Information and Communication Technologies Institute
Carnegie Mellon | PORTUGAL
AN INTERNATIONAL PARTNERSHIP

Undergraduate Internship Program
Summer ‘16

João Caramujo
Instituto Superior Técnico – Universidade de Lisboa
{joao.caramujo@tecnico.ulisboa.pt}

This report provides a summary of my work during the 10-week internship at Carnegie Mellon University (CMU). From May to July I was hosted by Professor Travis Breaux and his research group from the Institute for Software Research (ISR) department.

The work I developed during the internship ended up being slightly different from our initial proposal as a better way to accommodate the work that was already done or taking place at the moment, but that could also be included more directly in my MSc thesis. Taking into account the data set previously annotated by the crowdworkers in a related project of Professor Travis, we decided that it could be interesting and useful to perform a multi-label classification task in which I was supposed to build a classifier to predict the type(s) of a given unknown sentence using such data set as training set. As opposed to what I was doing back in Portugal (only multi-class classification using the words of a sentence as features), Professor Travis introduced me to the concept of *grounded analysis* (i.e., apply specific contextual knowledge to some task) and how it could be used to perform machine learning problems (e.g., classification). The grounded analysis led to the generation of small set of features that would then be used to train the classifier since these features would be responsible for identifying a type among the others. Each feature was comprised of multiple linguistic patterns using part-of-speech tags (e.g., adverbs, nouns, etc.). We also looked for existing dependencies between pronouns (e.g., We/we) and most common verbs of each type (e.g., type – “Collection”, verb – “collect”) as a complimentary strategy to identify the type of a sentence. Preliminary results yield an accuracy (i.e., correct predictions over all the sentences) of 54% for our classifier which was promising all things considered. The wrong predictions may be due to various factors: for example, the sentences can be ambiguous to humans or the annotations might be wrong. Regardless, one possible solution could be the removal of the sentences from the data set which we know beforehand are ambiguous (e.g., sentences that contain the verb “access” are not annotated with a specific type and it is really hard to come up with clear distinguishable patterns) and acknowledge that such sentences need to be dealt separately. The testing of this and other solutions is still being carried out but hopefully we will be able to get results fit for publication.

As for the work environment, I can only praise Professor Travis and his team. Everyone was friendly, outgoing and down to earth. I was invited to attend the weekly meetings where I was able to participate and keep everyone up to speed about my work. There were also weekly reading groups where we would read research papers aloud to stimulate the discussion as a way to improve their quality before submission.

During my free time I was able to visit the Carnegie Mellon museums and the remaining main attractions of Pittsburgh. I also had the chance to work out regularly at the new CMU fitness center that had opened to the public a couple of weeks prior to my arrival. In addition, I met up with other Portuguese scholars and we went to dinner several times.

Lastly, I believe this internship was definitely an enriching experience for my career as a researcher. It helped me to gain better insight about several different technical subjects that can shape my field of study but more important I could learn a new approach on the scientific research process in Computer Engineering. The methodologies for producing quality research were significantly different from the ones I used in my university and I intend to introduce this interesting way of doing research to my fellow colleagues in Portugal.