Le Pham Minh Duc

Un-polarizing news in social media platform

Master’s thesis of mathematical information technology

March 15, 2019

Author:Le Pham Minh Duc (Lê Phạm Minh Đức)

Contact information: [miduleph@student.jyu.fi](mailto:miduleph@student.jyu.fi), minhduc1993@yahoo.com

Supervisors: Oleksiy Khriyenko (oleksiy.o.khriyenko@jyu.fi)

Title: Un-polarizing news in social media platform

Työn nimi:

Project: Master’s thesis

Study line: Web Intelligence and Service Engineering

Page count: x+y (x = page count without appendices; y = page count of appendices)

Abstract: Abstract…

Keywords: Keywords…

Suomenkielinen tiivistelmä: Abstract in Finnish…

Avainsanat: Keywords in Finnish…

Preface

This is where you can write a preface for your thesis. The preface is optional so you don’t have to write it if you don’t want. Also, the glossary on the next page is optional. List of figures and tables are required if there is at least one figure or table in the thesis.

**The most important guideline** for the use of this template: look from the title list which style has been used to which section of the document and use similar styles also in the future when modifying the document. Use always styles, never do formatting outside styles.

If you notice bugs or mistakes in the template, please inform them to Antti Ekonoja ([antti.j.ekonoja@jyu.fi](mailto:antti.j.ekonoja@jyu.fi)).

Jyväskylä, March 15, 2019

Le Pham Minh Duc (Lê Phạm Minh Đức)

Glossary

NLP Natural language processing

DCOM Distributed Component Object Model  
More explanation…

C++ Shouldn’t need any explanation…

List of Figures

Figure 1. Service architecture **Error! Bookmark not defined.**

Figure 2. Test figure 25

List of Tables

Table 1. Example of the table 26

Contents

[1 INTRODUCTION 1](#_Toc3060154)

[1.1 Problems overview 1](#_Toc3060155)

[1.2 Proposed solution and research questions 1](#_Toc3060156)

[1.3 Research method 3](#_Toc3060157)

[1.4 Thesis structure 3](#_Toc3060158)

[2 HYPOTHESIS AND TECHNOLOGIES 4](#_Toc3060159)

[2.1 Original hypothesis 4](#_Toc3060160)

[2.2 Evaluation criteria 4](#_Toc3060161)

[2.3 Required technologies 5](#_Toc3060162)

[2.3.1 Natural language processing and its sub-domains 5](#_Toc3060163)

[2.3.2 Open Information Extractor / Semantic triple 6](#_Toc3060164)

[2.4 Technologies used in this work 6](#_Toc3060165)

[2.4.1 Stanford CoreNLP 6](#_Toc3060166)

[2.4.2 Node.js 8](#_Toc3060167)

[2.4.3 Version control system, Git and GitHub 9](#_Toc3060168)

[3 UN-POLARIZING ALGORITHM 11](#_Toc3060169)

[3.1 Overall solution/algorithm architecture 11](#_Toc3060170)

[3.1.1 Article annotation pipeline 11](#_Toc3060171)

[3.1.2 Article matching pipeline 13](#_Toc3060172)

[3.2 The first algorithm using Sentiment Analysis and Name Entity Recognition 13](#_Toc3060173)

[3.2.1 Web content processor 13](#_Toc3060174)

[3.2.2 Stanford Core NLP Annotator 14](#_Toc3060175)

[3.2.3 Sentiment analysis 15](#_Toc3060176)

[3.2.4 Named entity recognizer 16](#_Toc3060177)

[3.2.5 Articles similarity calculation using only Named Entity Recognition 19](#_Toc3060178)

[3.3 Un-polarizing algorithm using Semantic triples 20](#_Toc3060179)

[3.3.1 Open information extraction 20](#_Toc3060180)

[3.3.2 Triplet pairs, common entities and common statement counts 23](#_Toc3060181)

[3.3.3 Providing the information to the user 24](#_Toc3060182)

[3.3.4 Limitation of the current system. 26](#_Toc3060183)

[4 RESULTS EVALUATION 27](#_Toc3060184)

[5 FUTURE WORKS 28](#_Toc3060185)

[5.1 Ontologies based entity relevance 28](#_Toc3060186)

[5.2 Word-net verb contradiction 28](#_Toc3060187)

[5.3 Cloud service design 29](#_Toc3060188)

[5.4 User interface and user experience design 29](#_Toc3060189)

[6 CONCLUSION 30](#_Toc3060190)

[Bibliography 31](#_Toc3060191)

[Appendices 32](#_Toc3060192)

[A Title of the first appendix 32](#_Toc3060193)

[B Title of the second appendix 32](#_Toc3060194)

# INTRODUCTION

## Problems overview

(*Should write like a lot more, with more references as well*) Ideological polarization has been a problem in our society for quite a long time. (that reference from 1986). With the rise of social media (citation here or not), it’s estimated that 66% of US citizen use social media as one source of news (citation from that web page), the amplification of ideological polarization has been increasing much faster than previously due to social endorsement, and other social media techniques that is used to keeps its user engaged (Sporh. 2017). This creates the echo-chamber effects that, by the design of social networks that only show what the users want to see, make the user even furthermore polarized into his own belief and makes him see the world wrongly, which may turn the user into some extremists that might be harmful for the society.

Scholars have been researching about this problem and solutions are proposed (many citations needed) but these solutions are either too impractical (e.g: needing the giant media companies to change their entire business models) or just way out of reach of the scholar scope (e.g: needing of the government’s intervention on the issue or people to stop using the services).

https://en.wikipedia.org/wiki/Echo\_chamber\_(media)

*References:*

*Boczkowski, P., Mitchelstein, E., & Matassi, M. (2017, January). Incidental news: How young people consume news on social media. In Proceedings of the 50th Hawaii international conference on system sciences.*

[*https://scholarspace.manoa.hawaii.edu/handle/10125/41371*](https://scholarspace.manoa.hawaii.edu/handle/10125/41371)

*Beam, M. A., Hutchens, M. J., & Hmielowski, J. D. (2016). Clicking vs. sharing: The relationship between online news behaviors and political knowledge. Computers in Human Behavior, 59, 215-220.*

*https://www.sciencedirect.com/science/article/pii/S0747563216300656*

## Proposed solution and research questions

The main goal of this thesis is to find the way to break the people’s echo chamber that is mostly caused by the effect of social medias only show the user what he/she wants to see. To combat this, we need to show the user the news from the other side of view. If he/she reads about the opening of a new coal mines help creating a few hundreds of new jobs for the area, he should also know that the new coal mines will cause a great damage to the environment and might cause some local wild-life to disappear.

On top of that, the service must be accessible and easy to use, as the reason of many people using social media as their main source of news as it’s so convenience to have one place to go to and can see both your friend’s status as well as news.

With that goal in mind, the main research question of the thesis is:

* **How to find articles with alternative (different) points of view to a given article?**

We will only attempt to find the news that is relevant to the article but also provide oversight from different point of view that the first article misses. We will not check if the news is credible (but we will try to only provide news from credible sources) or if it is true, we simply provide the user different articles from many points of views about the relevance topic so that he/she can choose to interpret it whatever way he/she wants to.

With the first question answered, we will address two additional support questions on deploying the news un-polarizing service for the mass to use:

* **What is the most convenience way to deliver the service for the user?**

If the service is too complicated to use, or requires too many unnecessary steps, the user will rarely use the service, if at all, which reduces the effectiveness of the system. We need a method that can deliver the alternative points of view to the user that is most convenience for him, for example: a fully automatic system that whenever the user reads a news about a topic, he also has a snippet information of other relevance articles about the topic.

* **How to engineer the service so that it is autonomous, up-to-date and scalable?**

As a news service, it must always catch up with the latest news to be relevant to use. The service needs to read and analyze articles to from various news source all the time so it can serve the user with the latest and most relevance news. Also, as a cloud service, we need to prepare ahead of the service, so that when there are more users, the service will be able to handle that.

## Research method

(*Some good part from the book*) Abc, test text.

## Thesis structure

(*I should write this part last, because there is still more things to change*) – Including this introduction and the problem overview, this thesis will contain five parts. The next part contains our hypothesis based on theoretical research, as well as the state of the current technology and the tools we choose to go forward with the practical prototype.

The third and biggest part, named Un-polarizing algorithm will describe our step to step practical implementation, the problems we faced along the way and the reasons for our implementation decision.

The next part will test the result of our prototype using real world articles with cross human check/validation to see the effectiveness of the solution.

Finally, conclusions for the thesis as well as possible future work and extension are given in the last chapter.

# HYPOTHESIS AND TECHNOLOGIES

## Our hypothesis

Our hypothesis for finding articles with different point of view is simple:

* Given two articles talking about a similar topic, if one article has a positive view on the situation and the other has a negative view regarding the same situation, they are considered to have different point of views.

An example different point of view is:

* Topic: The US’s war in Iraq.
* First article main point: The US’s war in Iraq is good and justified because Saddam Hussein is a dictator and the people living under his reign are suffering.
* Other article main point: The US’s war in Iraq is bad because it furthermore destabilizes the region and the main intention of waging war was because of oil, not for humanitarian purpose.

We think that: if someone read an article, it would be interesting and beneficial for him to also see other articles about the same topic but from a different point of view, as having multiple view angles on a subject make the reader more informed about a problem and less likely to be affected by propaganda as well as reducing the effect of echo-chamber of social media platform, which is the news source of many people nowadays.

However, terms like “similar subject” or “positive/negative views” are quite abstract and, as there is no universally defined rule for finding these characteristics in different articles, we need to define our own rule for finding “Article similarity” and “Positive/negative views”. This leads to our secondary and third hypothesis:

* Two articles are considered to have similar topic if they both contains a good number of similar named entities. For example, given three articles: A, B and C. Article B will be considered “more similar” to A than C to A if the number of similar named entities between B and A is bigger than the number between C and A, and vice versa. (more on chapter **3.2.4** and **3.2.5**).
* An article is considered to have a positive or negative view on a subject can be determined by the sentiment value of such article. (more on chapter **3.2.3**)

However, after the implementation of our first prototype based on these hypotheses, we were not happy with the results (more on chapter **2.3.2** and **2.3.5**). With insights from the previous prototype, we came up with the second set of hypotheses that utilize new technologies and more complex algorithms that we hoped to answer our research question better:

* If two articles state different facts, they are providing contradiction information and the reader should know about both of them.
* If there is no contradiction information between the comparing article and our knowledge corpus, we should suggest the most similar articles to our user. “Article similarity” now is calculated by the number of similar named entity in the two articles, as well as the context the named entity appears in (more on chapter **3.3.1** and **3.3.2**).

Due to the limitation of current technologies (more on chapter **3.3.4**) and our time constraint, we were not able to fully develop and evaluate the second set of hypotheses as much as the first sets. Still, the later prototype provided a lot of interesting information for our evaluation and we believed that we proceeding to the correct direction with these hypotheses and with more resource poured into this work, it could contribute a huge role for the society.

## Evaluation criteria

As we discussed in the previous chapter, term like “subject similarity” or “different point of views” are abstract term, so, there is no concrete way to evaluate these characters. We could not find any statistic or equation to evaluate the results of our algorithm. We could only judge the output by using our common sense and opinion’s survey.

To test the rigidity of our algorithm, we gathered a dataset of 78 articles (+ 8 non-readable by the web content parser), consist of three main themes:

* Muslim in Europe: 24 articles
* Muslim in Asia: 39 articles
* Asians in Europe: 17 articles

We then went through each of them and decided which set of articles are more relevant to each other and which pair of articles that contain opinion from different point of view. Once we finished annotating the database, we do the evaluation by comparing our annotation to the results from the algorithm.

With articles spanning in three different main categories that are also related to each other, for each them, there will be some positive hits (related articles) as well as false negatives: news/documents that share similar set of entity and keyword but convey different fields and are not related at all (for example: sports and politics). With these “traps”, we want to test if our algorithm can truly return the relevant information and how close the suggestion is to our annotation.

Finally, since our algorithm can read through the whole article thoroughly, it might discover interesting information that we missed as we only skimmed quickly through our database. Skimming is a realistic behavior though, since most people only read the title or consume through each news source quickly (Gabielkov, Ramachandran, Chaintreau & Legout, 2016).

*Ref:*

*Maksym Gabielkov, Arthi Ramachandran, Augustin Chaintreau, Arnaud Legout. Social Clicks: What and Who Gets Read on Twitter?. ACM SIGMETRICS / IFIP Performance 2016, Jun 2016, Antibes Juan-les-Pins, France. 2016*

## Required technologies

### Natural language processing

(*Consider rename the title*) (*Lots of references in this, this is like a wildly researched fields*) Named entity recognizer and sentiment analysis are two of the many sub-tasks covered by Natural language processing (NLP) technology, *which is an area of computer science and artificial intelligence concerned with the interactions between computers and human (natural) languages, how to program computers to process and analyze large amounts of natural language data. (Source from wiki- more information needed).*

For just named entity recognizer only and sentiment analysis only, there exists multiple tools that do not require NLP. For example, a simple named entity recognizer (*citation needed)* consists of a dictionary of many individual nouns and a lookup function to match the result from the document and the dictionary. Similarly, a basic sentiment analysis can be made by assigning a sentiment value to each of the adjectives in the sentence/paragraph and then calculate the sum of these values as the sentiment value for the sentence/paragraph. However, these methods are quite barebone and quite usually not really correct since their rules are quite flawed and there are many ways for them to misinterpret the true sentiment of the sentence. Sophisticated NLP tools with “sentiment analysis” supported have a more complex ways of defining the sentiment value

Named entity recognizer and Sentiment analysis can be easily created without NLP using a rule-based system. However, these basic tools are not good and sophisticated NLP tools are much better because they have a more complex system that take many other things into consideration.

### Open Information Extractor / Semantic triple

(*What is semantic triple, why it is important and a critical part of this thesis.*) – After our first implementation of the algorithm, we realized that just Sentiment analysis and NER was not enough. To make the result more interesting and more useful to the user, we have to use Open Information Extraction, which, unfortunately, was not as widely developed as the two previous features since they are much more complicated.

*https://nlp.stanford.edu/software/openie.html*

*https://nlp.stanford.edu/pubs/2015angeli-openie.pdf*

[*https://getd.libs.uga.edu/pdfs/hooge\_david\_c\_200705\_ms.pdf*](https://getd.libs.uga.edu/pdfs/hooge_david_c_200705_ms.pdf)

[*https://en.wikipedia.org/wiki/Semantic\_triple*](https://en.wikipedia.org/wiki/Semantic_triple)

*https://github.com/dair-iitd/OpenIE-standalone*

## Technologies used in this work

### Stanford CoreNLP

Developed by the researchers at Stanford University from 2006, released as a free and open source software in 2010, with updates still being developed and released nowadays (Manning, Surdeanu, Bauer, Finkel, Bethard, & McClosky, 2014), Stanford CoreNLP is a Java (or JVM based) annotation pipeline framework for most of the common Natural Language Processing (NLP) steps like Named Entity Recognition (NER) (Finkel, Grenager & Manning. 2005), Sentiment Analysis (Socher, Perelygin, Wu, Chuang, Manning, Ng & Potts, 2013) and Open Information Extraction (OpenIE) (Angeli, Premkumar & Manning, 2015). We used Stanford CoreNLP to process raw web-based article text into annotated data and properties, ready for our “un-polarizing” algorithm. Detailed information on the role and usage of Stanford CoreNLP in our work will be presented in later chapters (chapter 3.1 and chapter 3.2.2) where we go in depth with our solution.

We chose Stanford CoreNLP as the foundation technology for our thesis because of two main reasons:

* It has all the services we needed integrated into one big package that will work well together. There are many tools that provide the necessary services (especially NER and sentiment analysis) for us, but each of them has different requirement for the input data as well as different output format. Using separated tools instead of just one require us to put time and effort into making them work together instead of focus on the main research question, which is the “un-polarizing” algorithm. We could argue that using a specialized tool for each of the task might provide better quality output, but our testing results does not show any significant different in the results outputted by these tools compared to Stanford NLP anyway (more on chapter 3.2.3, chapter 3.2.4 chapter 3.3.1).
* It is free and open-source, with full access to source code that can be installed and run locally. Having every cog in our machine (or solution) fully available is important, as the private and close-source service are subjected to changes or shut down at any moment, which, is problematic. Having our algorithm run well and not depending on services we do not control is important not only us, now, but also for when other researchers want to try or test or improve our solution, now, for 10 years from now.

*Ref:*

*Manning, Christopher D., Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pp. 55-60.*

*Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005), pp. 363-370*

*Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher Manning, Andrew Ng and Christopher Potts. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. Conference on Empirical Methods in Natural Language Processing (EMNLP 2013)*

*Gabor Angeli, Melvin Johnson Premkumar, and Christopher D. Manning. Leveraging Linguistic Structure For Open Domain Information Extraction. In Proceedings of the Association of Computational Linguistics (ACL), 2015.*

### Node.js

Even though most of the works done in this report are prototype code to demonstrate and test our hypothesis, we want to continue working on our “Un-polarizing algorithm” after this thesis work is completed. We our final goal is to produce a product for people all around the world to use and thus, help creating a better society. With that in mind, we want to choose a programming language that is capable producing quality and stable code base for longevity, performant and highly scalable, but also flexible enough for changes in our prototype development.

Node.js comes to mind as the perfect candidate for our requirements as its multi-paradigm nature and its giant ecosystem of libraries (Tilkov & Vinoski, 2010) allows quickly creation, testing and modification of our prototype with little overhead cost. Several benchmarks also prove the superior performance of a Nodejs web system when compare to other popular technologies like PHP and Python (Lei, Ma & Tan, 2014), which shows the potential of node.js for longevity and development of industrial application.

*Ref:*

*S. Tilkov and S. Vinoski, "Node.js: Using JavaScript to Build High-Performance Network Programs," in IEEE Internet Computing, vol. 14, no. 6, pp. 80-83, Nov.-Dec. 2010.*

*http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5617064&isnumber=5617049*

*K. Lei, Y. Ma and Z. Tan, "Performance Comparison and Evaluation of Web Development Technologies in PHP, Python, and Node.js," 2014 IEEE 17th International Conference on Computational Science and Engineering, Chengdu, 2014, pp. 661-668.*

*http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7023652&isnumber=7023510*

### Version control system, Git and GitHub

A version control system (VCS) is “a tool that tracks different versions of software or other content” (Loeliger, Jon, & McCullough, 2012). Using VCS is considered as one of software development best practices, even just for a personal project (Spinellis, 2005). As we are creating a software prototype to evaluate our hypothesis and algorithm, it is best to follow these principles and to use a VCS for our project. These principles are later on, proved to be quite helpful as, throughout the course of our prototype development, we found ourselves utilizing many features of VCS such as source-code backup, code synchronization between different computers, progression roll-back and, finally, through the commit messages: a diary/documentation system.

“Git” is a free “Decentralized version control system” that has a clean internal design, performs quickly and efficiently, enforces accountability (Loeliger, Jon, & McCullough, 2012), and is the VCS we chose to use for this thesis. “Git” was created in 2005 by Linus to help developing the Linux kernel as other VCSs system at that time had limitations and flaws that would make them not a viable solution. These reasons make “Git” not only a good solution to applied to our works, but also make it one of the mostly used VCS nowadays in both public and private sectors.

GitHub was chosen as our hosting service for the project as it was one of the biggest Git supporting services (hence, the name) and is free. All our software prototype, coding history, instructions and documentations are kept on GitHub and are freely available to view, access and execute at any moment from any computer anywhere in the world. An “url” to the project are provided at the end of this report.

*Ref:*

*D. Spinellis, "Version control systems," in IEEE Software, vol. 22, no. 5, pp. 108-109, Sept.-Oct. 2005,*

[*http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=1504674&isnumber=32260*](http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=1504674&isnumber=32260)

*Loeliger, Jon, and Matthew McCullough. Version Control with Git: Powerful tools and techniques for collaborative software development. " O'Reilly Media, Inc.", 2012.*

# UN-POLARIZING ALGORITHM

## Overall solution architecture overview

To answer the main research question of: **“How to find articles with alternative (different) points of view to a given article?”**, we developed a prototype called the “Un-polarizing algorithm”, containing two main parts:

* Article annotation pipeline: to process the natural language text from the article to machine readable format and save them to a local database for comparing later.
* Article matching pipeline: compare a given article to all the annotated articles in our database and find the most appropriate articles with relevance information with different point of views.

### Article annotation pipeline

Here is the overall architecture of the Article annotation pipeline.

Local

database

Web

content

processor

Stanford

Core

NLP

Sentiment

Value

Article

url

Semantic

Triples

Named

Entities

Sentiment analysis

NER

filtering

Semantic

triples

Filtering

Annotated

article

Article annotation pipeline

Figure 1. Test figure

Boxes and lines with different colors represent different parts of process:

* Cyan **box and line**: open information, freely available on the internet
* Red **box and line**: Stanford Core NLP service.
* Green **box and line**: our own custom logic, implemented for this works
* Blue **box**: the article annotation process.
* Black **box**: the database to store the annotated articles
* Dashed box and line: process and algorithm implemented in the second phase prototype, things that we didn’t plan to implement initially.

To annotate an article, we read its content using its url, and then pre-process the website’s data so that it can be parsed for the NLP task (more on 3.2.1).

Then, we push the news text into the Stanford Core NLP (more on part 3.2.2). From the information received from the Core NLP engine, we use 3 sets of annotation: the sentiment analysis (more on part 3.2.3), the named entity recognition (more on part 3.2.4) and the semantic triples (more on part 3.3.1). Each dataset then will be furthermore analyzed/filtered using various equation we will described in later part of this thesis to make it ready for the articles matching solution. Then, the results will be recombined into one annotated object and saved into our local database.

Note that the Semantic triples and other processes related to it are only utilized in the second prototype, where we weren’t satisfied with the results we have from the initial implementation.

### Article matching pipeline

Figure 1. Test figure

Local

database

Web

content

processor

Article

Annotation

pipeline

Articles scoring

and ranking based

on comparing article

Comparing

url

Most relevance articles

from different point of views

Articles matching pipeline

To find articles with different point of view to a given article, first, we need to run the article’s url through our annotation pipeline to extract the necessary information for the scoring and comparing tasks. After that, we calculate a relevance score for every article stored in our database and rank them based on this score (more on 3.3.2). Finally, we display the highest scored articles to the users as they are the most relevance. We weren’t able to implement the different point of view check in our prototype, but we have some good theoretical claims to do it.

## The first algorithm using Sentiment Analysis and Name Entity Recognition

### Web content processor

The article on the internet are presented inside a web page, with just not only the article itself, but with tons of other unnecessary things like html tag for formatting, images and captions, links to other things on their website and advertisement text.

To strip away all the unnecessary function, we have a web content parser module which utilize a similar technique to reader mode on Firefox or Safari which will automatically strip away all the non-article part in the web content. However, this is not enough since it only strips the advertisement and related news, the core article and the html tag around it still persists, which, to solve, I wrote a smaller module to automatically remove all the html tag as well as the image and the caption, which is not really accurate since each website have different layouts and ways to present content, make parsing out the content really not effective, since, for example, some website, when they end the sentences or the caption of the image, the don’t add the “.” (dot) or sometimes they make multiple dots, which make finding out the article contents with proper sentences are quite a problem in most of the article, which make the overall annotating result worse.

One effective way of ensuring that the sentences forwarded into the CoreNLP annotator are correct is to use a tools call SMMRY, an article summarization tools, which will read through the article and gives out the sentences that it thinks contains the most important information of the article. This tool is quite effective for our case as it first, strip away all of the garbage contents like html, tag and about this website … which make the article annotation work correctly. One downside of this is that it will not return the whole article, only part of it (the more important part) so there might be possible information that will be lost during the stripping of the content. Still, I think it’s better to use this SMMRY tool instead of my own web content parser because it’s not very good and accurate.

When all the data-collecting and parsing is completed, we push the parsed data into our annotation pipeline.

### Stanford Core NLP Annotator

(*What is the role of the Stanford Core NLP in our program and why we need the local db*) - After extracting the text part from the web article, we push the text into the Stanford Core NLP engine. Stanford Core NLP have many annotators, like NER, Sentiment, lemma … (*find more*), we can specify which annotator we want to run through to save processing power. After receiving the annotated result from the Core NLP, we will run the result through our special filter for sentiment, ner and openie to retain the only relevance information for our un-polarize algorithm.

The filtered annotation result (let’s call them core feature) will be saved into the local database for future comparison of the un-polarizing algorithm. The use of the local database for the core-results is necessary because the processing time of each article is quite long, around 30 seconds each, so, to compare one article to other hundred to find the most relevance one, we must process them beforehand so that our results are readily available for future accessing.

Used annotators: tokenize, ssplit, ner, openie, sentiment, depparse, pos

### Sentiment analysis

(*How do we use the sentiment analysis, how it’s not working and how we decided to just not use it*) – This is our very first idea upon figuring out a solution to this problem. Our initial hypothesis was simple, if two articles talking about one problem, for example: The U.S President, Donald Trump, if one article talk about him with a positive sentiment, and the other with a negative sentiment, then they are from the different point of view and we should suggest the other article to our user.

Sentiment analysis services are offered by many big companies like IBM, Google, Microsoft, to some other smaller startups and a lot of open-source library exists as well. However, this method proves to be completely useless for our use-case (at least using the Stanford’s NLP lib). Because practically, sentiment analysis is only good for single sentences only, without taking into account the paragraph or the article. It is mostly used for analyzing customer reviews for products or customer feedbacks.

Furthermore, the Stanford NLP sentiment analysis result are quite limited, only 5 values, from very negative to normal to very positive, and work on the context of a single sentence only. Google’s service proves to be a bit more useful with sentiment scale from 0 to 100, but still work on a single sentence context. IBM’s service is the fanciest with scale of 100 for sentiment but also some other adjectives like “Anger”, “Nice”, “Happiness”, … which seems to be the most suitable for our case if we decided to go with this.

However, as this is a public research with source code fully available for people to use on their own, we don’t want to tie ourselves to private services that might either be unavailable someday, or restrict our access because of reasons, we decided to go with Stanford NLP, the open-sourced solution from other academia.

We calculate the sentiment value for article using Stanford NLP using the weighted average sentiment value of each sentence in the article. In which, we calculate the average value for the whole thing, but longer sentence (by characters count) will have a bigger weight than short sentence.

* **V** is the sentiment value of the article.
* is the sentiment value of sentence *i*
* is the length in number characters of sentence *i*

The sentiment analysis hypothesis later on proves to be almost useless, as the sentiment value of the sentence/paragraph have very little correlation to the content in the articles. Which mean, even if the paragraph talks about the killing of Yemen people, it might still have a normal or positive sentiment (*TODO: find example*), because journalism is usually supposed to give provide information in the most neutral way, so it’s hard to find any correlation between them. Furthermore, it’s kind of easy to fool the system, using word like nice, good, or similar to that, to make it have a higher sentiment value.

In short, sentiment analysis proves to be almost useless.

This is flaw since it can’t detect sarcasm and is flawed.

http://blog.conceptnet.io/posts/2017/how-to-make-a-racist-ai-without-really-trying/

Sentiment from text analysis are bias.

*https://researchportal.bath.ac.uk/en/publications/semantics-derived-automatically-from-language-corpora-necessarily*

### Named entity recognizer

(*How do we use the NER, the initial solution/algorithm use NER. Briefly the result of the one using only NER*) – Named entity recognizer (NER) is also presented in our original idea. With the sentiment value used to judge the \**difference in point of view*\* between two articles, we use named entity recognizer (NER) to find the relevance articles to the one the user is reading and then, compare the sentiment result to suggest it to him.

The default NER of Stanford NLP contains a lot of unnecessary information that bloat the return result for our algorithm so much. Many information (named entities) detected are quite un-relevance to our case, for example, he/she, year, number and some proposition text like Mister, Miss … We have to write a filter for the NER results to get the relevance information only. For the user convenience, we also split the result into two categories: abstract entities and discrete entities.

The two groups contain:

|  |  |
| --- | --- |
| Discrete entities | Abstract entities |
| PERSON | RELIGION |
| LOCATION | NATIONALITY |
| ORGANIZATION | TITLE (job title) |
| MISC | IDEOLOGY |
| CITY | CAUSE\_OF\_DEATH |
| STATE\_OR\_PROVINCE |  |
| COUNTRY |  |

Stanford Core NLP work on a single sentence basis only, which mean the meaning of the next or previous sentence in the paragraph doesn’t affect the annotation of one sentence in anyway, which is a bad thing for us since we want to work on a rather macro way, instead of micro based way like this. Thus, after filtering out the un-needed result, we perform the entities counting, in which we count the number of each time an entity has appeared in the article, as well as calculate the sentiment value or the average sentiment value of that entity, and finally, save all the result to the local database.

|  |  |
| --- | --- |
| (*Remove the table border later*) – Initially, the annotated value stored for one article in our database look like the part on the right:  All article will be stored as a JavaScript object, in one big single database.json file.  Each annotation object contains:   * Meta data about the articles: url, title * Annotated title, which contains sentiment value, length, and entities appearance * Annotated contents of the article, which contains information on all the entities, their appearance and their average sentiment value.   With these annotations’ information stored, we implemented a matching algorithm to suggest relevance articles to the one the use is viewing (more information on part **3.3.1 Articles similarity calculation using only NER**). However, this approach is not good enough since the sentiment analysis doesn’t work as well as expected, so this in effect just return bunch of articles talking about the same thing. | Figure n: Example of an annotated article stored in our database |

With the article annotations stored in our database, we can now suggest news with different point of view to any given article with our un-polarizing algorithm.

### Articles similarity calculation using only Named Entity Recognition

(*What a crap result we got with NER – something to cite here, I don’t know, some people must have probably researched about this*) – Our initial idea was really simple, find articles that talk about similar topic, and then sort them by the different in sentiment value, so that if two articles talk about similar topic but with vastly different sentiment value, they will have different point of view and it worth showing to the users. This approach however, as pointed out in part 3.2.2-Sentiment Analysis, is not effective since the Sentiment analysis results are all over the place and have no meaningful contribution to the algorithm at all. However, after finish writing the Similarity calculation, we found that even just suggesting the articles with relevance topics can provide lots of interesting information from many different points of view already, without considering the sentiment value.

The goal here is to not find the most similar articles, but the most relevance one, in which, we defined the relevance <*do some research here!*> as articles sharing both similar and different contents (entities). Two articles, if deemed relevance, should have half of their contents talking about similar things, and the other half talk about different things. We calculate the relevance score between two articles using this equation:

In which:

* X is the relevance score
* is the number of unique similar entities. Unique means that each entity entrance is only count once, even if they appear multiple times in both articles.
* is the number of unique different entities, summed from both articles.

As shown in the equation, when the two articles are absolute relevance (by our definition above), X should be 0, and the less relevance the two articles are, the bigger the X value will be. So, to find the list of the most relevant articles, we find the article with the smallest X value (because X is the absolute value so 0 will be smallest).

In the code, we call the -<*similarityModule.findSimilarArticles*> - this is the function we called. The input is one single “url” and the output is the list of relevance articles with their annotated information displayed so that the user know why we suggest these.

(*End and hook for the next part*) - This equation provides good results, but however, we weren’t happy with the result, or more exactly, what we get from the solution. The result did give us some other articles to the solution, but what we have is just article’s titles, and some mentioned keywords. We felt that we could do better by analyzing the article furthermore and find different opinions/facts on a sentence basis, not just by keyword counting/sentiment analyzing.

## Un-polarizing algorithm using Semantic triples

### Open information extraction

(*How openie improve our result, how do we use it and store it information*) – To furthermore improve the matching algorithm as to find even more relevance information for the user, we use Semantic Triples. The idea now is not only just match article that talks about the same things, but to read the sentences of the article and then see if there are any other articles that is also mention the same thing: ie: Trump hates brown people and Trump push brown people. This can in turn provide much more relevance information as two articles, at one point, talk about the same thing, so the user can not only see what each article says, but they can also have the power to see the content which the subject is talked about. To do such marvelous task, we use OpenIE from Core NLP.

Open information extraction from Core NLP works pretty good out of the box. Inputting one sentence into it and it gives you a lot of statements constructed from that sentence to work on. These statement however are too many and too noisy as some of the statement are quite useless information (he is president) or some are shortened version of other statement (USA hates muslim people vs USA hates muslim people all over the world).



Picture n: example of result from Core NLP

*TODO: Need example of this*

To make use of the OpenIE data, we do a three-step filtering.

First, triplets with the relation word that is not a verb and not the verb “be” is also removed. This make sure that all the non-meaningful statement (example needed!) will be removed from our result <*tripletMeaningfulProcessor.filterOpenieResult*>

Second, we remove all the triplets that are just shortened version of others, this remove quite a bit of them (needs some number value, to see the effect) *<tripletTrimmer.trimShorterTriplets>*

Last, combined with the named entities analyzed from 3.2.4, all the statement that doesn’t have an entity mentioned will also be removed, since the triplets without any meaningful entities mentioned will be useless in term of information for us anyway <*coreFeatureExtractor\_getContainingEntities*>

* Find example for each of these cases to see which result is removed and which is saved to see the impact, also, maybe some number on how much is removed for example.

After these three-step filtering, the annotated data of the article is saved to our local database as a JavaScript object in json format, same process as part 3.2.4, only with different data.

|  |  |
| --- | --- |
| This is a snippet of the annotation data stored in our database. Each entry contains:   * Meta data about the article: url and title * Array of annotated information about the content of the article, split down to a sentence level.   Each data-sentence contains:   * Full text content of the sentence * Triplets exists in the sentences and their information.   Each triplet in the sentence annotation contains:   * Subject, relation and object text. * Full text content of the triplet (combine subject, relation and object) * Containing entities.   We made the decision to store as much information as possible like the full sentence text as well as the triplet’s text so that when we return the un-polarized result to the user, they can see all the reason that leads to the decision to show them the results and can judge the comparisons for themselves. | Figure n: Example of an annotated article stored in our database (current version) |

### Triplet pairs, common entities and common statement counts

(*What are these 3 above stuffs and our sorting equation*) – With the semantic triples annotation implemented, we hoped to find some common or opposite statements from different articles that we normally could not normally find if we just read through the titles and skim through the contents. We use this equation to calculate the relevance of any given article to the source article:

In which:

* is the relevance score, the higher X is, the more relevant the article is to the comparing one.
* is the number of triplet pairs appear in both articles. Two triplets are considered a triplet pair when they share two common entities, on both the subject and object part. For example, if one statement says <*United Kingdom negotiates deal with the European Union*> and the other statement says <*European Union rejects the deals from United Kingdom*> they are considered a pair because they both talk about <*European Union*> and <*United Kingdom*> in theirs subject and object part.
* is the number common unique entities appear in both articles. If two news piece talk about an entity, it is count (but only once, even if it appears multiple times)
* is the number of common statements appear in two articles. If two statements (semantic triples) talk about a similar entity, this number goes up.
* are the constant weight value for each of these variables. In our prototype, we set =1. We use these values because entity pairs are quite rare, as most of the articles in our database, when compared to the rest of the other articles, doesn’t exist a pair at all. Maybe test with a bigger database?

General rating of the information here? I don’t know, maybe we add more data or stuffs like that?

### Providing the information to the user

Since the main purpose of our solution is to provide the user more information so that they can make a better judgement of themselves, we feel that it is important that the we should also provide as much information as possible. So, for our un-polarizing result, we will give the user the list of the most relevance articles to the one he wants to check, as well as other information that we use to come up with the conclusion, so that he can see the full picture himself, knows the reason we come up with the result, and now, being informed, can fully know the news about the situation or subjects.

|  |  |
| --- | --- |
| The return result for the user is a list of relevant articles and their annotated data, each contains:   * Meta data: general information about the two articles, containing their urls, titles and the number of entity pairs, common entities count and common statement count. * Entities pair data: we feel like this is the most interesting information so if is exist, it’s should be shown. * Entities data: all the rest information of the entities, which contains what the entity is as well as in what context it appears in, in both the source and target articles. | Figure n: Example of an annotated article stored in our database (current version) |

### Limitation of the current system.

* No verb processing in the Semantic Triple: Our current implementation only utilizes the “subject” and the “object” part of the semantic triples. The relation part, which, based on our filters, will always contains at least one action word. This information could be important as it can help us to find contradiction between different statement, thus, gives the user an even better view of the topic.
* No negation checking: current Stanford Core NLP system doesn’t detect negation in their Open Information Extraction yet, so we might miss some semantic triples from the articles. Especially, when the verb processing is implemented, missing the negative triples could lead to missing some of the contradict information.
* Computational drawback: processing an article on my computer (i5-6700HQ) took around 10 second to process one article. It’s not a big problem because 10 seconds isn’t too long but should be noted since it’s not instant.
* Needs huge database, to works. A single topic (ie immigrant) should have around 50 articles to be able to generate good ground truth of information, and the bigger the database is, the longer the computational time it takes. We can build the database to be bigger by automatically fetch news articles from source like Google news, but to solve the computational problem for big database, we need more research on how big companies like Google, Amazon or IBM deal with it.

# RESULTS EVALUATION

(*Detailed information from our article base, how many is close to the point, how many useful information can we get from that*) –

# FUTURE WORKS

## Ontologies based entity relevance

(*How using ontologies, can help finding similar words/entities, similar to the ontologies relevance in Chinh’s thesis*) – The current system finds entities pair base totally on their word-to-word similarity. Using ontologies, we could find and link together entities that are relevant to each other (ie, Muslim and Christian, as both are religion), thus, making the system smarter and able to find more information.

## Word-net verb contradiction

(*With the triplet pairs implemented, we could find contradiction between the triplets. Using word-net to find verb that have similar meaning/or opposite meaning*) – For processing the relation verbs, using wordnet (<https://wordnet.princeton.edu/>) or similar tool, so we can find words that have similar meaning, close meaning or opposite of each other, thus, making the system able to find more connection/contradiction between different statements within different articles.

## Cloud service design

Local

database

Web

content

processor

Article

annotator

Requested URL

News gatherer

News

suggestions

module

Suggested

articles

Figure 1. Test figure

With the algorithm ready, we need to automatically get our data somehow. Fortunately, there are a lot of news APIs available, for example, Google News or many other things, just one simple APIs and it can give you all the thing you want. It’s also good for evaluation later, because these APIs allow you to search for query by word, date and time, which will be useful to compare the results between our stuffs and theirs.

## User interface and user experience design

# CONCLUSION

Hope you enjoyed the text...

In the bibliography the recommendable style is Chicago. You can also use other styles: the main thing is that the styles of the bibliography and referring technique are **consistent** in the whole thesis.

|  |  |  |  |
| --- | --- | --- | --- |
| Word | Year | Magnitude | Example |
| example | [1700,2000] | [1,10] | example |
| example | [1950,2000] | [1,106] | example |
| example | [1995,2000] | [10–6,106] | example 1, example 2 |

Table 1. Example of the table

Bibliography

Spohr, Dominic. "Fake News and Ideological Polarization: Filter Bubbles and Selective Exposure On Social Media." *Business Information Review* 34.3 (2017): 150-160.

Meikäläinen, Matti. *Tietotekniikan perusteet.* Helsinki: WSOY, 2000.

Santanen, Jukka-Pekka. "Opinnäytteiden kirjoittaminen, lyhyt oppimäärä." 2000. http://users.jyu.fi/~santanen/info/kirjoittamisesta.html (accessed 5.10.2012).

Writer, John. "Name of the article." *Computer*, no. 2/2000 (2000): 31-32.

Appendices

1. Title of the first appendix
2. Title of the second appendix