A HUMAN READABLE TOPIC CLUSTERING – An Applied Case Study on Natural Language Condensation

Paul Greiner, M.A.
Friedrich-Alexander-Universität Erlangen-Nürnberg
Department Germanistik und Komparatistik
Professur für Korpuslinguistik

Concept sketch · last modification: spring 2018

> If you know the enemy and know yourself, you need not fear the result of a hundred battles. If you know yourself but not the enemy, for every victory gained you will also suffer a defeat. If you know neither the enemy nor yourself, you will succumb in every battle.
>
> – Sun Tzu, *On the Art of War*, about 500 B.C.

Introduction

Knowledge has always translated to power and it always will. And knowing about the ideas and opinions of other people has always been one of the most important aspects to be considered by decision makers all through human history. Be it thoughts about potential uprising in the heads of peasants against an aristocratic ruler, the attitude of voters towards the promises of a currently campaigning politician or the reputation of brand names and products with potential customers on today's global scale markets. The opinions of people have always mattered and will continue to be an indispensable resource for everyone trying to establish or increase a lead over political or economic competition.

For a long time, the problem of gathering information has been the decisive factor in having a head start in terms of knowledge. This is due to the fact that actually having to ask people for opinions, recording answers and interpreting results in a concise and structured way is an incredibly costly, if not even impossible endeavour to be accomplished by manual labor in an analog world. But this situation has changed drastically with the emergence of digitally interconnected devices for personal use and the accompanying, all-embracing spread of the internet. These

developments give market and opinion researchers unprecedented possibilities alongside unforeseen challenges.

This work will explore some of those challenges and present at least one applicable solution in abstract, well founded definition as well as actual, comprehensively documented implementation, complete with extensive evaluation. Insights will be applicable for scientific and economic purposes alike.

Motivation

Today, enormous amounts of digitally stored data are available online and people willingly share their thoughts on social networks as well as via blog and microblogging services or other means specifically set up to encourage the explicit statement and exchange of opinions, e.g. in the form of product reviews. What is more, architecture and spread of the internet make it easy for market researchers to deliberately gather attitudes and opinions on a specific theme and its individual aspects in the form of carefully composed, well-directedly deployed online polls.

As established before, the information available is, profanities aside, of huge interest for an extensive number of enterprises. The decisions and resulting acts of such entities might have strong influence on economics as well as politics on a potentially global scale and therefore possibly affect the everyday life of virtually everybody. Consequently, analyses need to be swift, precise and – for mainly ethical reasons – objective. Specific approaches aside, results need to be extensive as well as comprehensible and – again for mainly ethical reasons – reproducible.

The hurdle to overcome at this point is, of course, that the most promising material does not come in a well-structured, easy-to-process and numerically analyzable format. Instead, quite the opposite is the case. The material at hand comes in the probably most frequently used, arguably least specified and potentially most ambiguous form of code: Natural language.

Goal

The goal of this work is the definition and implementation of a framework denoted *Natural Language Condensation* (NLC), the purpose of which is the automated analysis of text collections derived from the sphere of market research online polls. This analysis will include sentiment detection, semantic clustering and the generation, respectively discovery, of easily interpretable labels for the constituents provided by preceding steps. Since extensive work has been done in the fields of sentiment detection and semantic clustering, the work at hand will focus on the development of label generation while only secondarily dealing with the other topics for contextual reasons.

The framework created in this manner will be subjected to conscientious fine-tuning of parameters and be exposed to extensive testing on real world data. Results will be thoroughly reviewed and evaluated against competing systems, thus proving the effectiveness of the system at hand.

In conclusion, this work will provide proposals for actual application in research as well as in industry and deliver thoughts on further progression of the subject.

Outline

Introduction: Compare motivational and introductory texts above. Subchapters will be (i) *Motivation*, (ii) *Goal*, (iii) *Problem Setting* and (iv) *Outline* or alternatively something like (iv-a) *Composition of this Document*. A concrete example on actual data for the illustration of NLC will be delivered in (iii), thus providing easy access to the uninformed reader.

Literature Review and Related Work: Give an extensive overview about existing literature and projects. Subchapters will be (i) *Historic Development*, (ii) *Overview about the Current Situation* and (iii) *Recent Related Work*. Motivate and justify which existing material will be utilized and which will be omitted. For development purposes tag entries for 'foundational' vs. 'inspirational'.

Methodology: Describe the architecture and functioning of the project at hand on theoretical as well as practical levels. Subchapters will be (i) *Theoretical Outline*, (ii) *Existing Foundations* and (iii) *Implementation* or rather (iii-a) *Documentation of the Project at Hand*. Discuss modules for sentiment detection and clustering in (ii), while (iii) deals with label generation, i.e. the core of this work.

Evaluation: Have a close look at the results, compare to existing projects and material and validate. Expect this to be a rather complicated endeavour for the lack of established conventions. Subjectivity in evaluation (not in creation) of results will be a problem. Make sure *Literature Review* points out these difficulties at an early stage. The proposal of applicable solutions will be an essential contribution.

Conclusion: Summarize methodology, results and evaluation in a concise way. Re-establish motivation and goal from introductory sections. Give examples for fields of application and describe potential future developments.

Introductory Reading

A very brief listing of the most fundamental literature. Items are roughly organized by the principles postulated in the paragraph about literature review above, with (f) denoting fundamental and (i) inspirational entries. With a number of extensive surveys, overview could be considered about complete.

Historic Development: Luhn (1958) (f/i); Edmundson (1969) (f)

Overview about the Current Situation: Das and Martins (2007) (f); Smith et al. (2017) (f); Gambhir and Gupta (2017) (f)

Recent Related Work: Mei, Shen, and Zhai (2007) (i); Lau et al. (2010) (i); Lau et al. (2011) (i); Greiner et al. (2013) (f); Evert et al. (2016) (f)

Bibliography:

Das, Dipanjan, and André F. T. Martins. 2007. "A Survey on Automatic Text Summarization." *Literature Survey for the Language and Statistics II Course at CMU* 4: 192–195.

Edmundson, H. P. 1969. "New Methods in Automatic Extracting." *J. ACM* 16 (2): 264–285.

Evert, Stefan, Paul Greiner, João Filipe Baigger, and Lang Bastian. 2016. "A Distributional Approach to Open Questions in Market Research." *Computers in Industry* 78: 16–28.

Gambhir, Mahak, and Vishal Gupta. 2017. "Recent Automatic Text Summarization Techniques: a Survey." *Artificial Intelligence Review* 47 (1): 1–66.

Greiner, Paul, Thomas Proisl, Stefan Evert, and Besim Kabashi. 2013. "KLUE-CORE: A Regression Model of Semantic Textual Similarity." In *Proceedings of the Second Joint Conference on Lexical and Computational Semantics (*SEM 2013)*, edited by Mona T. Diab, Timothy Baldwin, and Marco Baroni, 181–186. Atlanta, GA: Association for Computational Linguistics.

Lau, Jey Han, Karl Grieser, David Newman, and Timothy Baldwin. 2011. "Automatic Labelling of Topic Models." In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, 1536–1545. HLT '11. Stroudsburg, PA, USA: Association for Computational Linguistics.

Lau, Jey Han, David Newman, Sarvnaz Karimi, and Timothy Baldwin. 2010. "Best Topic Word Selection for Topic Labelling." In *Coling 2010: Posters*, 605–613. Beijing, China: Coling 2010 Organizing Committee.

Luhn, H. P. 1958. "The Automatic Creation of Literature Abstracts." *IBM J. Res. Dev.* 2 (2): 159–165.

Mei, Qiaozhu, Xuehua Shen, and ChengXiang Zhai. 2007. "Automatic Labeling of Multinomial Topic Models." In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 490–499. KDD '07. New York, NY, USA: ACM.

Smith, Alison, Tak Yeon Lee, Forough Poursabzi-Sangdeh, Jordan Boyd-Graber, Kevin Seppi, Niklas Elmqvist, and Leah Findlater. 2017. "Evaluating Visual Representations for Topic Understanding and Their Effects on Manually Generated Labels." *Transactions of the Association for Computational Linguistics* 5: 1–15.