Linguistic Explorations in Real-valued Syntactic Word Vectors (RSV)

UPPSALA UNIVERSITET

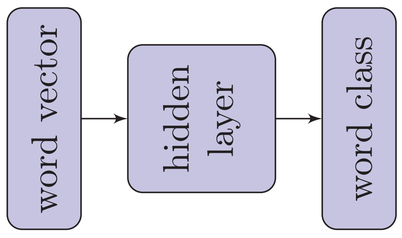
Marc Tang and Ali Basirat

Introduction

- **Word Embeddings**
  - Representing words as elements of a vector space
  - Providing a continuous representation of words
  - Useful for machine learning

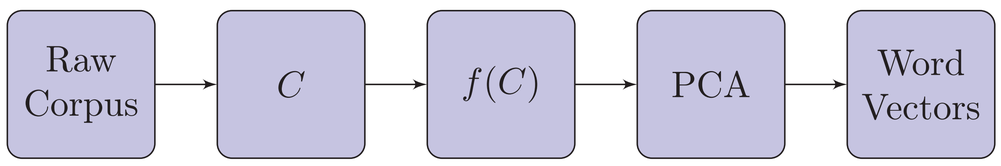
- **Noun Classes**
  - Grammatical gender
  - Common versus proper
  - Mass versus count

- **Linguistic Information in Word Embeddings**
  - Meaningful to NLP tasks
  - Unclear what type of syntactic and semantic information is in embeddings
  - A neural network to explore linguistic information in embeddings



  - The performance of this classifier is considered as an indicator of the presence of information

- **Real-valued Syntactic Word Vectors (RSV)**
  - Build a co-occurrence matrix
  - Perform a power transformation with $p \approx 0.14$
  - Perform PCA on the transformed matrix

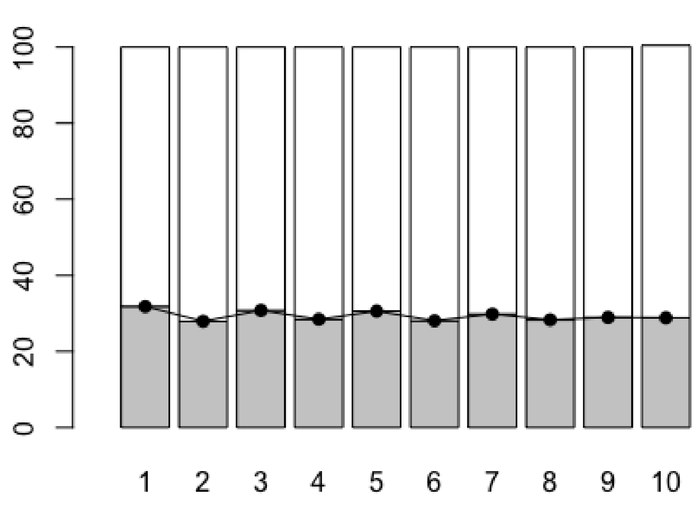


The process of constructing word vectors. C is a co-occurrence matrix

Research Question and Experimental Setup

- **Research Question**: Can RSV provide sufficient information for a *perfect* prediction of grammatical gender in Swedish?

- **Experimental Setup**
  - Language: Swedish
  - Word classes: Grammatical gender (uter - neuter)
  - Information about grammatical gender is collected from SALDO
  - Raw corpus: Språkbanken + Swedish Wikipedia (# $\geq$ 100)



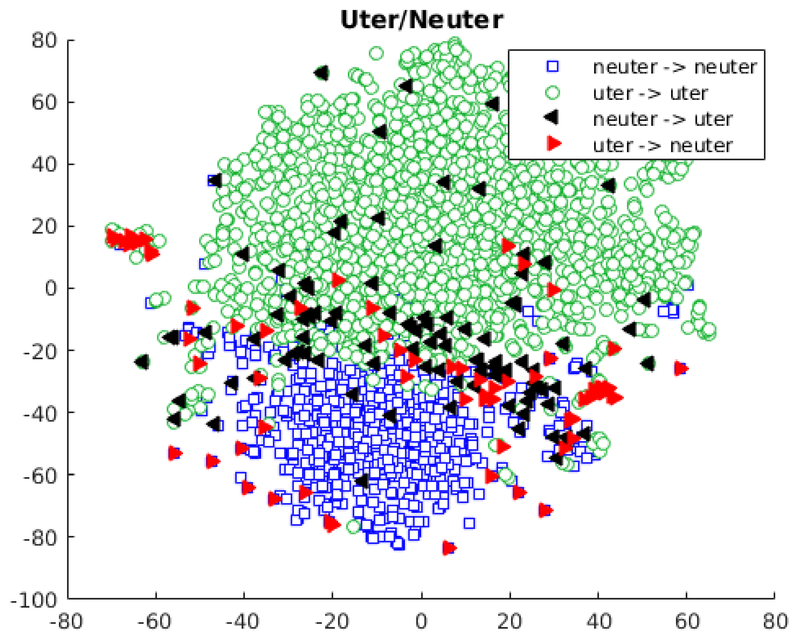
Distribution of uter (white) and neuter (gray) nouns with regard to frequency. The y-axis indicates the total ratio. The x-axis represents the nouns of the corpus partitioned into ten groups by their descending frequency

Results

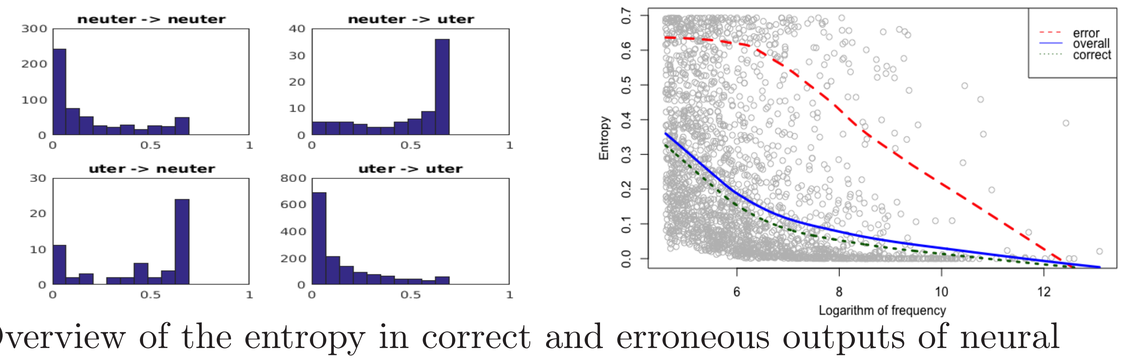
- **Overall Performance**

	Precision	Recall	F-score
Neuter	88.70%	84.16%	86.37%
Uter	93.34%	95.40%	94.36%
Overall	91.98%	92.12%	92.03%

The performance of neural network on grammatical gender prediction



tSNE representation of the word vectors classified by neural network according to their grammatical gender



Overview of the entropy in correct and erroneous outputs of neural network with regard to grammatical gender. *Left*: the y-axis indicates the amount of words from the test set, whereas the x-axis refers to the entropy

Discussion

- **Error Analysis**

Category	Quantity	Ratio	Example
Noise	17	9.94%	
dictionary/corpus	11	6.43%	tidsplan
proper name	6	3.51%	rosengård
Bare noun	44	25.73%	
abstract noun	10	5.85%	fjärilsim
fixed usage	12	7.02%	pistolhot
mass	22	12.87%	fosfat
Polysemy	110	64.33%	
different gender	10	5.85%	vad
different POS	100	58.48%	kaukasiska
Total	171	100%	

Errors of the neural network in the test set

- **Conclusion**
  - To a large extent, RSV can capture the information about grammatical gender of Swedish nouns.
  - The information about grammatical gender in RSV word vectors is only weakly influenced by the frequency of the words.
  - The errors are mostly related to the noise in the raw corpus or cases of polysemy.