The Construction and Evaluation of Word Space Models

Yves Peirsman, Simon De Deyne, Kris Heylen & Dirk Geeraerts

KULeuven
Quantitative Lexicology and Variational Linguistics

Overview

Introduction

Word Space Models

- model semantic similarity between two words as distributional similarity in a corpus.
- are often evaluated against (Euro)WordNet measures of semantic similarity.

Introduction

Word Space Models

- model semantic similarity between two words as distributional similarity in a corpus.
- are often evaluated against (Euro)WordNet measures of semantic similarity.

Questions

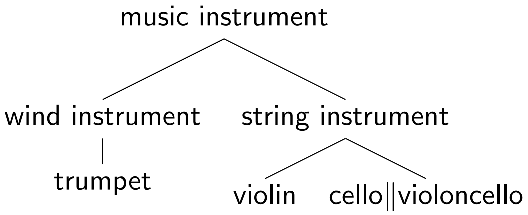
- Is the evaluation of the (Euro)WordNet measures reliable?
  - $\Rightarrow$ Evaluation against 5,000 human intra-category similarity judgements
- Is (Euro)WordNet a good Gold Standard for the similarity judgements of Word Space Models?
  - $\Rightarrow$ Evaluation of both approaches on same data

Overview

EuroWordNet similarity measures

Dutch EuroWordNet

A lexical database of noun synsets and their taxonomical relations.



Semantic similarity $\sim$ closeness in tree.

EuroWordNet similarity measures

Inverse Path Length

$$d_{PL}(w_1, w_2) = min(len(w_{1i}, w_{2j})) \tag{1}$$

$$s_{IPL}(w_1, w_2) = \frac{1}{d_{PL}(w_1, w_2)} \tag{2}$$

Leacock and Chodorow

Normalization by tree depth:

$$s_{LC}(w_1, w_2) = -log\frac{d_{PL}(w_1, w_2)}{2D} \tag{3}$$

EuroWordNet similarity measures

Wu and Palmer

Influence of depth of lowest shared hypernym:

$$s_{WP}(w_1, w_2) = \frac{2 \times depth(w_l)}{d_{PL}(w_1, w_l) + d_{PL}(w_2, w_l) + 2 \times depth(w_l)} \quad (4)$$

Jiang and Conrath

Combination with corpus statistics:

$$d_{JC}(w_1, w_2) = IC(w_1) + IC(w_2) - 2 \times IC(w_l) \quad (5)$$

EuroWordNet similarity measures: Evaluation

Rubenstein and Goodenough, Miller and Charles

- 65, 30 word pairs
- mostly inter-category: gem - jewel, food - fruit, cord - smile

Ruts et al.

- $>$ 5,000 human judgements
- musical instruments, fruit, birds, fish, clothing, etc.
- intra-category judgements: piano – guitar, pigeon – sparrow

EuroWordNet similarity measures: Evaluation

Category	*n*	IPL	WP	LC	JC
Professions	377	.32	.20	.22	.41
Fruit	406	.07	.11	.005	.25
Vegetables	325	.29	.25	.28	.27
Insects	253	.08	-.06	-.02	.24
Kitchen Utensils	465	.46	.25	.36	.37
Clothing	378	.25	.05	.11	.31
Musical Instruments	276	.68	.70	.67	.51
Reptiles	78	.49	.09	.27	.44
Sports	105	.53	.45	.50	.39
Fish	120	.44	.27	.37	.37
Vehicles	351	.49	.55	.48	.44
Birds	300	-.01	-.05	-.03	.19
Weapons	153	.39	.22	.30	.38
Tools	325	.50	.49	.50	.03
Mammals	351	.11	.10	.08	.29
average	**284**	**.34**	**.24**	**.28**	**.33**

EuroWordNet similarity measures: Evaluation

Category	n	IPL	WP	LC	JC
Professions	377	.32	.20	.22	.41
Fruit	406	**.07**	.11	.005	**.25**
Vegetables	325	.29	.25	.28	.27
Insects	253	**.08**	-.06	-.02	**.24**
Kitchen Utensils	465	.46	.25	.36	.37
Clothing	378	.25	.05	.11	.31
Musical Instruments	276	**.68**	.70	.67	**.51**
Reptiles	78	.49	.09	.27	.44
Sports	105	.53	.45	.50	.39
Fish	120	.44	.27	.37	.37
Vehicles	351	.49	.55	.48	.44
Birds	300	-.01	-.05	-.03	.19
Weapons	153	.39	.22	.30	.38
Tools	325	**.50**	.49	.50	**.03**
Mammals	351	.11	.10	.08	.29
average	284	.34	.24	.28	.33

EuroWordNet similarity measures: Evaluation

Professions	377	.32	.20	.22	.41
Kitchen Utensils	465	.46	.25	.36	.37
Clothing	378	.25	.05	.11	.31
Musical Instruments	276	.68	.70	.67	.51
Sports	105	.53	.45	.50	.39
Vehicles	351	.49	.55	.48	.44
Weapons	153	.39	.22	.30	.38
Tools	325	.50	.49	.50	.03
average	304	.45	.36	.39	.36

Fruit	406	.07	.11	.005	.25
Vegetables	325	.29	.25	.28	.27
Insects	253	.08	-.06	-.02	.24
Reptiles	78	.49	.09	.27	.44
Fish	120	.44	.27	.37	.37
Birds	300	-.01	-.05	-.03	.19
Mammals	351	.11	.10	.08	.29
average	262	.21	.10	.13	.29

EuroWordNet similarity measures: Evaluation

Correlation with human judgements

- ranges from nonexistent to pretty high.
- depends on the detail of the category in the taxonomy.
- is inconsistent across and within similarity measures.

⇒ Is (Euro)WordNet a valuable Gold Standard for other approaches?

Overview

Word Space Models

Distributional hypothesis

Semantically similar words occur in similar contexts.

Word Space Models

model the similarity between two words on the basis of their distributional similarity in a corpus.

- distributional information is stored in *context vectors*.
- semantic similarity is operationalized as similarity between two context vectors

Evaluation

Word Space Models are often evaluated against (Euro)WordNet measures.

Word Space Models

Corpora

Success of Word Space Models depends on size and type of corpus

- TwNC: 300m words of Dutch newspaper articles
  - ⇒ reasonable amount of data, high quality
- Web corpus: 700m words of web material, specifically compiled for Ruts et al's categories
  - ⇒ large amount of data, quality uncertain

Word Space Models

Corpora

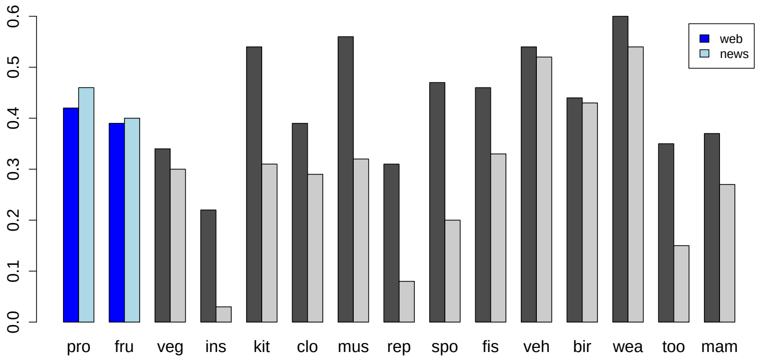
Success of Word Space Models depends on size and type of corpus

- TwNC: 300m words of Dutch newspaper articles
  - $\Rightarrow$ reasonable amount of data, high quality
- Web corpus: 700m words of web material, specifically compiled for Ruts et al's categories
  - $\Rightarrow$ large amount of data, quality uncertain

Other parameters

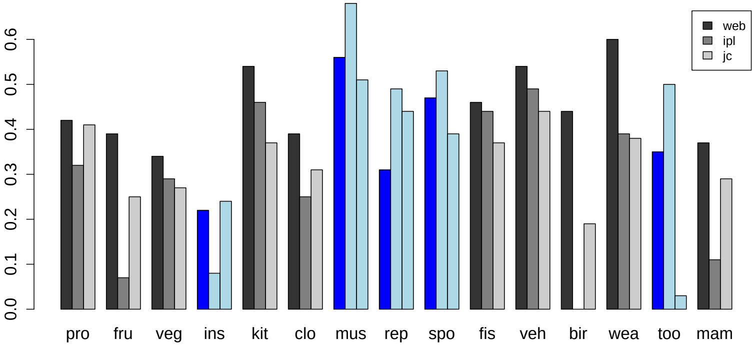
- context size: 2 words on either side of target
- weighting: log-likelihood between target and feature
- cut-off: 2

Word Space Models



⇒ The web corpus (.43) generally outperforms the news corpus (.31).

Word Space Models



⇒ The web corpus regularly outperforms EuroWordNet.

Overview

Introduction

EuroWordNet

Word Space Models

Conclusions

Conclusions

Evaluation of computational approaches to semantics

- (Euro)WordNet evaluated against small set of human judgements.
- Word Space Models against (Euro)WordNet

Problems

- Intra-category human judgements give a totally different picture.
- Word Space Models often outperform EuroWordNet.

Conclusions

The moral of the story

Computational models of semantic similarity that are meant to mirror human judgements are best evaluated against such human judgements directly.

Conclusions

The moral of the story

Computational models of semantic similarity that are meant to mirror human judgements are best evaluated against such human judgements directly.

See also:

Kris Heylen, Yves Peirsman, Dirk Geeraerts and Dirk Speelman
Modelling Word Similarity:
an Evaluation of Automatic Synonymy Extraction Algorithms.
15h40, Fez 1

For more information:

http://wwwling.arts.kuleuven.be/qlvl

yves.peirsman@arts.kuleuven.be