COMPNEWS: LINKING NAMED ENTITIES IN AUSTRALIAN NEWS TEXT

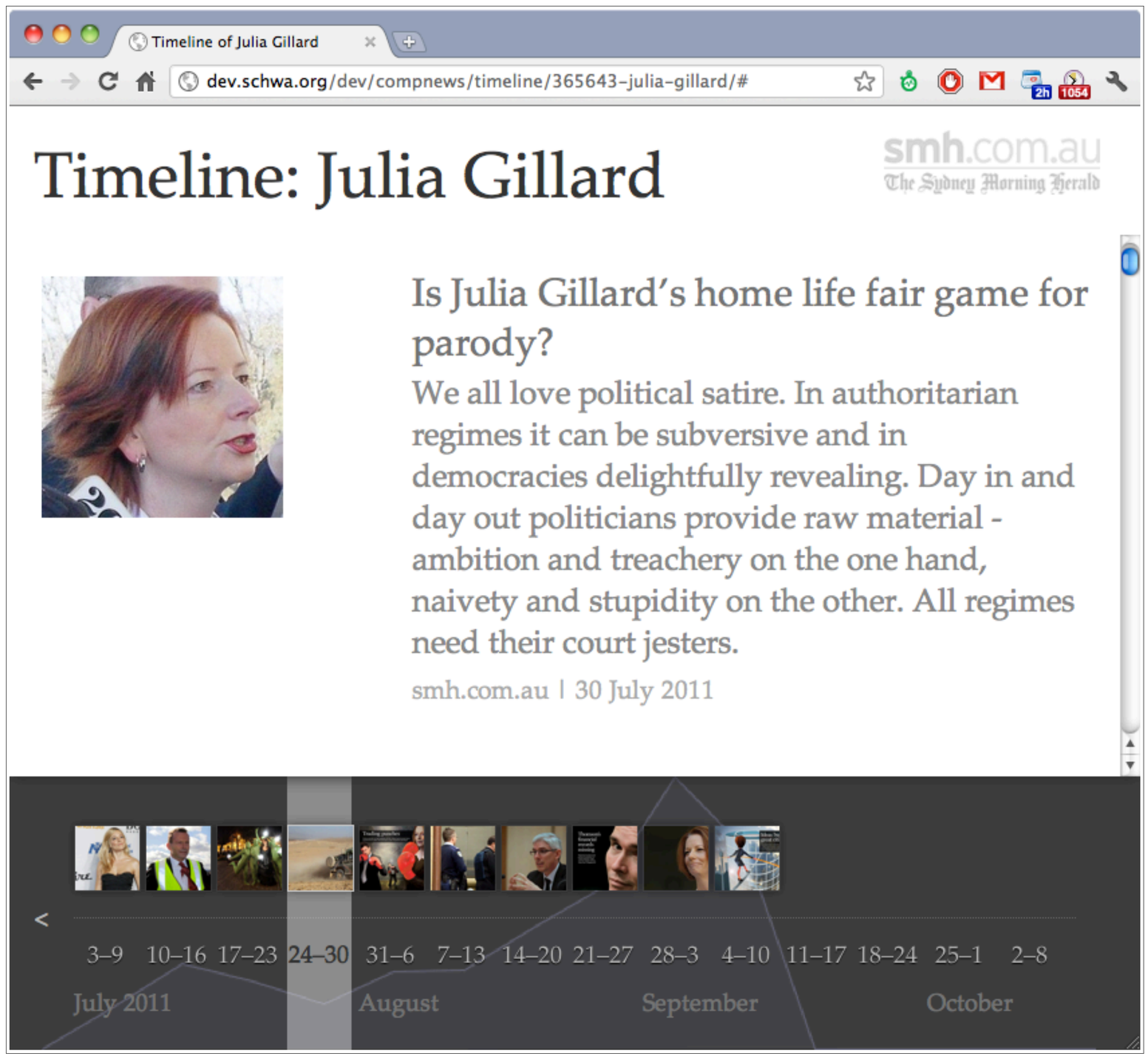
Will Radford – Supervisors: James Curran and Ben Hachey

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

1. Computable News

News is about **named entities** (NEs): the people, places and organisations found in news story text. The *Computable News* project overlays stories with *structured* entity information that can be exploited by news applications to engage with readers in novel ways.

- **What stories does an entity appear in?**
- **What other entities is this person related to?**
- **What quotes has this person said?**
- **What is the background for this story?**
- **Can we integrate entity research into editing tools so journalists can post rich stories more quickly?**



2. Named Entity Linking

Named Entity Recognition (NER) is the problem of identifying NEs in text. Named Entity Linking (NEL) attempts to resolve entity mentions to a knowledge base (KB) and addresses the problem of name ambiguity. For example, in the sentence: "John Howard will continue to work with the Seven Network."

- **Which** John Howard **are we talking about?**
  - **Former Prime Minister?**
  - **Australian Actor?**
  - **An employee of the Seven Network who is not in our KB?**

3. Linking to Wikipedia

Wikipedia is a large, free and dynamic online encyclopedia that has many useful features for use as a KB for NEL since many articles are about NEs:

- **Redirect pages** are valuable alternative aliases for NEs.
- **Disambiguation pages** explicitly flag ambiguous NEs.
- **Link structure** between articles provides contextual information.

4. Approach

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

- **Extract NEs from the query document:** East Timor, Australian, Dili, ABC , Ministry of Education.
- **Search Wikipedia for possible candidates for each NE:** the acronym ABC could refer to a number of different NEs.
- **Disambiguate the candidates:** select the candidate for each mention that best matches the *other* NEs found in the document.

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

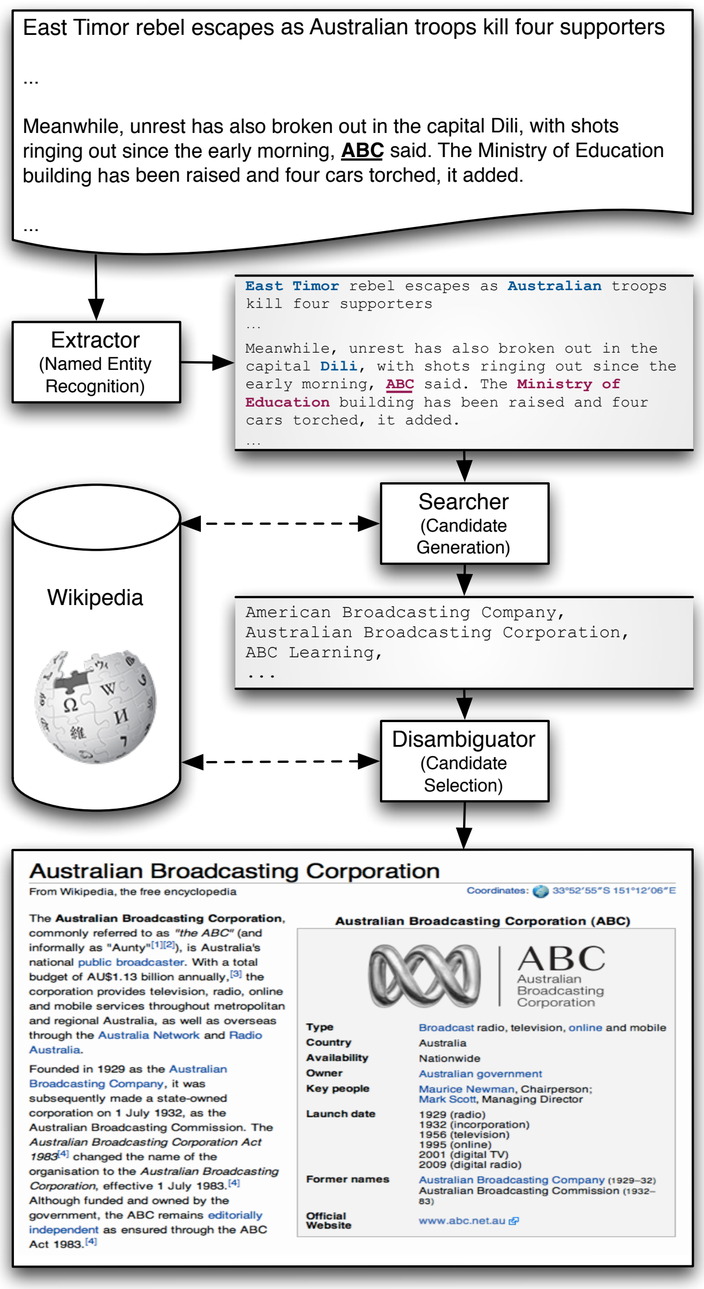


Figure 1: NEL *Workflow*

5. Evaluation

We designed a task and tool allowing users to read Sydney Morning Herald (SMH) stories, *annotate* NEs in the text and *link* them to the correct Wikipedia page or NIL if there was no page.

The evaluation in Table 1 shows the how well our system identified and linked each NE mention identified by the annotators. This is compared with the Wikipedia Miner system [Milne and Witten, 2008] that addresses the related task of *wikification*:

System	Precision	Recall	F-score
Milne & Witten based	32.57	30.29	31.38
Our system	65.48	61.45	**63.40**

Table 1: NEL *evaluation over 727* SMH *stories*

6. Industry Collaboration

The *Computable News* project is a cooperative project between The Capital Markets Cooperative Research Centre, Fairfax Media and The University of Sydney with clear research and commercial goals. The team includes academic staff, three postdoctoral researchers, three PhD students and Honours students.

References

Silviu Cucerzan. Large-scale named entity disambiguation based on Wikipedia data. In *Proceedings of EMNLP-CoNLL*, 2007.

David Milne and Ian H. Witten. Learning to link with Wikipedia. In *Proceedings of CIKM*, 2008.