E-mail Data Extraction Using Multilayered Network ~~and binding network~~

10-701 Course Project Haw-Shiuan Chang (hawshiuc) TA: Jayant Krishnamurthy

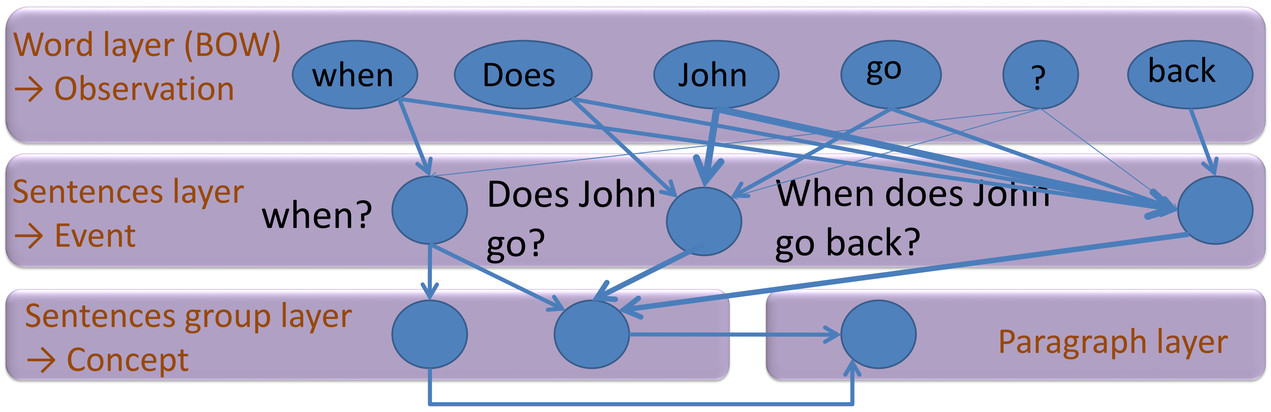
Goal: Unsupervised data extraction

- Can we get more information from an e-mail than bag of word?
- Dealing with an article, human learn more abstract features (say, meaning of sentences or paragraphs). Can we do that on machine?

Method (different from the midway report):

- Assumption:

If two words appear in one sentence, they must be somehow related.



Algorithm:

- For each sentence

  For layer(i), activate layer(i+1), clamp off units with activation < **t**

  if(∃unit j in layer(i+1) covers 80% new sets in layer(i))

  update w_j from i to i+1 // freq words have less weights

  else (Reuse weights like human being)

  clamp on a new unit and update its weights

  end // we keep the (activation value of last sentence)*0.2

  // meaning of coverage is the number of words biased by weights

- For each paragraph, we treat each "sentence group" as a "word"

Application:

Data extraction, Point indicator, (Object recognition?)

Drawback:

Very time consuming

Experiments:

- Classify folder task on Enron e-mail dataset.
- According to time line, first train 100/200 e-mails then test next 100 e-mails in one person's mailbox[1].
- On each layer (including BOW)
  - Normalized 1NN classifier
  - 1NN classifier
  - Naïve Bayes classifier

Results (error rate):

Train 100, t=0.05	BOW	Sentences	Sentences group	Paragraph
Normalized-1NN	0.89	0.86 -> **0.76**	0.86 -> 0.80	**0.84** -> 0.85
1NN	0.77	**0.74**	0.83	0.92
Naïve Bayes	0.78	**0.76**	0.83	0.89
Dimension	6093	12109	14172	**2113**

- Note that minimal error rate is 0.5

Train 200, t=0.1	BOW	Sentences	Sentences group	Paragraph
Normalized-1NN	**0.68**	0.70	0.71	0.72
1NN	0.66	**0.62**	0.66	0.72
Naïve Bayes	0.73	**0.69**	0.70	0.79
Dimension	7998	20102	22911	**3227**

- Note that minimal error rate is 0.22
- Note that decrease t will increase the computational time and performance, so only the results from the smallest t are shown here

Reference:

[1] Ron Bekkerman, Andrew McCallum, Gary Huang: Automatic Categorization of Email into Folders: Benchmark Experiments on Enron and SRI Corpora