Processament del Llenguatge Oral i Escrit (GCED)
Project Proposal, Class of 2020

March 25, 2020

1 Brief Introduction

It is known that Knowledge Data Base is one of the main form to represent structured information, as well as, ontologies. The reason why a lot of companies have recently started to use this DB format is due to the wide range of industrial applications, such as question answering systems, recommender systems, etc.

Thus, our team thinks that tackling a problem around the theme of Natural Language Generation (NLG) with Knowledge Data Base is a topic that current data scientists must be familiar with. Not only is it important for us to know how to apply Deep Learning techniques to real world problems, but also to have a first glimpse of how how complex human language generation can be.

2 Problem and Dataset

We might consider the following datasets:

- **WebNLG**: Creating Training Corpora for NLG Micro-Planners[1]
- **SemEval-2010 task 8 datase**: Multi-Way Classification of Relations Between Pairs of Entities

These datasets contains Resource Description Frameworks (RDF), i.e. triplets of the form (subject, predicate, object), and their corresponding text. An interesting NLG task would be to generate high-quality text from a given set of RDFs.

3 References

- The WebNLG Challenge: Generating Text from RDF Data
- Deep Graph Convolutional Encoders for Structured Data to Text Generation
- Triple-to-Text: Converting RDF Triples into High-Quality Natural Languages via Optimizing an Inverse KL Divergence
- Neural data-to-text generation: A comparison between pipeline and end-to-end architectures

4 Team Working on this project

- Bergés Lladó, David
- Cantenys Sabà, Roser
- Creus Castanyer, Roger
- Domingo Roig, Oriol

[1] This is the most frequent dataset used for this task.