Language and Culture

Sunmoo Yoon (Columbia University)
Misagh Falahi (UCLA)
Ehsan Ebrahimzadeh Houlasou (UCLA)
Kathleen Brennan (University of Hawai'i, Manoa)
Kristoffer Nielbo (Aarhus University)
Peter Leonard (Yale University)

Introduction

Language is dually privileged as a site of cultural analysis. It has served as a proxy for human cultural production for hundreds of years before computation, and it emerged as the first and still most easily-parsed vector for computation in the 1940s as typified by the work of Roberto Busa.[1] In the current computational moment, it may be useful to think of text as the first *model* of human culture in the sense of an imperfect abstraction: the limited dimensionality and stability of written language enabled cultural output to be quantified and processed using the technologies of the day.

In the decades since World War II, advances in algorithms and computation have pushed the boundaries of text analysis, as well as enabled analysis of non-textual material. The present moment is characterized by the increasing sophistication of text mining methods, and the increasing interest in data mining visual, auditory, three-dimensional, and motion-media cultural artifacts.

In this white paper we describe some of the advances in language analysis presented during the Culture Analytics Long Program with an emphasis on Workshop IV: Mathematical Analysis of Expressive Forms: Text Data. We identify three trends that represent significant advances in computational text analysis: 1) positioning **domain expertise** as the driver of text analysis; 2) recognizing that **nuanced layers of meaning** and information are tied to text external concepts; and 3) tracking **cultural dynamics over time and space**. We then use the current state-of-the-art in text analysis as a yardstick for evaluating the progress of emergent non-textual forms of culture analytics, including auditory, visual and moving-image.

Theoretical Components and Emblematic Work

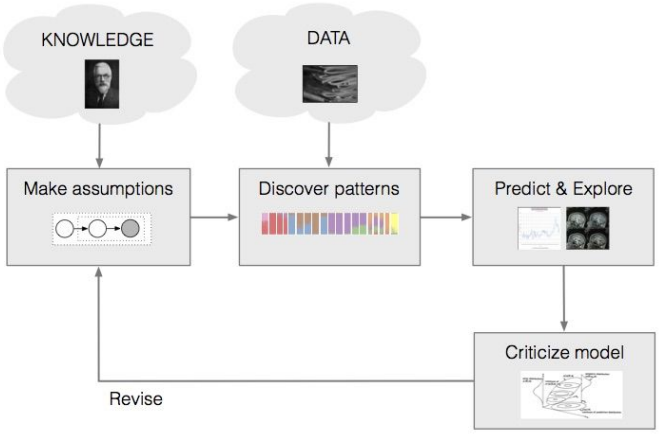
The work of Jesuit scholar Roberto Busa in the 1940s is commonly thought to be the originary moment of computational text analysis for the humanities. Busa's work to create an electronic corpus of Thomas Aquinas demonstrated the enormous potential in applying

[1] Busa, Roberto A. 2004. Foreword: Perspectives on the digital humanities. *A companion to digital humanities*: 187-188.

computational methods to language at scale. Grounded in the medieval monastic tradition of hand-composed biblical concordances, this early effort at electronic text processing also reinforced the centrality of *written* language as the privileged abstract model of human culture.

Concordances (whether on vellum or punchcard) are, by their nature, exponents of the "bag-of-words" model of language processing.[2] As computational power increased throughout the era of mainframe, mini- and micro-computer, and as decentralization led to democratization of access and eventual personal ownership, more sophisticated approaches emerged. From the early bag of words approaches of the 1960s and 70s to LSA[3] in the 1980s and the development of probabilistic methods[4] such as LDA[5] in the 1990s, one can trace an arc of increasing sophistication in method, an exponential increase in text data, and awareness of the importance of context and domain expertise. More recent techniques for word embedding have foregrounded the importance of context and allowed for more meaningful intervention by domain experts.

During Workshop IV, David Blei, one of the creators of LDA, discussed his recent work[6] with historians to demonstrate the research pipeline he uses when developing and applying text mining to large corpora:



[7]

[2] Harris, Zellig S. "Distributional structure." *Word* 10.2-3 (1954): 146-162.
[3] Dumais, Susan T. "Latent semantic analysis." *Annual review of information science and technology* 38.1 (2004): 188-230.
[4] Blei, David M, and Pedro J Moreno. "Topic segmentation with an aspect hidden Markov model." *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval* 1 Sep. 2001: 343-348.
[5] Blei, David M, Andrew Y Ng, and Michael I Jordan. "Latent dirichlet allocation." *the Journal of machine Learning research* 3 (2003): 993-1022.
[6] Blei, David M. "Build, compute, critique, repeat: Data analysis with latent variable models." *Annual Review of Statistics and Its Application* 1 (2014): 203-232.
[7] Blei, David M. "Understanding History Through Topic Modeling." Presentation at Workshop IV: Mathematical Analysis of Cultural Expressive Forms: Text Data during the Culture Analytics Long Program at the Institute for Pure and Applied Mathematics, UCLA, CA, May 23-27, 2016.

In this model, Blei breaks the process into stages, which both clarifies it and delineates steps in which one might collaborate with scholars across different disciplines. Blei hopes that the creation of this research pipeline will allow scholars to examine questions from many fields by iterating through the process quickly.

There were also a number of critical presentations on Natural Language Processing (NLP) and sentiment analysis. At the heart of the automated analysis of cultural expressive forms is the development of mathematical models with sufficient levels of sophistication for understanding human language. At a broader level, these of models for the human mind at a functional level recognize language as an encoding of concepts in the human mind.

Early attempts in the 1950s tackled the problem of NLP through an axiomatic approach, defining a large set of rules governing the structure of the language representations. The advent of statistical learning techniques changed perspectives on NLP dramatically which, in turn, made it possible to take advantage of the large amounts of machine actionable data generated by humans, particularly in recent years.

Common to many NLP applications, whether sentiment analysis, topic modeling, or word embedding, is an interest in tracking and modeling the temporal development and regional variation in texts. How for instance does thematic structure depend on specific historical and external events, and how does the content of social media differ between different cultural and linguistic groups?

Central to all these models is the notion of context. Distributional semantics models treat context as what governs the co-occurrence of the words within the text, in different levels of granularity;[8][9] while framing semantics models aims at capturing the real world situation that the text refers to.[10] [11] [12] Denotational and model-based semantics models are another effort to tie natural language understanding more closely to the real world.[13]

Work in NLP, both as a way to analyze text and as a potential model for understanding the brain, has been combined with new approaches to sentiment analysis. Understanding sentiment is increasingly seen as crucial to a broader comprehension of a text. This extra layer of information about sentiment can lead to extraction of complicating or even contradictory meanings. There have been many projects aiming to quantify the sentiment of a piece of text. Most of the more robust algorithms currently available take a lexicon-based approach to sentiment analysis[14], using unigrams and bigrams to evaluate positivity and negativity of a sentence. There is an immense need for algorithms that use the structure of the sentence to

---

[8] Levy, Goldberg, “Dependency-Based Word Embeddings”, ACL 2014.
[9] Mikolov, Sutskever, Chen, Corrado, Dean, “Distributed representations of words and phrases and their compositionality” NIPS 2013.
[10] Punyakanok, Roth, Yih, “The Importance of Syntactic Parsing and Inference in Semantic Role Labeling”, Computational Linguistics 34 (2), 257-287.
[11] Banarescu, et. al. “Abstract Meaning Representation for Sembanking.” Proceedings of the 7th Linguistic Annotation Workshop & Interoperability with Discourse, pages 178–186, Sofia, Bulgaria, August 8-9, 2013.
[12] Täckström, Ganchev, Das , “Efficient Inference and Structured Learning for Semantic Role Labeling”, Transactions of the Association for Computational Linguistics, vol. 3 (2015), pp. 29-41.
[13] Hermann, et. al. “Teaching Machines to Read and Comprehend”, **arXiv:1506.03340v3**
[14] Liu, Bing. "Sentiment analysis and opinion mining." *Synthesis lectures on human language technologies* 5, no. 1 (2012): 1-167.

evaluate its sentiment and do not stop at independent analysis of words. The sentiment of a sentence becomes complicated very quickly as the size of the sentence grows just over a few words. Accordingly, it would be impossible to fully understand that sentiment without the use of the sentence structure. Moreover, sentiment analysis tools should dive deeper into the context of a sentence and not judge it individually and separately from the body of text from which it is taken, the contexts in which that text is embedded, or the language and/or medium in which it is written. Context is the key to understanding subtleties of a piece of text. Without this contextual understanding, it is difficult if not impossible to determine whether the sentiment attached to a sentence (or utterance) should be interpreted as positive, negative, ambivalent, not to mention whether that expression can also be read, for example, as sarcastic.

During Workshop IV, Matthew Jockers (University of Nebraska, Lincoln) discussed his new Sentiment Analysis tool[15] [16] which attempts to capture the emotional register of novel plots. During the tool's development, Jockers confronted challenges including the fact that existing sentiment dictionaries were built for microblogging rather than narrative fiction, and the need to negotiate the nuances of sentiment at the level of the individual word and in blocks of text. He met these challenges by developing a sentiment dictionary optimized for fiction, and tagging individual words based on the context of 3-sentence windows.

It has broadly made sense to focus on language as written text: permanence, large corpora, machine readability, etc. We are now at a moment when we are becoming increasingly capable, from a technological perspective, to capture and analyze objects from a variety of media. If we define language as communication between individuals and groups then we can see how expressive forms such as film, visual art, and music can also be considered within a broader understanding of language and culture.

Dan Jurafsky's "ketchup theory of innovation"[17] offers a new model for understanding innovation in and between communities. In this model, innovation occurs through processes of borrowing from other countries or disciplines. Jurafsky based this model on the history of ketchup, and demonstrated how the word ketchup captures not only the history of a specific condiment but also the interrelation of cultures across the globe in the eighteenth and nineteenth centuries.

Overlaps with other aspects of CA

The projects, approaches and questions surfaced here resonated with several other areas of the CA Long Program. The problem of **surfacing relevant content** in enormous digital databases which are too large to browse is both an Archival and User Experience challenge (cf White Paper: Cultural Archives; White Paper: Interfaces and User Experience Design for Culture Analytics), as well as a promising domain for machine learning (cf White Paper: Cultural Phenomena at Multiscale and Multiresolution). The domain can be addressed computationally through "recommendation engines" that use both explicit metadata and latent signals to lead researchers to relevant content. In many cases, the user's own intentions can be inferred and

[15] *Syuzhet* [https://github.com/mjockers/syuzhet]
[16] Jockers, Matthew. 2015. Package 'syuzhet'.
[17] Jurafsky, Dan. *The language of food: A linguist reads the menu*. WW Norton & Company, 2014.

these intents can be fed into a recommendation engine as signal. These tools can also make large digital archives more accessible to the general public.

Main Challenges

Baseline Tools: For language to truly function as a model of human culture, all written languages need to have equal access to baseline processing and preparation tools. These include optical character recognition (OCR), tokenization (word separation), part-of-speech tagging, removal of function words[18] and other crucial tasks. Currently, support for written language is concentrated on Western European languages with a primacy for high-status varieties of English. Fundamental problems in tokenization of Chinese-Japanese-Korean, not to mention Thai and Sanskrit, for example, remain to be solved. Within Western European languages, most tools lose efficacy earlier than the nineteenth century, leaving millennia of human language unavailable for analysis. The same problem applies to very modern idioms, such as culturally-specific slang and informal registers that predominate on social media such as Twitter and similar microblogging services.

Distant reading tools such as NLP and sentiment analysis tools still need improvements in capturing a more accurate and consistent set of results. NLP tools have to take into account the context of text and topic-specific forms of language. Sentiment measures also need to be in concordance to the meaning of text and the context it is written in. This introduces the greater challenge of modeling context in written text and quantifiable measures to analyze it.

Data Sources: A relatively small set of data sources (Twitter, Flickr, Wikipedia, Google Books) serve as the training basis and baseline datasets for an enormous percentage of Culture Analytics tools. These datasets have their own representational biases, based on the specific historical conditions of their creation, curation, and distribution. More diverse sources are needed to ensure that Culture Annalytics tools and methods are truly effective, rather than just 'overfitting' to the small fraction of human culture that has been digitized and is easily accessible.

Cultural dynamics: Several branches of mathematics have already developed techniques that offer great potential for the study of cultural change over time. Culture Analytics needs to identify these techniques and adapt them to large-scale text analysis as well as other aspects of the analysis of culture at scale (see Arrowhead Problems).

Further Incorporating Domain Expertise: We must continue furthering the incorporation of domain expertise in all phases of the data-mining research process.[19] In particular, domain expertise should be emphasised in the application phase in which questions about how the

[18] Jurafsky, Daniel, and James H Martin. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.
[19] Fayyad, Usama, Gregory Piatetsky-Shapiro, and Padhraic Smyth. 1996. From data mining to knowledge discovery in databases. *AI magazine* 17, no. 3: 37.

results of a particular analysis will be used come into play.[20] This issue was highlighted during a working group project on public health issues.

Beyond Textual Models: Within the broad area of work on culture and language text analysis is still the primary focus. There are many good reasons for this status quo, but there is much to be learned from thinking of language more broadly to include artifacts in addition to written text. Moving beyond textual models (cf Workshop 1: Beyond Text) necessarily implies moving beyond the agreed-upon representation of text as ascii (later Unicode). In many cases, signal processing approaches must be used to identify constituent elements of visual artworks, or sonic performances. These might serve as the rough equivalents to words, sentences and paragraphs in the textual model. One main challenge associated with shifting to other media is the need to deal with even higher-dimensional data. Analyzing such data will likely require more machine learning and deep learning approaches.[21] [22]

Future Directions

As text mining and related approaches continue to develop and become even more nuanced, there must be equal focus on considering language more broadly. We saw this movement in a number of presentations during Workshop IV. Jurafsky challenged us to rethink our model of innovation through the evolution of ketchup not only by examining traces in historical texts, but also through a consideration of flavor. Tanya Clement posed the question, “What would a Poetry 101 class look like if it was all sound recordings rather than texts on a page?” During the entire Culture Analytics Long Program, we were asked to consider culture beyond text, and that led to a number of fruitful conversations on how we can further the longstanding practice of treating the written word as a privileged model of human culture, while also advancing the analysis of models in sonic, visual, spatial and other dimensions. Much of the work presented at the workshop captured the early work of larger projects. Discussing these early-stage projects in an interdisciplinary environment creates more meaningful opportunities for collaboration.

---

[20] Glorot, Xavier, Antoine Bordes, and Yoshua Bengio. 2011. Domain adaptation for large-scale sentiment classification: A deep learning approach. *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*.

[21] Weigel, Van B. 2002. *Deep Learning for a Digital Age: Technology's Untapped Potential To Enrich Higher Education.*. Jossey-Bass, 989 Market Street, San Francisco, CA 94103-1741.

[22] Ngiam, Jiquan, Aditya Khosla, Mingyu Kim, et al. 2011. Multimodal deep learning. *Proceedings of the 28th international conference on machine learning (ICML-11)*.