Shahid Razzaq
CSE 574 – Artificial Intelligence II – WINTER 08
Project Proposal

Members:

Shahid Razzaq

Topic:

Timeline Extraction of Noun Phrase Entities

Proposals:

Timelines provide an easy and concise way to comprehend a series of events related to a given entity, as opposed to several paragraphs of written text. Temporal information on the internet about entities can thus be extracted from existing plain text or structured data (infoboxes). Wikipedia, owing to its community based approach to manage data, provides for a very accurate description of events related to (noun phrase) entities. Thus the idea is to train a temporal data extractor based on Wikipedia source and use it to complete or fill in additional data from other sources on the web.

The output of the timeline extraction will be to produce tuples of time and event that cannot be found on Wikipedia, for a given (noun phrase) entity. This data can then be used to create a visual timeline but that is out of the scope of this project.

The Timeline Extractor will execute the following tasks (in order):

- **Entity Category/Class Recognizer:** for a given noun phrase, find out the category to which it belong e.g. Barak Obama -> a politician
- **General Events/Milestones related to Category:** for the category found above, crawl numerous pages on Wikipedia for that category and find the most likely events that are related to entities in the category e.g. for the example of a politician, events would be 'getting elected', 'carry out rallies', 'vote on/against bill' etc
- **Timeline Supplementer Step 1:** For a given noun phrase entity, find the missing events data that corresponds to the General Events/Milestones for the entities category, by extracting data from the web e.g. if Barak Obama's tenure as Senator is missing, find it from the web.
- **Timeline Supplementer Step 2:** for a given noun phrase entity, search the web by using the general events/milestone data for the entity, and find new event/timeline information.

Training Data: Wikipedia

Test Data: Manual testing i.e. community driven edit/update of information extracted from the timeline extractor

Success will be evaluated by displaying two timelines (visual horizontal/vertical bar calibrated with date information and annotated with event data), one for the original Wikipedia article and a second from the output of the Timeline Extractor.

Schedule:

Feb 20 (Milestone 1): Experiment with different entity classifiers and present the classifier of choice.
Experiment with learners for finding the General Events/Milestones for entity categories.

Feb 29 (Milestone 2): Present the learner of choice for finding General Events/Milestones for categories.
Complete the implementation of the classifier and learner that generates a list of General Events, given an entity (noun phrase), using Wikipedia.

Mar 7 (Milestone 3): Complete the implementation of the information extractor that crawls the web to complete timeline data.

Mar 12 and 14: Project presentations, demos, etc.

Mar 21: Final Report