Initial framework

PAUL WARING
School of Computer Science, University of Manchester

1. PRINTER-FRIENDLY PAGES

A significant number of the popular news sites have 'printer-friendly' pages, which contain the same text as the standard pages but with most (if not all) of the navigation and images stripped. Occasionally sites also use different spreadsheets (with the `media="print"` attribute) to achieve a similar effect when the user opts to print a page. If we could follow these links/rendering options to obtain pages stripped of navigation, images, advertising etc. (which we are not interested in), this would probably make analysing the text of the page significantly easier as the majority of distractions have been removed automatically.

2. DISTINCT ENTITIES WITH THE SAME NAME

Additional information about a named entity (e.g. a person or company) can often be found either immediately preceding the entity's name or between two commas immediately following the first mention of the entity. For example, 'Lucy Neville-Rolfe, *Tesco's executive director of corporate and legal affairs*, said: [...]'. This information can be used in two different ways. First of all, it allows us to differentiate between references to two or more distinct entities which happen to share the same name. In addition, this information can help us map references which refer only to a description of an entity's role back into a proper name. For example, if we see a reference to 'Tesco's CEO' , we should be able to map this back to 'Terry Leahy'.

3. ANCHORING TEMPORAL REFERENCES

In most event descriptions, the majority of references[1] to instances or periods of time are relative, using phrases such as 'yesterday', 'next week' or 'two years ago'. In order to understand what these temporal references mean, we need to *anchor* them to a specific absolute date (and possibly time, if available) within the text. In the majority of cases, the temporal references within the text are relative to the time the event description was created – this is usually explicitly mentioned near the top or bottom of the text (top for most news stories, bottom for many blog entries).

3.1 Unanchored temporal references

Sometimes a previous event will be referred to without any anchor point which would allow us to resolve the relative time references. For example, 'Guardian journalists put a series of questions to Tesco over a period of nearly four months.' We have no way of knowing when this four month period took place, based on

[1] 75% according to a pilot study by Mani et al. [2003]

the text of the story. We could assume that it was the four months immediately preceeding the publication of the article, but there is no guarantee that this was actually the case.

3.2 Future events

Often stories will discuss events which are going to occur in the future, usually as a result of an announcement that an event is coming up. Indications of this can be found in temporal references involving phrases such as 'from' or 'until', and descriptions of events such as `<entity> is planning to do <something>`.

4. LOCATIONS

A small but significant number of event descriptions do not contain any references to locations within the text. This is usually because the event is not tied down to any particular area – e.g. the effects of the so-called 'credit crunch'. There are two ways in which this issue could be tackled:

(1) Ignore the location attribute for the event (i.e. assign it a null value). This means that we cannot link this event to any others based on location, but this limitation is not necessarily a problem.

(2) Try and determine a broad location based on other attributes of the text. For example, a story which mentions the Liberal Democrats and the British Bankers Association, and which is published on a UK-based web site, can have its location set to 'United Kingdom' based on these factors.

5. REPORTING ENTITY

Most reports of events mention the name of the entity which has authored the report, often with an affiliating entity (e.g. journalist and newspaper), either near the top or bottom of the text. However, a problem arises as to whether or not to include this entity in the description of the event. Some cases are more clear-cut than others – for example, if a description of an event mentions 'I', this will generally be a reference to the author of the text (especially so on blogs, which often follow a first-person narrative format) and therefore should be included, although care would need to be taken in instances where this pronoun is being used within quotes and therefore may apply to someone other than the author of the text. However, in other cases it is less obvious. How can we tell if someone is merely reporting an event, in which case we should not consider them to be one of the entities associated with it, or if they actually took part and are therefore associated with the event?

6. RELEVANT ENTITIES

Although a number of entities may be mentioned in the description of an event, we probably want to concentrate only on the most relevant ones, in order to avoid creating links to events which are unrelated. The identification of the most relevant entities can generally be achieved by frequency and positional analysis, as these entities are usually mentioned multiple times and towards the beginning of the event description (often within the title, if present).

Paul Waring – pwaring@cs.man.ac.uk

7. LOCATION OF RELEVANT INFORMATION

Although some event descriptions can run on for several pages, the most important and relevant information can usually be found, as one might expect, near the top of the first page – with the exception of the anchor time and the author, which may instead lie at the bottom of the text. The title of the page is also a useful indicator, as in many cases it will feature on of the entities associated with the event, as well as 'what happened', often in the form `<entity x> <verb> <entity y>`, e.g. 'Pirates seize French yacht'.

8. SITE-SPECIFIC INFORMATION

As might be expected, there are differences in the ways in which specific sites describe events. These can generally be split into two areas: structure and content.

8.1 Structure

Specific sites deal with the structure of the page in different ways – for example, the New York Times has the date followed by the title and the author, whereas CNN has the title followed by story highlights and the author information is located at the bottom of the page. Having a site-specific method of extracting this information for popular sites might be useful in order to ensure that we obtain the most accurate information possible.

8.2 Content

The actual text used to describe events also differs from site to site, although there are often large chunks of common information across sites. In particular, there are subtle differences in the ways in which US and UK authors report events, although since the majority of event attributes are described in the same way (e.g. 'London' is going to be referred to by the same name in both), this may not be a major problem.

REFERENCES

Mani, I., Schiffman, B., and Zhang, J. 2003. Inferring temporal ordering of events in news. In *NAACL '03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*. Association for Computational Linguistics, 55–57.