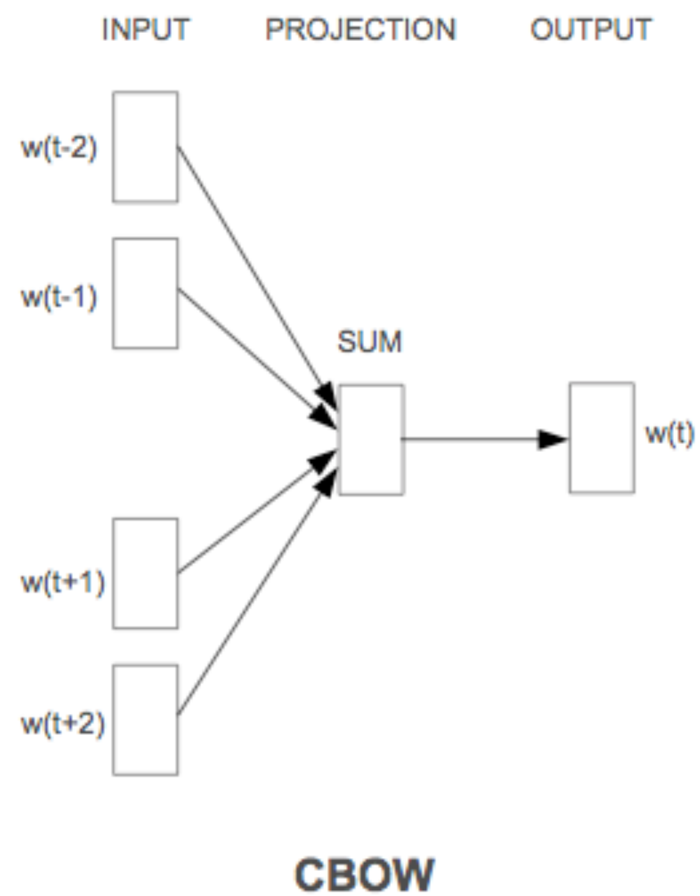
Machine Learning for Language Processing
ACS 2015/16
Stephen Clark
L7: Word Embeddings

UNIVERSITY OF CAMBRIDGE

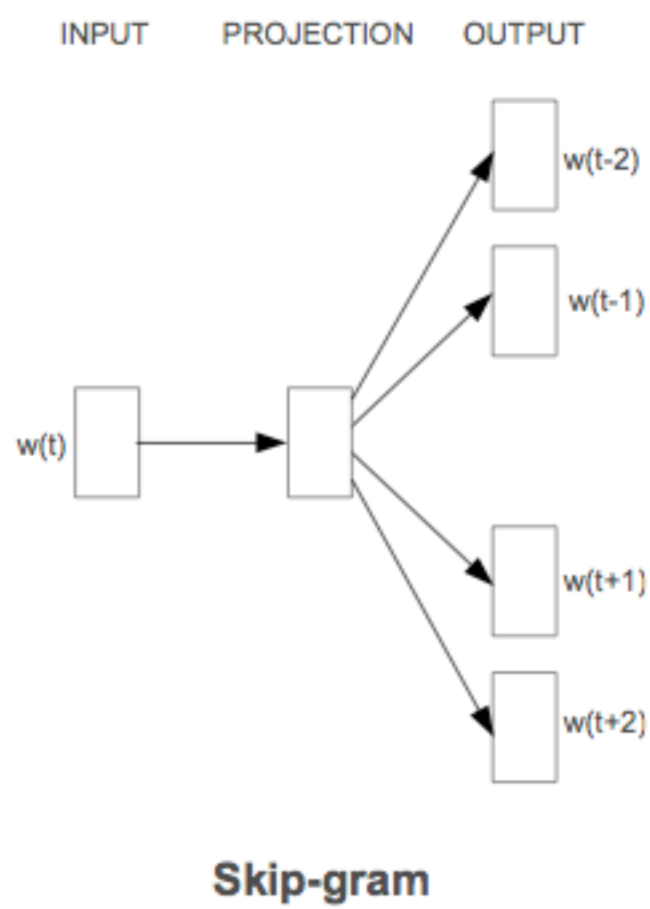
Neural Distributional Models



Continuous bag of words model, from Mikolov et al. 2013

UNIVERSITY OF CAMBRIDGE

Neural Distributional Models



Skip-gram model; picture taken from Mikolov et al. 2013

UNIVERSITY OF CAMBRIDGE

Skip-Gram "Language Modelling"

$$\arg\max_{\theta} \prod_{w \in Text} \prod_{c \in C(w)} p(c|w;\theta)$$

where $C(w)$ is the set of contexts for each word w

$$\arg\max_{\theta} \prod_{(w,c) \in D} p(c|w;\theta)$$

where D is the set of word, context pairs

UNIVERSITY OF CAMBRIDGE

Parameterisation of Skip-Gram

$$p(c|w,\theta) = \frac{e^{v_c \cdot v_w}}{\sum_{c' \in C} e^{v_{c'} \cdot v_w}}$$

where v_c and $v_w \in R^d$ are vector representations for c and w and C is the set of all possible contexts

Negative Sampling

$$\arg\max_{\theta} \prod_{(w,c)\in D} p(D=1|c,w;\theta) \prod_{(w,c)\in D'} p(D=0|c,w;\theta)$$

$$= \arg\max_{\theta} \prod_{(w,c)\in D} p(D=1|c,w;\theta) \prod_{(w,c)\in D'} (1-p(D=1|c,w;\theta))$$

$$= \arg\max_{\theta} \sum_{(w,c)\in D} \log p(D=1|c,w;\theta) + \sum_{(w,c)\in D'} \log\,(1-p(D=1|c,w;\theta))$$

where $D=1$ when (c,w) is from the data and $D=0$ when not and D' is a set of negative word, context pairs

UNIVERSITY OF CAMBRIDGE

Negative Sampling

$$= \arg\max_{\theta} \sum_{(w,c)\in D} \log \frac{1}{1+e^{-v_c \cdot v_w}} + \sum_{(w,c)\in D'} \log \left(1 - \frac{1}{1+e^{-v_c \cdot v_w}}\right)$$

$$= \arg\max_{\theta} \sum_{(w,c)\in D} \log \frac{1}{1+e^{-v_c \cdot v_w}} + \sum_{(w,c)\in D'} \log \left(\frac{1}{1+e^{v_c \cdot v_w}}\right)$$

$$= \arg\max_{\theta} \sum_{(w,c)\in D} \log \sigma(v_c \cdot v_w) + \sum_{(w,c)\in D'} \log \sigma(-v_c \cdot v_w)$$

where $\sigma(x) = \frac{1}{1+e^{-x}}$

UNIVERSITY OF CAMBRIDGE

Sampling Details

For each $(w, c) \in D$ we construct k samples $(w, c_1), \ldots (w, c_k)$ where each c_j is sampled from the unigram distribution$^{\frac{3}{4}}$

The contexts are taken from a window of size N around the target word: $w_{i-N}, \ldots, w_{i-1}, w_{i+1}, \ldots, w_{i+N}$ where N is sampled uniformly between 1 and N for each word

words appearing less than M times are discarded

UNIVERSITY OF CAMBRIDGE

Linguistic Regularities?

$$\overrightarrow{\mathrm{KING}} - \overrightarrow{\mathrm{MAN}} + \overrightarrow{\mathrm{WOMAN}} = \overrightarrow{\mathrm{QUEEN}}$$

Taken from Mikolov et al. 2013

UNIVERSITY OF CAMBRIDGE

Evaluation

- Semantic Relatedness

love	sex	6.77
tiger	cat	7.35
tiger	tiger	10.00
computer	internet	7.58
plane	car	5.77
doctor	nurse	7.00
professor	doctor	6.62
smart	stupid	5.81
stock	phone	1.62

Baroni et al., *Don't count, predict!*

UNIVERSITY OF CAMBRIDGE

Evaluation

- Synonym Detection (TOEFL)

You will find the office at the main **intersection**.
(a) place (b) crossroads (c) roundabout (d) building

Baroni et al., *Don't count, predict!*

UNIVERSITY OF CAMBRIDGE

Evaluation

- Concept Categorization

Concept categorization Given a set of nominal concepts, the task is to group them into natural categories (e.g., *helicopters* and *motorcycles* should go to the *vehicle* class, *dogs* and *elephants* into the *mammal* class). Following previous art, we tackle categorization as an unsupervised clustering task.

Baroni et al., *Don't count, predict!*

UNIVERSITY OF CAMBRIDGE

Evaluation

- Selectional Preferences

Selectional preferences We experiment with two data sets that contain verb-noun pairs that were rated by subjects for the typicality of the noun as a subject or object of the verb (e.g., *peo-ple* received a high average score as subject of *to eat*, and a low score as object of the same

Baroni et al., *Don't count, predict!*

Evaluation

- Analogy

Analogy While all the previous data sets are relatively standard in the DSM field to test traditional count models, our last benchmark was introduced in Mikolov et al. (2013a) specifically to test predict models. The data-set contains about 9K semantic and 10.5K syntactic analogy questions. A semantic question gives an example pair (*brother-sister*), a test word (*grandson*) and asks to find another word that instantiates the relation illustrated by the example with respect to the test word (*granddaughter*). A syntactic question is similar, but in this case the relationship is of a grammatical nature (*work–works*, *speak*... *speaks*). Mikolov

Baroni et al., *Don't count, predict!*

UNIVERSITY OF CAMBRIDGE

Results

- Baroni et al. report very strong results for the “predict” over the “count” vectors
- But see Levy and Goldberg (NIPS, 2014) for a more nuanced picture

Baroni et al., *Don’t count, predict!*