Disambiguating False-Alarm Hashtag Usages in Tweets for Irony Detection

Hen-Hsen Huang[1], Chiao-Chen Chen[1], and Hsin-Hsi Chen[12]
[1]Department of Computer Science and Information Engineering,
National Taiwan University, Taipei, Taiwan
[2]MOST Joint Research Center for AI Technology and All Vista Healthcare, Taipei, Taiwan
ACL 2018

Agenda

- Issue of false-alarm self-labeled data
- Dataset
- Disambiguation of hashtag usages
- Irony Detection
- Conclusions

Self-Labeled Data

- Large amount of self-labeled data available on the Internet are popular research materials in many NLP areas.

- Metadata such as tags and emoticons given by users are considered as labels for training and testing learning-based models.

- The tweets with a certain types of hashtags are collected as self-label data in a variety of research works.

  - Sentiment analysis

  - Stance detection

  - Financial opinion mining

  - **Irony detection**

Irony Detection with Hashtag Information

- It is impractical to manually annotate the ironic sentences from randomly sampled data due to the relatively low occurrences of irony.

- Alternatively, collecting the tweets with the hashtags like #sarcasm, #irony, and #not becomes the mainstream approach.

@Anonymous doing a great job... #not What do I pay my extortionate council taxes for? #Disgrace #OngoingProblem http://t.co/FQZUUwKSoN

@Anonymous doing a great job... What do I pay my extortionate council taxes for? #Disgrace #OngoingProblem http://t.co/FQZUUwKSoN

False-alarm Issue

- The reliability of the self-labeled data is an important issue.
- Misused Hashtag
  - Not all tweet writers know the definition of irony

BestProAdvice @Anonymous More clean OR cleaner, never more cleaner. #irony

Hashtags Functioning as Content Words

- A hashtag in a tweet may also function as a content word in its word form.

- The removal of the hashtag can change the meaning of the tweet, or even make the tweet grammatically incomplete.

The #irony of taking a break from reading about #socialmedia to check my social media.

Research Goal

- Two kinds of unreliable data are our targets to remove from the training data for irony detection.
  - The tweets with a misused hashtag
  - The tweets in which the hashtag serves as a content word,
- Compared to general training data cleaning approaches, our work leverages the characteristics of hashtag usages in tweets.
- With small amount of golden labeled data, we propose a neural network classifier for pruning the self-labeled tweets, and train an ironic detector on the less but cleaner instances.

Dataset

- The ground-truth is based on the dataset released for SemEval 2018 Task 3.

- The hashtag itself has been removed in the SemEval dataset.
  - The hashtag information, the position and the word form of the hashtag (i.e., not, irony, or sarcasm), is missing.

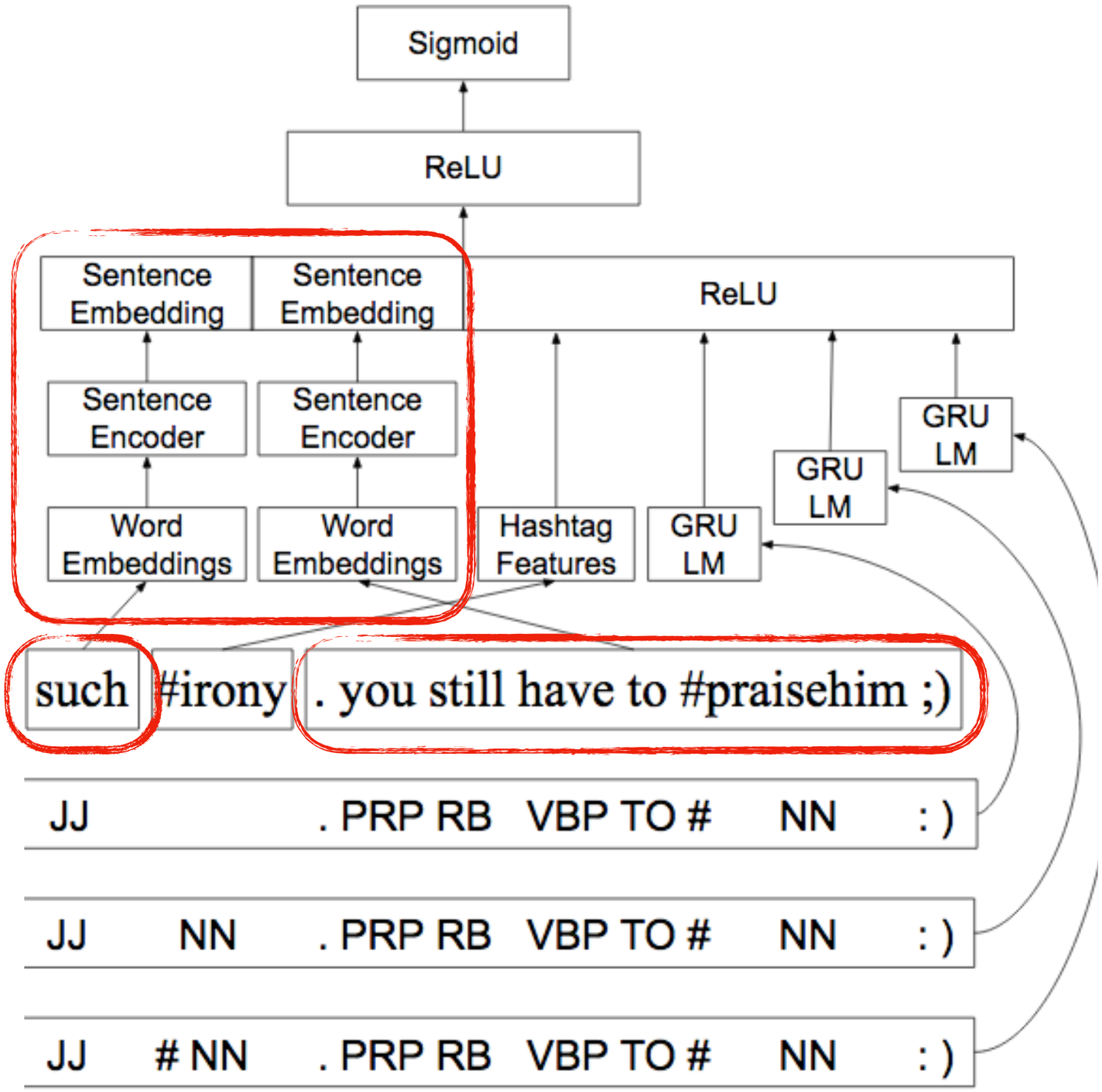
- We recover the original tweets by using Twitter search.

Hashtag	False-Alarm	Irony	Total
#not	196	346	542
#sarcasm	46	449	495
#irony	34	288	322
Total	276	1,083	1,359

Disambiguation of Hashtags

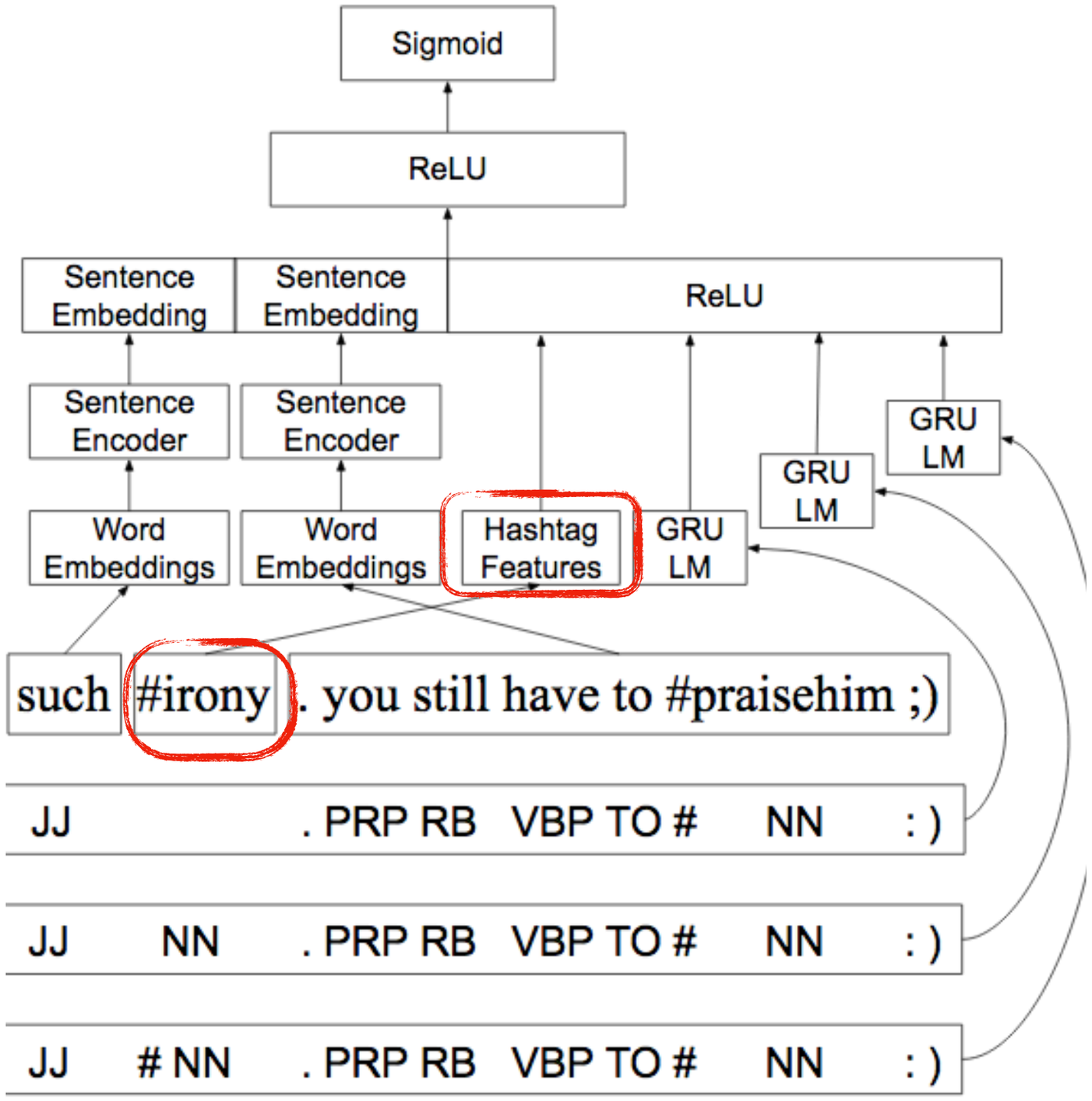
- **Word sequences** of the context preceding and following the targeting hashtag are separately encoded by neural network sentence encoders.

- CNN
- GRU
- Attentive GRU



Disambiguation of Hashtags

- **Handcrafted features**
  - Lengths of the tweet in words and in characters.
  - Type of the target hashtag
  - Number of all hashtags in the tweet.
  - If the targeting hashtag is the first/last token in the tweet.
  - If the targeting hashtag is the first/last hashtag in the tweet
  - Position of the targeting hashtag



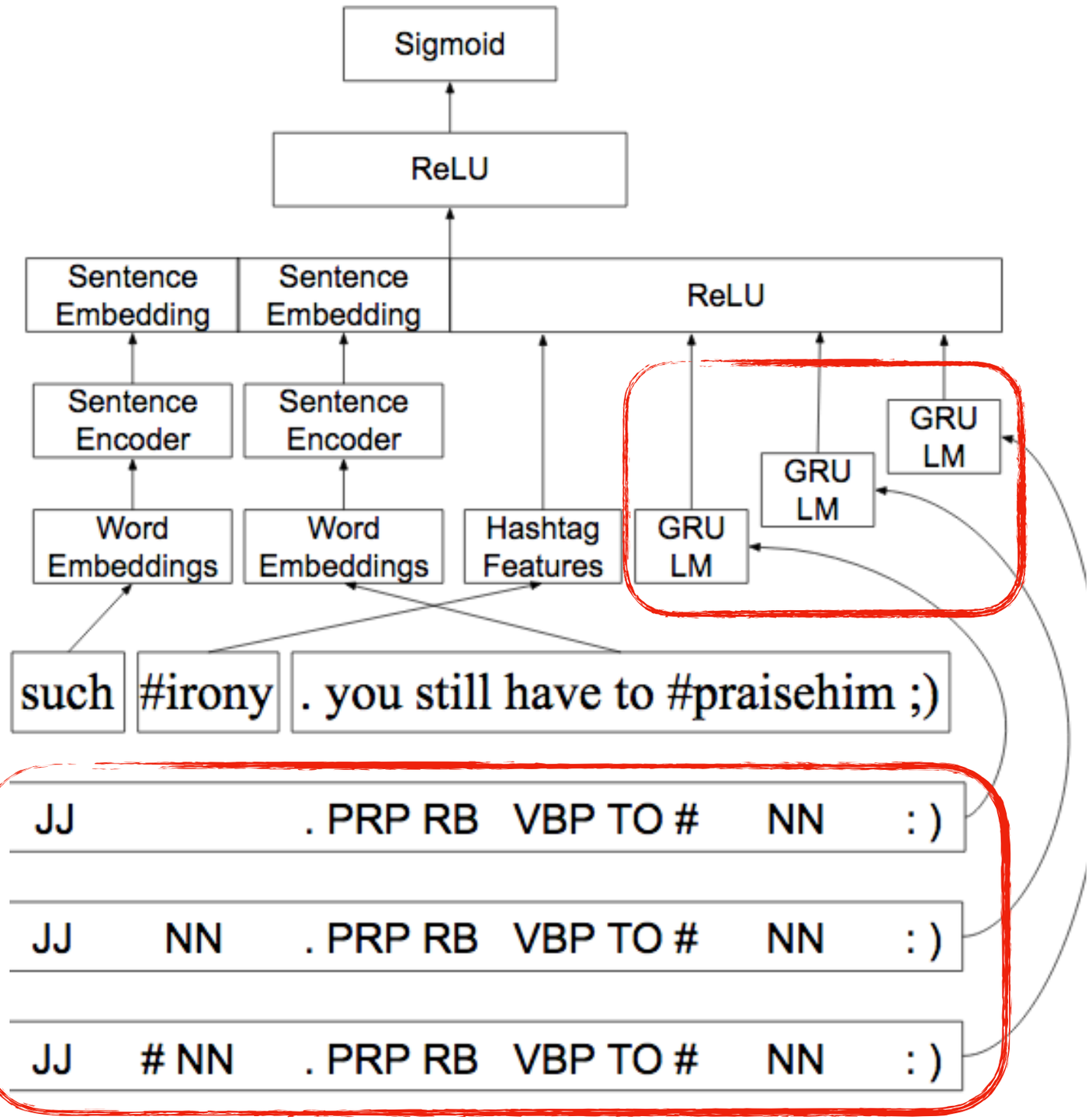
Disambiguation of Hashtags

- A tweet will be more grammatical complete with only the hash symbol removed if the hashtag is also a content word.

- On the other hand, the tweet will be more grammatical complete with the whole hashtag removed since the hashtag is a metadata.

- **GRU-based language model on the level of POS tagging** is used to measure the grammatical completeness of the tweet with and without the hashtag.

  - Remove the whole hashtag removed.
  - Remove the hash symbol # only.
  - The original tweet.



Results of Hashtag Disambiguation

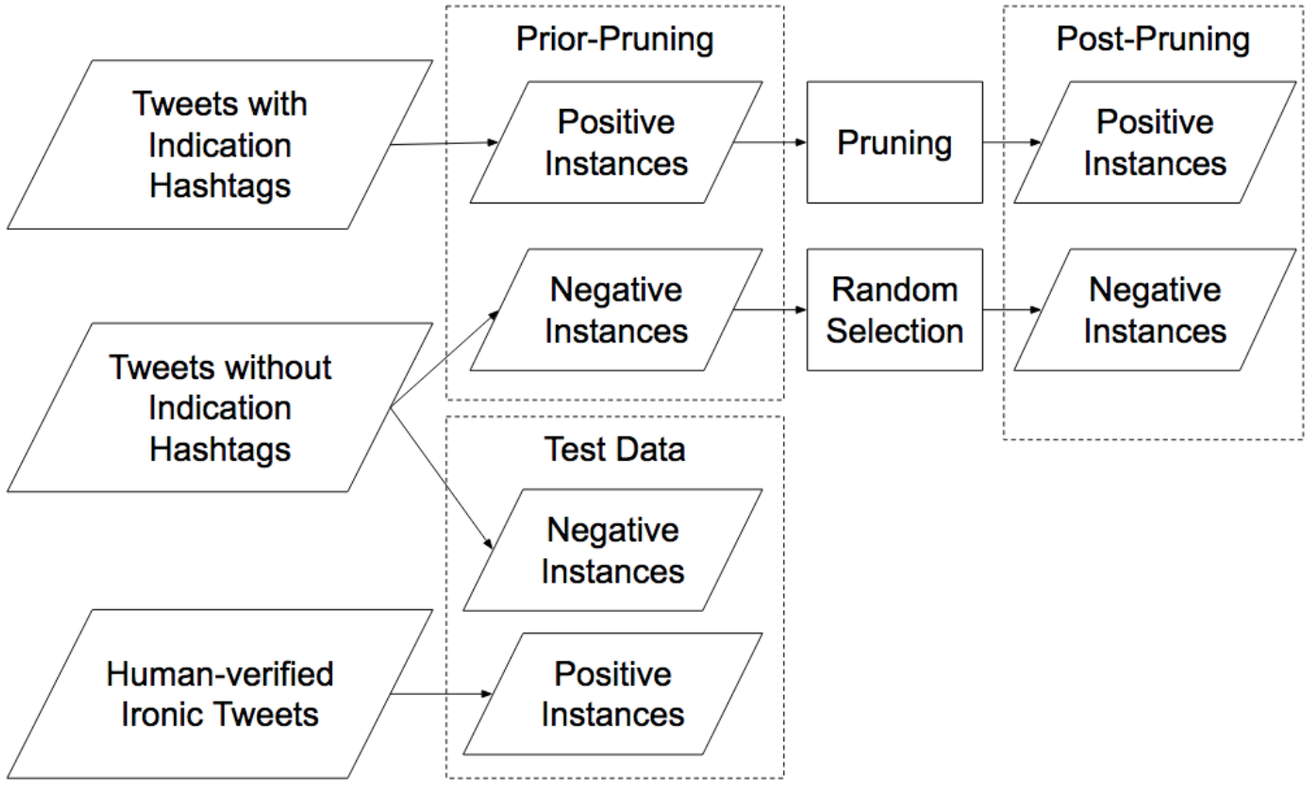
- By integrating various kinds of information, our method outperforms all baseline models no matter which encoder is used. The best model is the one integrating the attentive GRU encoder, which is significantly superior to all baseline mod- els ($p < 0.05$), achieves an F-score of 88.49%.

- The addition of language model significantly improves the performance ($p < 0.05$).

Model	Encoder	Precision	Recall	F-score
LR	N/A	91.43%	75.81%	82.89%
CNN	N/A	89.16%	56.97%	69.52%
GRU	N/A	90.75%	77.01%	83.32%
Att. GRU	N/A	87.97%	79.69%	83.63%
Our Model	CNN	90.35%	83.84%	86.97%
Our Model	GRU	90.90%	78.39%	84.18%
Our Model	Att.GRU	90.86%	86.24%	88.49%
Without LM	Att.GRU	88.17%	80.52%	84.17%

Training Data Pruning for Irony Detection

- We employ our model to prune self-labeled data for irony detection.
  - A set of tweets that contain indication hashtags as (pseudo) positive instances
  - A set of tweets that do not contain indication hashtags as negative instances.
- Our model is performed to predict whether it is a real ironic tweet or false-alarm ones, and the false-alarm ones are discarded.



Results on Irony Detection

- We implement a state-of-the-art irony detector, which is based on attentive-RNN classifier, and train it on the prior- and the post-pruned training data.

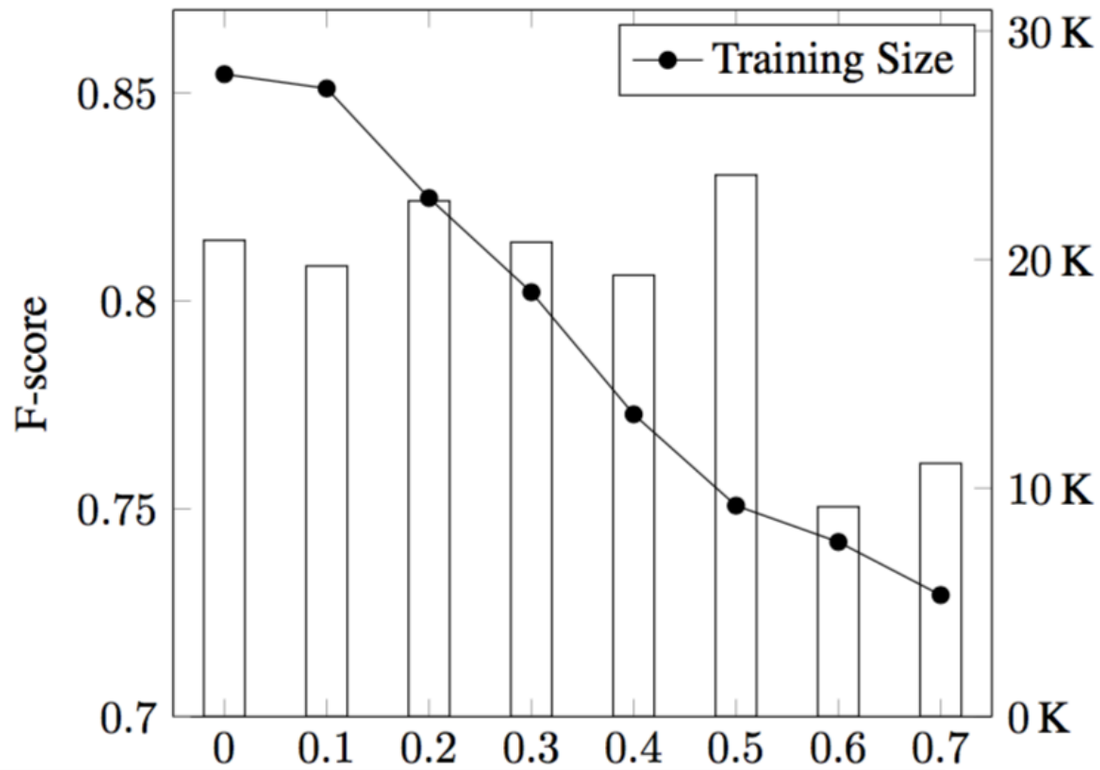
- The irony detection model trained on the less, but cleaner instances significantly outperforms the model that is trained on all data ($p < 0.05$).

- The irony detector trained on the small genuine data does not compete with the models that are trained on larger amount of self-labeled data.

Data	Size	Precision	Recall	F-score
Prior-Pruning	28,110	79.04%	84.05%	81.46%
Post-Pruning	9,234	80.83%	85.35%	83.03%
Human Verified	2,166	86.35%	66.70%	75.26%

Different Threshold Values for Data Pruning

- We can sort all self-labeled data by their calibrated confidence and control the size of training set by adjusting the threshold.
  - The higher the threshold value is set, the less the training instances remain.
- The best result achieved by the irony detector trained on the 9,234 data filtered by our model with the default threshold value (0.5).
- This confirms that our model is able to select useful training instances in a strict manner



The bullet symbol (•) indicates the size of training data, and the bar indicates the F-score achieved by the irony detector trained on those data.

Conclusions

- We make an empirically study on an issue that is potentially inherited in a number of research topics based on self-labeled data.

- We propose a model for hashtag disambiguation. For this task, the human-verified ground-truth is quite limited. To address the issue of sparsity, a novel neural network model for hashtag disambiguation is proposed.

- The data pruning method is capable of improving the performance of irony detection, and can be applied to other work relied on self-labeled data.