Jonathan Mandl 211399175
Danielle Hodaya Shrem 208150433

Part 5 – Convolution-based sub-word units

Architecture

In this section, we used convolution-based subword units. Our model is a simple feed-forward neural network with a single hidden layer with 250 neurons and tanh activation.

We used the following hyperparameter configuration for the NER task:

- Learning rate: 0.001
- Epochs: 7
- Batch size: 64

For the POS task, we used the following hyperparameter configuration:

- Learning rate: 0.001
- Epochs: 3
- Batch size: 64

We implemented the method from the Ma and Hovy paper where each word's representaton is its pretraiend word embeddings concatenated with a max-pooled vector resulting from a cnn over the word's character embeddings. We started out with the 30 filters and a window size of 3 which were the parameters from the paper. Character embeddings were initialized in the same method in the paper, with uniform samples from $[-\sqrt{(3/\text{dim})}, +\sqrt{(3/\text{dim})}]$, where we set dim = 30.

Prediction Perfromance

Our POS-tagging model reaches an accuracy of 0.96—identical to Part 4. On the NER task, it improves accuracy from 0.80 up to 0.82 compared to Part 4.

Doubling the number of filters and increasing the window size from 3 to 5 had no significant effect on either POS or NER performance. In contrast, halving the number of filters caused a slight drop in accuracy—from 0.82 to 0.81 on NER, and from 0.96 to 0.956 on POS.

Learned filters

To investigate what each character-CNN filter was learning for POS and NER, we ran the entire training set through our models and, for each filter, extracted the top 50 trigrams (contiguous 3-character sequences) with the highest filter activations across all its feature maps across and training examples. Since the network predicts the label of the center token in each window, we only considered trigrams of the center token.

In the NER analysis, this revealed filters informative for the task of entity recognition, such as:

- “oxf”, the prefix of Oxford (an Organization), which is in the training set.
- “ham”, which occurs in place names like Durham, Hamburg, and Bahamas (Locations) in our training set
- “tsb”, as in Pittsburgh (a Location) – a very common location enity in the training set
- “n.j” – location entity in our training set (abbreviation to New jersey)

These filters can be used to identify common entities in the training set

Our POS analysis returned filters important for the task of Part of Speech tagging, such as:

- ‘es<PAD>’ – ending for plural of 3rd person singular
- ‘why’, ‘who’ – question words
- ‘saw’ – common verb
- ‘nly’ - common ending which appears in adverbs like ‘only and ‘finally’
- ‘r.<PAD>’ period at end of common abbreviations (e.g. “Mr.”, “Dr.”)

These filters can help distinguish between different parts of speech by focusing on specific subword patterns.

Overall, we observed that the POS model’s filters tend to capture subword patterns related of different parts of speech (e.g., morphological endings and function words), whereas the NER model’s filters focus on subword fragments specific to entities (e.g., “oxf” for Oxford or “n.j” for New Jersey).