Natural Language Processing (almost) from Scratch

1 Main Idea

The paper [1] attempts to train a generic single learning system for multi-task learning. The taks include Part-of-Speech (POS) tagging, chunking (CHUNK), Named Entity Recognition (NER) and Semantic Role Labeling (SRL). The authors intend to acheive this without hand-engineering task-specific features, and instead rely on a large amount on unlabeled data. They also wish to avoid baselines that have been created using differently labeled data.

2 Background

- The state-of-the-art (SoTA) system for POS tagging uses bidirectional sequence decoders (Viterbi algorithm) and maximum entropy classifiers to determine, which among a set of pre-defined tags, can be attributed to a token.
- Chunking is essentially the same as POS-tagging, but for phrases instead of single words. SoTA for chunking uses pairwise SVM-classifiers, for which the features were word-contexts. Matrix SVD based methods have also been successfull.
- For NER, the SoTA is a linear model combined with Viterbi decoding, where the features include the tokens themselves, the POS tags, CHUNK tags, suffixes and prefixes.
- SRL is similar to obtaining an entity-relation model from unstructured data (text). SoTA on SRL are pasrse trees, CHUNK and POS tags, voice, types of verb etc. in combination with context-window classifiers.

3 Method

- The system used by the authors is a simple MLP architecture with minimal pre-processing and no task-specific engineered features.

- The first layer extracts word-level features and the second layer extracts sentence-level features. The architecture uses the equivalent of an embedding/lookup layer that models the language by learning dense representations of words. Sentence level embedding is learnt by using convolutions.
- There could be multiple lookup tables, with the feature vector for a word being the concatenation of all the lookup tables entries.
- For some tasks, the training objective is a multi-class softmax probability and for others, the objective is to collectively maximize the probability of the entire sequence rather that the prediction at each step individually.

4 Observations

- One pertinent question is whether multi-task learning is only effective if the sub-tasks are not completely orthogonal to each other.

References

[1] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(Aug):2493–2537, 2011.