

INTERNSHIP REPORT

PRESENTED TO:
RISE

Technische Universität Darmstadt
Sherbrooke University

BY
Mathieu Germain, intern
Computer Science

1 September 2009

Table of Contents

1. 1	Where I Worked	3
1. 1.1	The University	3
2. 1.2	The Department	3
2. 2	My Work	3
1. 2.1	Title, Freedom and The Project	3
2. 2.2	Environment, Tools and Resources	4
3. 2.3	Change of plans	4
4. 2.4	Results	5
3. 3	Skills and knowledge	5
1. 3.1	Skills	5
1. 3.1.1	Missing	5
2. 3.1.2	Useful.....	6
3. 3.1.3	Learned	6
2. 3.2	Knowledge of the world	6
4. 4	Goals and Thoughts.....	6
1. 4.1	Interpersonal	6
2. 4.2	Goals.....	7
3. 4.3	Reflections	7

1 Where I Worked

1.1 The University

Darmstadt ("TU Darmstadt," or simply, "TUD") is a rather big, very focused place. The German school system has developed such that there are three types of universities, and TUD is of the technical variety. (The other two are "Hochschulen", schools which generally specialize in fields like media studies and digital design, and "Fachhochschulen", where students pursue degrees similar to the ones found elsewhere in the world in applied science areas.) Being a technical university means that TUD has a huge focus on sciences and engineering, and there have in fact been several discoveries of note here: Darmstadtium, atomic number 110, was discovered here, for instance. TU Darmstadt is approximately 30,000 students, and there are myriad opportunities for them. Student-run clubs thrive here, and there have been a few events for students to attend put on by the university itself. There was a day for discovering sports and games from countries around the world, and there are clubs for things from Jugger (a sport based on an Australian movie) to chess (where the board is 4 meters long and the pieces reach from the ground to one's thigh).

1.2 The Department

The Computer Science (Informatik) department in TUD is contained entirely within one building on the campus, and it is divided into several areas representing the different sorts of research that go on here. There is an area for the robotics researchers, one for the pervasive computing researchers, et cetera. The lab that I am a part of, the Ubiquitous Knowledge Processing Lab, is a group of approximately two dozen people led by Professor Iryna Guyrevich. The participants in the group and its projects range from bachelor's students to post-doctoral researchers, with representation from every group in between. The structure of the group is fairly flat; in general, the PhD candidates and post-doctoral students report directly to Professor Guyrevich, with each one having perhaps one (and none more than two) younger students reporting to them. There is no problem, though, with approaching Iryna with a question on one's work; she is knowledgeable and interested by the research of everyone in the group, and always willing to help.

2 My Work

2.1 Title, Freedom and The Project

Working as a "Guest Researcher" in the Ubiquitous Knowledge Processing (UKP) Lab at TU Darmstadt, I was tasked with investigations into automatic summarization of multiple documents and especially the MMR (Maximal Marginal Relevance) algorithm. My supervisor, Joachim

Caspar, is researching this area, and my work coincided well with his. There was no real task for the summer other than improving existing methods; in research, the work is often open-ended, and it was up to me to choose how I wished to approach it.

2.2 Environment, Tools and Resources

The accepted process here in the UKP lab is to spend a few weeks reading papers, thereafter taking some time to become familiar with the framework used (it is called the Apache Unstructured Information Management Architecture, or UIMA), immediately followed by jumping into the implementation of the project. After that follows evaluation of the work done, and all writings and presentations of the work.

The papers that I read dealt with the small amount of research that has so far been done in this area, which is not a great deal. Those papers have been mostly influenced by the work Goldstein and Carbonell. I have the feeling that not much work have been done in that field since it's beginnings almost 10 years ago. Which led me to try to implement their systems before trying new approaches of my own.

The systems in place in this lab, mostly written for use with UIMA, were a bit of a challenge to learn. They are written to abstract information about a text away from it in the form of "annotations," which can be anything from "Token" to "SenseID" to "Keyphrase Candidate." The manner in which this is done is somewhat convoluted, and there is little documentation available. Apache's official website for UIMA provides a quick overview (and tutorial) on how this sort of thing is to be processed, and there is some in-code documentation for the files written by the UKP lab that describe briefly how they work. However, it was difficult to get the hang of coding around the processes that would be more intuitive in raw Java code. I spent a lot of time trying to understand the utility of XML files that point at Java files, and I was never able to discern it. There were not many bugs associated with the code in the UKP's SVN repository, but there were some difficulties associated with getting it up and running the first time.

The work is all closed-source, and signing Non-Disclosure Agreements (NDAs) and applying for endless passwords was quite frustrating. There were a few complications related to accessing multiple documents at a time (the way that the UIMA framework is built, it is designed to fully process one document, then move on to the next, rather than holding all documents at a single stage of processing simultaneously), but they were easily enough resolved by either asking Joe or, when I ventured into areas that he had not yet explored, by asking the UIMA-users mailing lists.

2.3 Change of plans

In order to facilitate my work, I eventually asked Joe if I would be permitted to write a better system in Java. He is also not a proponent of UIMA, and he agreed that there were better ways to accomplish my task than by using it. I spent some time rewriting the system to be easier to use. This involved thought about design patterns and best practices. When there is already a

legacy system in place, often the task is more involved with trying not to break what is already there, whereas by starting afresh with a blank slate I was given the chance to create a system without all the cruft that is present in UIMA. Once I began working on the project within my own framework, everything went very smoothly, and I managed to improve over results even my supervisor had gotten.

2.4 Results

The results of my testing were comparable, and even slightly better than, those achieved by my supervisor and the authors of the papers I read during the research phase of my work. I developed a system that combines many techniques previously used in MMR (as well as tasks involved with TF*IDF which is used a lot in NLP (Natural Language Processing)) with great efficacy.

3 Skills and knowledge

I learned a lot from my experience in Germany, both related to my field and not. The most personal developments were not related to computer science, but were instead a reflection of my newfound wanderlust. I learned a number of things about computer science that I would prefer not to remember (not everyone who is a computer scientist is a programmer, which can lead to some interesting code and discussions at times), but there were also things concerning applying computer science to various different fields.

3.1 Skills

Personally, I see computer science as a tool for investigating things in other fields, and this internship reinforced that belief. But I never lost faith in computer scientists; the folks who worked in this lab were dedicated to their work, and I was impressed with their work ethic.

3.1.1 Missing

I never learned enough math to read the papers that I needed to read in order to do the basic research for this internship. This was really unfortunate; all of the formulae that described the algorithms I needed to implement for my experiments were described in a language that I had a difficult time grasping. Valkyrie is a math major, so I was lucky to have her there to explain some things to me when I did not understand them. I wish that math could be expressed in a more universal sort of way, since it is the universal language. Notations are difficult to understand when one has not been previously exposed to them.

3.1.2 Useful

I had to synthesize a lot of the skills that I had previously acquired in my field; doing research on natural languages requires an understanding not only of computer science and programming, but also it requires that one be able to think critically of how it is that a human might process language himself. Parts of speech are not a trivial thing to discern: deciding whether "land" is a noun or a verb depends heavily on context, and explaining context to a computer is more difficult than one might expect. I spent a lot of time thinking about what it is that makes a good summary good; is it the sophistication of the words? Is it the density of numerical information? Is it the proximity of the dates referenced? Is it something else? Critical thinking skills were a key part of this task.

3.1.3 Learned

I certainly developed my English skills during this internship. Valkyrie and Titilayo are both native speakers, and since my German is limited to what I learned during the two week language course in Berlin at the beginning of the summer, my supervisor spoke to me always in English. Everyone in the UKP lab speaks English at a near-perfect level, as well. There is at least one doctoral researcher who does not speak German at all. Not many people in this area know French, either, so I did not have any choice but to evolve in this respect.

3.2 Knowledge of the world

Being forced out of my comfort zone was a great learning experience. Before I had never even been on a plane, and now I have spent 4 months flying to places like Athens and London. During my travels, I realized how small the world really is. Québécois are everywhere, and they are always excited to see someone from home. I saw a lot of new places, and a lot of new faces, and I learned how the world works. I developed a new sort of people-skills for relating to people with vastly different views on life, the universe, and everything. I felt previously that I was good at meeting people and seeing things from a different point of view, but there were some things that just caught me off guard about Europe's more open culture. At any rate, I learned a new acceptance of vastly different belief systems.

4 Goals and Thoughts

4.1 Interpersonal

My coworkers were Valkyrie, a girl from the United States, and Titilayo, a girl from Nigeria but who currently attends school in the United States. We three worked shoulder-to-shoulder in the downstairs lab, and we developed quite a sense of camaraderie between us. Not only were we all three there to solve each other's UIMA troubles and other coding problems, but we discussed

also the deeper meanings of natural language processing and cognitive science. We traveled together to places like Vienna and the Swiss Alps, and we cooked dinner and watched movies together, as well. It was reassuring to have a group of people with whom I could associate without feeling guilty for not knowing German.

4.2 Goals

More or less, I feel that I attained my goals. I got a taste of what it is like to do research in a school environment. Not having a direction is very liberating, but it is also a sort of burden; when no one is telling me what to do, there is a huge array of things that I am free to try. However, since I had no previous experience in this field, there was some amount of flailing before I caught my stride and was able to make reasonably directed decisions on what I should explore next. Having the chance to learn about the scientific process as it relates to computers was great; working in industry and working in research/academia are apparently very different tasks, and I am glad that I have now had the chance to try my hand at both of them.

I improved my English. There are still some errors that I make consistently (for instance, I almost always say "I did <past tense of verb>" instead of "I did <present tense of verb>"), but I asked Valkyrie and Titilayo to correct me whenever they thought it was appropriate, and now I think that I have improved measurably. The fact that I had to use English to communicate my ideas all the times was a great motivator. I knew that if I wanted to talk to them about any sort of topic that required conscious thought, I would have to gain some measure of charisma in English. Necessity is the mother of invention, and I think that I learned a lot from being thrown into the fire in this way.

I learned a bit more about math, but I am uncertain how long those skills will stay with me (they seem to be only useful for reading papers). There are so many strange notations that vary between mathematicians between different areas of the world, and even between mathematicians in different areas of math. There is a lot of breadth to study to understand fully even a narrow topic in mathematics, and I am afraid that I simply did not have time to scratch the surface. I learned what it was that I needed to learn, but perhaps someday I will go back and permanently learn what it is that links those isolated skills together.

As far as UIMA goes, I was hoping to get better at using it for processing, but that did not work out as well as I might have hoped. I certainly have experience with it now, but I have joined the anti-UIMA crusade by creating my own tools for document summarization. I have never thought that bad tools are worth learning, even if they are popular, and it seemed much easier and more efficient to me to write something new than to try to understand all of the crazy things that the UKP lab had already written for UIMA.

4.3 Reflections

The amount of things that I learned from experiencing a new culture were greater than the number of things that I learned from the internship itself. I had new experiences in researching, as I

mentioned, but the real personal growth came from the interpersonal and/or international reactions that I had with all the new people I met.

The power of what one can learn outside the classroom is intense; once one knows how to learn, one can learn from any situation. I feel like living in Europe and seeing all sorts of famous sights that I had previously only read about in textbooks and on the Internet has taught me more than I possibly could learn from sitting in front of a computer screen and thinking. I gained a lot of inspiration from seeing the differing backgrounds and paths of the people I met, and I have found myself in several situations that I imagine I never should have been in if I had not met the people I had met.

Overall, the experience was positive, definitely. I had the sad realization that Europe is really, really expensive (especially during this economic downturn), but it was worth every cent that I spent. The sponsorship from the DAAD was a great help, and the chance to learn a new language and experience a new culture was invaluable. I do not know where my life will ultimately lead me, but I can be certain that it will have been shaped, at least somewhat, by my summer in Germany.