Applications using word embeddings

Friday, June 8, 2018 1:58 PM

Sentiment classification

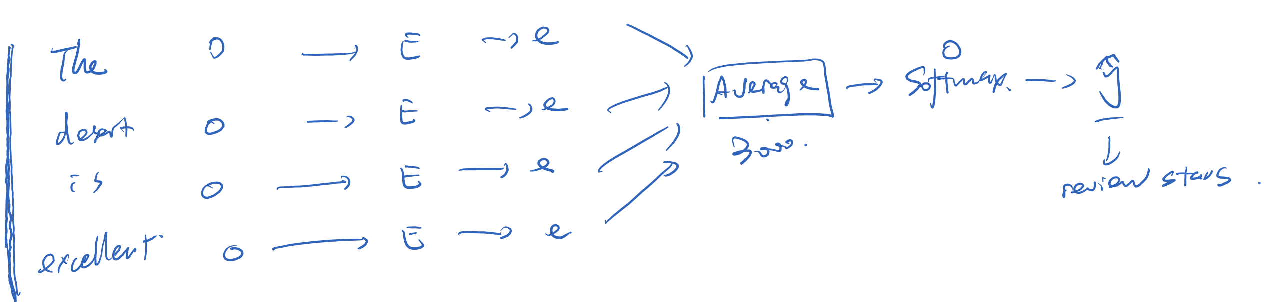
challenge : Don't have a huge labeled dataset.

x → y.

The dissert is excellent ★★★★.

slow servic ★★

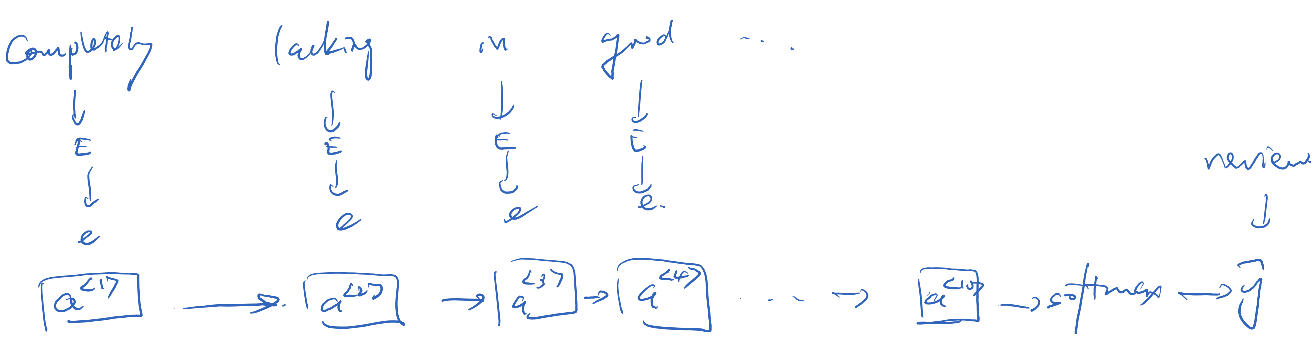
Challenge : you might not have a huge dataset.



problem : ignore order of word.

Completely lacking of ~~good~~ food good ~~~~ not good

Solution : RNN



Because word embedding can be trained from a larger dataset.

this will still do a good job.

Debiasing word embedding

这bias不是技术的bias.

Deminishing bias. [gender, ethnicity ... bias].

Take as doctor | mother as nurse.

Deminishing ...

Father as doctor | mother as nurse.

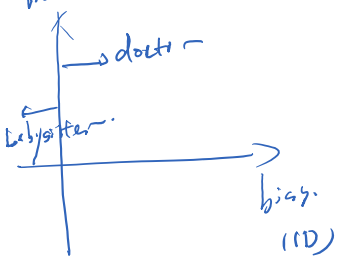
Man programer | woman homemaker)

Word embedding can reflect gender ethnicity ... bias in the text used to train the model.

procedure

1. Identify bias direction

non-direction (good).



$\begin{bmatrix} e_{he} - e_{she} \\ e_{male} - e_{female} \end{bmatrix}$ get direction gender direction.

2. **Neutralization**: for each word that is not directional, project to get rid of bias.

how to decide which word to neutralize? train a classifier to tell which word is definitional

gender is a part by def.

e.g. a directional word grandmother/grandfather

3. Equalize pairs.

grandmother — grandfather.

girl — boy.

The paper

Bolukbasi et al. 2016 Man is to computer program ...