Multi-task Pairwise Neural Ranking for Hashtag Segmentation

Mounica Maddela, Wei Xu and Daniel Preoțiuc-Pietro

Hashtag Segmentation
- Unsegmented hashtags are difficult to interpret.
- Task: Break a hashtag into its constituent words.

#songsongaddafisitunes → songs on Gaddafi’s itunes

Challenges
(C1): Difficult to identify rare tokens.
(C2): Multiple segmentations look promising.

C1: Diverse set of features
- Language model probabilities with Good Turing and modified Kneser-Ney smoothing.
- Linguistic features like word length, Wikipedia entities, Urban Dictionary, word-shapes etc.

C2: Segmentation as Pairwise Ranking

1. Extract top-k potential candidates using language model.
2. Given two candidates (s_a, s_b), predict the better segmentation of the two.

\[ L_{MSE} = \frac{1}{m} \sum_{i=1}^{m} (g^{(i)}(s_a, s_b) - g^{(i)}(s_a, s_b))^2 \]

\[ g^*(s_a, s_b) = \text{sim}(s_a, s^*) - \text{sim}(s_b, s^*) \]

3. Combine pairwise scores and rerank

Multi-task Learning
Auxiliary task: Check whether the hashtag h is multi-word.

Probability that h is multi-word

\[ L_{multi-task} = \lambda_1 L_{MSE} + \lambda_2 L_{BCE} \]

\[ L_{BCE} = \frac{1}{m} \sum_{i=1}^{m} \left[ \log(w^{(i)}_h) + (1 - L^{(i)}) \cdot \log(1 - w^{(i)}_h) \right] \]

Experiments
- 2518 train + 629 development hashtags.
- Test: STAN_6 (1,108) and STAN_7 (9,447 hashtags).

Twitter Sentiment Analysis
- SemEval 2017 - Sentiment Analysis in Twitter
- 40,000 train + 9,669 validation + 3,384 test tweets

Code and Data
https://github.com/mounicam/hashtag_master

Conclusion
- New state-of-the-art for hashtag segmentation.
- Helps with downstream tasks