Multi-task Pairwise Neural Ranking for Hashtag Segmentation

Mounica Maddela¹, Wei Xu¹, Daniel Preotu-Pietro²
¹ Department of Computer Science and Engineering, The Ohio State University
² Bloomberg LP

{maddela.4, xu.1265}@osu.edu dpreotu@bloomberg.net

Abstract

Hashtags are often employed on social media and beyond to add metadata to a textual utterance with the goal of increasing discoverability, aiding search, or providing additional semantics. However, the semantic content of hashtags is not straightforward to infer as these represent ad-hoc conventions which frequently include multiple words joined together and can include abbreviations and unorthodox spellings. We build a dataset of 12,594 hashtags split into individual segments and propose a set of approaches for hashtag segmentation by framing it as a pairwise ranking problem between candidate segmentations.¹ Our novel neural approaches demonstrate 24.6% error reduction in hashtag segmentation accuracy compared to the current state-of-the-art method. Finally, we demonstrate that a deeper understanding of hashtag semantics obtained through segmentation is useful for downstream applications such as sentiment analysis, for which we achieved a 2.6% increase in average recall on the SemEval 2017 sentiment analysis dataset.

1 Introduction

A hashtag is a keyphrase represented as a sequence of alphanumeric characters plus underscore, preceded by the # symbol. Hashtags play a central role in online communication by providing a tool to categorize the millions of posts generated daily on Twitter, Instagram, etc. They are useful in search, tracking content about a certain topic (Berardi et al., 2011; Ozdikis et al., 2012), or discovering emerging trends (Sampson et al., 2016).

Hashtags often carry very important information, such as emotion (Abdul-Mageed and Ungar, 2017), sentiment (Mohammad et al., 2013), sarcasm (Bamman and Smith, 2015), and named entities (Finin et al., 2010; Ritter et al., 2011). However, inferring the semantics of hashtags is non-trivial since many hashtags contain multiple tokens joined together, which frequently leads to multiple potential interpretations (e.g., lion head vs. lionhead). Table 1 shows several examples of single- and multi-token hashtags. While most hashtags represent a mix of standard tokens, named entities and event names are prevalent and pose challenges to both human and automatic comprehension, as these are more likely to be rare tokens. Hashtags also tend to be shorter to allow fast typing, to attract attention or to satisfy length limitations imposed by some social media platforms. Thus, they tend to contain a large number of abbreviations or non-standard spelling variations (e.g., #iloveu4eva) (Han and Baldwin, 2011; Eisenstein, 2013), which hinders their understanding.

The goal of our study is to build efficient methods for automatically splitting a hashtag into a meaningful word sequence. Our contributions are:
- A larger and better curated dataset for this task;
- Framing the problem as pairwise ranking using novel neural approaches, in contrast to previous work which ignored the relative order of candidate segmentations;
- A multi-task learning method that uses different sets of features to handle different types of

<table>
<thead>
<tr>
<th>Type</th>
<th>Single-token</th>
<th>Multi-token</th>
</tr>
</thead>
<tbody>
<tr>
<td>Named-entity (33.0%)</td>
<td>#lionhead</td>
<td>#toyota_prius</td>
</tr>
<tr>
<td>Events (14.8%)</td>
<td>#oscars</td>
<td>#ipv6_summit</td>
</tr>
<tr>
<td>Standard (43.6%)</td>
<td>#snowfall</td>
<td>#epic_fall</td>
</tr>
<tr>
<td>Non-standard (11.2%)</td>
<td>#sayin</td>
<td>#ilove4eva</td>
</tr>
</tbody>
</table>

Table 1: Examples of single- (47.1%) and multi-word hashtags (52.9%) and their categorizations based on a sample of our data.

¹Our toolkit along with the code and data are publicly available at https://github.com/mounicam/hashtag_master
• Experiments demonstrating that hashtag segmentation improves sentiment analysis on a benchmark dataset.

Our new dataset includes segmentation for 12,594 unique hashtags and their associated tweets annotated in a multi-step process for higher quality than the previous dataset of 1,108 hashtags (Bansal et al., 2015). We frame the segmentation task as a pairwise ranking problem, given a set of candidate segmentations. We build several neural architectures using this problem formulation which use corpus-based, linguistic and thesaurus based features. We further propose a multi-task learning approach which jointly learns segment ranking and single- vs. multi-token hashtag classification. The latter leads to an error reduction of 24.6% over the current state-of-the-art. Finally, we demonstrate the utility of our method by using hashtag segmentation in the downstream task of sentiment analysis. Feeding the automatically segmented hashtags to a state-of-the-art sentiment analysis method on the SemEval 2017 benchmark dataset results in a 2.6% increase in the official metric for the task.

2 Background and Preliminaries

Current approaches for hashtag segmentation can be broadly divided into three categories: (a) gaze- teer and rule based (Maynard and Greenwood, 2014; Declerck and Lendvai, 2015; Billal et al., 2016), (b) word boundary detection (Çelebi and Öztür, 2017, 2016), and (c) ranking with language model and other features (Wang et al., 2011; Bansal et al., 2015; Berardi et al., 2011; Reuter et al., 2016; Simeon et al., 2016). Hashtag segmentation approaches draw upon work on compound splitting for languages such as German or Finnish (Koehn and Knight, 2003) and word segmentation (Peng and Schuurmans, 2001) for languages with no spaces between words such as Chinese (Sproat and Shih, 1990; Xue and Shen, 2003). Similar to our work, Bansal et al. (2015) extract an initial set of candidate segmentations using a sliding window, then rerank them using a linear regression model trained on lexical, bigram and other corpus-based features. The current state-of-the-art approach (Çelebi and Öztür, 2017, 2016) uses maximum entropy and CRF models with a combination of language model and hand-crafted features to predict if each character in the hashtag is the beginning of a new word.

Generating Candidate Segmentations. Microsoft Word Breaker (Wang et al., 2011) is, among the existing methods, a strong baseline for hashtag segmentation, as reported in Çelebi and Öztür (2017) and Bansal et al. (2015). It employs a beam search algorithm to extract k best segmentations as ranked by the n-gram language model probability:

$$Score^{LM}(s) = \sum_{i=1}^{n} \log P(w_i|w_{i-N+1} \ldots w_{i-1})$$

where $[w_1, w_2 \ldots w_n]$ is the word sequence of segmentation s and N is the window size. More sophisticated ranking strategies, such as Binomial and word length distribution based ranking, did not lead to a further improvement in performance (Wang et al., 2011). The original Word Breaker was designed for segmenting URLs using language models trained on web data. In this paper, we reimplemented and tailored this approach to segmenting hashtags by using a language model specifically trained on Twitter data (implementation details in §3.6). The performance of this method itself is competitive with state-of-the-art methods (evaluation results in §5.3). Our proposed pairwise ranking method will effectively take the top k segmentations generated by this baseline as candidates for reranking.

However, in prior work, the ranking scores of each segmentation were calculated independently, ignoring the relative order among the top k candidate segmentations. To address this limitation, we utilize a pairwise ranking strategy for the first time for this task and propose neural architectures to model this.

3 Multi-task Pairwise Neural Ranking

We propose a multi-task pairwise neural ranking approach to better incorporate and distinguish the relative order between the candidate segmentations of a given hashtag. Our model adapts to address single- and multi-token hashtags differently via a multi-task learning strategy without requiring additional annotations. In this section, we describe the task setup and three variants of pairwise neural ranking models (Figure 1).
hash tag \( h \) | #songsonghaddafisitunes \\
---|---
segmentation \( s^* \) | songs on ghaddafi's itunes (i.e. songs on Ghaddafí's iTunes) \\
candidate segmentations \( s \in S \) | songs on ghaddafis itunes songs on ghaddafis tunes songs on ghaddafi is itunes song song haddafis itunes songsong haddafistunes (and...) \\

Table 2: Example hashtag along with its gold and possible candidate segmentations.

### 3.1 Segmentation as Pairwise Ranking

The goal of hashtag segmentation is to divide a given hashtag \( h \) into a sequence of meaningful words \( s^* = [w_1, w_2, \ldots, w_n] \). For a hashtag of \( r \) characters, there are a total of \( 2^r - 1 \) possible segmentations but only one, or occasionally two, of them \( (s^*) \) are considered correct (Table 2).

We transform this task into a pairwise ranking problem: given \( k \) candidate segmentations \( \{s_1, s_2, \ldots, s_k\} \), we rank them by comparing each with the rest in a pairwise manner. More specifically, we train a model to predict a real number \( g(s_a, s_b) \) for any two candidate segmentations \( s_a \) and \( s_b \) of hashtag \( h \), which indicates \( s_a \) is a better segmentation than \( s_b \) if positive, and vice versa. To quantify the quality of a segmentation in training, we define a gold scoring function \( g^* \) based on the similarities with the ground-truth segmentation \( s^* \):

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

We use the Levenshtein distance (minimum number of single-character edits) in this paper, although it is possible to use other similarity measurements as alternatives. We use the top \( k \) segmentations generated by Microsoft Word Breaker (§2) as initial candidates.

### 3.2 Pairwise Neural Ranking Model

For an input candidate segmentation pair \( \langle s_a, s_b \rangle \), we concatenate their feature vectors \( s_a \) and \( s_b \), and feed them into a feedforward network which emits a comparison score \( g(s_a, s_b) \). The feature vector \( s_a \) or \( s_b \) consists of language model probabilities using Good-Turing (Good, 1953) and modified Kneser-Ney smoothing (Kneser and Ney, 1995; Chen and Goodman, 1999), lexical and linguistic features (more details in §3.5). For training, we use all the possible pairs \( \langle s_a, s_b \rangle \) of the \( k \) candidates as the input and their gold scores \( g^*(s_a, s_b) \) as the target. The training objective is to minimize the Mean Squared Error (MSE):

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

where \( m \) is the number of training examples.

To aggregate the pairwise comparisons, we follow a greedy algorithm proposed by Cohen et al. (1998) and used for preference ranking (Parakhin and Haluptzok, 2009). For each segmentation \( s \) in the candidate set \( S = \{s_1, s_2, \ldots, s_k\} \), we calculate a single score \( \text{Score}_{PNR}(s) = \sum_{s \neq s_j \in S} g(s, s_j) \), and find the segmentation \( s_{\text{max}} \) corresponding to the highest score. We repeat the same procedure after removing \( s_{\text{max}} \) from \( S \), and continue until \( S \) reduces to an empty set. Figure 1(a) shows the architecture of this model.

### 3.3 Margin Ranking (MR) Loss

As an alternative to the pairwise ranker (§3.2), we propose a pairwise model which learns from candidate pairs \( \langle s_a, s_b \rangle \) but ranks each individual candidate directly rather than relatively. We define a new scoring function \( g' \) which assigns a higher score to the better candidate, i.e., \( g'(s_a) > g'(s_b) \), if \( s_a \) is a better candidate than \( s_b \) and vice-versa. Instead of concatenating the features vectors \( s_a \) and \( s_b \), we feed them separately into two identical feedforward networks with shared parameters. During testing, we use only one of the networks to rank the candidates based on the \( g' \) scores. For training, we add a ranking layer on top of the networks to measure the violations in the ranking order and minimize the Margin Ranking Loss (MR):

\[
L_{MR} = \frac{1}{m} \sum_{i=1}^{m} \max(0, 1 - \hat{l}_{ab} P_{ab})
\]

\[
P_{ab} = \begin{cases} 
1 & g'(s_a) > g'(s_b) \\
-1 & g'(s_a) < g'(s_b) \\
0 & \text{otherwise}
\end{cases}
\]

where \( m \) is the number of training samples. The architecture of this model is presented in Figure 1(b).

### 3.4 Adaptive Multi-task Learning

Both models in §3.2 and §3.3 treat all the hashtags uniformly. However, different features address different types of hashtags. By design, the linguistic features capture named entities and multi-word hashtags that exhibit word shape patterns,
such as camel case. The ngram probabilities with Good-Turing smoothing gravitate towards multi-word segmentations with known words, as its estimate for unseen ngrams depends on the fraction of ngrams seen once which can be very low (Heafield, 2013). The modified Kneser-Ney smoothing is more likely to favor segmentations that contain rare words, and single-word segmentations in particular. Please refer to §5.3 for a more detailed quantitative and qualitative analysis.

To leverage this intuition, we introduce a binary classification task to help the model differentiate single-word from multi-word hashtags. The binary classifier takes hashtag features $h$ as the input and outputs $w_h$, which represents the probability of $h$ being a multi-word hashtag. $w_h$ is used as an adaptive gating value in our multi-task learning setup. The gold labels for this task are obtained at no extra cost by simply verifying whether the ground-truth segmentation has multiple words. We train the pairwise segmentation ranker and the binary single- vs. multi-token hashtag classifier jointly, by minimizing $L_{MSE}$ for the pairwise ranker and the Binary Cross Entropy Error ($L_{BCE}$) for the classifier:

$$L_{multitask} = \lambda_1 L_{MSE} + \lambda_2 L_{BCE}$$

$$L_{BCE} = -\frac{1}{m} \sum_{i=1}^{m} \left[ l^{(i)} \log(w_h^{(i)}) + (1 - l^{(i)}) \log(1 - w_h^{(i)}) \right]$$

where $w_h$ is the adaptive gating value, $l \in \{0, 1\}$ indicates if $h$ is actually a multi-word hashtag and $m$ is the number of training examples. $\lambda_1$ and $\lambda_2$ are the weights for each loss. For our experiments, we apply equal weights.

More specifically, we divide the segmentation feature vector $s_a$ into two subsets: (a) $s_{ab}^{KN}$ with modified Kneser-Ney smoothing features, and (b) $s_{ab}^{GL}$ with Good-Turing smoothing and linguistic features. For an input candidate segmentation pair $(s_a, s_b)$, we construct two pairwise vectors $s_{ab}^{KN} = [s_a^{KN}; s_b^{KN}]$ and $s_{ab}^{GL} = [s_a^{GL}; s_b^{GL}]$ by concatenation, then combine them based on the adaptive gating value $w_h$ before feeding them into the feed-forward network $G$ for pairwise ranking:

$$\hat{g}(s_a, s_b) = G(w_h s_{ab}^{GL} + (1 - w_h) s_{ab}^{KN})$$

We use summation with padding, as we find this simple ensemble method achieves similar performance in our experiments as the more complex multi-column networks (Ciresan et al., 2012). Figure 1(c) shows the architecture of this model. An analogue multi-task formulation can also be used for the Margin Ranking loss as:

$$L_{multitask} = \lambda_1 L_{MR} + \lambda_2 L_{BCE}.$$  \hspace{1cm} (5)

### 3.5 Features

We use a combination of corpus-based and linguistic features to rank the segmentations. For a candidate segmentation $s$, its feature vector $s$ includes the number of words in the candidate, the length of each word, the proportion of words in an English dictionary\(^3\) or Urban Dictionary\(^4\) (Nguyen et al., 2018), ngram counts from Google Web 1TB corpus (Brants and Franz, 2006), and ngram probabilities from trigram language models trained on the Gigaword corpus (Graff and Cieri, 2003) and

\(^3\)https://pypi.org/project/pyenchant
\(^4\)https://www.urbandictionary.com
1.1 billion English tweets from 2010, respectively. We train two language models on each corpus: one with Good-Turing smoothing using SRILM (Stolcke, 2002) and the other with modified Kneser-Ney smoothing using KenLM (Heafield, 2011). We also add boolean features, such as if the candidate is a named-entity present in the list of Wikipedia titles, and if the candidate segmentation $s$ and its corresponding hashtag $h$ satisfy certain word-shapes (more details in appendix A.1).

Similarly, for hashtag $h$, we extract the feature vector $h$ consisting of hashtag length, ngram count of the hashtag in Google 1TB corpus (Brants and Franz, 2006), and boolean features indicating if the hashtag is in an English dictionary or Urban Dictionary, is a named-entity, is in camel case, ends with a number, and has all the letters as consonants. We also include features of the best-ranked candidate by the Word Breaker model.

### 3.6 Implementation Details

We use the PyTorch framework to implement our multi-task pairwise ranking model. The pairwise ranker consists of an input layer, three hidden layers with eight nodes in each layer and hyperbolic tangent ($\tanh$) activation, and a single linear output node. The auxiliary classifier consists of an input layer, one hidden layer with eight nodes and one output node with sigmoid activation. We use the Adam algorithm (Kingma and Ba, 2014) for optimization and apply a dropout of 0.5 to prevent overfitting. We set the learning rate to 0.01 and 0.05 for the pairwise ranker and auxiliary classifier respectively. For each experiment, we report results obtained after 100 epochs.

For the baseline model used to extract the $k$ initial candidates, we reimplemented the Word Breaker (Wang et al., 2011) as described in §2 and adapted it to use a language model trained on 1.1 billion tweets with Good-Turing smoothing using SRILM (Stolcke, 2002) to give a better performance in segmenting hashtags (§5.3). For all our experiments, we set $k = 10$.

### 4 Hashtag Segmentation Data

We use two datasets for experiments (Table 3): (a) $STAN_{small}$, created by Bansal et al. (2015), which consists of 1,108 unique English hashtags from 1,268 randomly selected tweets in the Stanford Sentiment Analysis Dataset (Go and Huang, 2009) along with their crowdsourced segmentations and our additional corrections; and (b) $STAN_{large}$, our new expert curated dataset, which includes all 12,594 unique English hashtags and their associated tweets from the same Stanford dataset.

#### Dataset Analysis. $STAN_{small}$ is the most commonly used dataset in previous work. However, after reexamination, we found annotation errors in 6.8% of the hashtags in this dataset, which is significant given that the error rate of the state-of-the-art models is only around 10%. Most of the errors were related to named entities. For example, #lionhead, which refers to the “Lionhead” video game company, was labeled as “lion head”.

#### Our Dataset. We therefore constructed the $STAN_{large}$ dataset of 12,594 hashtags with additional quality control for human annotations. We displayed a tweet with one highlighted hashtag on the Figure-Eight (previously known as Crowd-Flower) crowdsourcing platform and asked two workers to list all the possible segmentations. For quality control on the platform, we displayed a test hashtag in every page along with the other hashtags. If any annotator missed more than 20% of the test hashtags, then they were not allowed to continue work on the task. For 93.1% of the hashtags, the workers agreed on the same segmentation. We further asked three in-house annotators (not authors) to cross-check the crowdsourced annotations using a two-step procedure: first, verify if the hashtag is a named entity based on the context of the tweet; then search on Google to find the correct segmentation(s). We also asked the same annotators to fix the errors in $STAN_{small}$. The human upperbound of the task is estimated at ~98% accuracy, where we consider the crowdsourced segmentations (two workers merged) as correct if at least one of them matches with our expert annotator’s segmentations.

<table>
<thead>
<tr>
<th>Data</th>
<th>num. of Hashtags (multi-token%)</th>
<th>avg. #char</th>
<th>avg. #word</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train</td>
<td>2518 (51.9%)</td>
<td>8.5</td>
<td>1.8</td>
</tr>
<tr>
<td>Dev</td>
<td>629 (52.3%)</td>
<td>8.4</td>
<td>1.7</td>
</tr>
<tr>
<td>Test</td>
<td>9447 (53.0%)</td>
<td>8.6</td>
<td>1.8</td>
</tr>
</tbody>
</table>

Table 3: Statistics of the $STAN_{small}$ and $STAN_{large}$ datasets – number of unique hashtags, percentage of multi-token hashtags, average length of hashtags in characters and words.

---

1 More specifically, 4.8% hashtags is missing one of the two acceptable segmentations and another 2.0% is incorrect segmentation.

6 https://figure-eight.com
Table 4: Evaluation results on the corrected version of STAN\textsubscript{small}. For reference, on the original version of STAN\textsubscript{small}, the Microsoft Word Breaker API reported an 84.6% F\(_1\) score and an 83.6% accuracy for the top one output (Çelebi and Özgür, 2017), while our best model (MSE+multitask) reported 89.8% F\(_1\) and 91.0% accuracy.

Table 5: Evaluation results on our STAN\textsubscript{large} test dataset. For single-token hashtags, the token-level F\(_1\) is equivalent to segmentation-level A@1. For multi-token cases, A@1 and F\(_1\) for the original hashtag baseline are non-zero because 11.4% of the hashtags have more than one acceptable segmentations. Our best model (MSE+multitask) shows a statistically significant improvement (p < 0.05) over the state-of-the-art approach (Çelebi and Özgür, 2017) based on the paired bootstrap test (Berg-Kirkpatrick et al., 2012).

5 Experiments
In this section, we present experimental results that compare our proposed method with the other state-of-the-art approaches on hashtag segmentation datasets. The next section will show experiments of applying hashtag segmentation to the popular task of sentiment analysis.

5.1 Existing Methods
We compare our pairwise neural ranker with the following baseline and state-of-the-art approaches:
(a) The original hashtag as a single token;
(b) A rule-based segmenter, which employs a set of word-shape rules with an English dictionary (Billal et al., 2016);
(c) A Viterbi model which uses word frequencies from a book corpus\(^7\) (Berardi et al., 2011);
(d) The specially developed GATE Hashtag Tokenizer from the open source toolkit,\(^8\) which combines dictionaries and gazetteers in a Viterbi-like algorithm (Maynard and Greenwood, 2014);
(e) A maximum entropy classifier (MaxEnt) trained on the STAN\textsubscript{large} training dataset. It predicts whether a space should be inserted at each position in the hashtag and is the current state-of-the-art (Çelebi and Özgür, 2017);
(f) Our reimplementation of the Word Breaker algorithm which uses beam search and a Twitter ngram language model (Wang et al., 2011);
(g) A pairwise linear ranker which we implemented for comparison purposes with the same features as our neural model, but using perceptron as the underlying classifier (Hopkins and May, 2011) and minimizing the hinge

---

\(^7\)Project Gutenberg [http://norvig.com/big.txt]

\(^8\)https://gate.ac.uk/
Tables 4 and 5 show the results on the STAN_{large} dev set. A denotes accuracy@1. While Kneser-Ney features perform well on single-token hashtags, GT+Ling features perform better on multi-token hashtags.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Single</th>
<th>Multi</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A MRR</td>
<td>A MRR</td>
<td>A MRR</td>
</tr>
<tr>
<td>Kneser-Ney</td>
<td>95.4</td>
<td>95.7</td>
<td>56.0</td>
</tr>
<tr>
<td>Good-Turing (GT)</td>
<td>91.4</td>
<td>93.5</td>
<td>85.9</td>
</tr>
<tr>
<td>Linguistic (Ling)</td>
<td>89.4</td>
<td>91.7</td>
<td>71.6</td>
</tr>
<tr>
<td>GT + Ling</td>
<td>92.4</td>
<td>93.9</td>
<td>86.2</td>
</tr>
<tr>
<td>All Features</td>
<td>91.1</td>
<td>93.1</td>
<td>89.0</td>
</tr>
</tbody>
</table>

5.2 Evaluation Metrics

We evaluate the performance by the top $k$ ($k = 1, 2$) accuracy (A@1, A@2), average token-level $F_1$ score ($F_1@1$), and mean reciprocal rank (MRR). In particular, the accuracy and MRR are calculated at the segmentation-level, which means that an output segmentation is considered correct if and only if it fully matches the human segmentation. Average token-level $F_1$ score accounts for partially correct segmentation in the multi-token hashtag cases.

5.3 Results

Tables 4 and 5 show the results on the STAN_{small} and STAN_{large} datasets, respectively. All of our pairwise neural rankers are trained on the 2,518 manually segmented hashtags in the training set of STAN_{large} and perform favorably against other state-of-the-art approaches. Our best model (MSE+multitask) that utilizes different features adaptively via a multi-task learning procedure is shown to perform better than simply combining all the features together (MR and MSE). We highlight the 24.6% error reduction on STAN_{small} and 16.5% on STAN_{large} of our approach over the previous SOTA (Çelebi and Özgür, 2017) on the Multi-token hashtags, and the importance of having a separate evaluation of multi-word cases as it is trivial to obtain 100% accuracy for Single-token hashtags. While our hashtag segmentation model is achieving a very high accuracy@2, to be practically useful, it remains a challenge to get the top one prediction exactly correct. Some hashtags are very difficult to interpret, e.g., #BTVSMB refers to the Social Media Breakfast (SMB) in Burlington, Vermont (BTV).

The improved Word Breaker with our addition of a Twitter-specific language model is a very strong baseline, which echoes the findings of the original Word Breaker paper (Wang et al., 2011) that having a large in-domain language model is extremely helpful for word segmentation tasks. It is worth noting that the other state-of-the-art system (Çelebi and Özgür, 2017) also utilized a 4-gram language model trained on 476 million tweets from 2009.

5.4 Analysis and Discussion

Feature Analysis. To empirically illustrate the effectiveness of different features on different types of hashtags, we show the results for models using individual feature sets in pairwise ranking models (MSE) in Table 6. Language models with modified Kneser-Ney smoothing perform best on single-token hashtags, while Good-Turing and Linguistic features work best on multi-token hashtags, confirming our intuition about their usefulness in a multi-task learning approach. Table 7 shows a qualitative analysis with the first column (○○○) indicating which features lead to correct or wrong segmentations, their count in our data and illustative examples with human segmentation.

<table>
<thead>
<tr>
<th>Kneser-Ney</th>
<th>Good-Turing</th>
<th>Linguistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>count</td>
<td>Example Hashtags</td>
<td></td>
</tr>
<tr>
<td>○○○</td>
<td>#omnomnom #BTVSMB</td>
<td></td>
</tr>
<tr>
<td>○○</td>
<td>#commbank #mamapedia</td>
<td></td>
</tr>
<tr>
<td>○</td>
<td>#wewanimlyf #winebarst</td>
<td></td>
</tr>
<tr>
<td>○○</td>
<td>#cgf09 #TechLunchSouth</td>
<td></td>
</tr>
<tr>
<td>○○</td>
<td>#cwcc #ecom09</td>
<td></td>
</tr>
<tr>
<td>○○</td>
<td>#LetsGoPens #epicwinn</td>
<td></td>
</tr>
<tr>
<td>○○</td>
<td>#prototype #newyork</td>
<td></td>
</tr>
</tbody>
</table>

Length of Hashtags. As expected, longer hashtags with more than three tokens pose greater challenges and the segmentation-level accuracy of our best model (MSE+multitask) drops to 82.1%. For many error cases, our model predicts a close-to-correct segmentation, e.g., #youknowyouuptooearly, #iseelondoniseefrance, which is also reflected by...
### 6 Extrinsic Evaluation: Twitter Sentiment Analysis

We attempt to demonstrate the effectiveness of our hashtag segmentation system by studying its impact on the task of sentiment analysis in Twitter (Pang et al., 2002; Nakov et al., 2016; Rosenthal et al., 2017). We use our best model (MSE+multitask), under the name **HashtagMaster**, in the following experiments.

#### 6.1 Experimental Setup

We compare the performance of the BiLSTM+Lex (Teng et al., 2016) sentiment analysis model under three configurations: (a) tweets with hashtags removed, (b) tweets with hashtags as single tokens excluding the # symbol, and (c) tweets with hashtags as segmented by our system, HashtagMaster. BiLSTM+Lex is a state-of-the-art open source system for predicting tweet-level sentiment (Tay et al., 2018). It learns a context-sensitive sentiment intensity score by leveraging a Twitter-based sentiment lexicon (Tang et al., 2014). We use the same settings as described by Teng et al. (2016) to train the model.

We use the dataset from the **Sentiment Analysis in Twitter** shared task (subtask A) at SemEval 2017 (Rosenthal et al., 2017). Given a tweet, the goal is to predict whether it expresses **POSITIVE**, **NEGATIVE** or **NEUTRAL** sentiment. The training and development sets consist of 49,669 tweets and we use 40,000 for training and the rest for development. There are a total of 4,840 tweets containing 12,128 hashtags in the SemEval 2017 test set, and our hashtag segmenter ended up splitting 6,975 of those hashtags present in 3,384 tweets.

#### 6.2 Results and Analysis

In Table 9, we report the results based on the 3,384 tweets where HashtagMaster predicted a split, as for the rest of tweets in the test set, the hashtag segmenter would neither improve nor worsen the sentiment prediction. Our hashtag segmenter successfully improved the sentiment analysis performance by 2% on average recall and \( F_1 \) comparing to having hashtags unsegmented. This improvement is seemingly small but decidedly important for tweets where sentiment-related information is embedded in multi-word hashtags.

---

### Table 8: Evaluation results on 500 random hashtags from the year 2019.

<table>
<thead>
<tr>
<th>Method</th>
<th>( F_1 )</th>
<th>( @1 )</th>
<th>MRR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word Breaker w/ Twitter LM</td>
<td>92.1</td>
<td>93.9</td>
<td>94.7</td>
</tr>
<tr>
<td>Pairwise neural ranker (MSE+multitask)</td>
<td><strong>94.6</strong></td>
<td><strong>95.6</strong></td>
<td><strong>96.7</strong></td>
</tr>
</tbody>
</table>

---

9We did not use the Stanford Sentiment Analysis Dataset (Go and Huang, 2009), which was used to construct the STANsmall and STANlarge hashtag datasets, because of its noisy sentiment labels obtained using distant supervision.
and sentiment prediction would be incorrect based only on the text (see Table 10 for examples). In fact, 2,605 out of the 3,384 tweets have multi-word hashtags that contain words in the Twitter-based sentiment lexicon (Tang et al., 2014) and 125 tweets contain sentiment words only in the hashtags but not in the rest of the tweet.

### 7 Other Related Work

Automatic hashtag segmentation can improve the performance of many applications besides sentiment analysis, such as text classification (Billal et al., 2016), named entity linking (Bansal et al., 2015) and modeling user interests for recommendations (Chen et al., 2016). It can also help in collecting data of higher volume and quality by providing a more nuanced interpretation of its content, as shown for emotion analysis (Qadir and Riloff, 2014), sarcasm and irony detection (Maynard and Greenwood, 2014; Huang et al., 2018). Better semantic analysis of hashtags can also potentially be applied to hashtag annotation (Wang et al., 2019), to improve distant supervision labels in training classifiers for tasks such as sarcasm (Bamman and Smith, 2015), sentiment (Mohammad et al., 2013), emotions (Abdul-Mageed and Ungar, 2017); and, more generally, as labels for pre-training representations of words (Weston et al., 2014), sentences (Dhingra et al., 2016), and images (Mahajan et al., 2018).

### 8 Conclusion

We proposed a new pairwise neural ranking model for hashtag segmentation and showed significant performance improvements over the state-of-the-art. We also constructed a larger and more curated dataset for analyzing and benchmarking hashtag segmentation methods. We demonstrated that hashtag segmentation helps with downstream tasks such as sentiment analysis. Although we focused on English hashtags, our pairwise ranking approach is language-independent and we intend to extend our toolkit to languages other than English as future work.

### Acknowledgments

We thank Ohio Supercomputer Center (Center, 2012) for computing resources and the NVIDIA for providing GPU hardware. We thank Alan Ritter, Quanze Chen, Wang Ling, Pravar Mahajan, and Dushyanta Dhyani for valuable discussions. We also thank the annotators: Sarah Flanagan, Kaushik Mani, and Aswathnarayan Radhakrishnan. This material is based in part on research sponsored by the NSF under grants IIS-1822754 and IIS-1755898, DARPA through the ARO under agreement number W911NF-17-C-0095, through a Figure-Eight (CrowdFlower) AI for Everyone Award and a Criteo Faculty Research Award to Wei Xu. The views and conclusions contained in this publication are those of the authors and should not be interpreted as representing official policies or endorsements of the U.S. Government.

### References


Appendix

A.1 Word-shape rules

Our model uses the following word shape rules as boolean features. If the candidate segmentation \( s \) and its corresponding hashtag \( h \) satisfies a word shape rule, then the boolean feature is set to True.

<table>
<thead>
<tr>
<th>Rule</th>
<th>Hashtag ( \rightarrow ) Segmentation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Camel Case</td>
<td>( \text{XxXxx} \rightarrow \text{Xxx+Xxx} )</td>
</tr>
<tr>
<td>Consonants</td>
<td>( \text{ccc} \rightarrow \text{ccc} )</td>
</tr>
<tr>
<td>Digits as prefix</td>
<td>( \text{ddwww} \rightarrow \text{dd+www} )</td>
</tr>
<tr>
<td>Digits as suffix</td>
<td>( \text{wwwdd} \rightarrow \text{www+dd} )</td>
</tr>
<tr>
<td>Underscore</td>
<td>( \text{www_www} \rightarrow \text{www + _ + www} )</td>
</tr>
</tbody>
</table>

Table 11: Word-shape rule features used to identify good segmentations. Here, \( X \) and \( x \) represent capitalized and non-capitalized alphabetic characters respectively, \( c \) denotes consonant, \( d \) denotes number and \( w \) denotes any alphabet or number.