Assignment 2: Sequence to sequence learning

What we saw this was a lot of structures for sequence learning in neural nets, and also some application of these structures. E.g. PoS tagging, word embeddings, machine translation, dialogue.

For this assignment, implement one of two styles of sequence to sequence learning.

Option 1: A tagger

Implement, using LSTM or another RNN, a tagger for part-of-speech tagging or named entity recognition. You can choose any dataset or language you like, but I recommend either:

1) Universal Dependencies data, English or Russian, for part of speech. www.universaldependencies.org
2) WikiAnn data, for English or Russian, for named entity recognition.

Implement the tagger in any framework of your choice.

Evaluation: for part-of-speech, measure token accuracy %, i.e. how many tokens get the right label; and *also*, how many sentences were labelled completely correctly (e.g. sentence accuracy %). For NER, use “conlleval.pl” or “conlleval.py” - you can find it on the web. Report strict entity F1, as well as per-entity accuracies; strict means that matches must be perfect.

Option 2: Language generation

Implement, using LSTM, either a machine translation system, or a dialogue generation system.

1) Machine translation - find something from OPUS http://opus.nlpl.eu, for one language pair.
2) Dialogue - you can use, for example, OpenSubtitles; this is film language.

Evaluate using BLEU and chrF https://github.com/m-popovic/chrF. I recommend you try using attention, and using Google Colab for doing the processing - don’t forget to save the output!

The analysis

What did the system do well? What kind of mistakes did it make? Find what the most common errors were, and give some examples of them - and examples of the tagger working correctly.

Assignment hand-in

Essay
Description of what you did and why, describing your general code, and the answers to all of the above questions in the work description. About 1000 words.

Code
Include your code. A link to a Colab notebook is best. Test the whole notebook first. I will run it myself using “reset and run all cells”.

How?
By Innopolis Moodle.

When?
The end of April 4. Good luck, and I hope you enjoy it!