Master Thesis Proposal: Name entity selection for second learner text using LSTM

Dana Dannélls
2018-10-15

1. **Goal**: Automatically select appropriate name entities to substitute with anonymized entities.

2. **Background**:
   In the SweLL infrastructure project (see: [https://språkbanken.gu.se/swe/swell_infra](https://språkbanken.gu.se/swe/swell_infra)) some manual approaches have been taken to replace name entities in learner texts with random strings. To ensure that the general flow of text will remain unchanged, the morphological form and imitation of a learner error have been replicated.

   However, manual replacement is time consuming work. To date there are no tools that exploit automatic methods, e.g. Named Entity Recognition (NER), for full- or semi-automatic learner text anonymization that aim to render personal and other sensitive information automatically while preserving the fluency of the text. For example changing all feminine names in the context to 'Eva' might result in a sentence like: “I have three sisters. Their names are Eva, Eva and Eva.”

3. **Project description**: In the project, the student(s) should examine whether recurrent neutral networks are suitable for rendering name entities in second learner text.

4. **Recommended knowledge and skills**: The project requires knowledge of LSTM-based recurrent neural network. Some knowledge of the Swedish language is a plus.

5. **Supervisors**: The work would be supervised by Dana Dannélls and Elena Volodina at Språkbanken and possibly others from the Department of Computer Science and Engineering.
6. References:
