From Raw Text to Universal Dependencies – Look, No Tags!

Miriam de Lhoneux, Yan Shao, Ali Basirat, Eliyahu Kiperwasser, Sara Stymne, Yoav Goldberg and Joakim Nivre

Word and Sentence Segmentation

We model joint sentence and word segmentation as a character-level sequence labeling problem in a Bi-RNN-CRF model.

Example:

La sede del condado es Ottawa. En BEXBIIEXKZJXBIIIIIEXBEXBIIIIETXBE

Most multiset tokens are transcribed using dictionary lookups. We use an attention-based LSTM encoder-decoder for the sparse cases.

Dependency Parsing

We use transition-based parsing with the arc-hybrid system and a dynamic oracle for training.

We use pseudo-projective parsing to handle non-projective dependencies.

We use feature representations learned by BiLSTMs with word and character embeddings but no part-of-speech tags.

Multilingual models with language embedding used for the low-resource setting.

Pre-Trained Word Embeddings

Three step word embedding method:

1. Build a co-occurrence matrix over universal features of words collected in the UD treebanks.

   Universal features: \( t_w, h_b, (t_w, d, h_b) \)

   \( t_w \): POS tag of word

   \( h_b \): POS tag of word's head

   \( d \): dependency relation between word and head

2. Normalize the data distribution along each universal feature.

3. Use the dominant singular vectors of the normalized data as word vectors.