Marking Time in Developmental Biology

Gail Sinclair and Bonnie Webber
School of Informatics
University of Edinburgh
Edinburgh EH8 9LW
c.g.sinclair@ed.ac.uk, bonnie@inf.ed.ac.uk

1 Introduction

In developmental biology, to support reasoning about cause and effect, it is critical to link genetic pathways with processes at the cellular and tissue level that take place beforehand, simultaneously or subsequently. While researchers have worked on resolving with respect to absolute time, events mentioned in *medical* texts such as clinical narratives (e.g. Zhou et al, 2006), events in developmental biology are primarily resolved relative to other events.

In this regard, I am developing a system to extract and time-stamp event sentences in articles on developmental biology, looking beyond the sentence that describes the event and considering ranges of times rather than just single timestamps.

I started by creating four gold standard corpora for documents, event sentences, entities and timestamped events (for future public release). These datasets are being used to develop an automated pipeline to (1) retrieve relevant documents; (2) identify sentences within the documents that describe developmental events; and (3) associate these events with the developmental stage(s) that the article links them with or they are known to be linked with through prior knowledge.

Different types of evidence are used in each step. For determining the relevant developmental stage(s), the text surrounding an event-containing sentence is an efficient source of temporal grounding due of its immediate accessibility. However, this does not always yield the correct stage and other sources need to be used. Information within the sentence, such as the entities under discussion, can also be used to help with temporal grounding using mined background knowledge about the period of existence of an entity.

2 Creation of Datasets

In creating the four new data sets mentioned above, I annotated 1200 documents according to relevance to murine kidney development. From 5 relevant documents, 1200 sentences were annotated as to whether they contained an event description. (Two annotators - one biologist, one computer scientist - achieved an inter-annotator agreement kappa score of 95%.) A sentence is considered a positive one if it contains a description of the following event types:

- molecular expression within tissue/during process/at stage X (molecular event)
- tissue process, i.e. what forms from what (tissue event)
- requirement of a molecule for a process (molecular or tissue event)
- abnormality in a process/tissue/stage (molecular or tissue event)
- negation of the above e.g. was not expressed, did not form, formed normally (molecular or tissue event).

A negative sentence is one that does not fall under at least one of the above categories.

In addition, 6 entities (*tissue, process, species, stage, molecule and event verb*) were annotated in 1800 sentences (1200 described above + 600 from

BioNLP 2007: Biological, translational, and clinical language processing, pages 197–198,
Prague, June 2007.

relevant documents not yet annotated at sentence level) and 347 entity-annotated positive event sentences were marked with their associated developmental stage.

Example: *At E11, the integrin α8 subunit was expressed throughout the mesenchyme of the nephrogenic cord.* Entities annotated: *E11*(stage), *integrin α8* (molecule), *expressed* (event verb), *mesenchyme of the nephrogenic cord* (tissue).

3 Evidence for Temporal Resolution

Developmental biology is not as concerned with the absolute time of events in a specific embryo as it is with events that generally happen under the same circumstances in developmental time. These are referred to with respect to *stages* from conception to birth. The evidence sufficient to resolve the developmental stage of an event sentence can come from many places. The two significant areas of evidence are *local context* (i.e. surrounding text) and *prior* (i.e. background) knowledge.

Local context can further be classified as:

- **explicit**: evidence of stage is mentioned within current (event) sentence,
- **previous sentence**: evidence is found in sentence immediately previous to current sentence,
- **following sentence**: evidence is found in sentence immediately following current sentence,
- **current paragraph**: evidence is found in paragraph containing current sentence but not in adjacent sentences,
- **referenced to figure**: evidence is found in figure legend referenced in current sentence.

Evidence Source	# Event Sentences
Explicitly Stated	48
Immed Prev Sentence	7
Following Sentence	1
Current Paragraph	19
Referenced Figure Legend	38
Within Figure Legend	43
Time Irrelevant	65
Prior Knowledge	126
Total	347

When local context does not provide evidence, **prior knowledge** can be used about when entities mentioned within the sentence normally appear within development. Event sentences can also be **irrelevant** of individual time ranges and apply to the whole of development. The table above shows the frequency with which each evidence type is used to resolve developmental stage.

4 Experiments

Event sentence retrieval experiments (using separate training and test data) resulted in a F-score of 72.3% and 86.6% for Naive Bayes and rule-based classification approaches respectively (relying upon perfect entity recognition). A baseline method (classifying all sentences as positive) achieves 58.4% F-score.

Experiments were also carried out to assign developmental stage to sentences already known to contain events. The baseline approach is to use the last mentioned stage in the text and any methods developed should score higher than this baseline. Rules were developed to assign developmental stage based on the knowledge gained from two fifths of the investigations into temporal evidence described above. The other three fifths were annotated after the rules had been defined. Precision scores for all 347 sentences can be seen in the following table with the *Naive* method representing the baseline and *Local* representing the use of rules.

Paper	Naive Prec.	Local Prec.
1	75.7	97.3
2	89.6	90.9
3	89.1	100
4	95.6	92.3
5	95.5	91.3
Average	89.1	94.5

Experiments are currently ongoing into exploiting the use of background knowledge of the developmental processes and tissues mentioned within event descriptions in order to assign developmental stage to events sentences not already assigned by the local context rules and to increase confidence in those stages already assigned.

References

L. Zhou, G. B. Melton, S. Parsons and G Hripcsak, A temporal constraint structure for extracting temporal information from clinical narrative, J Biomed Inf 39(4), Aug 2006, 424-439