Lydia: A System for Large-Scale News Analysis*
(Extended Abstract)

Levon Lloyd, Dimitrios Kechagias, and Steven Skiena

Department of Computer Science,
State University of New York at Stony Brook,
Stony Brook, NY 11794-4400
{lloyd, dkechag, skiena}@cs.sunysb.edu

1 Introduction

Periodical publications represent a rich and recurrent source of knowledge on both current and historical events. The *Lydia* project seeks to build a relational model of people, places, and things through natural language processing of news sources and the statistical analysis of entity frequencies and co-locations. *Lydia* is still at a relatively early stage of development, but it is already producing interesting analysis of significant volumes of text. Indeed, we encourage the reader to visit our website (*http://www.textmap.com*) to see our analysis of recent news obtained from over 500 daily online news sources.

Perhaps the most familiar news analysis system is *Google News* [1], which continually monitors 4,500 news sources. Applying state-of-the-art techniques in topic detection and tracking, they cluster articles by event, group these clusters into groups of articles about related events, and categorize each event into predetermined top-level categories, finally selecting a single representative article for each cluster. A notable academic project along these lines is Columbia University's *Newsblaster* [2,4,8], which goes further in providing computer-generated summaries of the day's news from the articles in a given cluster.

Our analysis is quite different from this. We track the temporal and spatial distribution of the entities in the news: who is being talked about, by whom, when, and where? Section 2 more clearly describes the nature of the news analysis we provide, and presents some global analysis of articles by source and type to demonstrate the power of *Lydia*.

Lydia is designed for high-speed analysis of online text. We seek to analyze thousands of curated text feeds daily. *Lydia* is capable of retrieving a daily newspaper like *The New York Times* and then analyzing the resulting stream of text in under one minute of computer time. We are capable of processing the entire 12 million abstracts of Medline/Pubmed in roughly two weeks on a single computer, covering virtually every paper of biological or medical interest published since the 1960's.

A block diagram of the *Lydia* processing pipeline appears in Figure 1. The major phases of our analysis are:

* This research was partially supported by NSF grants EIA-0325123 and DBI-0444815.

M. Consens and G. Navarro (Eds.): SPIRE 2005, LNCS 3772, pp. 161–166, 2005.

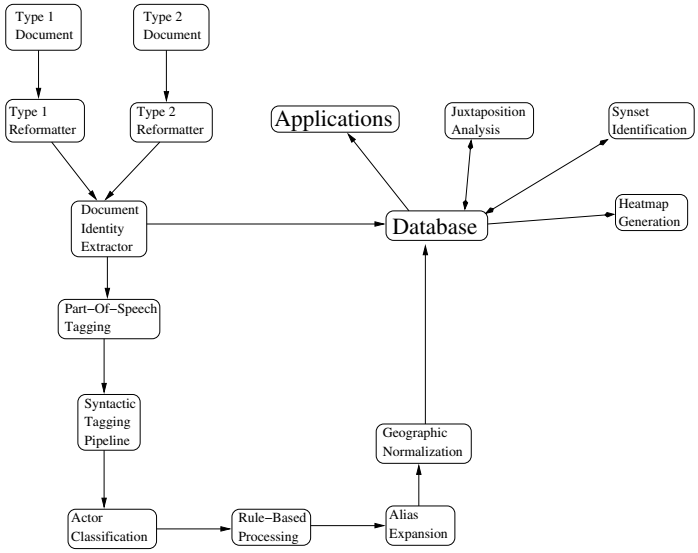

Fig. 1. Block diagram of the *Lydia* Pipeline

- *Spidering and Article Classification* – We obtain our newspaper text via spidering and parsing programs which require surprisingly little customization for different news sources.
- *Named Entity Recognition* – Identifying where *entities* (people, places, companies, etc.) are mentioned in newspaper articles is a critical first step in extracting information from them.
- *Juxtaposition Analysis* – For each entity, we wish to identify what other entities occur near it in an overrepresented way.
- *Synonym Set Identification* – A single news entity is often mentioned using multiple variations on their name. For example, *George Bush* is commonly referred to as *Bush*, *President Bush* and *George W. Bush*.
- *Temporal and Spatial Analysis* – We can establish local biases in the news by analyzing the relative frequency given entities are mentioned in different news sources. To compute the sphere of influence for a given newspaper, we look at its circulation, location, and the population of surrounding cities. We expand the radius of the sphere of influence until the population in it exceeds the circulation of the newspaper.

2 News Analysis with Lydia

In this section, we demonstrate the juxtapositional, spatial, and temporal entity analysis made possible by *Lydia*. We again encourage the reader to visit (*http://www.textmap.com*) to get a better feel of the power of this analysis on contemporary news topics.

Table 1. Top 10 Juxtapositions for Three Particular Entities

Martin Luther King		Israel		North Carolina	
Entity	Score	Entity	Score	Entity	Score
Jesse Jackson	545.97	Mahmoud Abbas	9,635.51	Duke	2,747.85
Coretta Scott King	454.51	Palestinians	9,041.70	ACC	1,666.92
"I Have A Dream"	370.37	West Bank	6,423.93	Wake Forest	1,554.92
Atlanta, GA	286.73	Gaza	4,391.05	Virginia	1,283.61
Ebenezer Baptist Church	260.84	Ariel Sharon	3,620.84	Tar Heels	1,237.39
Saxby Chambliss	227.47	Hamas	2,196.72	Maryland	1,029.20
Douglass Theatre	215.79	Jerusalem, ISR	2,125.96	Raymond Felton	929.48
SCLC	208.47	Israelis	1,786.67	Rashad McCants	871.44
Greenville, SC	199.27	Yasser Arafat	1,769.58	Roy Williams	745.19
Harry Belafonte	190.07	Egypt	1,526.77	Georgia Tech	684.07

Except where noted, all of the experiments in this paper were run on a set of 3,853 newspaper-days, partitioned among 66 distinct publications that were spidered between January 4, 2005 and March 15, 2005.

Juxtaposition Analysis. Our mental model of where an entity fits into the world depends largely upon how it relates to other entities. For each entity, we compute a significance score for every other entity that co-occurs with it, and rank its juxtapositions by this score. Table 1 shows the top 10 scoring juxtapositions (with significance score) for three popular entities. Some things to note from the table are:

- Many of the other entities in *Martin Luther King*'s list arise from festivities that surrounded his birthday.
- The position of *Mahmoud Abbas* at the top of *Israel*'s list reflects his ascent to the presidency of the *Palestinian National Authority*.
- The prominence of other universities and basketball terms in the *North Carolina* list reflects the quality and significance of the UNC basketball team.

There has been much work [5,6] on the similar problem of *recommender systems* for e-commerce. These systems seek to find what products a consumer is likely to purchase, given the products they have recently purchased. Our problem is also similar to the word collocation problem[7] from natural language processing. The goal there is to find which sets of two or more words occur close to each other more than they should by chance.

Developing a meaningful juxtapositionness function proved more difficult than anticipated. First, we discovered that if you simply use raw article counts, then the most popular entities will overly dominate the juxtapositions. Care must be taken, however, when punishing the popular entities against spurious juxtapositions dominated by the infrequently occurring entities. Our experience found that the popular scoring functions appearing in the literature [3] did not adequately correct for this problem.

To determine the significance of a juxtaposition, we bound the probability that two entities co-occur in the number of articles that they co-occur in if occurrences where generated by a random process. To estimate this probability we use a Chernoff Bound:

$$P(X > (1+\delta)E[X]) \leq (\frac{e^{\delta}}{(1+\delta)^{(1+\delta)}})^{E[X]}$$

where δ measures how far above the expected value the random variable is. If we set $(1+\delta)E[X] = F =$ number of co-occurrences, and consider X as the number of randomized juxtapositions, we can bound the probability that we observe at least F juxtapositions by calculating

$$P(X > F) \leq (\frac{e^{\frac{F}{E[X]}-1}}{(\frac{F}{E[X]})^{(\frac{F}{E[X]})}})^{E[X]}$$

where $E[X] = \frac{n_a n_b}{N}, N =$ number of sentences in the corpus, $n_a =$ number of occurrences of entity a, and $n_b =$ number of occurrences of entity b, as the juxtaposition score for a pair of entities. We display $-\log$ of this probability for numerical stability and ranking.

Spatial Analysis. It is interesting to see where in the country people are talking about particular entities. Each newspaper has a location and a circulation and each city has a population. These facts allow us to approximate a *sphere of influence* for each newspaper. The *heat* an entity is generating in a city is now a function of its frequency of reference in each of the newspapers that have influence over that city.

Figure 2 show the heatmap for *Washington DC* and *Phoenix*, in the news from over 500 United States news sources from April 11–May 30, 2005. The most intense heat for both city-entities focuses around their location, as should be expected. *Washington DC* generates a high level of interest throughout the United States. There is an additional minor concentration in the Pacific Northwest, which reflects the ambiguity between *Washington* the city and the state.

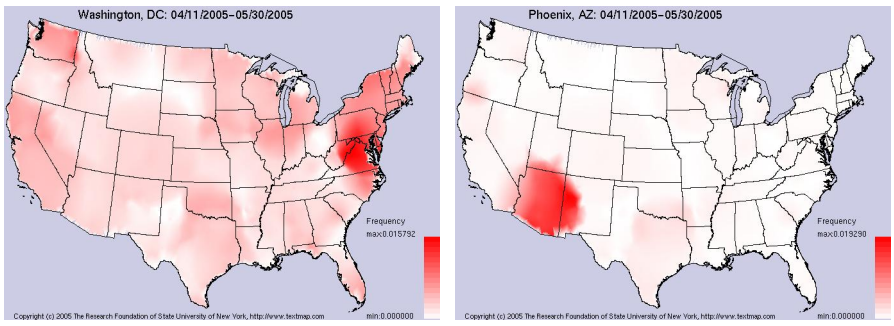

Fig. 2. Geographic News Distribution of two Spatially-Sensitive Entities

Table 2. Most Overrepresented Entities in Three Important U.S. Newspapers

San Francisco Chronicle		Chicago Tribune		Miami Herald	
Entity	Score	Entity	Score	Entity	Score
Gavin Newsom	10.84	Chicago, IL	8.57	Miami, FL	10.26
San Francisco, CA	10.56	Richard Daley	7.06	South Florida	9.53
Bay Area	8.44	Joan Humphrey Lefkow	5.20	Fort Lauderdale, FL	8.76
Pedro Feliz	5.36	Aon Corp.	4.69	Cuba	8.09
BALCO	5.29	Salvador Dali	4.54	Caracas	7.02
Kimberly Bell	5.02	Wrigley Field	4.42	Florida Marlins	6.91

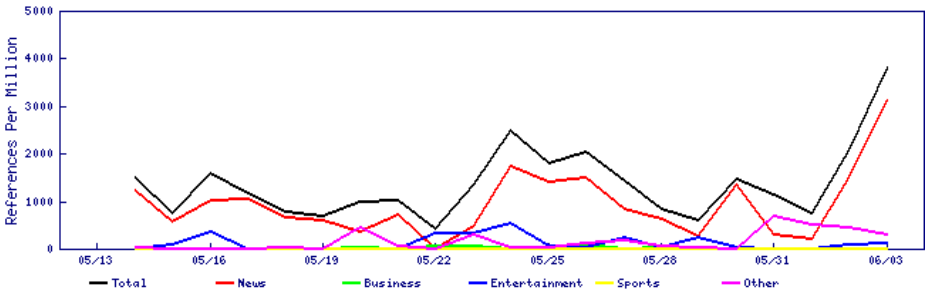

Fig. 3. Reference Frequency Time-Series for *Michael Jackson*, partitioned by article type

An alternate way to study relative geographic interest is to compare the reference frequency of entities in a given news source. Table 2 presents the most overrepresented entities in each of three major American newspapers, as scored by the number of standard deviations above expectation. These over-represented entities (primarily local politicians and sports teams) are all of stronger local interest than national interest.

Temporal Analysis. Our ability to track all references to entities broken down by article type gives us the ability to monitor trends. Figure 3 tracks the ebbs and flows in the interest in *Michael Jackson* as his trial progressed in May 2005. Note that the vast majority of his references are classified as news instead of entertainment, reflecting current media obsessions.

3 Conclusions and Future Work

We have presented the basic design and architecture of the *Lydia* text analysis system, along with experimental results illustrating its capabilities and performance. We are continuing to improve the entity recognition algorithms, particularly in synset construction, entity classification, and geographic normalization.

Future directions include dramatically increasing the scale of our analysis, as we anticipate moving from a single workstation to a 50-node cluster computer in the near future. With such resources, we should be able to do long-term

historical news analysis and perhaps even larger-scale web studies. We are also exploring the use of the *Lydia* knowledge base as the foundation for a question answering system, and extracting semantic labels for explaining juxtaposition relationships.

Acknowledgments

We thank Alex Kim for his help in developing the pipeline, Manjunath Srinivasaiah for his work on making the pipeline more efficient, Prachi Kaulgud for her work on markup and web interface design, Andrew Mehler for his Bayesian classifier and rules processor, Izzet Zorlu for his web interface design, Namrata Godbole for her work on text markup and spidering, Yue Wang, Yunfan Bao, and Xin Li for their work on Heatmaps, and Michael Papile for his geographic normalization routine.

References

1. Google news. http://news.google.com.
2. R. Barzilay, N. Elhadad, and K. McKeown. Inferring strategies for sentence ordering in multidocument news summarization. *Journal of Artifical Intelligence Research (JAIR)*, 17:35–55, 2002.
3. W. Frakes and R. Baeza-Yates. *Information Retrieval: Data Structures and Algorithms*. Prentice-Hall, 1992.
4. V. Hatzivassiloglou, L. Gravano, and A. Maganti. An investigation of linguistic features and clustering algorithms for topical document clustering. In *Proceedings of the 23rd ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 224–231, Athens, Greece, 2000.
5. W. Hill, L. Stead, M. Rosenstein, and G. Furnas. Recommending and evaluating choices in a virtual community of use. In *Proceedings of ACM Conference on Human Factors in Computing Systems(CHI'95)*, 1995.
6. T. Malone, K. Grant, F. Turbak, S. Brobst, and M. Cohen. Intelligent information-sharing systems. *Communications of the ACM*, 30:390–402, 1987.
7. C.D. Manning and H. Schutze. *Foundations of Statistical Natural Language Processing*. MIT Press. Cambridge, MA, 2003.
8. K. McKeown, R. Barzilay, D. Evans, V. Hatzivassiloglou, J. Klavans, A. Nenkova, C. Sable, B. Schiffman, and S. Sigelman. Tracking and summarizing news on a daily basis with columbia's newsblaster. In *Proceedings of HLT 2002 Human Language Technology Conference*, San Diego, California, USA, 2002.