LensingWikipedia: Parsing Text for the Interactive Visualization of Human History

Ravikiran Vadlapudi
rvadlapu@cs.sfu.ca
Simon Fraser University

Maryam Siahbani
msiahban@cs.sfu.ca
Simon Fraser University

Anoop Sarkar
anoop@cs.sfu.ca
Simon Fraser University

John Dill
dill@cs.sfu.ca
Simon Fraser University

Abstract

Extracting information from text is challenging. Most current practices treat text as a bag of words or word clusters, ignoring valuable linguistic information. Leveraging this linguistic information, we propose a novel approach to visualize textual information. The novelty lies in using state-of-the-art Natural Language Processing (NLP) tools to automatically annotate text which provides a basis for new and powerful interactive visualizations. Using NLP tools, we built a web-based interactive visual browser for human history articles from Wikipedia.

1 Introduction

Information visualization techniques aim to translate abstract information into a visual form to get a new insight and have been successfully done so for a wide range of tasks. However, text visualization remains challenging, mostly due to the complex underlying grammatical structure of natural language.

Visually highlighting key terms and relations among them to gain new insights is a common practice. The downside is it completely ignores linguistic structures. To allow analysts to search for entities related to specific events requires a sophisticated language analysis. For this purpose, state-of-the-art Natural Language Processing (NLP) algorithms can be used to identify both entities and relationships among such entities in text. Visualization can then show underlying entity relationships enabling discovery of hidden information. Here, we propose a novel way to visualize high-level textual information leveraging linguistic structures extracted by NLP algorithms.

To demonstrate this process, we chose human history articles from Wikipedia describing who did what to whom, when and where. This *who-did-what-to-whom* structure is defined as predicate-argument structure. We use a wide-coverage semantic parsing approach called *Semantic Role Labeling* (SRL) [3] to automatically extract this linguistic information. Linguistic annotations of text are represented in three connected interactive visualizations: *R-graph*, *timeline* and *map*. We call this system LensingWikipedia (`lensingwikipedia.cs.sfu.ca`).

2 Data Processing

We used web pages from Wikipedia summarizing about 2000 years of human history. English Wikipedia contains about 2000 URLs which are natural language summaries of important events in each year or decade in human history (e.g., Fig. 1 is a web page with about 85 events from the 1470's). In all, Wikipedia URLs gave us 20,000 events, with each event described by one to several sentences. We use the SRL approach to generate a *predicate argument structure* for each sentence (section 2.1) and then extract temporal and spatial information for each event (section 2.2). This NLP-based automatic annotation process provides a view into the data that expresses who did what to whom, when and where.

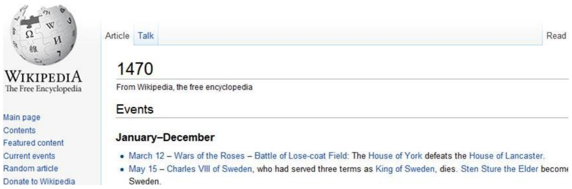
2.1 Predicate-Argument Structures by SRL

To extract predicate-argument structures, we use an SRL approach based on large-scale statistical machine learning [1, 3] based on a semantic role data set annotated by linguistic experts called the Proposition Bank corpus (PropBank) [4]. The following is an example of a PropBank style SRL annotation of a sentence.

> **Input**: *In the opener, Sony Corp. would agree to buy Columbia Pictures Entertainment Inc. in a transaction valued at close to $ 5 billion.*
> **Semantic role labeling's output**:
> A0 (Buyer): *Sony Corp.*
> Pred V (Buy): *buy*
> A1 (Thing Bought): *Columbia Pictures Entertainment Inc.*

The SRL tool provides the predicate argument structure for sentences in the text such as *The House of York defeats the House of Lancaster* where *defeat* is the predicate with arguments *The House of York* (*arg0*) and *the House of Lancaster* (*arg1*). The semantic role labels (*arg0* and *arg1*) are transformed automatically into text (*arg0:'entity_victorious'* and *arg1:'entity_defeated'*) that is readable by the user using PropBank. This involves learning a mapping between abstract semantic role labels and verbose descriptions. This task is harder than it seems, because the verbose label depends on the sense of the verb. For instance, 'get' might have 'receiver' as verbose label for 'arg0', but it might also have 'instigator' for another sense of the verb (get across). We have worked on many different models to solve this task achieving an accuracy of 92%.

Each sentence might have multiple predicates, each with multiple arguments. We use only the first two arguments (*arg0* and *arg1*). The two most frequent predicates in the data are *found* and *defeat*.



1470

From Wikipedia, the free encyclopedia

Events

January–December

- March 12 – Wars of the Roses – Battle of Lose-coat Field: The House of York defeats the House of Lancaster.
- May 15 – Charles VIII of Sweden, who had served three terms as King of Sweden, dies. Sten Sture the Elder become Sweden.

Figure 1: One URL from Wikipedia about the decade 1470.

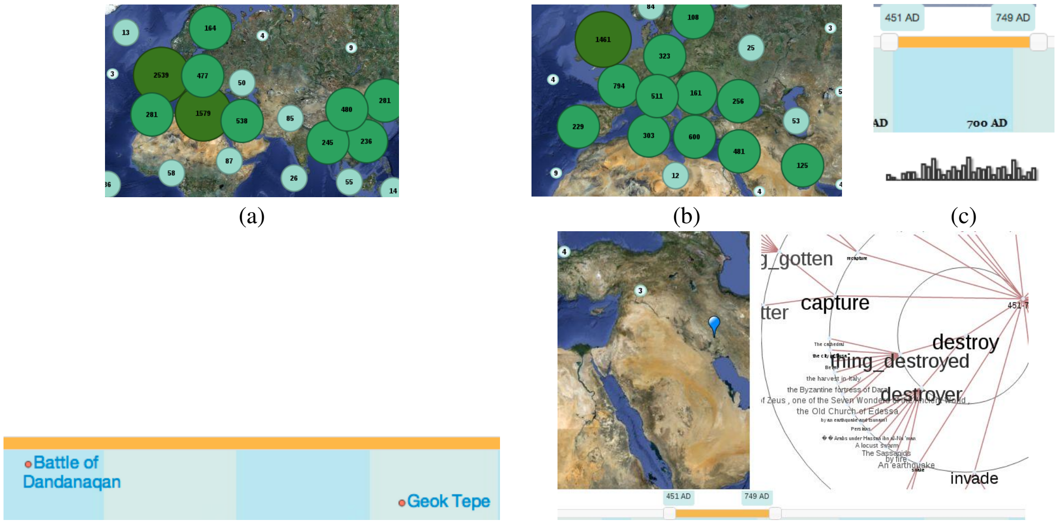
2.2 Temporal and Geographical Identification

Our main idea was to look for events we could situate in time and space. From the semantic parse and the URL we extract information such as the date when the event occurred and the main actors of the event as expressed in the different predicates associated with each event. For event geo-location, we used human annotated version of these 2000 URLs with geo-location data from a separate project: `http://www.ragtag.info/2011/feb/2/history-world-100-seconds/`.

For instance, we converted the event:

> Portuguese sailors reach Mina de Ouro on the Gold Coast (present-day Ghana) and explore Cape St. Catherine, two degrees south of the equator. Mina de Ouro becomes the chief center for the gold trade and a major source of revenue for the crown.

into the following representation for the predicate *become* as part of this event. Each event could contribute several such entries in our transformed data-set.



(a) (b) (c)

(d) (e)

Figure 2: Visualizations of events in time and space. Cluster colour scheme in the map view (a)&(b) is sequential as per colorbrewer[2] depending on number of events in each region.

```
”arg0”:        ”Mina de Ouro”,
”arg1”:        ”the chief center for the gold ..”,
”event”:       ”become”,
”latitude”:    5.5499977999999999,
”longitude”:   -0.249999,
”roleArg0”:    ”Agent”,
”roleArg1”:    ”entity_changing”,
”title”:       ”Ghana”,
”year”:        1471
```

Figure 3: The final output from the natural language processing pipeline combined with the temporal identification and geo-location. There is an additional description field which includes the text of the event from Wikipedia which is quoted in the text.

3 Visualization

We collectively visualize location information, temporal information and predicate-argument information about events using three connected visualizations: geographical view (map), a temporal view (timeline) and a view of the relationship between the predicates and various entities in the text (R-graph). Geo-location data allowed situating events on an interactive map (for which we used the Google Maps API) shown in Fig. 2(a). Clusters [1] indicate number of events in a region depicted by color and size. Zooming to region of interest (Fig. 2(b)) would load only region sepcific events. The temporal identification of each event allowed us to use an interactive timeline shown in Fig. 2(c) (we should emphasize that we are using information extracted from semi-structured text on Wikipedia for this). Finally, we are able to show individual types of predicates, such as *destroy*, and their arguments, such as *destroyer*, on Rgraph as shown in Fig. 2(e). The three views: map, timeline and R-graph are all inter-linked to enable a joint exploration of the information in the underlying text data. The natural language descriptions of arguments permits the user to search for intuitive terms such as ’thing_bought’ rather than abstract semantic role labels such as the ’arg1’ of predicate ’buy’ which is what is produced by our SRL system.

Using the joint exploration via map, timeline and R-graph, the user can identify individual events of interest which are then shown in the timeline view - Fig. 2(d). Brushing the event title shows the original sentence(s) from Wikipedia for that event and clicking on it takes the user to the Wikipedia page for that event.

4 Observation and Conclusion

Some advantages of LensingWikipedia are: a) Focusing on a location with single click reveals a summary of its history from Wikipedia. b) It is potentially useful for Wikipedia editors to monitor Wikipedia coverage and add missing important events. c) Easy exploration of events, e.g., to find out how often a specific country was engaged in ’wars’, for example, would require selecting events like ’conquer’ or ’attack’. And zooming out all the way would reveal all countries engaged in ’wars’. Additional information about the distribution of such events across time is provided by the *Bar graph* below the timeline indicating active and passive time frames. d) To list out all ’invaders’ of a specific location requires just two clicks, selecting a location and the ’invader’ role on SRL view.

LensingWikipedia shows that text transformed using NLP tools allows a powerful exploration of textual documents. The key idea was to extract who-did-what-to-whom information from text which was captured using semantic role labelling and URL information. We observed that our visual browser for Wikipedia provided valuable insights which could be easily obtained with a few clicks.

Acknowledgements

This work was partially funded through a Collaborative Research and Development (CRD) grant from NSERC, Canada based on a generous contribution from The Boeing Company and AeroInfo Systems.

References

[1] D. Gildea and D. Jurafsky. Automatic labeling of semantic roles. *Comput. Linguist.*, 28(3):245–288, Sept. 2002.

[2] M. Harrower and C. A. Brewer. Colorbrewer.org: An online tool for selecting colour schemes for maps. *Cartographic Journal*, 40(1):27–37, 2003.

[3] Y. Liu and A. Sarkar. Experimental evaluation of LTAG-based features for semantic role labeling. In *Proceedings of the (EMNLP-CoNLL)*, pages 590–599, jun 2007.

[4] M. Palmer, D. Gildea, and P. Kingsbury. The proposition bank: An annotated corpus of semantic roles. *Comput. Linguist.*, 31(1):71–106, Mar. 2005.

[1] Clusters generated by MarkerClusterer.js redundantly encodes density information as both size and color. We plan to replace it with choropleth style encoding.