Naïve but effective NIL clustering baselines – CMCRC at TAC 2011

Will Radford [†‡] **Ben Hachey** [‡◇] **Matthew Honnibal** [◇] **Joel Nothman** [†‡] **James R. Curran** [†‡]

[†]School of Information Technologies
University of Sydney
NSW 2006, Australia

[‡]Capital Markets CRC
55 Harrington Street
NSW 2000, Australia

[◇]Centre for Language Technology
Macquarie University
NSW 2109, Australia

{wradford,james,joel}@it.usyd.edu.au honnibal@gmail.com bhachey@cmcrc.com

Abstract

This paper describes the CMCRC systems entered in the TAC 2011 entity linking challenge. We used our best-performing system from TAC 2010 to link queries, then clustered NIL links. We focused on naïve baselines that group by attributes of the top entity candidate. All three systems performed strongly at 75.4% B^3 F1, above the 71.6% median score.

1 Introduction

Named Entity Linking (NEL) is the task of grounding entity mentions to the appropriate knowledge base (KB) node or NIL if the link target is not found in the KB. NEL has been included in the TAC KBP track in 2009 and 2010 and has yielded important datasets and a variety of approaches.

The 2011 TAC KBP task specification saw two important changes from 2010. Firstly, participants were required to cluster NIL-linked queries that reference the same entity, despite it not being in the KB. Secondly, the evaluation was expanded to include a B^3 (Bagga and Baldwin, 1998) clustering metric adjusted to account for NIL-links.

We participated in the English NEL task with wiki text and decided to concentrate on NIL clustering, rather than improve linking. We first use our best TAC 2010 system (Radford et al., 2010) to link all queries. Any NIL-assigned queries are clustered, creating distinct NIL IDs. These are combined with the KB-assigned queries for output.

We focused on naïve clustering techniques that group NIL queries by attributes of their top entity candidates. All three systems performed at 75.4% B^3 F1, above the 71.6% median, but below the 84.6% best score of the 21 teams.

2 Review: CMCRC at TAC 2010

We presented three systems at TAC KBP 2010 and used the best of them as a preliminary linker this year. Our 2010 system CMCRC 1 used a whole-document approach to each query. The context document was tokenised and tagged with NEs using the C&C tools (Curran et al., 2007).

We retrieve a list of candidates from the KB for each NE mention. We first identify coreference chain heads within the context document. These are longer and hopefully more specific and canonical forms of the NE. We expand acronyms: "ABC" → "Australian Broadcasting Corporation". We also match substrings of other NEs: "Mr Obama" → "Barack Obama", taking a gazetteer of person titles into account. The head of the coreference chain is then searched in a Solr[1] index containing the following fields of a Wikipedia dump:[2] article titles and redirects, redirects, titles and redirects for disambiguation pages, and bold terms in disambiguation pages.

Each candidate list is ranked using three main context sources: contexts and categories from the whole document (Cucerzan, 2007), a filter that removes unreliable aliases and a re-ranker that uses the in-link graph for candidates across all NEs in the document. These are explained in more detail in Sections 5 and 6 of our notebook paper last year (Radford et al., 2010).

[1] http://lucene.apache.org/solr/

[2] Generated on 2010/07/30

We compare the query string with the NE mentions and the top-ranked candidate is checked against the TAC KB. If it exists in the TAC KB we return the entity ID. A NIL link can thus be generated in two ways: choosing a top-ranked candidate that does not exist in the TAC KB, returning *no* candidate matches for the query NE. This is similar to the approach taken in the top-ranked system in TAC KBP 2009 (Varma et al., 2009). In experiments after the official competition, this approach scores 84.4% over the TAC 2010 evaluation data. This is equivalent to second place in the overall rankings after the LCC system (Lehmann et al., 2010), which scored 86.8% and well above the next system at 81.7%. Our approach is unsupervised and competitive with the best heuristic-based system from TAC 2010 - another LCC system at 85.8%.

3 NIL clustering

Figure 1 describes how, once queries have been linked, NIL queries are filtered and clustered separately, to be recombined for output. The following paragraphs set out the three techniques we used to cluster NIL queries.

CMCRC 1: Term The most naïve clustering technique is to group queries by term. For example, two queries "ABC" would be given the same NIL ID.

CMCRC 2: Coref In-document coreference improved our linking performance in TAC 2010 experiments and it forms the basis for our second approach. We group by the coreference head, with expansion of US state name acronyms. Error analysis of the TAC 2009 and 2010 queries suggested that these were common errors, so we built a gazetteer from the Wikipedia page.[3]

CMCRC 3: KB Our final approach takes advantage of the fact that disambiguation is done with respect to the full Wikipedia dump (see Section 2). Our preliminary phase linked 1,263 of the 2,250 queries to NIL. Of these NIL queries, 56% (711) were linked to a page in the larger KB and 44% (552) had no candidate. We group the former by the page title, backing off to query term (CMCRC 1) for the latter.

[3] http://en.wikipedia.org/wiki/List_of_U.S._state_abbreviations

```
# Input:
# * a list of queries
# * a linker
# * a clusterer
entity_qs = {}
nil_qs = {}
# Link the queries.
for q in queries:
  entity_id = linker.link(q)
  if entity_id == "NIL":
    nil_qs[q] = entity_id
  else:
    entity_qs[q] = entity_id
# Cluster the NILs.
clusterer.cluster(nil_qs)
# Output all queries.
for q in queries:
  if q in nil_qs:
    q.entity_id = nil_qs[q]
  else:
    q.entity_id = entity_qs[q]
return queries
```

Figure 1: Algorithm sketch for integrating NEL with NIL clustering

System	Micro	B^3 P	B^3 R	B^3 F1
CMCRC 1 - Term	77.9	74.2	**76.4**	75.3
CMCRC 2 - Coref	77.9	75.0	75.5	75.3
CMCRC 3 - KB	77.9	**75.2**	75.6	**75.4**

Table 1: Results over the TAC 2011 evaluation data

4 Results

Table 1 shows how the systems perform over the TAC 2011 evaluation data, using the TAC 2010 micro-averaged accuracy and TAC 2011 B^3 metrics. All three systems perform above the median performance of 71.6% across the 44 runs from 21 teams, but below the best score of 84.6%.

CMCRC 1 groups queries by term, which improves recall at the cost of precision – especially problematic considering that ambiguous queries are a feature of TAC datasets. CMCRC 3 uses the extra information from the larger KB for a gain in precision, but relies on a TAC-style setup where the target KB is a subset of the larger KB.

We also investigated agglomerative and k-means clustering, using features from the context documents, such as tokens, the top candidate from *other* mentions in the document. These did not improve upon the naïve baselines.

The two-stage approach is not ideal since it does not provide any avenues for correcting linking errors. For example, a KB query may be assigned NIL, but this model does not allow the query to be merged back amongst queries the system assigned a KB node (we only cluster NIL queries). We believe that the linking process should be more closely integrated with clustering for better performance.

5 Conclusion

We have presented several naïve yet effective baseline systems for NEL and NIL-clustering. All three score at 75.4% B^3 F1. This is above the median performance of 71.6% across TAC 2011 submissions.

References

Amit Bagga and Breck Baldwin. 1998. Entity-based cross-document coreferencing using the vector space model. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, pages 79–85, Montreal, Quebec, Canada.

Silviu Cucerzan. 2007. Large-scale named entity disambiguation based on Wikipedia data. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 708–716, Prague, Czech Republic.

James Curran, Stephen Clark, and Johan Bos. 2007. Linguistically motivated large-scale NLP with C&C and Boxer. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion*, pages 33–36, Prague, Czech Republic.

John Lehmann, Sean Monahan, Luke Nezda, Arnold Jung, and Ying Shi. 2010. LCC approaches to knowledge base population at TAC 2010. In *Proceedings of the Text Analysis Conference*, Gaithersburg, MD, USA.

Will Radford, Ben Hachey, Joel Nothman, Matthew Honnibal, and James R. Curran. 2010. Document-level entity linking: CMCRC at TAC 2010. In *Proceedings of the Text Analysis Conference*, Gaithersburg, MD USA.

Vasudeva Varma, Praveen Bysani, Kranthi Reddy, Vijay Bharat, Santosh GSK, Karuna Kumar, Sudheer Kovelamudi, Kiran Kumar N, and Nitin Maganti. 2009. IIIT Hyderabad at TAC 2009. In *Proceedings of the Text Analysis Conference*, Gaithersburg, MD, USA.