L665 - ML for NLP, Deep Learning: Assignment 2

Damir Cavar

02/06/2019

I recommend using TensorFlow with Keras for the following tasks.

1 Task

Points: 15

Read the paper by Helmut Schmid "Part-of-Speech Tagging with Neural Networks" (`http://www.aclweb.org/anthology/C94-1027`). Build a similar or improved neural network system and replicate his results. In general, use the resources from NLTK as a corpus for supervised Neural Network training and development. Use for example the Brown corpus or the Penn Treebank section, extract the language models, as for example N-gram model of Part-of-Speech (PoS) tags, tokens and PoS-tags, and map them to input vectors, define and train the model to generate a probability vector for PoS-tags for every token.

You are free to decide:

- which NN-architecture with what kind of properties to use,
- what kind of features to vectorize how, and how to train the network,
- how to evaluate the result.

You should use the freely available resources in the NLTK data collection, as for example the Brown corpus, the Penn Treebank section (see treebank in NLTK data).

Will will offer a price for the best solution, where we will measure the accuracy using a different corpus with a similar tagset like Brown and Penn treebank.

Compared to Schmid's implementation, describe your improvement or decisions for a different design to improve the accuracy.

2 Task

Points: 15

We want to classify text using NNs to identify the goal or Dialogue Act in a conversation or chaterbot implementation. Read the following paper:

- Pavel Král and Christophe Cerisara. 2012. Dialogue act recognition approaches. *Computing and Informatics* 29, 2 (2012), 227–250. (download: `http://www.cai.sk/ojs/index.php/cai/article/download/82/66`)

You can also consult for example the following papers:

- Hamed Khanpour, Nishitha Guntakandla, and Rodney Nielsen. 2016. Dialogue Act Classification in Domain-Independent Conversations Using a Deep Recurrent Neural Network. In COLING.

- Ryuichiro Higashinaka, Kenji Imamura, Toyomi Meguro, Chiaki Miyazaki, Nozomi Kobayashi, Hiroaki Sugiyama, Toru Hirano, Toshiro Makino, and Yoshihiro Matsuo. 2014. Towards an open-domain conversational system fully based on natural language processing. In COLING. 928–939.

Use the Switchboard Dialog Act Corpus from:

- `http://compprag.christopherpotts.net/swda.html`

Read about the annotations of the Dialog Acts in this document and follow the instructions and hints.

- `http://compprag.christopherpotts.net/swda.html#tags`

Imagine that we only want to train a network that distinguishes between *greeting*, *goodby*, and *request* (or order, or content). Consider simplifying the tags and try to generate a sub-corpus. Use the Python code and tools to process the data. Generate a training corpus from the corpus resources listed above, if you want, create your own resources from other corpus or entirely on your own.

Consult this tutorial as an example:

- `https://machinelearnings.co/text-classification-using-neural-networks-f5cd7b8765c6`

Generate the corpus and data set to train and evaluate a network that recognizes the different types of Dialogue Acts.

Describe your approach, results, and ideas for improvement.