English pages hit multiple times are simply added.

We used our new multi-lingual dictionaries in the same way as for our English-specific entity linking submission, by focusing on highest-scoring entries, with a simple NIL-clustering strategy (Chang et al., 2011, §1, §4.1).[1] Although it is easy to implement, the GOOG dictionary offers a fairly weak baseline, scoring some two-and-a-half points below the median entry in this year's competition (see Table 1).

KB	MicroAve	B^3F_1	2011 System
GOOG	69.7	65.0	`Stanford2-3`
English EXCT→LNRM	71.4	66.0	`Stanford2-1`
`median`		67.5	
EXCT→LNRM	**74.5**	**69.5**	`Stanford2-2`
`highest`		78.8	

Table 1: Stanford2 results for cross-lingual entity linking.

[1] This time, however, we correctly assigned a unique NIL identifier to each distinct string mention (strategy N1), instead of accidentally making each NIL unique to the query (strategy N2).

English EXCT→LNRM Dictionary

Next, we made several improvements to our core English dictionary, which is based primarily on the anchor-texts of web-links: both internal inter-Wikipedia links and external links from the web into the English Wikipedia (Agirre et al., 2009, §2.2). First, we created a new view of external, non-Wikipedia links into the English-Wikipedia, according to the August 2nd, 2011 Google web crawl. And second, we introduced a number of additional relevant boolean features, to augment the raw counts (in the end, we did not use these features in our submission).[2] Thus, our new, still-monolingual dictionary contained more and fresher string-to-article mappings than the original version from 2009.

We used the refreshed English dictionary in our standard cascading way, going by exact matches (EXCT) whenever possible and falling through to more forgiving matching (LNRM) if needed (Chang et al., 2011, §2.1). This new monolingual dictionary scored one point higher than GOOG, but still one-and-a-half points lower than the median entry, in the cross-lingual evaluation (see Table 1).

Cross-Lingual EXCT→LNRM Dictionary

Finally, we created the cross-lingual dictionary by incorporating a new kind of information: anchor-texts from non-Wikipedia web-pages into non-English-Wikipedia pages covered by our cross-mapper. This gave us a stream of indirect web-counts, as if their anchor-texts had come from direct links to the corresponding canonical English Wikipedia pages. To counter-balance this additional weight of web-links, we also created a new view of inter-English-Wikipedia links, according to the same crawl, complementing the information from the 2008/9 Wikipedia dumps that closely resemble the KB but may have now become stale.

Our new, multi-lingual dictionary performed significantly better than its monolingual counterpart, scoring two points higher than the median entry in the 2011 cross-lingual entity linking competition (see Table 1). We believe that it offers a surprisingly strong baseline, considering that it uses neither context nor any knowledge specific to Chinese.

Further Monolingual Evaluation

To get a better sense of the multi-lingual dictionary's quality, we tested it on all three English evaluation sets, using both exact lookups and our usual cascade of dictionaries. The new dictionary scored well above the median and not far below the highest entry on the 2009 evaluation set (see Table 2*a*); higher than the highest entry that did not access Wikipedia pages associated with KB nodes in inference in 2010 (see Table 2*b*); and again higher than the median but lower than the highest entry among *no-wiki-text* submissions in 2011 (see Table 2*c*).[3] Exact lookups were slightly — but consistently — worse than our cascade strategy.

[2]We intend to explain all features in a sister paper (Spitkovsky and Chang, 2012) that is to accompany the public release of our cross-lingual dictionaries and newest associated components.

[3]Note that our approach would qualify as *no-wiki-text*, since it does not make use of the text of the Wikipedia article in question (though it may use anchor text from other Wikipedia pages).

		KB MicroAve	B^3F_1
a) 2009	`median`	71.1	
	EXCT	79.4	64.9
	EXCT→LNRM	**79.5**	**65.2**
	`highest`	82.2	
b) 2010	*no-wiki-text*-`median`		63.5
	`median`		68.4
	no-wiki-text-`highest`		77.9
	EXCT	82.3	78.8
	EXCT→LNRM	**82.9**	**79.5**
	`highest`		86.8
c) 2011	*no-wiki-text*-`median`		52.1
	EXCT	70.0	67.4
	EXCT→LNRM	**71.2**	**68.6**
	no-wiki-text-`highest`		71.4
	`median`		71.6
	`highest`		84.6

Table 2: Results for all three English entity linking tasks.

Conclusions

We described Stanford's knowledge-base population system for the cross-lingual entity linking task. Our multilingual dictionary uses neither context nor language-specific knowledge, yet performs better than the median scorers on all available TAC-KBP entity linking evaluation sets. Despite its simplicity, the static dictionary presents a surprisingly strong baseline to the research community, as well as possibly a useful platform for developing more sophisticated context-sensitive, machine learning approaches. We are currently in the process of publicly releasing this resource and related data (Spitkovsky and Chang, 2012).

Acknowledgments

This work was carried out in the summer of 2011, while both authors were employed full time at Google Inc., over the course of the second author's internship. We would like to thank our advisors at Stanford University, Dan Jurafsky and Chris Manning, for their continued help and support. We are also grateful to the other members of the original Stanford-UBC TAC-KBP entity linking team: Eneko Agirre and Eric Yeh; our initial (monolingual) dictionary for mapping strings to Wikipedia articles was conceived and constructed during that collaboration, in the summer of 2009.

We thank the task organizers for their effort.

References

E. Agirre, A. X. Chang, D. S. Jurafsky, C. D. Manning, V. I. Spitkovsky, and E. Yeh. 2009. Stanford-UBC at TAC-KBP. In *TAC*.

A. X. Chang, V. I. Spitkovsky, E. Yeh, E. Agirre, and C. D. Manning. 2010. Stanford-UBC entity linking at TAC-KBP. In *TAC*.

A. X. Chang, V. I. Spitkovsky, E. Agirre, and C. D. Manning. 2011. Stanford-UBC entity linking at TAC-KBP, again. In *TAC*.

H. Ji, R. Grishman, H. T. Dang, K. Griffitt, and J. Ellis. 2010. Overview of the TAC 2010 Knowledge Base Population track. In *TAC*.

P. McNamee and H. Dang. 2009. Overview of the TAC 2009 Knowledge Base Population track. In *TAC*.

V. I. Spitkovsky and A. X. Chang. 2012. A cross-lingual dictionary for English Wikipedia concepts. In *LREC*.