Recruiting @ Stanford – Is there free food? & Generate that subject line

Leo Mehr
leomehr@stanford.edu
Stanford University | CS229 | Spring 2019

Problem Statement

Stanford Computer Forum Recruiting Emails

1. Is there free food at this event?
2. Generate subject line

Dataset

- Stanford Computer Forum recruiting emails
- Every email since April 2007
- 6123 total emails, 16MB of text, 464k lines
- 80:10:10 split for train/dev/test
- test = 2015 and May 2019 emails

Tasks + Evaluation Metrics

emails $E = \{e^{(1)}, e^{(2)}, ..., e^{(n)}\}$

subject $s^{(i)}$

body $b^{(i)}$

has_food $f^{(i)} \in \{0, 1\}$

Metrics:
1. Accuracy
2. Precision

gen_subject $\tilde{s}^{(i)}$

Metrics:
1. Exact Match (EM)
2. Precision*
3. F1*

*using bag of words

gen_subject

Model	Split	F1	Precision	
subject-bigram	dev	0.097	0.108	(EM 0)
	test	0.112	0.124	
body-unigram	dev	0.085	0.111	
	test	0.107	0.145	
body-bigram	dev	0.102	0.159	
	test	**0.118**	**0.184**	

Language model $p(s_{j+1}^{(i)} \mid s_1^{(i)}, ..., s_j^{(i)}, b^{(i)})$

Subject bigram $p(s_{j+1}^{(i)} \mid s_j^{(i)})$

Body unigram $p(s_{j+1}^{(i)} \mid b^{(i)}) \approx \prod_{k=1}^{\text{len}(b^{(i)})} p(s_{j+1}^{(i)} \mid b_k^{(i)})$

Body bigram $p(s_{j+1}^{(i)} \mid s_j^{(i)}) \cdot p(s_{j+1}^{(i)} \mid b^{(i)}) \approx p(s_{j+1}^{(i)} \mid s_j^{(i)}) \cdot \prod_{k=1}^{\text{len}(b^{(i)})} p(s_{j+1}^{(i)} \mid b_k^{(i)})$

Subject bigram
- REMINDER Facebook Tech Talk Wed 10 13 30pm in Gates Building room 219
- Recruiting Digest For Oct 26 12pm in Gates 104

Body unigram
- Tech Info Digest Stanford Intern Digest 30pm 202 in of Info
- Gates Building and Disney
- LinkedIn Gates Gates

Body bigram
- FINAL REMINDER Computer Forum
- Reminder Uber Info Session 30pm in Gates 104
- REMINDER RSVP by March

has_food

NBC = Naive Bayes Classifier

$$p(s^{(i)} \mid f^{(i)}) \approx \prod_{j=1}^{\text{len}(s^{(i)})} p(s_j^{(i)} \mid f^{(i)})$$

SVM = Support Vector Machine

$$\ell(\theta) = \max\left(0, 1 - f^{(i)}(\theta^T s^{(i)})\right) + \lambda\|\theta\|^2$$

Model	Split	Accuracy	Precision
NBC	dev	0.893	0.905
	test	**0.878**	0.876
SVM	dev	0.902	0.904
	test	0.861	**0.885**

NBC test correct examples
- (-) Recruiting Digest for the week of February 2
- (-) REMINDER RSVP by 2/8: Pinterest Company Tour on Friday, 2/13
- (+) FINAL REMINDER: Spokeo Info Session, February 2 @ 5pm - 6pm @ Gates\n Building, room 104
- (+) SAVE THE DATE: SAP Info Session, April 9, 6:30pm - 7:30pm, Gates\n Building, room 104

NBC test incorrect examples
- (true label +) Reminder: RSVP by February 26: Workday, SF Company Tour on Friday,\n March 3
- (+) LinkedIn Data Infra Intern outreach program
- (-) SAVE the DATE: Chopper Trading Info Session - Wed, January 16 @ 6:30pm\n\tin Gates 104

NBC Predictive Features

Food Unlikely	Score	Score	Food Likely
08	-10.58	-3.274	reminder
0pm	-10.58	-3.192	104
100	-10.58	-3.067	info
100k	-10.58	-3.041	gates
103	-10.58	-3.041	session
106a	-10.58	-3.024	in

SVM Predictive Features

No Food	Score	Score	Food
from	-1.154	1.335	joblunch
drop	-0.938	1.046	coming
group	-0.902	0.974	company
counseling	-0.793	0.866	lunch
fairs	-0.757	0.866	dinner
oracle	-0.757	0.829	splunk

has_food supervision signal

- Cannot hand-label 6000+ emails
- Wrote a labeling function to infer $f^{(i)}$ from $b^{(i)}$
- Randomly sampled 100, labeled by hand to compare
- Labeling function has:
  - 97% accuracy, 98% precision, 96% recall
- Well-balanced dataset, 51.3% positive labels

Discussion

- In has_food, NBC and SVM performed similarly well
  - Interesting to compare the most/least predictive features between the two. Illustrates how they function differently.
  - A consistent drop from dev to test perf, potentially illustrating overfitting to the dev set or a slightly unrepresentative test set.
  - Error analysis uncovered issues with my labeling function!
- gen_subject is a hard task. All 3 models perform poorly, EM all 0.
  - With a precision of nearly 20%, body-bigram performed best, but recall was very poor. Examples illustrates that the bigram models are believable, while unigram is nonsense.
  - F1 scores quite bad. Words such as "Reminder" boost scores.

Future Work

- Use Neural Networks!
  - LSTM Networks with Attention capture long-range dependencies
  - Seq-to-seq architecture for text summarization
  - Pre-trained word embeddings such as ELMO or GloVE
- Incorporate supplemental data, such as event attendance counts and how many students read an email vs. delete it

References

- Michele Banko, Vibhu O Mittal, and Michael J Witbrock. Headline generation based on statistical translation. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pages 318–325. Association for Computational Linguistics, 2000.
- Mark Dredze, Hanna M Wallach, Danny Puller, and Fernando Pereira. Generating summary keywords for emails using topics. In *Proceedings of the 13th international conference on IntelLigent user interfaces*, pages 199–206. ACM, 2008.
- Bryan Klimt and Yiming Yang. The enron corpus: A new dataset for email classification research. In *European Conference on Machine Learning*, pages 217–226. Springer, 2004.
- Stanford "Recruiting" Email list. https://mailman.stanford.edu/mailman/listinfo/recruiting.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002.
- Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, 2016.
- Jiwei Tan, Xiaojun Wan, and Jianguo Xiao. From neural sentence summarization to headline generation: A coarse-to-fine approach. In *IJCAI*, pages 4109–4115, 2017.