EXPLOITING WIKIPEDIA FOR NAMED ENTITY LINKING

Will Radford – Supervisors: James Curran and Ben Hachey

ə-lab, School of Information Technologies
wradford@it.usyd.edu.au

1. Introduction

Named Entity Recognition (NER) is the problem of identifying names of people, places, organisations – "entities" – in text. Named Entity Linking (NEL) addresses practical issues of leveraging the extracted information, in particular:

- How do you identify *which* entity in a database a particular name refers to?
- If you have two identical strings that don't link to a database entity (NIL), do they link to the same entity?

2. TAC Entity Linking Task

The NEL task has been explored as a shared task within a Text Analysis Conference (TAC) track and frames the problem as follows.

Participants are given:

- Knowledge Base Entities – Wikipedia pages that have Infoboxes
- Source documents – News and web text
- Queries – a search term (NE mention) and document containing the term

Systems must take each query and decide whether it refers to an entity in the Knowledge Base or not, returning that entity's *id* or NIL .

3. Related Areas

- NE Disambiguation
- Cross-document coreference
- Record Linkage
- Web Person Search

4. Wikipedia

Wikipedia is a large, free and dynamic online encyclopedia that has many useful features for NEL:

- Redirect pages – alias pages are valuable for searching alternative aliases for entities
- Disambiguation pages – provides more context for ambiguous entities
- Link structure – the intra-wiki link graph structure is helpful for disambiguation

5. Approach

We believe that NEL information about *all* NE mentions in a document can be used to effectively disambiguate the query term [Cucerzan, 2007]. Figure 1 shows the process and main components of our system.

- Extract NEs from the query document: `East Timor`, `Australian`, `Dili`, `ABC` (the query term), `Ministry of Education`
- Search Wikipedia for possible candidates for each NE: the acronym ABC could refer to a number of different entities
- Disambiguate the candidates: select the entity for each mention that is most consistent given the *other* entities.

In our example, some mentions are reasonably unambiguous: `East Timor`, `Australian` and `Dili` are easily linked to the country, nationality and city entities. The mention `ABC` is more ambiguous, but the selection of other entities makes `Australian Broadcasting Corporation` the most consistent match.

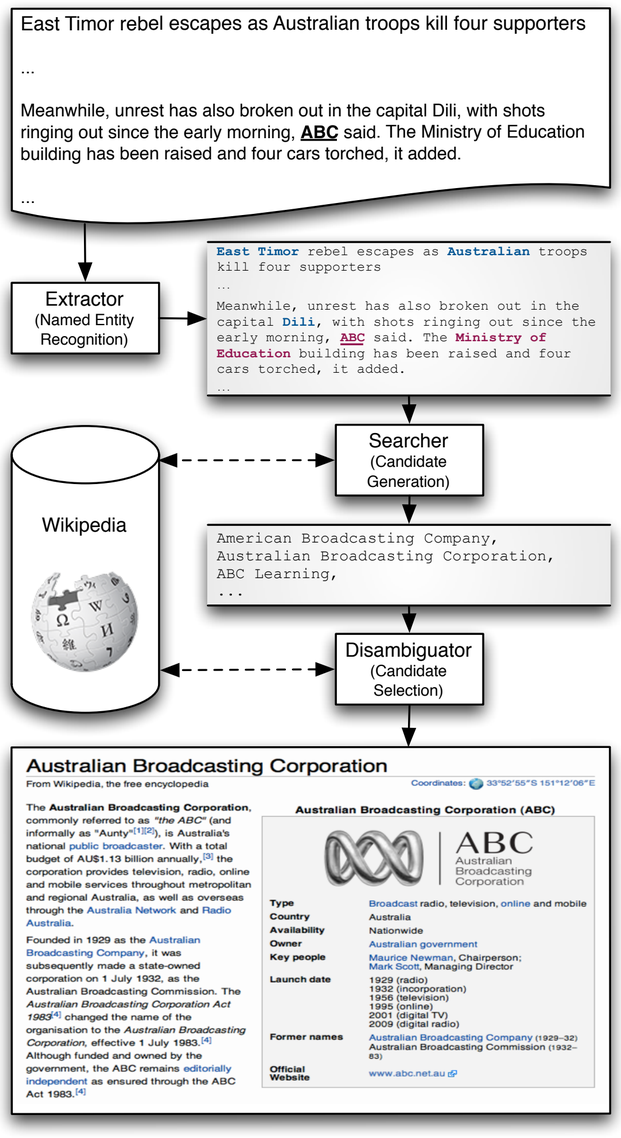

Figure 1: NEL *Workflow*

Our team submitted three systems:

A used a disambiguator that maximised agreement in textual context and category content

B as per A but with heuristics for high-precision Wikipedia search

C used a Cosine similarity system that targeted textual matches between query document and entity page text

6. Results

Systems are evaluated using accuracy scores, the percentage of correct queries, calculated over all, NIL and entity queries.

At this point, only limited results of the TAC 2010 competition are available. Table 1 shows our systems' performance compared to the maximum and median accuracies of all teams.

Statistic	Overall	ORG	GPE	PER
Maximum	86.80	85.2	79.57	96.01
Median	68.36	67.67	59.75	84.49
A	**81.91**	76.67	73.97	95.07
B	80.93	74.27	74.37	94.14
C	77.73	77.87	63.28	92.01

Table 1: TAC *10 test system statistics over 16 teams with a total of 46 runs*

7. Future Work

- Graph models over Wikipedia
- Clustering NIL entities
- Global models for NEL

8. Linking the news

We used our NEL system to link news text mentions to Wikipedia pages, allowing automatic aggregation of stories that relate to particular entities (see Figure below).

- Related entities – a "cloud" of entity names, the larger the name, the stronger the statistical relation.
- Related entity trends – the most related entities are plotted on a timeline with circle size and height reflecting the frequency and score.
- Recent stories – a list of stories that mention the entity. Clicking on a title displays the story text with hyperlinked entities.

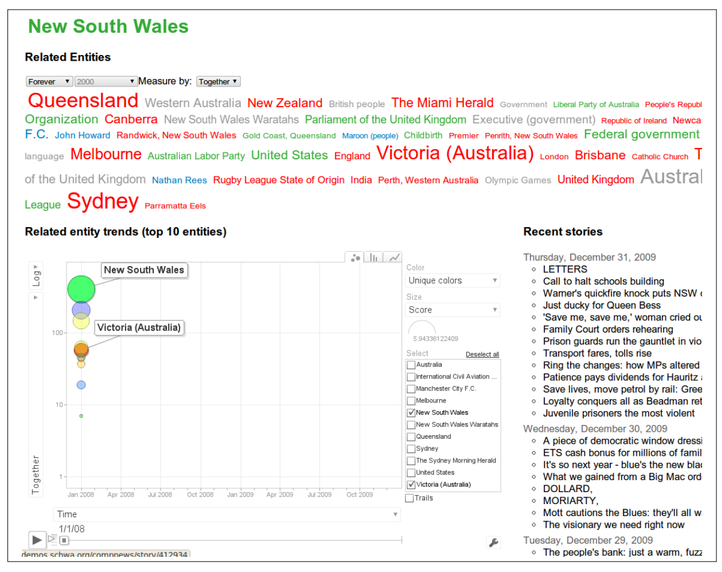

Figure 2: *Entity page for* `New South Wales`

References

Silviu Cucerzan. Large-scale named entity disambiguation based on Wikipedia data. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 708–716, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/D/D07/D07-1074`.

9. Acknowledgements

This is joint work with my supervisors and ə-lab colleagues Joel Nothman and Matt Honnibal.

This research was supported by the Capital Markets CRC.