MAD-NLP: Multi-Aspectual Diagnostics for NLP systems

With a rapid development of AI solutions that incorporate machine learning models, it becomes certain that in the near future people will interact with machines in everyday situations. Recently, Google presented its intelligent AI assistant that cannot be differentiated from human. It can be used for goal-oriented interactions like booking table in a restaurant. Although presented applications do not make any substantial decision yet, it should not be ruled out that in the near future their significance will grow.

With that being said, it is crucial to understand the level of stability of such systems' behavior. For example, question answering systems are not immune to changes in input text, meaning that minor perturbations such as introducing punctuation into asked question can lead to wrong answer. We can hypothesize that it can cause serious issues in health, legal and other crucial sectors, as natural language processing is extensively used in chatbot systems, which often assist people by answering their questions. The problem is: is it safe to make decisions based on answers from virtual assistant that are not stable and easily influenced? People making serious decisions based on the output of an automated system have the right and even obligation to ask what these predictions are based on.

Even more disturbingly, model predictions can be influenced on purpose, by deliberate injection of adversarial data. Adversarial data are examples sent into the system for prediction, which are perturbed in a way imperceptible to humans but which at the same time can easily fool deep neural networks in the testing or deployment stage. Crafting appropriate adversarial examples enables the attacker to provoke the system to produce a desired response. This way major security problems can be caused, for example if the system wrongly identifies a known terrorist as a regular citizen or gives a positive credit ranking to an individual with obvious history of fraud.

To address these issues, I am introducing a methodology which will make real my claim: **every model must be tested for robustness next to regular performance scores.**

Objective 1: I will solve the problem of analysis and strengthening of robustness in NLP systems. My methodology will allow to interpret robustness in its many faces called *aspects*, showing places where current model intelligence is lacking. In contrast to current research, my methodology will be model-flexible and comprehensive. I will analyze models of various NLP tasks: question answering (QA), sentiment analysis, machine translation (NMT), and others.

Objective 2: As the next step, I will create a method to automatically generate *adversarial* and *overstable* examples for any input text. I define *adversarial* examples as examples minimally perturbed which change model decision (although model decision should stay the same). On the other hand, *overstable* examples are examples which are maximally perturbed but do not change model decision (although they should). The method will work in step-wise manner to show the increasing level of perturbance. The method will eliminate the necessity of human intervention in the process of generating vulnerability-inducing examples.

Objective 3: While the first part and second part of research will give us insight into the models' "intelligence", the third part will use the methodology from both previous parts to analyze model robustness with regard to model performance.

Objective 4: The last part of research will focus on creating ways to correct the discovered gaps in robustness. I will propose a method of increasing robustness which will be influenced by findings from Objective 3.

The models which I will train for my robustness research (quesion answering, sentiment analysis, machine translation, and others) will use state-of-the-art deep learning techniques for their respective domains: recurrent neural networks, encoder-decoder models, convolutional neural networks. The universal aim of my work is that all created code and datasets will be made public to ensure reproducibility and wide application of my research.

In summary, my project will contribute to the area of NLP by making algorithms more understandable and robust. Their performance will be measured in a comprehensive, unified way. They will also be able to undergo automatic corrections of robustness. This will contribute to the comfort of individual users and companies, who will be able to decide their level of trust to a model.