SYDNEY_CMCRC at TAC 2013

Glen Pink [†] **Will Radford** [†‡] **Will Cannings** [†‡] **Andrew Naoum** [†]
Joel Nothman [†‡] **Daniel Tse** [†‡] **James R. Curran** [†‡]
[†]ə-lab, School of Information Technologies
University of Sydney
NSW 2006, Australia
[‡]Capital Markets CRC
55 Harrington Street
NSW 2000, Australia

{gpin7031,wradford,anaoum,joel,dtse6695,james}@it.usyd.edu.au
me@willcannings.com

Abstract

We use a supervised whole-document approach to English Entity Linking with simple clustering approaches. The system extends our TAC 2012 system (Radford et al., 2012), introducing new features for modelling local entity description and type-specific matching as well type-specific supervised models and supervised NIL classification. Our rule-based clustering takes advantage of local description and topics to split NIL clusters. The best system uses supervised entity linking and local description type clustering and scores 70.5% B^3+ F1 score. Our KB clustering score is competitive with the top system at 72.1%.

1 Introduction

Named Entity Linking (NEL) grounds a Named Entity (NE) mention to its corresponding knowledge base (KB) entry, or NIL if absent. The TAC KBP NEL task includes the further task of clustering NIL mentions that refer to the same entity.

Departing from our previous English NEL submissions (Radford et al., 2010; Radford et al., 2011; Radford et al., 2012), we adopt a supervised approach todisambiguation for NEL. We incorporate local entity description features for precise disambiguation.

A learnt disambiguator was able to improve our system performance in development. Additionally, we experiment with varying the model for the two primary components of the task: we benefit from conditioning the disambiguator on the query's NE type, learning one model for people and one for other NEs; and experiment with learning a model to explicitly decide between the top candidate or NIL was unsuccessful.

We continue to experiment with more sophisticated NIL clustering, exploiting local description types and a topic model.

2 Data Preprocessing and Resources

We continue (see Radford et al., 2012) to link against the Wikipedia dump from April 2012[1]. Entity aliases are extracted from article titles, redirects and titles of disambiguation pages that link to the article. These are normalized and indexed in Apache Solr[2]. The article wiki markup is processed to allow for lookup of the text, inlinks, outlinks and categories for a given title. We calculate statistics over the graph of links between entities including: *entity prior* – the number of links to an article normalized by total number of articles – and *reference probabilities*, the conditional probability of linking to an entity given a particular alias (i.e. $p(\text{entity}|\text{alias})$). We predict and store the NE type of the entity based on several features of its article (Nothman et al., 2013).

Seeking high recall of candidates for disambiguation, we derive additional aliases as follows.

2.1 Crosswikis Aliases

We use the Crosswikis dataset (Spitkovsky and Chang, 2012) to provide a wider set of entity aliases drawn from pages outside the Wikipedia article graph. These anchor texts of incoming links, should obtain higher recall with some noise. We apply a

[1] http://dumps.wikimedia.org/
[2] http://lucene.apache.org/solr/

similar normalization process as the Wikipedia anchors; targeted redirect URLs are resolved to their canonical article title. We exclude aliases that reduce to an empty string after normalization, appear less than three times in the dataset, or had a reference probability less than 5%.

2.2 Generated Aliases

From the set of redirects for each entity in Wikipedia, we extract common transformation rules from an entity name to its corresponding redirect as detailed in Radford et al. (2012). Rules include: the removal of name titles, prefixes, suffixes and middle initials; the abbreviation and removal of organisation suffixes; and the abbreviation and removal of state and country names. We then apply the curated list of rules to all applicable entities in our Wikipedia dataset. The resulting generated aliases are added to the Solr index for their entity.

3 Named Entity Linking

While the TAC task focuses on a particular name per document, we disambiguate all entity mentions in the query documents. Thus we may leverage relationships between candidate entities as disambiguating context to the query. The following subsections detail our approach.

3.1 Candidate Generation

Given a document, we tokenize and extract NEs using the C&C Tools (Curran and Clark, 2003) with a four-class (PER, ORG, LOC, MISC) model trained on approximately 1600 Australian newswire stories from 2009. We then resolve the queried name to one of the extracted NEs, creating a dummy NE if no full or partial match could be found.

3.1.1 In-document Coreference

We identify *chains* of NEs using simple coreference rules. NEs are sorted by length with longest first and each is processed in turn to find the best coreference match. The matching algorithm normalizes the NE for case and removes titles such as Mrs. Exact matches to previous NEs are preferred (i.e. Ms Gillard or Gillard matches Gillard), then non-uppercase unigram suffix matches (i.e. Gillard matches Julia Gillard), then non-uppercase unigram prefix matches (i.e. Julia matches Julia Gillard), then acronym matches where the initial upper-case characters (excepting stopwords) of the NE (i.e. DoJ matches Department of Justice). Since we are coreferring NEs, these rules do not handle nominal or pronominal coreference. We noticed worse NER performance at the beginning of sentences where capitalized words were misidentified as NEs so we add aliases missing their initial token for any sentence-initial NEs.

3.1.2 Query Expansion

We apply rules to extract more in-document evidence when searching for candidate entities. We maintain a list of backoff queries to apply if there are no hits for the first query. We exclude any single word NE mentions that are substrings of the longest NE, since we assume they are less specific. If the NE had been resolved to the query name, the name is added to a backoff list since there is not always perfect correspondence between names and NEs.

Several other expansion rules are applied to futher expand queries for state (location) aliases, organizational suffixes and bureaucratic organizations, see Radford et al. (2012).

3.1.3 Search

The expanded query is used to search the index using the Wikipedia, Crosswikis and generated aliases. The top 100 results are boosted by their *entity prior* (inlink count) with title and redirect matches weighted (weight = 100) more than disambiguation redirects, crosswiki and generated aliases (weight = 10).

3.2 Supervised Linking

Once candidate entities have been retrieved for each coreference chain, we use a sequence of processing components to extract features for each candidate.

We drew the design of our supervised features from two TAC 11 systems – Anastácio et al. (2011), including LDA features, and Zhao et al. (2011), Wikipedia link structure features, which are representative of features typically used in supervised entity linking.

We learn a regression model to assign a score to a candidate that indicates whether it should be linked to the query chain. During training, we take entity

candidates for the query chain, extract features and learn weights from them.

We reimplemented our supervised system from 2012 using the `scikit-learn` (Pedregosa et al., 2011) linear regression model and L2-regularization. We find linking to be a delicate machine learning problem. One issue is the instability of instance generation as a different search strategy can change the instances for learning and classification. The new implementation trains on the top three candidates ranked by our previous unsupervised model to ensure that the model is presented feasible examples.

3.2.1 Existing Features

The features used in our 2012 system, and carried over to this year are as follows.

Wikipedia Link Structure

- *Reference probability, Entity prior.* *Reference probability* is calculated for the entity and the longest NE in the chain. The *entity prior* is also used.

Entity Title-Chain Similarity

- *Alias cosine similarity.* The maximum character bigram cosine similarity between the mentions in the coreference chain and all of the candidate aliases for an entity.
- *Entity title dice similarity, Entity title cosine similarity.* As above with a variant that uses cosine distance.
- *Entity title begins/ends with query, query begins/ends with entity title.* Whether the entity article title begins/ends with a substring of the query name, or vice versa.
- *Entity title is substring of query, query is substring of entity title.* Whether the entity article title subsumes the query name, or vice versa.
- *Entity title edit distance/Jaro-Winkler distance.* Levenshtein distance or Jaro-Winkler distance computed between the entity article title and query name.
- *Article-query document cosine similarity.* Cosine similarity between the term vectors of the entity article and the query document.
- *Acronym match.* Whether the query is an acronym of the entity title.

Entity Type

- *Entity type matches.* Whether NER on the query string yields the same NE type as the candidate Wikipedia page's predicted NE type.
- *Mention preceded by locative P and is location.* Whether the query string is of type LOC, and is preceded by a locative preposition.

Topic Modelling

We trained an LDA model using the Vowpal Wabbit online machine learning toolkit[3], with training parameters $k = 100$ (the number of topics), $\alpha = 1$, $\rho = 0.1$, on documents from TAC 09 queries and the Wikipedia articles from April 2012.

- *Topic similarity.* The Hellinger distance between the predicted topic distribution of the query document and entity article, both using stemmed tokens.

3.2.2 New Features

For 2013 we added the following features.

Entity Chain Features

- *chain type combination.* Where there are multiple entity types in the chain for an entity, these are added as features with the candidate.
- PER *name match.* True if any PER mentions in the chain match article titles not including disambiguation parenthesis or commas.

Local Description

Matching mention context to an entry's `wikitext` field, or candidate Wikipedia article, is a fundamental part of many linking systems. This is often modelled as a cosine similarity between the candidate text and query context, either the whole document, token window around the mention (Bunescu and Paşca, 2006) or sentences in a coreference chain. While this presents a strong baseline, we are interested in extracting precise entity information for disambiguation. We use part-of-speech and named entity tag patterns to

[3] `http://hunch.net/~vw`

NE	Description	Example
LOC	Location	LOC, California
	Type	the city of LOC
ORG	Location	ORG in London
	Sponsor	Rupert Murdoch's ORG
	Type	ORG television
PER	Age	PER, 50
	Location	PER of Germany
	Type	shooting guard PER

Table 1: Types of local description for each NE mention.

extract typed local description for entities. Table 1 shows the entity types for the mention and type of description we aim to extract. For example, we might create a rule `per-type-left-np`, and capture noun phrases to the left of a PER mention that might contain type information, such as Former Prime Minister John Howard. These rules are naïve, as the POS and NE tagging may be wrong, or the rules may simply have captured non-description text. Our goal is to restrict the mention context rather than extract information, so extra noise should not be as damaging as for slot filling.

We compare each description to different field of the Wikipedia article: first sentence, first paragraph, first section, section title, title, categories, infobox values, tokens. We create a feature for each description, for each field, indicating the proportion of description unigrams or bigrams that are found in the field text. If Prime Minister is found in the first sentence of the article for a candidate, John Howard, we create the features indicating a match, unigram and bigram match: `per-type-left-np` (value=1), `per-type-left-np-ug` (value=0.6) and `per-type-left-np-bg` (value=0.5). Where a description has been extracted and a field does not match, we would add a feature `per-type-left-np-no-match`, as this may indicate negative matching evidence (i.e. Local dentist John Howard). These features are used in the supervised model and are also used in clustering as described below, to split clusters where queries have contradictory attributes. This technique is similar to Mann and Yarowsky (2003), who extract and use biographic facts for unsupervised person disambiguation.

3.3 Training Data

In our experiments, we use query data from past TAC years (2009, 2010, 2010 evaluation, 2011) as training data, and 2012 as a development (DEV) set. We add 2012 as training data for some runs for the final evaluation.

3.4 Separate PER Classifier

For some runs, we use two separate classifiers based on the query entity type. We trained a separate PER (trained on PER queries) and non-PER models. Entities having the longest mention in their chain labelled as PER by NER use the PER model, else they use the non-PER model.

3.5 NIL Classifier

We also experiment with an additional NIL classifier for some runs, which considers the top candidate returned by the supervised linking process and determines if that entity should instead be linked as NIL. Due to time constraints, we did not have time to appropriately experiment with this classifier, but have included the results.

4 NIL Clustering

We follow our previous rule-based clustering approach (Radford et al., 2011), implementing three methods: *basic*; *local*; and *topic*. All three methods first cluster the NIL entities on attributes in the of the mention. Any NIL mentions which share a TAC Entity ID are clustered together; followed by any mentions which have the same wiki title (since we link to a larger Wikipedia snapshot, these may be NILs w.r.t. the TAC KB); canonical term; and finally cleaned term. *Local* and *topic* split NIL clusters by a query attribute – a local description type and the most prominent LDA topic (from a model built over the query documents) respectively. The model is used in a effort to characterize the broad domain of a mention query. If all members have the attribute set and there are exactly two distinct values in the cluster, we split the cluster.

5 Results

As NIL clustering is performed after supervised linking, every run is a combination of a linking configuration and a NIL clustering configuration. Table 2

ID: Linking / Clustering	Acc	All	KB	NIL	NW	WB	DF	PER	ORG	GPE
Highest	81.0	72.1	72.4	72.0	80.1	67.3	63.3	75.8	73.7	73.1
1: supervised + local + PER / basic	**80.9**	**70.5**	**72.1**	**68.5**	**77.0**	**63.3**	⋆**63.3**	**73.4**	**67.6**	**70.3**
2: supervised + local + PER / local	**80.9**	70.3	**72.1**	68.2	76.8	**63.3**	⋆**63.3**	73.3	67.3	**70.3**
3: supervised + local + PER / topic	**80.9**	68.9	**72.1**	64.8	75.6	60.4	62.0	71.8	64.4	69.9
4: supervised + NIL / basic	72.4	54.5	44.5	64.1	60.1	52.0	46.3	55.3	54.8	53.4
5: supervised + NIL / topic	72.4	53.0	44.5	60.9	58.9	49.5	45.1	53.5	52.8	52.7
Median	74.9	58.3	54.3	60.2	63.7	54.5	46.2	61.7	59.3	52.9

Table 2: Accuracy and B^3+ F1 scores over TAC 13 data. All systems use base supervised features. '+ local' adds local description, '+ PER' a separate PER classifier, and '+ NIL' an additional NIL classifier trained on 2011 data. Top scores within our systems is marked using bold font and a ⋆ marks a competition high score.

ID	Acc	B^3+ F1	B^3+ KB	B^3+ NIL
Cucerzan (2012)	76.6	73.0	68.5	78.1
Radford et al. (2012)	72.2	66.5	65.6	67.5
1	77.5	73.4	67.2	80.4
2	77.5	73.6	67.2	80.7
3	77.5	73.9	67.2	81.3
4	58.4	52.8	25.5	83.3
5	58.4	53.5	25.5	84.8

Table 3: Linking accuracy and B^3+ F1 over the TAC 12 dataset. We include results for our and the best TAC 12 system.

lists the runs that we submitted to TAC 13, and the linking and clustering results with the highest and median results presented for comparison. Runs 1-3 use TAC 09, 10, 10-eval, 11 and 12 as training data. Runs 4-5 use TAC 09, 10, 10-eval as training data, and 11 as training for the NIL classifier.

Our best system uses the supervised linker with local description features and a separate PER classifier, with the basic configuration used for NIL clustering. The micro-averaged accuracy is 80.9%, 0.1% below the top accuracy and at 70.5% B^3+ F1 it performs well over the median score, and is 1.6% below the highest score. Reflecting the higher priority we place on linking over clustering, our KB score is even more competitive at 72.1%, 0.3% below the highest score.

Our system scored at the highest score for discussion forum queries (DF) showing the applicability of our approach of different domains. Future work would be to leverage features of the internal thread and post structure as opposed to treating a discussion forum document as one flat document.

We present a further analysis of the system for Run 1 in Table 4, which compares results for domain and type pairs. The high results for NW reflect our focus on this particular domain, and that more work needs to be done on the newer TAC domains. Further analysis of discussion forum structure, as above, is required, especially in regards to ORG entities, as the ORG-DF score is substantially lower than other scores. We expect discussion forum text to be less edited and more informal than newswire. In TAC 13, queries in this segment are primary sports teams, which can be difficult in more formal domains, and remain difficult when mentioned informally: Nawlins for New Orleans Saints, Piston's for Detroit Pistons. Despite these terms existing in Wikipedia redirects, linking mentions in limited context remains a significant challenge for NEL.

Table 3 lists the linking and clustering results on TAC 12 for the runs we submitted to TAC 13, as well as the highest TAC 12 system result, showing the impact of the new features, modelling and clustering in our system. Runs use TAC 09, 10, 10-eval and 11 as training data, with runs 4-5 using 11 for the NIL classifier only. We exceed our own highest score by 7.4% , and the highest system score by 0.9%. Topic splitting performed poorly in TAC 13 despite being

domain	type	count	B^3+ F1
PER	NW	348	81.5
ORG	NW	368	77.2
GPE	WB	6	74.1
GPE	NW	418	73.0
GPE	DF	379	67.2
PER	WB	130	66.4
PER	DF	208	64.0
ORG	WB	207	60.9
ORG	DF	126	48.0

Table 4: B^3+ F1 for Run 1 by document domain and entity type, as well as the counts for these across TAC 13 queries. Sorted by B^3+ F1.

the best performer over TAC 12, likely due to the LDA hyperparameters being tuned to TAC 12 and not being robust enough to appropriately model the TAC 13 evaluation document text.

6 Conclusion

Our systems in TAC 13 explore supervised whole-document approaches to NEL with naïve, local description type and topic clustering. Our best system uses supervised entity linking with a separate PER model and local description type clustering and scores 70.5% B^3+ F1 score. Our KB clustering score is competitive with the top system at 72.1%.

References

[Anastácio et al.2011] Ivo Anastácio, Bruno Martins, and Pável Calado. 2011. Supervised learning for linking named entities to knowledge base entries. In TAC (TAC, 2011).

[Bunescu and Paşca2006] Razvan Bunescu and Marius Paşca. 2006. Using encyclopedic knowledge for named entity disambiguation. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 9–16, Trento, Italy.

[Cucerzan2012] Silviu Cucerzan. 2012. MSR system for entity linking at TAC 2012. In TAC (TAC, 2012).

[Curran and Clark2003] James Curran and Stephen Clark. 2003. Language independent ner using a maximum entropy tagger. In Walter Daelemans and Miles Osborne, editors, *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 164–167.

[Nothman et al.2013] Joel Nothman, Nicky Ringland, Will Radford, Tara Murphy, and James R. Curran. 2013. Learning multilingual named entity recognition from Wikipedia. *Artificial Intelligence*, 194:151–175, January.

[Pedregosa et al.2011] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

[Radford et al.2010] 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 2010*, Gaithersburg, MD USA, November. National Institute of Standards and Technology.

[Radford et al.2011] Will Radford, Ben Hachey, Matthew Honnibal, Joel Nothman, and James R. Curran. 2011. Naive but effective NIL clustering baselines – CMCRC at TAC 2011. In TAC (TAC, 2011).

[Radford et al.2012] Will Radford, Will Cannings, Andrew Naoum, Joel Nothman, Glen Pink, Daniel Tse, and James R. Curran. 2012. (Almost) Total Recall – SYDNEY_CMCRC at TAC 2012. In TAC (TAC, 2012).

[Spitkovsky and Chang2012] Valentin I. Spitkovsky and Angel X. Chang. 2012. A cross-lingual dictionary for English Wikipedia concepts. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, Istanbul, Turkey, May.

[TAC2011] 2011. *Proceedings of the Text Analysis Conference 2011*, Gaithersburg, MD USA, November. National Institute of Standards and Technology.

[TAC2012] 2012. *Proceedings of the Text Analysis Conference 2012*, Gaithersburg, MD USA, November. National Institute of Standards and Technology.

[Zhao et al.2011] Yu Zhao, Weipeng He, Zhiyuan Liu, and Maosong Sun. 2011. THUNLP at TAC KBP 2011 in Entity Linking. In TAC (TAC, 2011).