COMPNEWS: QUOTATION EXTRACTION AND ATTRIBUTION

Tim O'Keefe – Supervisors: Irena Koprinska and James Curran

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

1. Quotes

Quotes are one of the main features of daily news stories, as they are often used by journalists to sum up the position that an individual holds on a certain topic. While certain quotes are quite memorable, there are countless others that could be interesting that we may have missed or forgotten. At present, news organisations do not consistently keep track of the quotes they have used in their news articles. As such there is no clear way to search through the quotes by an individual, short of a laborious search through countless news articles. As part of the *Computable News* project we have been investigating automatic quote extraction and attribution.

2. Task

The basic task of the quotes stream of the *Computable News* project is to use an automated system to find all of the quotes in a news article, and to attribute those quotes to the person who said them. The task can be divided into three steps:

1. **Quote Extraction** - Identify all of the direct quotes in the document of interest
2. **Named Entity Recognition** - Find all the mentions of named entities and any pronominal references to those entities in the document
3. **Quote Attribution** - Attribute each quote found in step 1 to an entity mention found in step 2

Some examples are shown below, with the appropriate mention highlighted in bold:

1. "The days when, you know, someone was a factional powerbroker behind the scenes should dictate the show, I think, should be stuck in the past," **Mr Rudd** said.
2. **He** said Mr Gates was aware of the team's progress. "We are very lucky to have him spending his money the way he does."
3. "We're doing everything we can to minimise the impact on our people," said a spokesman, **Scott Whiffen**
4. A **senior minister** asked yesterday: "What is it about Dickson?"

3. Approach

For quote extraction our system uses a simple rule-based approach that looks for text between quotation marks. To find named entities we were able to use the system built for the Named Entity Linking (NEL) part of the *Computable News* project, which gives us the extra advantage of being able to disambiguate the mentions of entities back to canonical representations of those entities in a knowledge base.

Finally, for attribution we have implemented two baseline rule-based approaches and are planning to implement a machine learning approach. The two baseline approaches are:

- **Nearest-to-quote** - Find the mention nearest to the quote in the same paragraph or any paragraph preceding the quote
- **Nearest-to-verb** - Find the mention nearest to a *reported speech verb* (said, exclaimed, etc) in the same paragraph or any paragraph preceding the quote

In the future we plan to cast the quote attribution task as a sequence labelling problem, which we can solve using one of several statistical learning techniques.

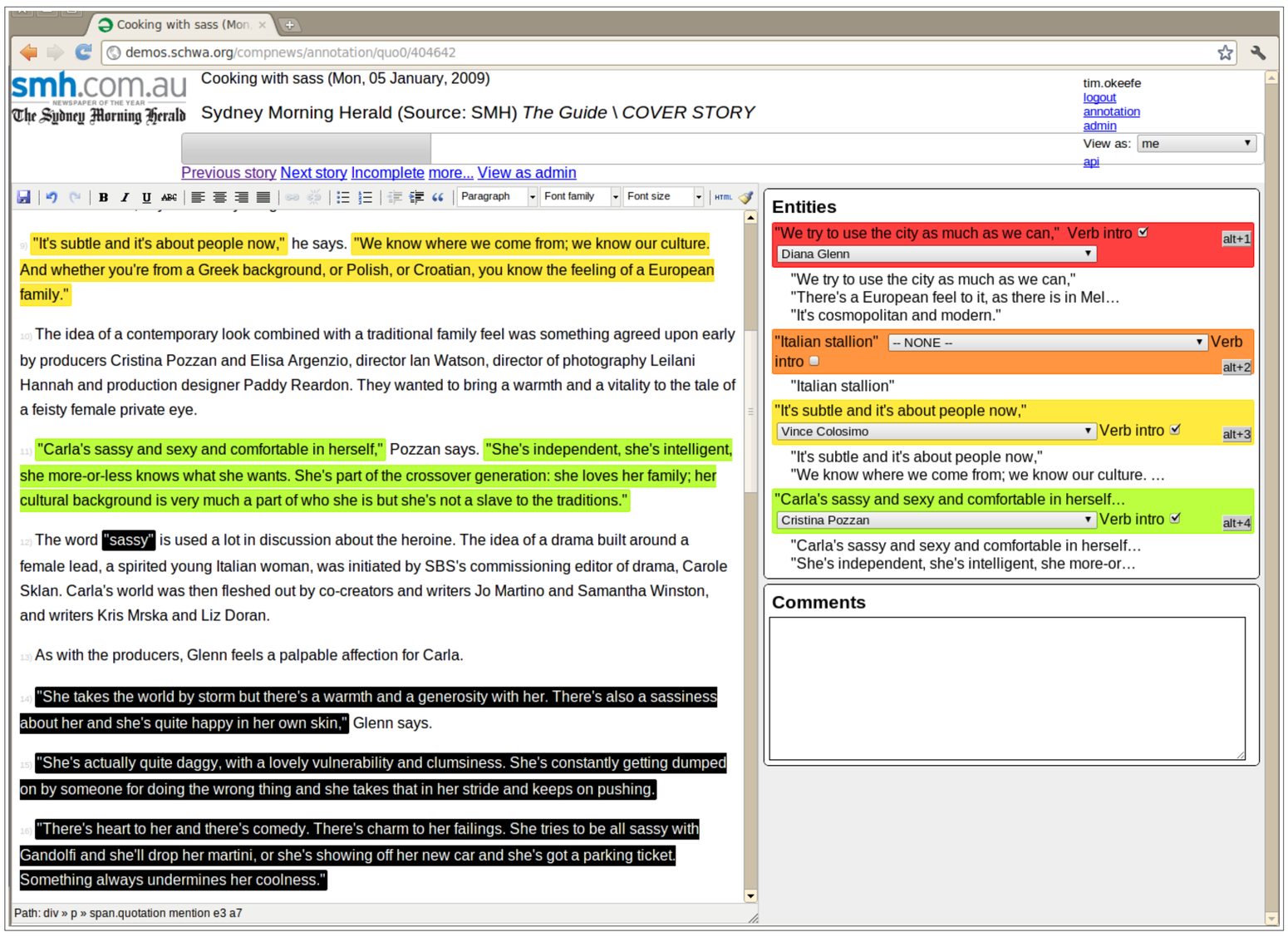

Figure 1: *The web-based annotation tool we used for marking named entities, quotes, and events*

4. Annotation

One of the most challenging aspects of research in Natural Language Processing (NLP) is that many tasks require annotated data in order to train machine learning algorithms and to determine how well they perform. Annotated data is usually very expensive and time-consuming to produce, and so for each task within NLP there are typically only a few data sets available. In the *Computable News* project, entity linking, event linking, and quote attribution all require annotated data, with different annotations required for each task. To be able to annotate a reasonable number of news documents for each of these tasks, we needed to develop an annotation process that was both cheap and versatile.

The process that we decided on was to employ non-expert annotators via the website Freelancer[1]. Freelancer allows people from all over the world to bid on small jobs, such as our annotation tasks. Crucially for us their bids come with an indication of their previous performance and their resumé is available. This allowed us to build partnerships with annotators who we could rely on to finish our tasks consistently, cheaply, and with a good level of quality.

Task	Documents	Words
Named Entities	2000	1,000,000+
Quotes	800	400,000+
Events	250	125,000+

Table 1: *Number of documents and words annotated for each task*

While Freelancer allowed us to annotate stories cheaply, it presented us with the challenge of building an annotation tool that was both easy-to-use and usable on many platforms. Additionally, the tool had to be versatile enough to allow us to produce different types of annotations, such as quotations and named entities.

To meet these goals we decided on a web-based tool, which we built with standard HTML and JavaScript, which can be seen in Figure 1. Using our annotation process we were able to cheaply produce a large quantity of annotations, as shown in Table 1.

5. Evaluation

Our rule-based extraction system was able to correctly identify 99.64% of the direct quotes in our corpus, with the small number of errors being the result of errors in the text, such as missing quotation marks.

For attribution, work on building the full machine learning approach is ongoing, so results for this are unavailable. We do have results for the baseline systems, which are shown in Table 2 below.

Nearest-to-quote	90.72%
Nearest-to-verb	91.16%

Table 2: *Quote attribution F-score over 3433 quotes*

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.

[1] http://www.freelancer.com