Description for the general public

Imagine that you have bought a set of LEGO bricks. It has a specific theme – it may be a model of a dinosaur or a sports car, but it can also be a place from a well-known movie, for example, a Harry Potter-themed castle - Hogwarts. Now, let's presume that you want to mix two sets. Suddenly you find out that your bricks are not working together – it is hard to connect them, and when you force it, the outcome is not what you have expected – instead of a dinosaur standing in the Hogwarts castle, you have an unidentifiable object.

The described picture shows the current Modular Deep Learning problems in Machine Translation (MT). MT has different aspects, like sets of bricks. These sets, known in the field as domains, are:

- topic (e.g. IT, medicine, law)
- level of formality (formal, casual)
- type of text (e.g. discourse, literal)
- style (e.g. narrative, expository)

While the specific modules that represent each element of the list (i.e. IT module, law module) work, the composed modules often produce unpredictable, incorrect translations – different from what we expect.

Machine Translation learns from the parallel corpora, i.e. pairs of sentences in two languages; however, it struggles with data. While high-resourced language pairs like English-German or English-French have a lot of available data, low-resource cases like Polish-Greek do not have enough, especially when considering corpora with a specific aspect. While we can learn a module that represents, for example, a formal tone, it is likely to fail if we try to apply it to a different language pair. Moreover, we cannot first apply a module for law text, then suddenly decide to add an informal tone and, at last, make it a conversational-style translation.

In this project, we will aim to achieve two primary goals. First, we want to comprehend why we cannot combine the two sets of bricks, i.e. why we cannot use multiple modules. We hypothesise that the problem is the language information encoded in the modules – we aim to remove that bias using Disentangled Representation Learning – a process of separating underlying factors that can be used to adapt the modules to focus solely on a specific task.. Second, we want to introduce modularity that reflects real-world scenarios, where a text can have multiple aspects, e.g. a law text about AI can use both law and IT-related modules. The results of our research will allow us to build a modular machine translation system that serves different domains depending on the end-users needs.

We hope to make the ~~LEGO bricks~~ Machine Translation modules fully compositional and allow you to successfully ~~put your dinosaur in the Hogwarts castle~~ translate a sentence for your personal use-case.