ENGINE Centre, Wrocław University of Technology
Wyb. Wyspiańskiego 27, building A1, room 203k
50-370 Wrocław, Poland
tel.: +48 71 320 3453, email: engine@pwr.wroc.pl

WT2.14: Universidad Carlos III de Madrid, Madrid, Spain

Report

Activities performed during the visit

in Madrid, Spain

period: 11.04.2015 – 09.05.2015

author: Łukasz Augustyniak

This project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no 316097 [ENGINE].

ENGINE Centre, Wrocław University of Technology
Wyb. Wyspiańskiego 27, building A1, room 203k
50-370 Wrocław, Poland
tel.: +48 71 320 3453, email: engine@pwr.wroc.pl

Personal Information

Mr. **Łukasz Augustyniak**, faculty member of **Wroclaw University of Technology, Poland** visited **Universidad Carlos III de Madrid, Madrid, Spain** in the period from **11.04.2015** to **09.05.2015** in order to carry out research and training activities in the field of **sentiment analysis, text mining, complex networks, role of text analytics in social network analysis, big data analysis**.

Information about Seminars

The seminar presentation was organized on **22.04.2015**

It was entitled:

Sentiment analysis - methods, algorithms and applications.

This project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no 316097 [ENGINE].

ENGINE Centre, Wrocław University of Technology
Wyb. Wyspiańskiego 27, building A1, room 203k
50-370 Wrocław, Poland
tel.: +48 71 320 3453, email: engine@pwr.wroc.pl

Description of scientific activities

(Please describe value added to the ENGINE project i.e. new knowledge, new skills with respect to the objectives of the project, the assigned common area of future cooperation with the partner, plans for common research, projects, publications and how it could be used in the scope of ENGINE)

During visit several meetings took place. I spoke with several researchers at Universidad Carlos III (UC3M) de Madrid and also at Universidad Politécnica de Madrid (UPM). I exchanged experiment and research ideas with Professor Esteban Moro (UC3M), Dr Jose Maria Alvarez Rodriguez (UC3M), Professor Jose Carlos Gonzalez (UPM), Professor Julio Villena Román and several PhD students. The most common research interest I had with Professor Jose Carlos Gonzalez and) Professor Julio Villena Román. We started collaboration as soon as it was possible. We agreed for research covering sentiment lexicon improvement, building the sentiment dataset based on publicly available data (e.g., datasets from contests such as RepLab[1], SemEval[2], Amazon Reviews Dataset - SNAP[3] and so on). Additionally, we have been started talks regarding to using machine learning approaches for sentiment analysis. This kind of methods are main part of my PhD dissertation.

It is worth to mention that Professor Jose Carlos Gonzalez is the founder of company called Daedalus[4], recently they change name to MeaningCloud[5]. Daedalus is a Spanish company specialized in the automatic extraction of meaning from all types of multimedia content. Daedalus applies semantic, language processing, speech recognition, and data and text analysis technologies to help customers to:

- Analyze and evaluate the impact of what is said in all kinds of social and traditional media (social networks, blogs, newspapers, radio, TV).
- Extract elements of meaning and context from all types of content and social conversations to enable a more focused and effective advertising.
- Enrich and customize all kinds of multimedia and multilingual content to better combine, distribute and monetize them.
- Extract information from financial documents and user-generated content to support risk management and investment decisions.
- Integrate and retrieve information from heterogeneous repositories.

Daedalus (MeaningCloud) was founded in 1998, they provide software tools that can be deployed both on-premise and in SaaS mode.

Thanks to Professor Jose Carlos Gonzalez I spent part of my time in Madrid as the internship in MeaningCloud. We were investigating how to evaluate the two version of MeaningCloud API (actual one and new which will be used by MeaningCloud client in near

[1] http://www.limosine-project.eu/events/replab2013
[2] http://alt.qcri.org/semeval2015/05/22
[3] http://snap.stanford.edu/data/web-Amazon.html
[4] http://www.daedalus.es/en/
[5] http://www.meaningcloud.com/

This project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no 316097 [ENGINE].

ENGINE Centre, Wrocław University of Technology
Wyb. Wyspiańskiego 27, building A1, room 203k
50-370 Wrocław, Poland
tel.: +48 71 320 3453, email: engine@pwr.wroc.pl

future). Afterwards, I learnt about the architecture, technologies and approaches used in their tool's engine, especially regarding to sentiment analysis. It was great opportunity to see the real system, which provides service for client around the world. The next step was improving the methods used in sentiment analysis engine. With employees of MeaningCloud (students of Universidad Politécnica de Madrid and Universidad Carlos III de Madrid), we investigated some misclassification examples from evaluation phase and propose extension for system. It was only brief analysis, the more complex and wider analysis probably be provided in near future.

During the visit, I implemented framework for:

1. Testing several approaches for opinion classification (supervised learning, lexicons-based, rule-based).
2. Improving lexicon-based/rule-based approaches with counting the measures improvement.
3. Improving the supervised learning approaches for sentiment analysis.
4. Merging various sentiment data sources, sentiment dataset creation.
5. Evaluating accuracy (with inclusion of several measures such as accuracy, precision, recall, f-measure) of sentiment analysis tools and methods.
6. Visualization of results and statistics regarding to data sources.

An article based on this experiments and framework is in writing phase.

I was also talking with Professor Esteban Moro regarding to topic of sentiment analysis variation and detecting hatred attitudes in text. This information could be used as a source for building complex networks and doing analysis on that. However, it was only the introduction to this topic.

I talked with Dr Jose Maria Alvarez Rodriguez on using sentiment analysis technics at GitHub. Dr Jose Maria Alvarez Rodriguez crawled last year during visiting our team at Wroclaw University of Technology several GitHub's projects – code, commits, issues and so on. This is great textual data source for further analysis. We decided that it could be area which we will investigate during next visit of Dr Jose Maria Alvarez Rodriguez in Wroclaw (probably this summer).

I agreed with Professor Jose Carlos Gonzalez and Professor Julio Villena Román that we will investigate using Word2Vec and Doc2Vec method as feature generator for natural language processing (not only for sentiment analysis). Word2Vec tools (e.g., google word2vec[6]) provides an efficient implementation of the continuous bag-of-words and skip-gram architectures for computing vector representations of words. These representations can be subsequently used in many natural language processing applications and for further research. This method takes a text corpus as input and produces the word vectors as output. It first constructs a vocabulary from the training text data and then learns vector representation of words. The resulting word vector file can be used as features in many natural language processing and machine learning applications. This is really promising

[6] https://code.google.com/p/word2vec/

This project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no 316097 [ENGINE].

ENGINE Centre, Wrocław University of Technology
Wyb. Wyspiańskiego 27, building A1, room 203k
50-370 Wrocław, Poland
tel.: +48 71 320 3453, email: engine@pwr.wroc.pl

area in which only a few experiment had been done around the world. The experiment based on this method is in planning phase.

This was really fruitful visit. I learnt a lot about sentiment analysis from supervised learning and rules-based points of view. The cooperation with Professor Jose Carlos Gonzalez and MeaningCloud company gave me great experience regarding to business point of view to research which I conduct. Professor Jose Carlos Gonzalez described me commercialization and intellectual property aspects in Spain brilliantly. I learnt a lot about these matters, which is presented in next section.

Information referring to the intellectual property

(the generally binding low in this area in the visited country and procedures of patenting);

In general most of the intellectual rights belongs to the university. However depending on the funding of the project (university resources vs. private resources) the shares might differ (from 25% to 50% for the researcher). If the researcher deeply believe in success of his idea then it is more profitable to invest his own resources. If he's not convinced that the idea will be successful then he may use the university's resources in exchange for lower shares.

If the project is made with the company then different utility models are applied. Most of the time, the another agreement regarding the intellectual property is needed.

All the agreements between researcher and university are managed by the internal offices (within Foundation and Technological park – described in next point). The offices have developed guidelines how to protect intellectual property, unfortunately the document is internal.

The more detailed presentation on the intellectual property topic, prepared by Jose Maria Alvarez Rodriguez, can be found at http://slides.com/josemariaalvarez/research-science-in-spain#/54

Professor Jose Carlos Gonzalez described great and really simple solution for researchers who is working or is founder of start-up. He is claiming special agreement for research which will be used in his company. This is really fair solution for both sides, university and researcher. The intellectual property rights are determined properly in this case.

Description of the cooperation between universities and industry

(how it is organized in partner's organization, the sources of funding, the opinions about drawbacks and strengths of existing solution).

The UC3M university is a public institution, hence there are some restrictions regarding cooperation with the business. There are couple of institutions which undertake specific actions to enhance the cooperation between university and industry:

This project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no 316097 [ENGINE].

ENGINE Centre, Wrocław University of Technology
Wyb. Wyspiańskiego 27, building A1, room 203k
50-370 Wrocław, Poland
tel.: +48 71 320 3453, email: engine@pwr.wroc.pl

1. Non-profit foundation which is responsible for managing and supporting private contracts between companies and academic staff (individuals or research groups). The main task of this foundation is to simplify administration.
2. Technological park which is a research institution responsible for creating links between university and business. The park provides also space and administrative help for start-ups.
3. Incubators, their role is really similar to polish equivalent. They help establish your own business and providing assistance for such "establishments".

Another very useful institution at the university is "The office of research and transfer", which consists of 10-20 people whose main task is to help researchers to apply for the funding. The office helps to prepare proposal for the grant in the administrative section (budget planning, etc.) and also makes sure that the proposal meets the call's criteria. Such a help allows the researchers to focus on the substantive scope of the proposal and significantly improves the proposal quality. The office's staff is specialized in different areas: local calls, national calls, European calls, etc.

The UC3M university also organizes 1-day courses on how to prepare project proposal.

The professors at the UC3M are pleased from the fact how the university is organized (especially from the office of research and transfer).

The funding are local, national and European calls, e.g.:

http://wayra.co/

https://www.centrodeinnovacionbbva.com/en

Other activities

REMARK: Apart from this information also a program of the visit and the presentation in electronic version should be given to the project office (please send all of them to Urszula.Markowska-Kaczmar@pwr.wroc.pl). Please respond to the points 1-5 for outgoing visit and points 1-3 for incoming visit. Point 6 is for extra activities that are not put in points 1-5.

This project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no 316097 [ENGINE].