Catalyst: Piloting Capabilities for more Transparent Text Analytics

Emergent Research Forum Paper

Fouad Zablith
American University of Beirut
fouad.zablith@aub.edu.lb

Bijan Azad
American University of Beirut
bijan.azad@aub.edu.lb

Ibrahim H. Osman
American University of Beirut
ibrahim.osman@aub.edu.lb

Abstract

The surge and value of unstructured text is attracting substantial research and industry attention. Subsequently we are witnessing novel techniques and algorithms that are performing increasingly sophisticated text mining tasks. However the majority of such techniques are opaque, making it hard to trace the provenance of the analytical task on hand. We propose Catalyst, a framework to automatically transform, enrich and expose text into a linked graph-based layer to enable more transparent processing and access to the text elements. In brief, Catalyst extracts text dependencies, performs sentiment analysis, detects semantic relatedness, and links the text elements into a semantic triple-store that enables an easy access to the text entities through direct query functionalities. We plan to evaluate the performance of Catalyst by processing a dataset of user reviews around the dimensions of an evaluation model deployed in the context of e-government services.

Keywords

Text analytics, user feedback, decision support, performance management, linked data, semantic web.

Introduction

Mining insights from text is increasingly gaining research momentum, despite the opacity of underlying processing algorithms and the lack of reasoning behind the specific outcomes. These challenges appear to limit the text mining utility by what we have come to call lack of analytics and reasoning "provenance" (i.e. the "what" and the "how" of the analysis process) (Varga and Varga 2016). To cope with today's surge of textual "big data" cradle, various algorithms and tools are being developed to support several text mining tasks. While such tools can be highly efficient, the underlying analytics process usually requires an initial training, and is often performed by a black-boxed algorithm. This approach can hinder the ability of making more informed data-driven decisions. Indeed, there are scholarly calls for more process "openness and visualization" (Abramson and Dohan 2015) in data analysis, and in better understanding and explicating its provenance.

In our research, we aim to answer the following research question: *how can we enable a more transparent analysis of text sources?* More specifically, *how can we automatically capture, store, enrich and consume the relations among the text elements involved in performing more open analysis?*

In order to expose provenance in text analysis, we adopt the "follow-your-nose" (W3C 2011) and "information seeking mantra" (Shneiderman 1996) approaches as a way to make text annotations and linkages more visible and explicable. We introduce in this work-in-progress paper, Catalyst, a framework and a tool that automatically converts text into a linked graph-based layer. Catalyst captures text dependencies, augments the text elements with sentiment levels, enables the computation of semantic relatedness to specified entities, and finally stores all the relations into a linked data repository. Such a repository is a key component to manipulate the extraction of the relevant text entities in a highly customizable manner, while displaying the level of details at various granularity levels. Effectively we are offering a system that fills some of the gaps in the existing text analysis and coding solutions (e.g., NVIVO

(Bazeley and Jackson 2013)). Our system provides a flexible and extensible toolset to automatically expose and enrich linkages among text elements. We plan to evaluate Catalyst on a large dataset composed of feedback reviews from users of e-government services.

Our research contributes to existing approaches in several ways. First, our proposed framework reuses existing open knowledge-bases as background knowledge and subsequently does not require a "training data-set." Second, we are taking steps to address the scholarly call for provenance-based analytics, by developing tools that keep traces and linkages visible back to the text sources. Third, we adapt the proven UI/UX heuristics of "follow-your-nose" and "information seeking mantra" to design and evaluate means of achieving greater provenance in text analytics.

Background and Related Work

Researchers are investing a lot of efforts aiming to generate insights that reside in text documents and transform them into "actionable knowledge" (Gopal et al. 2011). Sebastiani (2002) highlights that the interest in machine learning approaches has increased in popularity to replace some of the tedious tasks performed by experts in the area of knowledge engineering. In this context, various techniques have been employed to process text documents. For example through machine learning, Support Vector Machines (SVM) were used to process customer reviews and product opinions (Cheung et al. 2003), and to categorize text based on a set of features (Joachims 1998). While such techniques are proving useful in various contexts, they often require training, and tend to use a black-boxed processing phase when generating the analysis results. In other words, the "provenance" (Varga and Varga 2016) is usually hard to trace and capture. In our framework, we aim to keep the processing of data as explicit as possible. Instead of using statistical and association-based techniques and training, we rely on existing knowledge bases as background knowledge to infer sentiment states and semantic relatedness. We aim to keep all the data layers created on top of the text accessible for future reference using linked data.

Other groups of researchers have been working on transforming text relations into linked data (Augenstein et al. 2012; Gangemi et al. 2016). In this context the tasks are mainly about enriching the text with contextual information through, for example, Named Entity Recognition (NER), word sense disambiguation, relation extraction between entities, and others. Compared to our work, instead of making sense of entities' types in text and their relations, we further focus on enriching the linkages in the text with sentiment polarity and semantic relatedness.

Approach Overview

To achieve our research objective, we propose a framework to automatically transform text elements into a linked query-able graph, enriched with sentiment and semantic annotations.

Text Preprocessing. The *Text Preprocessing* module of the framework is the first step through which text documents are loaded. The objective of this module is to identify the dependency relations in the text sources and their part-of-speech (POS). We rely on the Stanford Natural Language Processing (NLP) tools (De Marneffe and Manning 2008) for the text preprocessing tasks. Stanford NLP provides software packages and algorithms for processing and manipulating text elements. Text documents are initially loaded into the *Dependency Relations Extractor* component, in which text dependencies are identified. Furthermore, POS are identified in the text through the *POS Extractor* module, which relies on the POS tagger features of the Stanford NLP.

Text Refinement. After extracting the text dependencies and part-of-speech, the text elements are passed to the *Text Refinement* module. The role of this module is to improve the aggregation and linkages among text elements. To achieve this, the *Lemmatizer* component, which is based on one of the Stanford NLP packages, is integrated in the framework to link together the different inflected modes of words. Having a unified view of terms will help in the analysis of the subsequent phases. Another component of the text refinement module is the *Attribute Extractor* that extracts the attributes of specific adjectives mentioned in the text sources. For example, if participants mentioned in feedbacks — how *fast* or *slow* a service is —

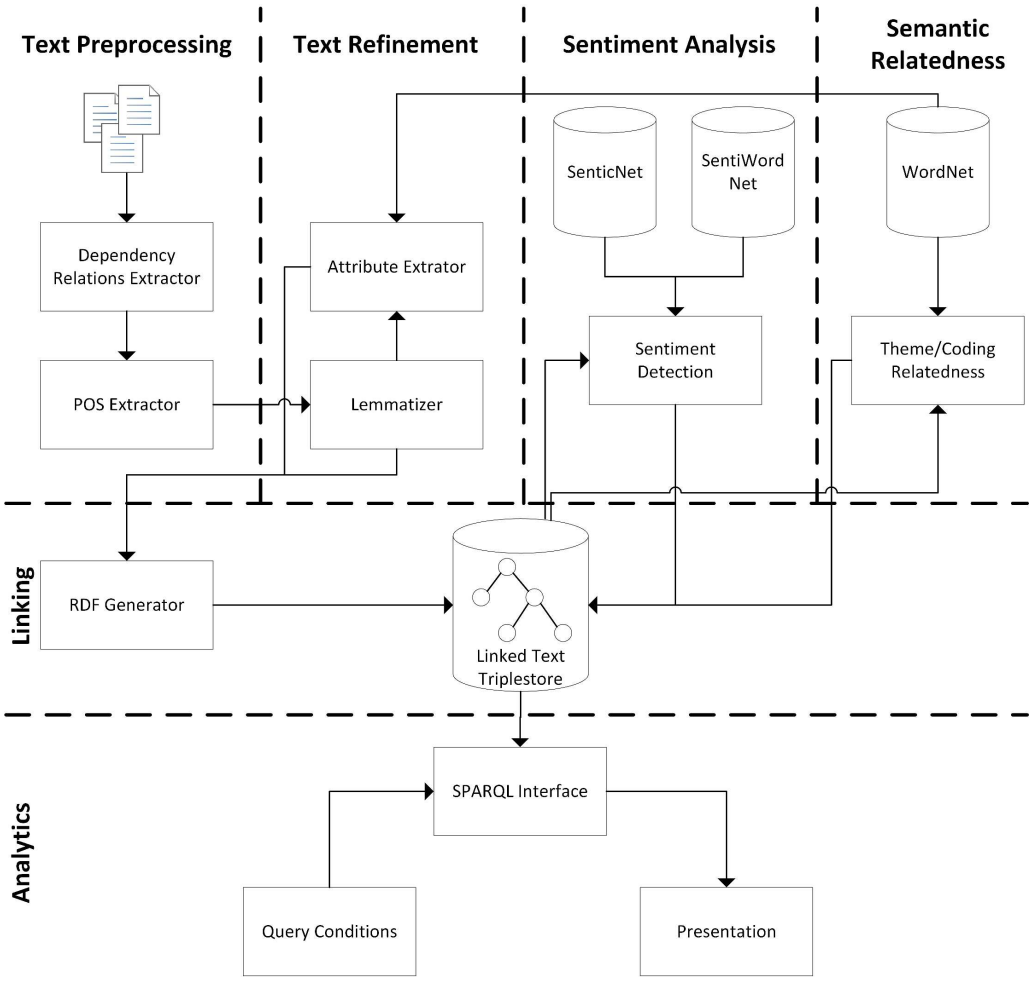

Figure 1. The Catalyst Framework

the corresponding attribute in this case would be *speed* for example. Attributes can serve as an indirect inference (or co-reference (Ding and Liu 2010)) of features mentioned in opinions and feedback.

Linking. Once the text fragments are processed, the objective of the *linking* module is to transform the text elements into a process-able and query-enabled layer. Our initial investigations highlight the transformation process of text into linked entities (Zablith and Osman 2015). The Resource Description Framework (RDF) generator module takes the processed text elements as input, and generates RDF output that is pushed to a triple-store. The Stanford Typed Dependencies are transformed and used as relations between the text segments and terms.

Sentiment Analysis. After transforming the text input into a linked graph, the objective of the *Sentiment Analysis* module is to enrich the text graph with polarity indicators. The module enables the enrichment of a customizable set of POS types. For example it is possible to limit the enrichment of only nouns or adjectives in the graph. This part relies on existing sentiment sources including SenticNet (Cambria et al. 2012) and SentiWordNet (Baccianella et al. 2010) to infer the polarity of terms in the text, and the polarity measures are linked back to the RDF graph of the text. One advantage of this representation is the ability to combine the influence of dependencies on the polarity of entities at the analysis level. For example a negation modifier in a sentence like, the service is "not" good, would invert the positive polarity of "good." Such manipulation can be done and inferred at the query level. The POS information extracted in the Text Preprocessing phase can also be used to match the right POS term to the one in the sentiment dictionaries.

Semantic Relatedness. Once the text graph is enriched with the sentiment measures, the role of the *Semantic Relatedness* module is to anchor the terms in the text to specific analytical dimensions. This module relies on WordNet (Fellbaum 1998), and the various existing relatedness measures available in the literature such as the Wu and Palmer similarity (Wu and Palmer 1994), among others. The relatedness measures are also added to the text graph in the triple store, to make it part of the subsequent analysis.

Analytics. After the transformation of text into a linked graph with accessible dependencies relations, sentiment indicators, and semantic relatedness, it is now possible to directly access and query this new layer using Simple Protocol and RDF Query Language (SPARQL) (Prud'Hommeaux and Seaborne 2008). Once the data is extracted, the output can be manipulated for further analytics reporting or visualization.

Evaluation Plans: Processing e-Government User Reviews

We are currently developing a web interface that implements the Catalyst framework. We have "collected, cleaned and glossed" a subset of 1,493 statements of user reviews about e-government services. In this context, we will use the Cost, Benefit and Risk, Opportunity Analysis (COBRA) model in (Osman et al. 2014) to analyze the text. One main difference between these variables is that the Cost and Benefit measures are tangible (e.g. time and money), while the Risk and Opportunity measures are intangible. The constructs of the model are as follows: Cost and Risk negatively impact user satisfaction, while Benefit and Opportunity positively impact it. We will evaluate the performance of the framework in terms of precision, recall and F-Measure, by comparing its automatic processing of the statements, to the classification performed by three coders working independently (this will enable us to detect the level of agreements), as positive, negative, and whether it points to something measurable or not. For example a user stated in her review that "*I am not satisfied with how fast my application was processed*." The coder is supposed to detect that this statement is classified as "*cost*," given that there is a measurable entity pointing to how "*fast"* the application went through, with a negative sentiment. When processed through Catalyst, the statement can be transformed into a graph (part of the graph is shown in Figure 2). In this case, the text was processed to see how related are the terms inside the statement to the concept "*measure*" in WordNet. The reasoning behind this is that quantifiable and tangible entities (e.g., time and money) would be semantically closer to "*measure*" as compared to other intangible entities such as design or communication opportunities.

It would be possible to extract all statements that are classified for example as *cost* through a query that imposes the following condition: "*segment.polarity < zero AND segment.measure > threshold"*, with further reasoning behind the classification output (e.g., to identify the text elements that led for the statement to be negative).

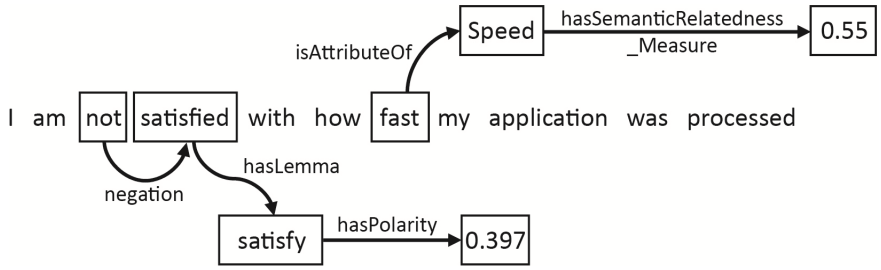

Figure 2. Example of a Statement Processed by Catalyst to Enhance Provenance Detection

Conclusion and Future Work

Today there is an increased need for more provenance enhanced text analytics. We discussed in this paper the feasibility of employing the "follow-your-nose" and "information seeking" mantra heuristics with the objective of making text analysis more transparent. We discussed Catalyst, our proposed framework to automatically transform, enrich and represent text into a linked graph-based layer that can be accessed and traversed through a semantic triple-store. We believe that preserving linkages among text entities, enriched with sentiment indicators and semantic relatedness would get us closer to having a more

traceable and enhanced provenance text analysis. Our framework can be extended in the future to include further features such as the ones described by Augenstein (2012), and reuse existing defined vocabularies (e.g., (Hellmann et al. 2013)) to improve the exchange of data with other systems. In our future steps we are planning to evaluate the framework by analyzing a dataset of users' feedback using a model deployed in the context of the evaluation of e-government services.

Acknowledgements

This work was supported by the University Research Board (URB) of the American University of Beirut.

References

Abramson, C. M., and Dohan, D. 2015. "Beyond Text Using Arrays to Represent and Analyze Ethnographic Data," *Sociological methodology* (45:1), pp. 272–319.

Augenstein, I., Padó, S., and Rudolph, S. 2012. "LODifier: Generating Linked Data from Unstructured Text," in The Semantic Web: Research and Applications, LNCS, Springer, pp. 210–224.

Baccianella, S., Esuli, A., and Sebastiani, F. 2010. "SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining," in LREC (Vol. 10), pp. 2200–2204.

Bazeley, P., and Jackson, K. 2013. Qualitative data analysis with NVivo, Sage Publications Limited.

Cambria, E., Grassi, M., Hussain, A., and Havasi, C. 2012. "Sentic Computing for social media marketing," *Multimedia Tools and Applications* (59:2), pp. 557–577 (doi: 10.1007/s11042-011-0815-0).

Cheung, K.-W., Kwok, J. T., Law, M. H., and Tsui, K.-C. 2003. "Mining customer product ratings for personalized marketing," Decision Support SystemsWeb Data Mining (35:2), pp. 231–243.

De Marneffe, M.-C., and Manning, C. D. 2008. "The Stanford typed dependencies representation," in Coling 2008: Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation, Association for Computational Linguistics, pp. 1–8.

Ding, X., and Liu, B. 2010. "Resolving object and attribute coreference in opinion mining," in Proceedings of the 23rd International Conference on Computational Linguistics, pp. 268–276.

Fellbaum, C. 1998. *Wordnet: An Electronic Lexical Database*, MIT Press.

Gangemi, A., Presutti, V., Reforgiato Recupero, D., Nuzzolese, A. G., Draicchio, F., and Mongiovì, M. 2016. "Semantic web machine reading with FRED," *Semantic Web* (Preprint), pp. 1–21.

Gopal, R., Marsden, J. R., and Vanthienen, J. 2011. "Information mining — Reflections on recent advancements and the road ahead in data, text, and media mining," Decision Support Systems (51:4), pp. 727–731.

Hellmann, S., Lehmann, J., Auer, S., and Brümmer, M. 2013. "Integrating NLP using linked data," in International Semantic Web Conference, Springer, pp. 98–113.

Joachims, T. 1998. "Text categorization with support vector machines: Learning with many relevant features," in European Conference on Machine Learning (ECML), Springer, pp. 137–142.

Osman, I. H., Anouze, A. L., Irani, Z., Al-Ayoubi, B., Lee, H., Balcı, A., Medeni, T. D., and Weerakkody, V. 2014. "COBRA framework to evaluate e-government services: A citizen-centric perspective," Government Information Quarterly (31:2), pp. 243–256.

Prud'Hommeaux, E., and Seaborne, A. 2008. "SPARQL query language for RDF," *W3C recommendation* (15).

Sebastiani, F. 2002. "Machine learning in automated text categorization," *ACM computing surveys (CSUR)* (34:1), pp. 1–47.

Shneiderman, B. 1996. "The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations," in Proc. of the IEEE Symposium on Visual Languages, USA, p. 336.

Varga, M., and Varga, C. 2016. "Visual Analytics: Data, Analytical and Reasoning Provenance," in Building Trust in Information, Springer, pp. 141–150.

W3C. 2011. "Linking patterns - Semantic Web Standards," (available at https://www.w3.org/2001/sw/wiki/Linking_patterns; retrieved February 28, 2017).

Wu, Z., and Palmer, M. 1994. "Verb semantics and lexical selection," in *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, Las Cruces, New Mexico: Association for Computational Linguistics, pp. 133–138.

Zablith, F., and Osman, I. H. 2015. "Linking Stanford Typed Dependencies to Support Text Analytics," in Proceedings of the 24th International Conference on World Wide Web Companion, International World Wide Web Conferences Steering Committee, pp. 679–684.