NLP
Project 1

TEAM 10 | HOTEL? TRIVAGO!

B03902101 楊力權 B03902101@NTU.EDU.TW
R05922038 黃郁庭 R05922038@NTU.EDU.TW
B03902086 李鈺昇 B03902086@NTU.EDU.TW

Outline

- Our best method on Kaggle
- Other methods

Outline

- Our best method on Kaggle
- Other methods

Aspect Extraction

- Why aspect extraction?

Train word2vector model

- polarity_review.txt
- aspect_review.txt
- test_review.txt

Similarity threshold

5 aspects need their own unique threshold, here are two ways below:

- Every threshold can be decided by calculate the average similarity of a whole sentence which contains the aspect in aspect_review.txt
- Every threshold can be decided by calculate the best aspect accuracy of aspect in aspect_review.txt

Test

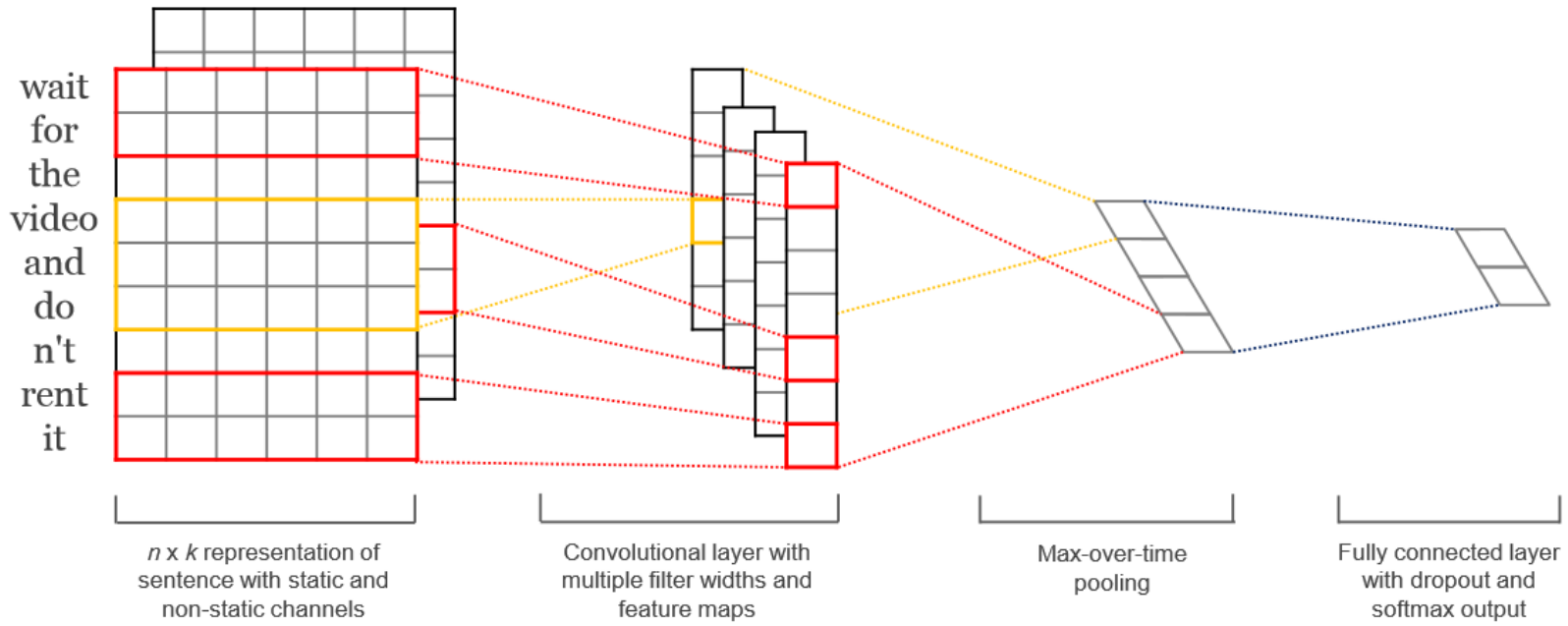
- Cut a sentence in phrases
- Use the word2vec model to check the word's similarity with 5 aspects [服務、環境、價格、交通、餐廳]
- If a word's (or the whole sentence's average) similarity is bigger than the aspect's threshold, we can assume that the sentence contains the aspect.

Convolutional Neural Networks

- 一個model預測一種aspect的positive or negative (共5個model)
- label：
  - 1 : positive
  - -1 : negative
  - 0 : unknown/none
- data：
  - 將review中的每個word的vector串接起來 (斷詞、remove stop words)
  - 限定每個review最多只能串接300個word (不足補0、超過不計)
  - EX.假設每個word的vector大小為1*150，則data的大小為300*150

Convolutional Neural Networks (cont.)

- Convolutional Neural Networks for Sentence Classification, Yoon Kim
  https://arxiv.org/abs/1408.5882



Convolutional Neural Networks (cont.)

- training set：需要先預測polarity_review.txt的aspect (258,003)
- validation set：Aspect_review.txt (200)
  - val_acc與Kaggle的public/private score差不多
  - 通常val_acc表現較好的model會出現在前5個epoch中

Convolutional Neural Networks (cont.)

Public	private	備註
0.81567	0.79402	Cnn-rand, 300*120, 使用epoch=8的model
0.81336	**0.80437**	Cnn-rand, 300*120, 使用val_acc最高的model
0.81452	0.79632	Cnn-non-static, 300*300, 使用epoch=8的model

	服務	環境	價格	交通	餐廳
Val_acc	0.82	**0.73**	0.83	0.82	**0.85**

Outline

- Our best method on Kaggle
- Other methods

Other methods - 1

- 前處理
  - 斷詞：jieba
  - Remove stop words：移除掉部分詞性(coreNLP)

    SP、BA、DER、DEV、DEC、DEG、MSP、ETC、DT、SB、 LB、LC、PU、URL、PN、CC、M、CD、P、AS、X、NT
- 找test_aspect.txt的aspect
  - Word2vec：每個review選出與5個aspect最相似的word，再自訂threshold篩選
- 預測test_aspect.txt的polarity
  - Linear Support Vector Classification：用polarity_review.txt訓練出來的模型去預測
- Public：0.73848/ Private：0.70081

Other methods - 2

- For each review R in polarity_review.txt, find the most **similar** review R' in aspect_review.txt, and assign **some of** the aspects of R' to R.
  - Similarity: cosine similarity between the averages of word vectors in the reviews
  - The most likely aspect by aspect term, or all the aspects with the same polarity
- Use classifiers to make predictions given a review R and aspect A:
  - Input: [averages of word vectors in R] concatenated with [one hot encoding of A]
  - Output: {-1, 0, +1}
- KNeighborsClassifier (0.724) and RandomForestClassifier (0.71) perform the best on polarity_review.txt as validation set, but both perform poorly on Kaggle public leaderboard.