Chapter 1
Academy and Company Needs: The Past and Future of NLP

Tiago Martins da Cunha
UNILAB, Brazil

ABSTRACT

This chapter presents a view of how the use of NLP knowledge might change the relation between universities and companies. Products from NLP analysis are expected in both ends of this at times not so reciprocal exchange. But history has shown the products developed by universities and companies are complementary for the development of NLP. The great volume of data the world is producing is requiring newer perspectives to provide understanding. These newer aspects found on big data may provide the comprehension of human language categorization and therefore possibly human language acquisition. But to process data more data need to be produced and not all companies have the time to dedicate for this task. This chapter aims to present through sharing literature review and experience in the field that partnerships are the most reliable resource for the cycle of knowledge production in NLP. Companies need to be receptive of the theoretical knowledge the university may provide, and universities must turn their theoretical knowledge for a more applied envionment.

INTRODUCTION

As a researcher in Computational Linguistics focusing on Machine Translation (MT) I had the opportunity to work for a project from the partnership between a mobile company and a University on the creation of a mobile personal assistant. This great interdisciplinary team had the opportunity to put some of the theoretical architecture of my doctoral thesis on hybrid MT in practical work. The architecture was naive, but the outcome was greater than expected.

With the compound use of statistical algorithm and the creation of rules based on the learning from the use of the prototype application the results were almost scary. The same architecture in two different mobile phones with two different team training the data with different context made the same system produce two different outcomes, almost as personality. But would this be the reach of singularity in Artificial Intelligence (AI)? Probably a lot more work would need to be done to be even talking about

DOI: 10.4018/978-1-7998-4240-8.ch001

it. But this research reached something great in a very short period of time due to its interdisciplinary approach and the high quality of its team.

This project opportunity along with the experience of a linguistic professor made me wonder what the future promises for NLP might. In the University, very differently from the experience with that mobile company project, the rhythm in each resources and data is produced is very different. Although it is those academic resources that lay the groundwork for companies' projects like the one mentioned. So, this got me thinking on how these different rhythms are related to the future of NLP.

Much has been said about the future of NLP. On applications, the popular interest in Bots and personal assistants bring science fiction closer to our everyday life. Personal assistants and conversational agents designed to tutor or auxiliar tasks are provided for many everyday activities. However, the range of understanding of popular conversationa agents and personal assistants is very limited. It is limited to a controlled language expectation. And the specific language frame within the expect context in not a controlled language.

The real-world discourse is very broad in meaning and context. Humans are designed or trained, depending on your theoratical belief, to understand language. Four years of our life is spent to master a mother language. Machines do not have this natural design. The levels of language understanding a unliterate human have is absurdly bigger than most of the complexes AI systems. The struggle to manage unstructured data is the real challenge in NLP researches. The less you struggle is the key to success in such researches. You may even find satisfaction on the implementation of computer readable resources that may provide the desired range of reachable language analysis for some NLP tools.

The volume of data is increasing everyday. Kapil, Agrawal & Khan (2016) affirm that the volume of data increased 45% to 50% in the last two years and will grow from 0,8 ZB to 35 ZB until 2020. And the biggest portion of them is text. So, many researchers aim their focus on techniques for analyzing this great volume of data. Big Data, it is called. Although this focus may be more than necessary, the effort may be given at the wrong end of the spectrum of information. Improving analysis must have reliable computational linguistic resources to get our control data from. Although a narrowed view may be given through analysis using probabilistic models to a variety of text, these resources may require a more subjective point of view. But how can these subjective analyses be implemented into such a technical field?

Well, that's what AI is all about. But not just a machine to machine analysis. By machine to machine, it is related to the use of evolutionary or genetic algorithms that produce stages of analysis not readable by humans. Understand, I'm not saying not to use statistical methods to analyze language data, but to produce readable stages so humans can shrive through. The key may lie on building hybrid methods. The use of statistical approaches to build rule-based engines that could be groomed by language experts.

However, the rules that have been mention here are not just syntactical or semantical, but cognitive as well. And not separated from each other either. Syntactical and semantical theories have been broadly used in NLP due to it extensive testing of their structured formats. The more these theories interact the more they may showed satisfying results. The limitations of syntactic and semantic systems have already showed themselves problematic. The accuracy of such approaches has reached a limit that many researchers have struggle to break through in broad context.

Many big corporation systems in spite their great success have reached a dead end on data analysis. The construction of specific context framework might not always be worth the work even though it may be the only solution for some of these big corporation problems. In the case of mobile applications, the problem now is not the specific context but the opposite.