A Method for Cross-Document Narrative Alignment of a Two-Hundred-Sixty-Million Word Corpus

Ben Miller
Departments of English
and Communication
Georgia State University
miller@gsu.edu

Jennifer Olive
Department of English
Georgia State University
jolive1@gsu.edu

Shakthidhar Gopavaram
Department of Computer Science
Indiana University
sgopavar@umail.iu.edu

Yanjun Zhao
Department of Computer Science
Troy University
yjzhao@troy.edu

Ayush Shrestha
IEEE Member
ayush.shrestha@gmail.com

Cynthia Berger
Department of Applied Linguistics
Georgia State University
cberger@gsu.edu

***Abstract*—Identifying similar narrative sections across longer documents would help identify key events within a corpus, enrich understanding of those events, provide a mechanism for organizing corpora according to their event content, and allow for bottom-up testing of theories of narrative. This paper proposes an automated method for narrative alignment across large textual corpora using techniques from natural language processing and similarity-based image segmentation. This method proceeds by segmenting each document into a series of events, constructs sequences of abstracted representations of those events, compares pairs of sequences to generate image matrices, segments the images, identifies similar segments to discover commonly occurring narrative units, and, finally, returns the source sentences to make the clusters of narrative similarity readable. Preliminary tests of elements of this method were conducted on a small heterogeneous corpus (< 100 documents) and a moderate heterogeneous corpus ($10k$ documents). Further implementation as described in this position paper is necessary to scale to the full $251k$ document corpus from which the moderate corpus was drawn.**

***Keywords*-Computational models of narrative, text mining, big data, computational linguistics**

I. Introduction

Journalism, historiography, and the first draft of history contained in witness statements and government documents rely upon the careful arrangement and deployment of short vignettes in larger narrative, ideological, and communicative frames. Frequently borrowed, one can find versions of compelling vignettes appearing broadly across a range of documents; for example, highly varying descriptions of the shooting in 2007 of Iraqi civilians by American mercenaries appeared throughout public and political media for a variety of purposes. Automatically finding these moments of narratological similarity from across a heterogeneous corpus would allow researchers to better understand how particular stories are utilized, how larger collective narratives are built through the deployment of frequently occurring vignettes, and how various genres treat similar narrative moments. Our preliminary work describes a method for identifying and clustering moments of narratological similarity from across a large heterogenous text corpus.

Identifying the appearance of these vignettes has been a core research area for language processing even before the ground-breaking, top-down script approach described by Schank and Abelson in [1]. Their approach provided descriptions of commonly occurring patterns of action and then sought out units of text that fit those logical patterns. Later work by Chambers and Jurafsky [2] proposed a bottom-up approach reliant on semantic role labeling for the learning of narrative schemas, or sequences of granular events and their actors that build into constrained, coherent, limited narrative units. The ability to recognize the deployment of a script or schema helps readers and researchers infer the assigned role of the actors in a story, perform searches using features of narrative rather than just strings or semantics, and attach implicit information to the explicit textual descriptions. One limit of "learning" approaches is that they are domain-specific and require that training examples occur frequently within a corpus. Frequency, particularly in

relation to implicit information, is not commensurate with importance. A method that can identify meaningful, significant narrative units with few occurrences, find these examples despite their varying representations, and flexibly extend or clip the narrative units' lengths would better facilitate research into the deployment of narrative vignettes across large humanities corpora.

Building upon our prior work in [6] and [3], the following short paper presents a method for cross-document non-fiction narrative alignment of big humanities data and a discussion of our preliminary experiments.

II. Method

In order to identify corresponding short narrative sequences from across a large corpus of documents that vary in length and are heterogeneous in regard to their authorship, time of composition, and purpose, the development of a multi-step, multi-tool process that could match information despite large semantic and syntactic variation was necessary. Processing our test corpus of humanities big data poses increasingly familiar challenges ranging from overutilization of all available storage and computational resources for an extended period of time to the simple impracticability of a method. As an example, our first iteration of the sequence comparison step required 8 days on a 12-core server to process just $2,000$ documents from the sub-corpus of $12,600$, or 5.75 minutes per document. What follows is a description of the corpus, the lessons learned during this preliminary work, and the revised method currently being implemented.

A. Corpus

The primary corpus used for this research was the $251,287$ non-fiction documents released in 2010 by Wikileaks under the Cablegate name. Consisting of approximately 260 million words, this corpus is a heterogeneous collection of cables sent from United States embassies, consulates, and other diplomatic missions between December 1966 and February 2010. Written by different authors to different purposes under different social and technological conditions, these cables provide a perspective on diplomatic history typically available only when a state fails. A cable was originally a kind of encrypted telegram transmitted over the secured cables that linked diplomatic offices [4]. As of 2008, cables were handled identically to all other kinds of diplomatic communication, rendering the term completely anachronistic [5]. The contents of cables varied from brief updates akin to text messages to full formal reports on stable and emerging diplomatic situations ranging in length from one paragraph and approximately 100 words (numerous examples such as 08MEXICO2891, "Demarche on U.S. Priorities for the Community of Democracies at UNGA Delivered") to 181 paragraphs and more than $18,000$ words (e.g. 10CHISINAU83, "Moldova: Tenth Annual Trafficking in Persons (TIP) Report"). Consisting of descriptions of individuals, regimes, and events, cables are digests of the first draft of history and their study is essential for understanding the relationship between events, global foreign policy, and historiography. Structurally, each cable is a text report preceded by a header. Genre conventions vary, but the reports are generally direct writing of observations and interpretations. The format requires a header indicate the cable's origin, destination, date, reference ID, classification, and subject. The body of each cable is separated into demarcated, sequentially numbered paragraphs. For our preliminary research, we constructed a sub-corpus of $12,600$ cables. The sub-corpus was drawn from the oldest cables and covers the period from December 1966 to August 2003.

B. Event Segmentation and Abstraction

Based upon prior experiments in non-fiction cross-document narrative alignment of a small homogenous corpus as reported in [6], the goal of our method was to align similar narrative units from across a corpus of non-fiction documents. The method begins with a set of documents in a database format each being segmented by events. Running in a JAVA implementation, the event segmentation tool EVITA [7] recognizes noun, adjective, and verbal events that are both punctual and progressive across eight event categories: reporting, perception, aspectual, intensional action, intensional state, state, and occurrence. Events are provided a sequential event ID number on a per-document basis and the keyword for each event is tagged. Our process continues by abstracting the identified keywords via a lookup to WordNet [8] for the keyword's direct hypernym using Lesk word sense disambiguation as implemented in [9]. This Lesk implementation provides a more accurate hypernym selection by utilizing a distributional semantic model of the probability distribution of word senses along with comparing the source word's part of speech and sentential context against those of the candidate sense definitions.

C. Hypernym Sequencing and Similarity Scoring

Following the segmentation, extraction, and abstraction processes, our method builds hypernym sequences

for each document. The sequences are of a length representative of the tripartite structure of narrative as described in [10]. Approximating the three elements of condition, event, and aftermath by looking across the boundaries of three sentences, our method's sequences are three times the average number of events per sentence on a corpus level. For the moderate Cablegate subcorpus, the average number of events per sentence was 3.63, yielding sequences of 11 keywords (11-grams). Each sequence is offset by one keyword from the prior sequence using a sliding window approach as documented in [6]. As an example, a document with 100 events will yield 90 11-grams using a sliding window offset of 1-gram. Single-term offset was implemented so that correspondences can be extended term-by-term and ensure a comprehensive search of the narrative feature space.

D. Image Segmentation and Clustering

After generating abstracted narrative sequences, multiple transformations and representations are necessary to, first, identify the recurring narrative patterns and the extent of those patterns, second, find the places within the corpus where those patterns occur, and, third, collect the occurring patterns from across the corpus and present the results. The method's first principal data structure is a $2D$ matrix that shows the similarity between two documents. Similarity is the measure of shared content between the two sequences on a scale from 0 indicating no correspondence to 11 indicating that all terms were shared between the two sequences. Comparisons are order independent and each term can only contribute once to the similarity score. To facilitate later segmentation and search, this matrix can be considered an image where the $x - axis$ contains the sequences from one document and the $y - axis$ the sequences from another document, and the intersection contains a value derived from the sequence values. The value of each sequence is represented as an array of colors, each representative of one gram in the 11-gram sequence.

The test corpus contained $16,770$ unique terms. Each term is represented as $h_t = \begin{Bmatrix} R_t \\ G_t \\ B_t \end{Bmatrix}$ for $1, 2, ..., n$ where h_t is the gram and $R, G,$ and B represent the color value. This method facilitates the computational and visual recognition of similarity at scale by indicating similar sequences of terms with similar sets of color values. Table I shows an example of two documents being compared, A and B with sentences A_1, A, A_M, and B_1, B, B_N where α_{ij} provides the similarity of sequence A_i and sequence B_j. This representation was chosen because it provides a compressed, SQL-compatible way of storing what are $(2((n)^2)) - n$ initial comparisons where n is the number of documents in the corpus. The 2 multiplier is necessary to provide for transpositions of the matrix' x and y axes so that patterns can be matched irrespective of the document comparison order. For the full corpus under address, 126.3 billion similarity matrices of maximal dimensions equivalent to $n-11+1$, or the number of events in the document segmented into an 11-gram sliding window, are necessary. The largest number of events in a document in the test corpus of $12,600$ documents has $2,779$ events and yielded $2,769$ sequences.

TABLE I
CROSS-DOCUMENT SIMILARITY MATRIX

	A_1	$A_{...}$	A_3
B_1	α_{11}	...	α_{1N}
$B_{...}$	...	...	...
B_M	α_{M1}	...	α_{MN}

Results of the similarity comparison replace the original matrix values in the array with the appended similarity score. The new general structure of the value of the intersection cell is $\alpha_{ij} = \{\{u_1, u_k, u_{11},\}, s\}$, where u holds either the color value of the gram shared between A_i and B_j or $null$ if there are fewer than 11 shared grams. Following population of the matrices with the combined color values and similarity scores, the arithmetic mean and standard deviation of the similarity scores across the corpus are used to threshold within each matrix. Elements below the threshold of significance are discarded, thereby segmenting the image matrix into regions of significant correspondence in a sea of null values. This thresholding process reveals patterns shared by at least two documents.

Immediately proximal positive values in the similarity matrices are grouped into arrays and used as extended search patterns for the clustering of similar narrative units from across the corpus. In this next step, pattern matching of non-null matrix elements serves to cluster the meaningful multi-document matches. Prior to this step, only the comparison of a pair of documents has been performed. This step uses that prior $1:1$ comparison to search through the corpus for the presence of the variable-length narrative vignettes represented by the matrix segments seen to repeat at least once in the corpus. This step proceeds by comparing two matrices,

a source b_{ij} and candidates, d_{pq}, to produce clusters, C_r, where r is the number of all significant matrix segments in the corpus.

Consider d_{pq} to be all similarity matrices after the thresholding step with the exclusion of one, b_{ij}. Every matrix will in turn occupy the b_{ij} position. For each b_{ij} all d_{pq} matrices are compared such that the search pattern and the matching matrix segments from the corpus are grouped into clusters, $C_1, , C_r$ where the domain of r is the number of matrix segments identified in b_{ij}. As more matrices occupy the b_{ij} position, r extends to incorporate those segments. Each element in the cluster array has as many values as the size of the original corresponding segment. The clusters are then compared and identical clusters are concatenated, aggregating the correspondences from one pairwise comparison with the correspondences from the other pairwise comparisons. At the conclusion of this process, what results is a set of clusters containing a visual representation of the abstracted grams representing narrative sequences found across the corpus. Finally, the source sentences corresponding to the meaningful patterns are returned. Sentences with at least one contributing term are considered to be a piece of the matched narrative unit, thereby allowing the algorithm to present narrative units that cross sentential boundaries.

III. Preliminary Discussion

Implemented elements of this method include event segmentation, keyword abstraction, and similarity comparison. Unimplemented elements include generating the color space, thresholding via similarity scores, and clustering of matching narrative segments. What preliminary tests of the implemented elements have shown is twofold. First, that the method can identify narrative units that are shared across more than two documents. And second, the processing of even relatively short documents is impracticable without parallelization, efficient data structures, efficient storage utilization, and reliance on a large number of simple computations. These preliminary experiments were most useful for guiding the development of the method. What further tests will investigate is if the method can identify clusters of narrative similarity from a highly heterogeneous corpus.

A. Evaluation of Event Segmentation and Hypernym Selection

The method strongly relies upon accurate event segmentation and language abstraction. The first is provided by EVITA. Per evaluations in [7], it has an F-measure of 0.801; however, it does generate some ambiguous results. In our test case, although many values only occurred once, (e.g. "zimbabwe-highlighting," "have_a_bun_in_the_oven"), only 57 of the $16,770$ distinct values were noise (e.g. "0lqqol-lowed," "01ca89f2.8020e1d," "-integrate"). Our language abstraction pipeline using DSM Lesk was evaluated against basic Lesk in [9]; basic Lesk returns the correct sense with an F-measure of 0.656 versus the DSM Lesk's score of 0.715. As our method is abstracting millions of event words, these percentages are significant.

B. Evaluation of Creation and Storing of Matrices

Pilot runs of this first draft of the method necessitated a reworking of the storage and format for the similarity matrices. Uncompressed storage of individual tab separated value (TSV) files quickly consumed $1.3TB$ of storage for fewer than $2,000$ comparisons across our test corpus of approximately $12,600$ documents. This occurred because the default block allocation size for Linux is $4K$, meaning that each TSV, no matter how little data it contained, could be no smaller than $4K$ in size. Given that the largest matrix was $2,769x2,760$ and that most were smaller than $500x500$, that file size constraint was an unexpected problem. Subsequent development will store the similarity matrices in SQL format.

C. Evaluation of Image Segmentation and Clustering

In our prior work on narrative alignment in small corpora, high similarity was manually recognized, and the method was evaluated against single-sentence as document LSA. Although LSA is the correct comparison because it effectively generates semantically similar clusters, the units need to be more commensurate. Rather than single sentence as document, the comparison should be against 3-sentence units. That evaluation framework using LSA would miss one of the core strengths of this proposed method: that it can segment for units of three or more sentences. Our method's use of similarity score for thresholding implements the simplest effective method of image segmentation, a necessary step for generating the search patterns. Many other common segmentation techniques such as the watershed method [11] require more complex computation. Thresholding is effective here, possibly because the image matrices are likely to possess a strong bimodal distribution at 0 and the mean.

IV. Conclusion and further work

The major shortcomings of the proposed method lay in its abstraction process, gram color selection, and efficiency. The implemented abstraction process ignores that

each event keyword originates from potentially different levels of the hypernym hierarchy. Consider the frequent event keyword, "reported." The first sense of that word's direct hypernym, "informed," has three additional hypernym levels. While another event keyword, "told," would in its second sense share the same direct hypernym, the more specific verb, "impart," requires two hypernyms to reach the same level of the hierarchy. In short, normalizing abstraction to the same hypernym degree would, we believe, generate better agreement and more accurate abstractions for the similarity scoring. It would also facilitate solutions to the color selection problem by reducing the set of distinct abstracted keywords. A second problem in that area is that the event segmentation process does tag items which are not words and are, instead, unique strings. These unique values do not contribute to the recognition of correspondence; though their uniqueness, if valid, does help distinguish otherwise similar sequences. An improved method may elide grams that only occur once in the overall corpus, thereby leading to higher and more varied similarity scores. The problem of efficiency affects both the computational and storage aspects of the method. To resolve the first, we are currently working on a Hadoop implementation of the segmentation and similarity algorithms. To resolve the second, the method moved from representing each similarity matrix as a TSV file to storing intermediary representations as a database. Ultimately, this proposed method for cross-document narrative alignment assists readers and researchers in understanding of how vignettes are deployed across a corpus, further develops cross-document narrative correlation, could further development of narrative search, facilitates bottom-up testing of narrative theory, and enriches research on narrative schema detection.

REFERENCES

[1] R. C. Schank and R. P. Abelson, *Scripts, plans, goals, and understanding: An inquiry into human knowledge structures*. Psychology Press, 2013.

[2] N. Chambers and D. Jurafsky, "Unsupervised learning of narrative schemas and their participants," in *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*. Association for Computational Linguistics, 2009, pp. 602–610.

[3] B. Miller, A. Shrestha, J. Olive, and S. Gopavaram, "Cross-Document Narrative Frame Alignment," in *6th Workshop on Computational Models of Narrative (CMN 2015)*, ser. OpenAccess Series in Informatics (OASIcs), M. A. Finlayson, B. Miller, A. Lieto, and R. Ronfard, Eds., vol. 45. Dagstuhl, Germany: Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2015, pp. 124–132. [Online]. Available: http://drops.dagstuhl.de/opus/volltexte/2015/5286

[4] B. Palmer, "Wikileaks has released thousands of confidential diplomatic cables. what's a cable, and why are we still using them?" *Slate*, Jan 2010.

[5] B. Bain, "State department will get smart," *FCW: The Business of Federal Technology*, 2008.

[6] B. Miller, J. Olive, S. Gopavaram, and A. Shrestha, "Cross-document non-fiction narrative alignment," *ACL-IJCNLP 2015*, p. 56, 2015.

[7] R. Saurí, R. Knippen, M. Verhagen, and J. Pustejovsky, "Evita: a robust event recognizer for qa systems," in *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2005, pp. 700–707.

[8] G. A. Miller, "Wordnet: a lexical database for english," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[9] P. Basile, A. Caputo, and G. Semeraro, "An enhanced lesk word sense disambiguation algorithm through a distributional semantic model," in *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. Dublin, Ireland: Dublin City University and Association for Computational Linguistics, August 2014, pp. 1591–1600. [Online]. Available: http://www.aclweb.org/anthology/C14-1151

[10] M. Bal, *Narratology: Introduction to the theory of narrative*. University of Toronto Press, 1997.

[11] F. Meyer, "Color image segmentation," in *Image Processing and its Applications, 1992., International Conference on*. IET, 1992, pp. 303–306.