

Linguistics 384

Bag of words exercise

Step 1: Read the following 5 pairs of English-Chinese sentences. The code (W_i) following each Chinese word is used to help you identify the same Chinese word in different Chinese sentences.

It is proven that it is healthy to celebrate birthdays.

事实(W1) 证明(W2) 庆祝(W3) 生日(W4) 是(W5) 健康(W6) 的(W7)。

Statistics show that those people who celebrate the most birthdays become the oldest.

统计数字(W8) 表明(W9) 那些(W10) 庆祝(W3) 生日(W4) 最多(W11) 的(W7) 人(W12) 变得(W13) 最老(W14)。

So celebrate your birthday every year!

所以(W15), 每年(W16) 都(W17) 庆祝(W3) 你的(W18) 生日(W4) 吧(W19)!

Cited from Dr. Birthday's PhD dissertation *Celebrate Your Birthdays to Get Healthier!*

引自(W20) 生日(W4) 博士(W21) 的(W7) 博士(W21) 论文(W22) 庆祝(W3) 你的(W18) 生日(W4) 来(W23) 变得(W13) 更(W24) 健康(W6)!

Dissertation will be published by Birthday Press.

论文(W22) 将(W25) 由(W26) 生日(W4) 出版社(W27) 出版(W28)。

Step 2: Using the bag of words method, calculate the probability of *celebrate* being translated into each of the possible candidate Chinese words based on the training data you read just now. Note that you should only consider the Chinese words W1, W2, etc. not the individual Chinese characters. There are extra rows in the table.

Code of Candidate Chinese Word	Frequency	Probability
W1	1	1/37
W2	1	1/37
W3	4	4/37 !
W4	5	5/37 !
W5	1	1/37
W6	2	2/37
W7	3	3/37
W8	1	1/37
W9	1	1/37
W10	1	1/37
W11	1	1/37
W12	1	1/37
W13	2	2/37
W14	1	1/37
W15	1	1/37
W16	1	1/37
W17	1	1/37
W18	2	2/37
W19	1	1/37
W20	1	1/37
W21	2	2/37
W22	1	1/37
W23	1	1/37
W24	1	1/37
Total	37	

W4 has the highest probability of alignment with *celebrate* according to the bag of words model. But note that W4 also occurs in the last sentence, which does not include the word *celebrate* in the corresponding English sentence and thus was ignored by the simple bag of word model. In fact, the translation of W4 turns out to be *birthday* so the correct alignment for *celebrate* is W3, which has the second highest probability.