Multi-task Pairwise Neural Ranking for Hashtag Segmentation

THE OHIO STATE UNIVERSITY

Mounica Maddela, Wei Xu and Daniel Preoţiuc-Pietro

Bloomberg
Engineering

Hashtag Segmentation

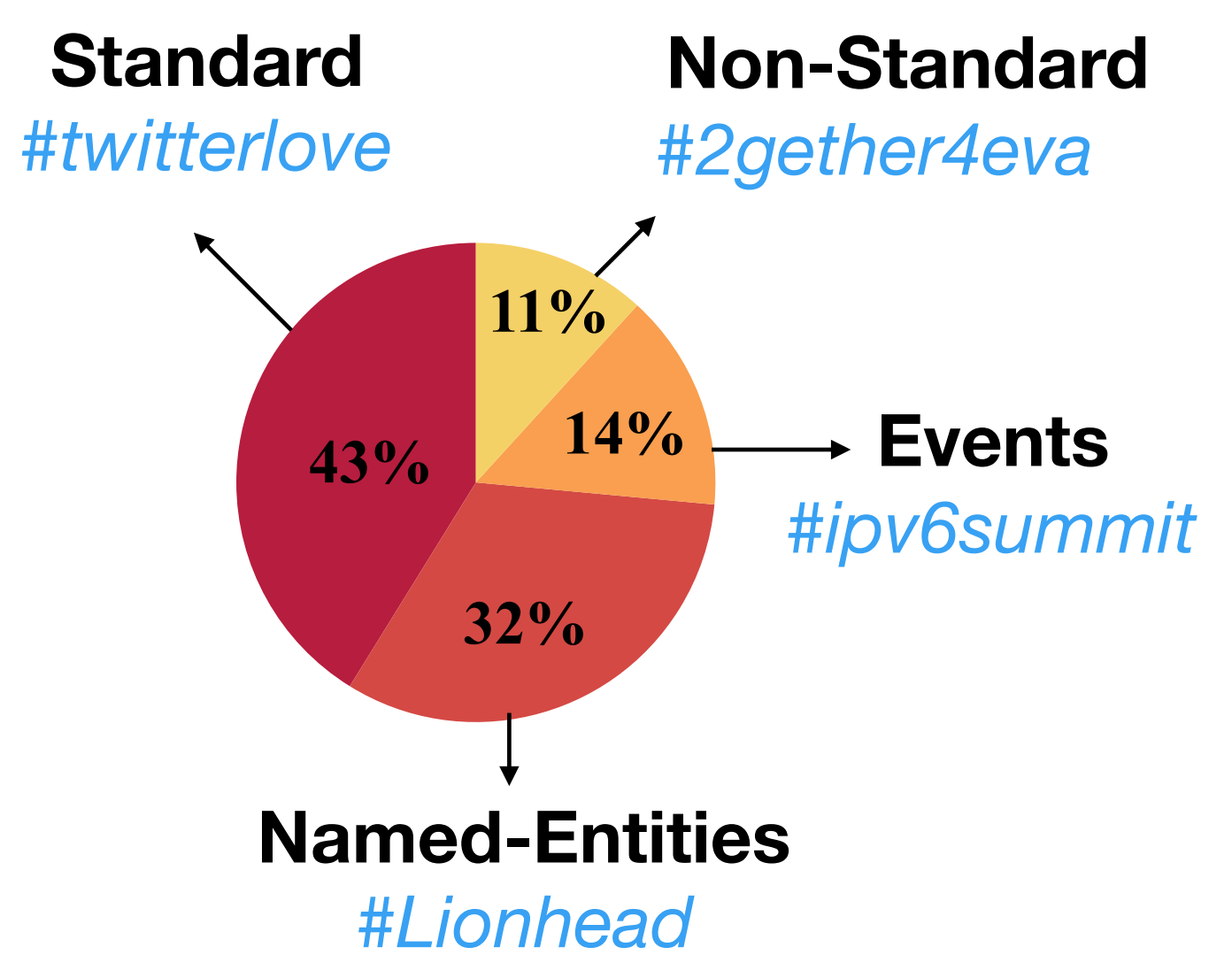
- Unsegmented hashtags are difficult to interpret.
- Task: Break a hashtag into its constituent words.

#songsongaddafisitunes → *songs on Gaddafi's itunes*

- **Challenges**

(C1): Difficult to identify rare tokens.

(C2): Multiple segmentations look promising.



BlackVelvet is the DJ for your #beatwittyparty. Whatcha wanna hear? d-_-b

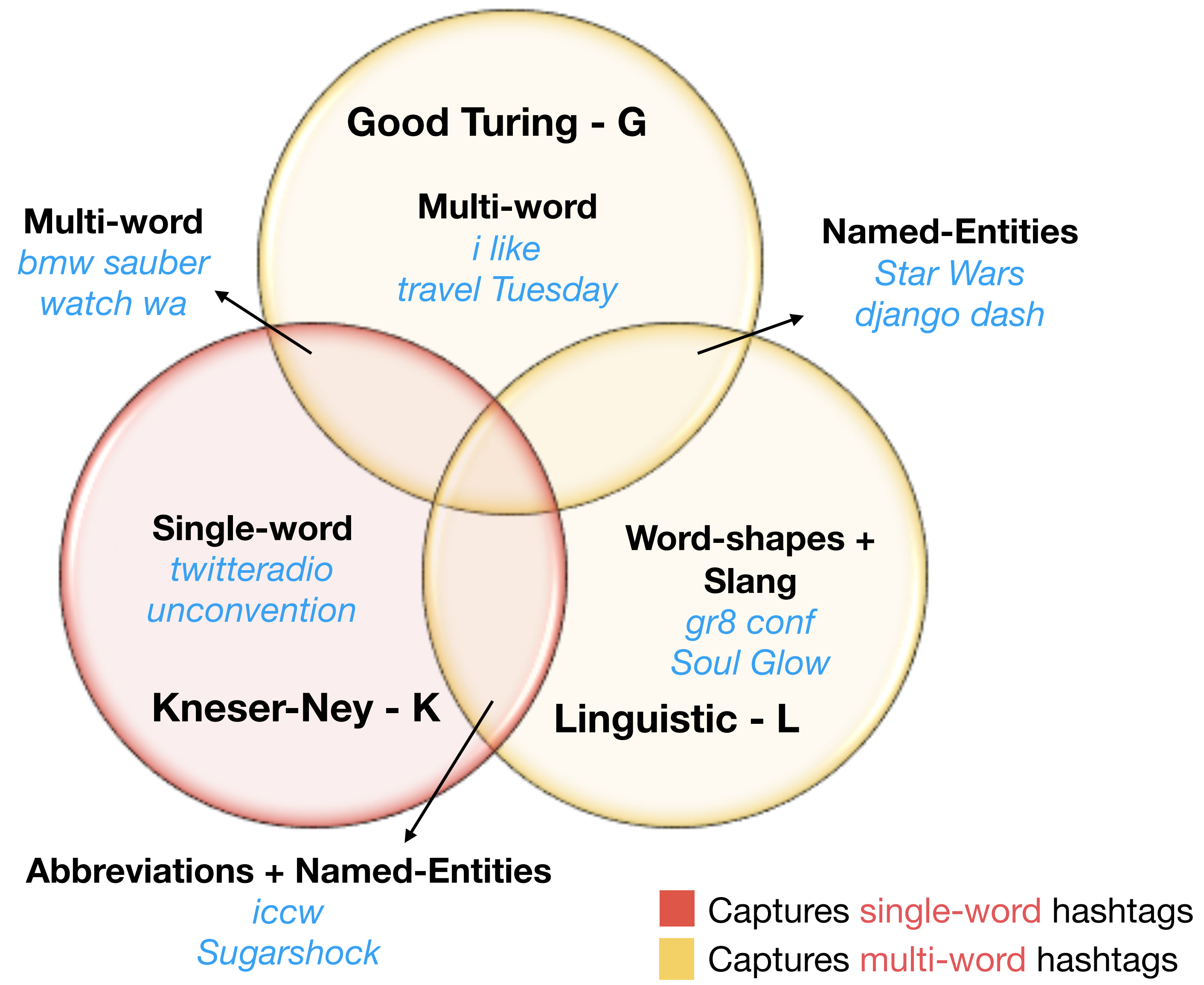
# bea twitty party ✓

# beat witty party ✗

# be at witty party ✗

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.



Good Turing - G

Multi-word
i like
travel Tuesday

Multi-word
bmw sauber
watch wa

Named-Entities
Star Wars
django dash

Single-word
twitteradio
unconvention

Word-shapes + Slang
gr8 conf
Soul Glow

Kneser-Ney - K

Linguistic - L

Abbreviations + Named-Entities
iccw
Sugarshock

🟥 Captures single-word hashtags

🟨 Captures multi-word hashtags

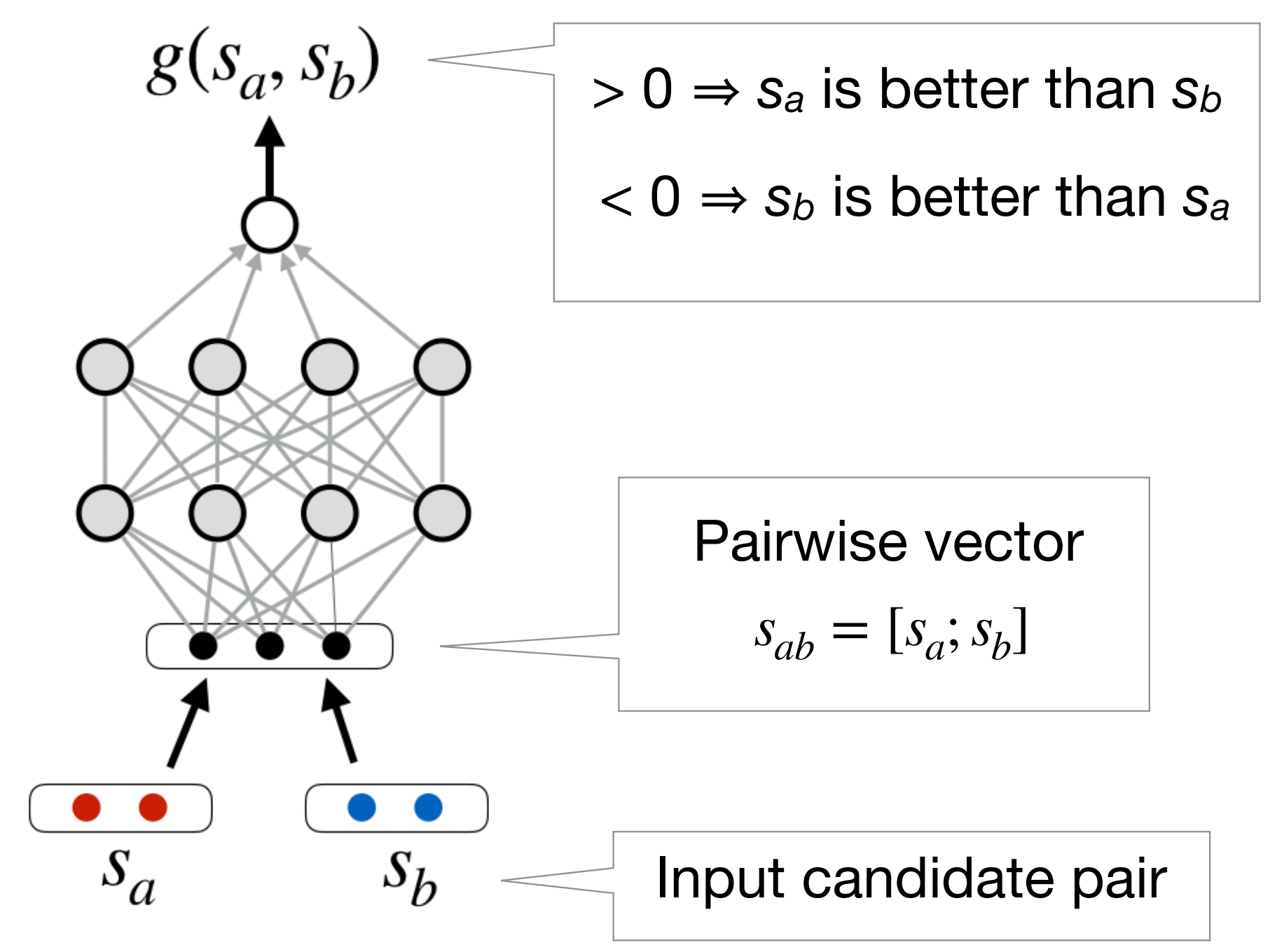
C2: Segmentation as Pairwise Ranking

h: #songsongaddafisitunes → *s*: songs on Gaddafi's itunes*

1. Extract top-k potential candidates using language model.

s₁: *song song addaf is itunes*
s₂: *songs on gaddafis itunes*
s₃: *songs on gaddaf is itunes*

2. Given two candidates (s_a,s_b), predict the better segmentation of the two.



> 0 ⇒ s_a is better than s_b

< 0 ⇒ s_b is better than s_a

Pairwise vector
$s_{ab} = [s_a; s_b]$

Input candidate pair

s₁: *song song addaf is itunes*
s₂: *songs on gaddafis itunes*

s₁: *song song addaf is itunes*
s₃: *songs on gaddaf is itunes*

s₂: *songs on gaddafis itunes*
s₁: *song song addaf is itunes*

s₂: *songs on gaddafis itunes*
s₃: *songs on gaddaf is itunes*

…

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

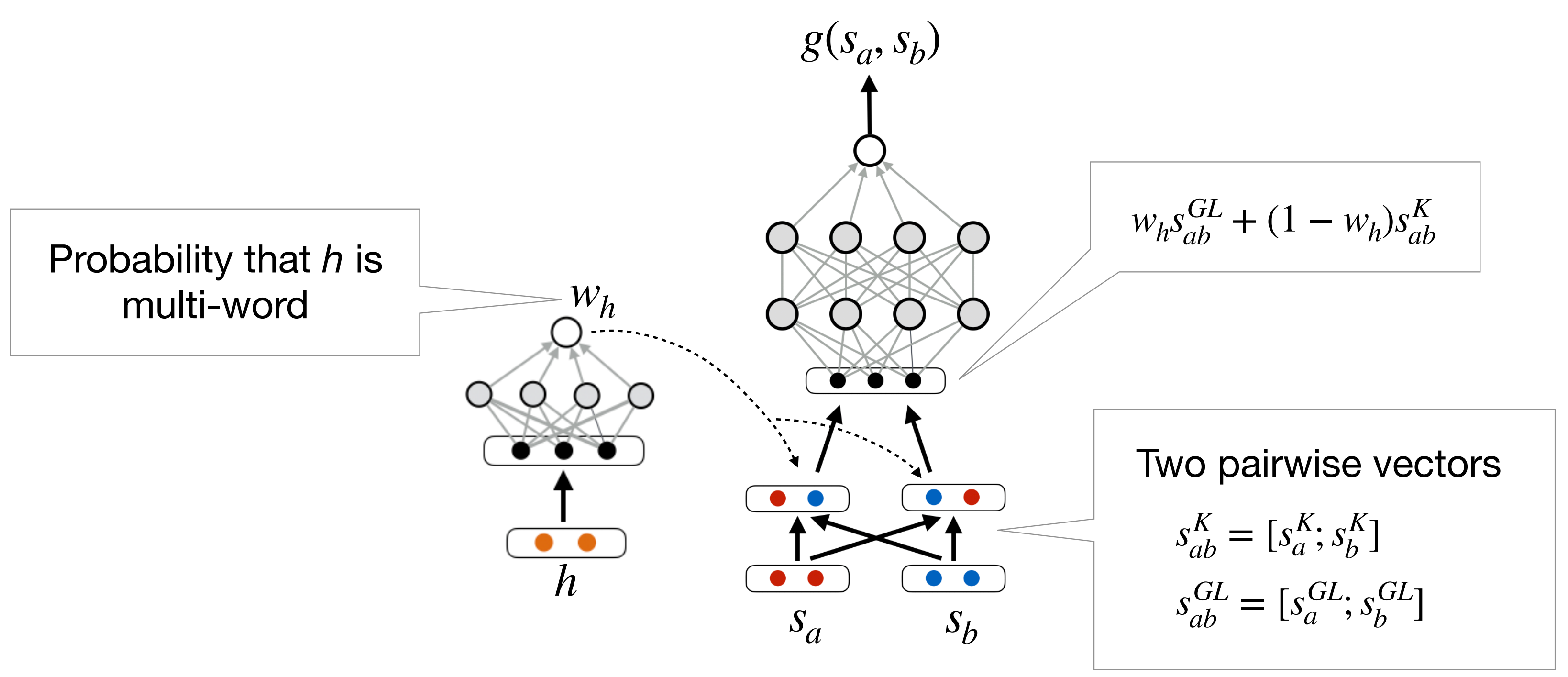
$$g^*(s_a, s_b) = sim(s_a, s^*) - sim(s_b, s^*)$$

3. Combine pairwise scores and rerank

s₂: *songs on gaddafis itunes*
s₃: *songs on gaddaf is itunes*
s₁: *song song addaf is itunes*

Multi-task Learning

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



Probability that *h* is multi-word

$w_h s_{ab}^{GL} + (1 - w_h) s_{ab}^{K}$

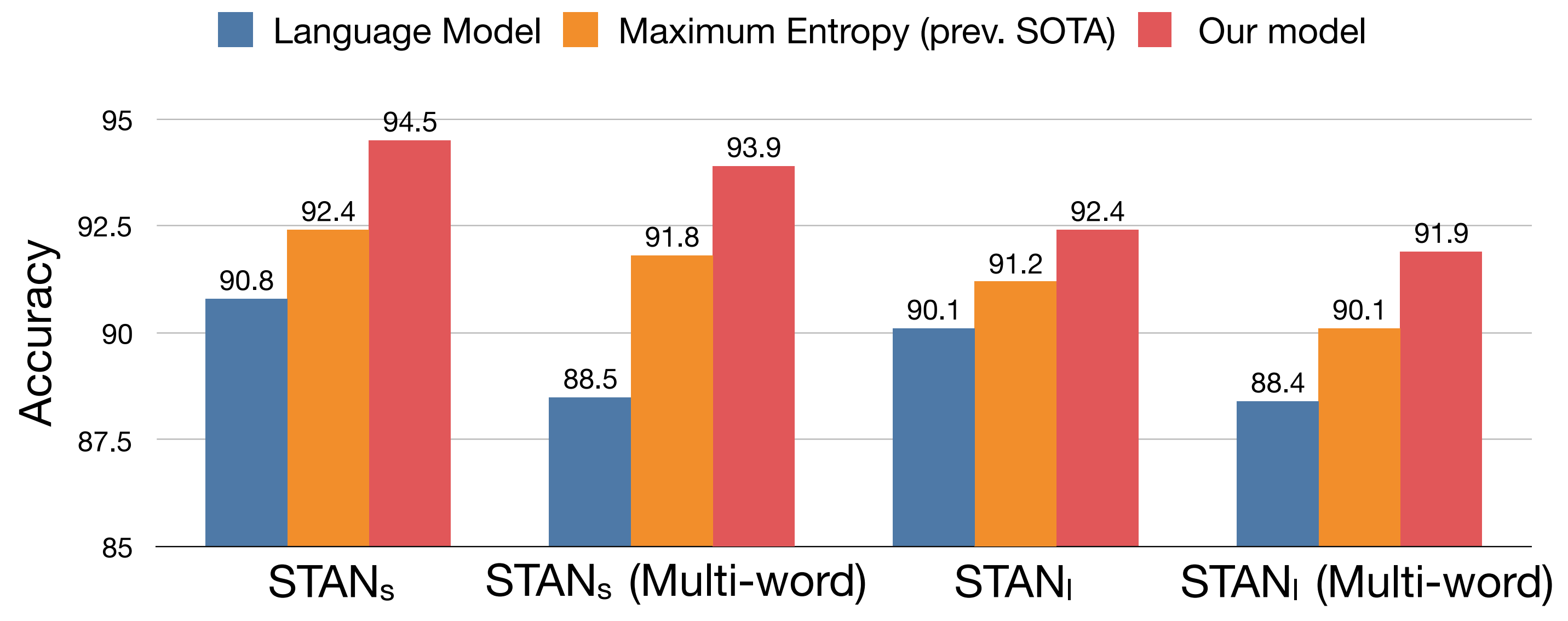
Two pairwise vectors
$s_{ab}^{K} = [s_a^K; s_b^K]$
$s_{ab}^{GL} = [s_a^{GL}; s_b^{GL}]$

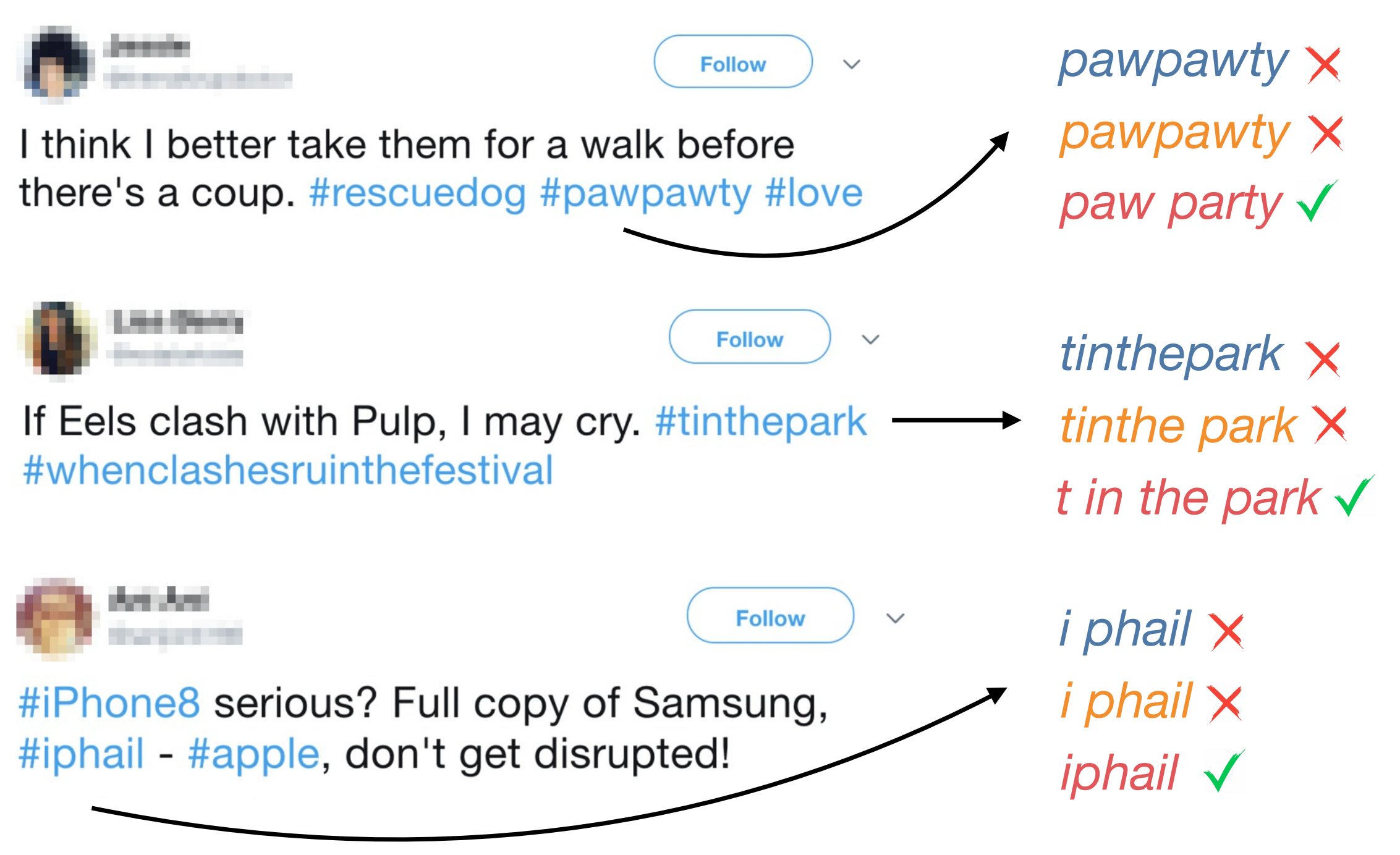
$$L_{multitask} = \lambda_1 L_{MSE} + \lambda_2 L_{BCE}$$

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

Experiments

- 2518 train + 629 development hashtags.
- Test: *STAN*ₛ (1,108) and *STAN*ₗ (9,447 hashtags).





I think I better take them for a walk before there's a coup. #rescuedog #pawpawty #love

pawpawty ✗
pawpawty ✗
paw party ✓

If Eels clash with Pulp, I may cry. #tinthepark #whenclashesruinthefestival

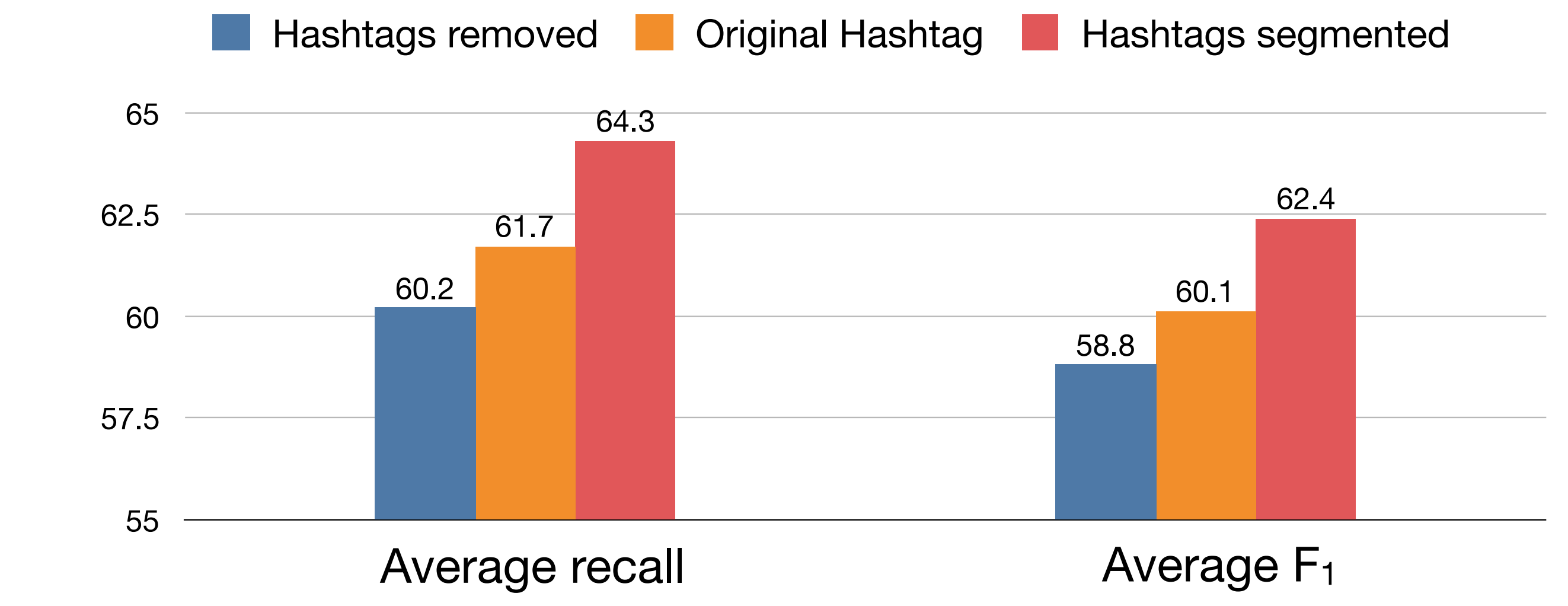
tinthepark ✗
tinthe park ✗
t in the park ✓

#iPhone8 serious? Full copy of Samsung, #iphail - #apple, don't get disrupted!

i phail ✗
i phail ✗
iphail ✓

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