Graphical View of Blog Content Using B2G

Sabine Bergler and **Jahnavi Dhananjaya**
ClaC Labs, Concordia University, Montreal
bergler@cse.concordia.ca, j_dhanan@encs.concordia.ca

Abstract

We present the simple idea that a graphical representation of subject-verb-object triples is useful for exploring blog texts, a preliminary implementation and a first level analysis of the positive and negative aspects of a naive implementation. We outline its potential for further development.

Introduction

The progress in natural language processing tools available has made text mining an active field of research even outside the natural language processing (NLP) community. With seamless integration in environments, application oriented tools developed by experts that use the NLP tools as black box components is on the rise. One such field that looks at textual data for new insights is epidemiology.

Social media in general, and blogs in particular, hold great promise for epidemiologists and public health oriented researchers. The expectation of plentiful data, uncensored and (hopefully) representative, is met: there are more blogs than can be surveyed manually. Automatic text mining systems are, by definition, limiting the insight into the context of the blogosphere and are thus not helpful in the initial phases of refining a research question or for the task of expanding and adapting protocols of information extraction.

Early stage epidemiological research requires to survey blogs as a new data source and to identify the types of information that could be expected to occur quickly and without great set-up costs.

We propose B2G, a lightweight blog text extraction and processing tool composed largely of stock open source software that is easily reconfigurable by the researcher. In particular, the researcher can input lists of words of interest and B2G will provide a graphic visualization of the associated subject-verb-object (SVO) triples in RDF format. While SVO triples are understood to be an insufficient representation of the content, the overload issue for a graphical user interface requires a severely abstracted, high-level view of the dataset with possibilities to inspect certain parts of the resulting graph in more detail.

Copyright © 2015, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

The main idea behind our baseline approach here is to give researchers a very general tool, not adapted or fine-tuned to any task and thus easy to deploy. We will outline some of the many shortcomings that can be addressed, speculating on the likely usefulness of the tool before and after different adaptation steps. In fact, the lightweight graphical and highly underspecified nature of the representation draws attention to some unexpected data the same way an unfortunate Google search query may not yield the desired results, but leads nevertheless to an interesting (and often quite lengthy) investigation of an unforeseen but serendipidous nature.

Background

Sophisticated text mining and text analysis systems exist, HealthMap, for instance, was used for analyzing the 2009 H1N1 virus outbreak (Brownstein et al. 2010). HealthMap is an aggregator that excels in collecting data from around the globe, from Google, Twitter, and many other resources, and displays maps that show geographic locations of reported cases, etc.

A nice overview over currently available tools for text mining on Medline is presented in (Jensen, Saric, and Bork 2006), which also critically compares different systems and the limited possibilities this comparison is based on.

At the other end of the spectrum, Stanford's Relation Extractor (Surdeanu et al. 2011) is a statistically trained (and thus retrainable) relationship extraction module, that can be adapted if an appropriate training corpus is available, that has exactly the relations of interest labelled as desired. In first-stage exploratory research, this is not usually so. But the explorer in the first stage of his project is more accepting of crude tools and is willing to carefully review the returns if, akin to panning for gold, valuable nuggets can be found this way. In fact, just by giving an unusual, non-standard view can shift attention to otherwise overlooked facts, an insight that has been operationalized in Bananaslug, The Long Tail Search Engine[1] that gives Google-like search with the added twist of allowing the searcher to select from a category of predefined random words (we chose Phonetic Alphabet, it selected 'x-ray') which are added to the search terms ('text mining' in our case). Thus instead of the same ranking from

[1] http://bananaslug.com

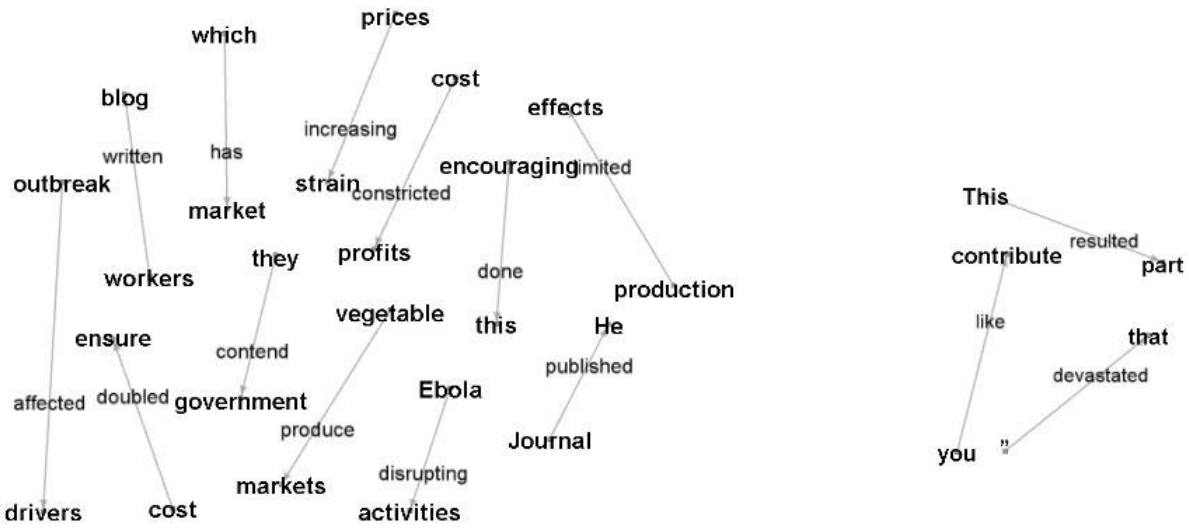

Figure 1: B2G representation of blog post

search to search, the added noise term will bring other material to the forefront (in our case, we discovered the Orange Data Mining environment (Demšar et al. 2013), by all accounts a much more evolved tool in the spirit of our overly simple B2G.) But highly evolved tools are not anathema to our point here, which is in fact that NLP tools can be insightfully applied even when their results are not tuned and far from perfect. NLP modules should be used more widely as an exploratory tool.

B2G

B2G (from blog to graph) has been put together with a goal of easy configuration and upward compatibility with a range of open source modules. We use the Gate environment (Cunningham 2002) to extract Blog text using the boilerplate removal library Boilerplate (Kohlschütter, Fankhauser, and Nejdl 2010), run the Annie (Bontcheva et al. 2002) text preprocessing suite provided through Gate (sentence splitting, tokenization) and the Stanford parser and dependency module (Klein and Manning 2003; de Marneffe, MacCartney, and Manning 2006). We export the Gate annotations of interest to RDF format outside Gate with an in-house mapping function. Another in-house system allows to identify a subset of the exported annotation that are then selected for viewing with Gephi (Bastian, Heymann, and Jacomy 2009), a open source graphical visualization tool.

The dependency relations extracted by the Stanford parser indicate grammatical relations as triples. Thus there is an `nsubj` relation between the subject noun phrase and the main verb. Similarly, objects of a main verb have a variety of dependency relations, depending on the grammatical realization of the object, for instance `dobj` for direct objects. However, a triple relates only the so called head features, a single word that acts here as a representative: for noun phrases this is usually the last noun. Thus for *On Monday, so very predictably, the little brown fox chased the big mouse.* we have `nsubj(chased, fox)` and `dobj(chased, mouse)`. While a most crude representation of the sentence, B2G combines the two relations and shows a graphical representation equivalent to `fox -- chased --> mouse`.

Figure 1 shows the resulting output for a blog on the effects of Ebola. We would like to address the many shortcomings by pointing out three in particular.

Head nouns are sometimes poor substitutes for the entire noun phrase The triple `resulted(This , part)` is uninformative, the full object noun phrase would have given a much better representation of the sentence *This has resulted in a large part of the harvest being left to rot or sold at throw-away prices.* In graph representations, too much information makes the graph unintelligible. We opt here to allow the user to click on a relation and display the extended triple. Figure2 shows the information added by displaying a compromise of so-called minNPs (including determiners and prenominal modifiers, but not prepositional phrases), demonstrating the tradeoff between overview and informativeness.

No accounting of other textual components leads to false impressions The triple `like(you , contribute)` suggests a factual statement, but the underlying text is *Contact us if you would like to contribute to the blog on anything from opinion pieces to project reflections, event reports or sharing a good lesson plan*. The omission of negation, conditionals or, as in this case, modality skews the result. The representation of complex linguistic structure quickly becomes hard to represent graphically. B2G will be developed to include such information.

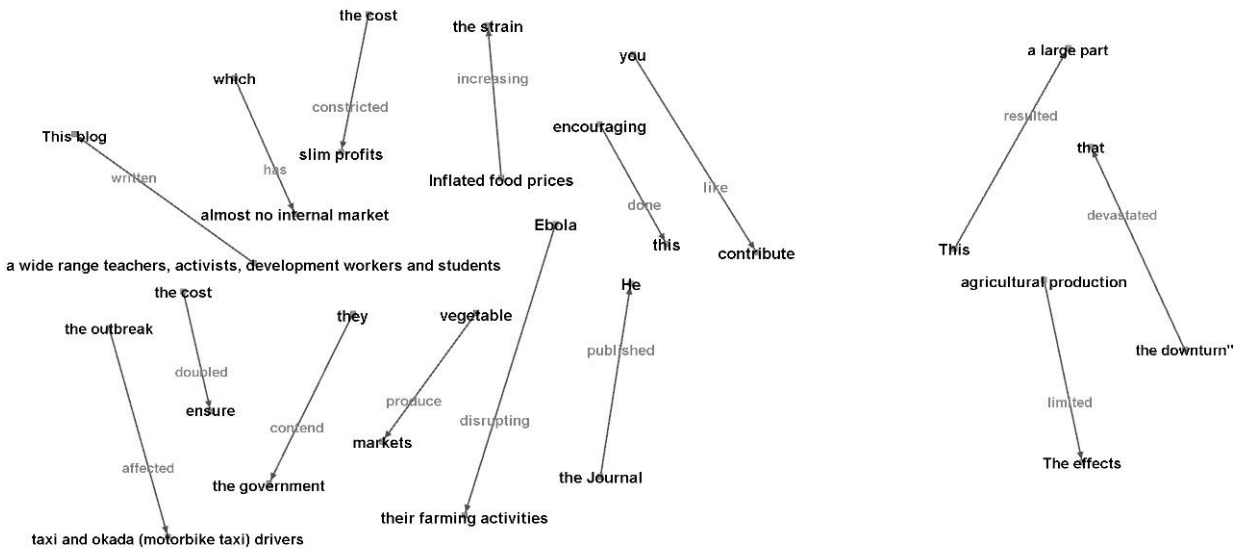

Figure 2: B2G graph displaying minNPs

Lack of honing the tool leads to spurious preprocessing errors The triple `devastated(that , ")` results from the closing quotation mark being mysteriously misinterpreted as a noun and presented as the head of the object NP of *devastated* in *Yvonne Aki-Sawyerr, a prominent Sierra Leonean businesswoman, believes that this can best be done by encouraging "maximised use of local supplies when DFID or UN award contracts, to support small and medium enterprises that have been devastated by the downturn"*. It is clear, however, that no matter how much the preprocessing is improved, there will be errors of this kind, but they are infrequent and don't invalidate the value of the other triples.

User fine tuning

The user has many options when using B2G: the tool can display triples for single sentences, single blogs, or collections of blogs. The user has the possibility to explore blogs by identifying relations of interest. In Figure1, an extensive list of causal relations was used. The system displays the SVO triples only for blog sentences that contain a verb from this list. The result can further be reduced to triples that, in addition, contain a subject or object from another list of terms of interest (possibly diseases or geographical names, not applied here). The guiding principle behind B2G is not one of providing the user with the top of the line text mining environment, but rather to provide an environment with hooks to attach and try out ideas that are still developing in an ad hoc fashion displayed by a baseline user interface. We believe this enables creative exploration by involving the researchers actively in the creation of their tools (over which they thus have almost full control). This is the true contribution of B2G and we claim it is one well adapted to the natural workflow of epidemiologists. The lists have to be supplied by the user and act as a filter on the displayed triples. The user interface makes it simple to link in new word lists and to compare results on given text encouraging the quick testing of ideas. A selected set of word lists which we assume to be of general interest are supplied to jump-start the process.

Conclusion and Future Work

B2G has been demonstrated to epidemiologists and generated good user engagement and feedback. We thus conclude that our initial hypothesis has been partially validated, namely that even a very simple pipeline of NLP tools without further grooming can yield interesting abstractions of blog texts for epidemiologists exploring blogs in an initial, exploratory research context.

Upon repeated use, the same users will likely object to some of the more obvious shortcomings and we intend to add more sophisticated modules little by little, starting with negation and modality, followed by identifying hedging environments, since we have experience in all of these (Rosenberg, Kilicoglu, and Bergler 2012; Rosenberg and Bergler 2012; Kilicoglu and Bergler 2010). Additional extentions to improve preprocessing quoted material and indicate reported speech (Krestel, Bergler, and Witte 2012), as well as certain preprocessing useful for life science articles as used for (Bergler et al. 2007). Each of these enhancements is to be added as a separate, independent module into the pipeline to be assessed by the users for its effectiveness to enhance the exploration experience. For each degree to which the output is tuned towards more complicated representation, the intuitive appeal may be lost. As one triple for another blog summarized so succinctly: `complicated(it, that)`.

Acknowledgements

This work has been supported by funding from NSERC, the Natural Sciences and Engineering Research Council of Canada and the help from Marc-Andre Faucher, Jonathan Villemaire-Krajden, and Canberk Ozdemir.

References

Bastian, M.; Heymann, S.; and Jacomy, M. 2009. Gephi: An open source software for exploring and manipulating networks. In *3rd International ICWSM Conference*.

Bergler, S.; Schuman, J.; Dubuc, J.; and Lebedev, A. 2007. BioKI, a general literature navigation system at TREC Genomics 2006. In *Proceedings of The Fifteenth Text REtrieval Conference (TREC 2006)*.

Bontcheva, K.; Cunningham, H.; Maynard, D.; Tablan, V.; and Saggion, H. 2002. Developing reusable and robust language processing components for information systems using GATE. In *13th International Workshop on Database and Expert Systems Applications*.

Brownstein, J. C. F.; Chan, E.; Keller, M.; Sonricker, A.; Mekaru, S.; and Buckeridge, D. 2010. Information technology and global surveillance of cases of 2009 H1N1 influenza. *New England Journal of Medecine* 362.

Cunningham, H. 2002. GATE, a general architecture for text engineering. *Computers and the Humanities* 36:223–254. http://gate.ac.uk.

de Marneffe, M.; MacCartney, B.; and Manning, C. 2006. Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC-06*.

Demšar, J.; Curk, T.; Erjavec, A.; Č. Gorup; Hočevar, T.; Milutinovič, M.; Možina, M.; Polajnar, M.; Toplak, M.; Starič, A.; Štajdohar, M.; Umek, L.; Žagar, L.; Žbontar, J.; Žitnik, M.; and Zupan, B. 2013. Orange: Data mining toolbox in Python. *The Journal of Machine Learning Research* 14(1).

Jensen, L.; Saric, J.; and Bork, P. 2006. Literature mining for the biologist: from information retrieval to biological discovery. *Nature Reviews Genetics* 7(2).

Kilicoglu, H., and Bergler, S. 2010. A high-precision approach to detecting hedges and their scopes. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning (CoNLL-2010)*.

Klein, D., and Manning, C. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting of the Association of Computational Linguistics*.

Kohlschütter, C.; Fankhauser, P.; and Nejdl, W. 2010. Boilerplate detection using shallow text features. In *3rd ACM International Conference on Web Search and Data Mining WSDM*.

Krestel, R.; Bergler, S.; and Witte, R. 2012. Modeling human newspaper readers: The fuzzy believer approach. *Natural Language Engineering* 20(2).

Rosenberg, S., and Bergler, S. 2012. Uconcordia: CLaC negation focus detection at *Sem 2012. In *Proceedings of *SEM 2012: The First Joint Conference on Lexical and Computational Semantics*.

Rosenberg, S.; Kilicoglu, H.; and Bergler, S. 2012. CLaC Labs: Processing modality and negation. In *Working Notes for QA4MRE Pilot Task at CLEF 2012*.

Surdeanu, M.; McClosky, D.; Smith, M.; Gusev, A.; and Manning, C. 2011. Customizing an information extraction system to a new domain. In *Proceedings of the ACL 2011 Workshop on Relational Models of Semantics*.