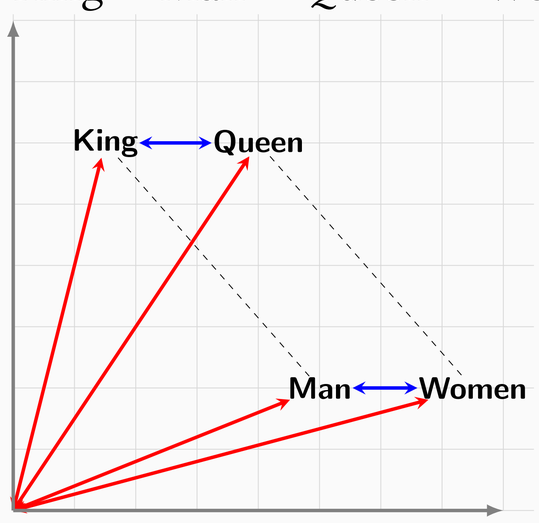
Multilingual Reliability and "Semantic" Structure of Continuous Word Spaces

Maximilian Köper, Christian Scheible, Sabine Schulte im Walde
Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart
{koepermn,scheibcn,schulte}@ims.uni-stuttgart.de

Analogies in Vector Spaces

Increasing attention is being devoted to continuous word vector representations (predict vector spaces). Mikolov et al. (2013) reported that a predict vector space seemingly encodes syntactic and semantic properties:

$$\overrightarrow{\text{King}} - \overrightarrow{\text{Man}} = \overrightarrow{\text{Queen}} - \overrightarrow{\text{Woman}} \quad (1)$$



Baroni et al. (2014) presented experiments for English where predict vectors outperform count vectors

➢ *But how reliable are they for other languages than English?*

Open Questions

Google-Analogy Dataset

	Relation	Example
Semantic	common-countries	Tokyo:Japan :: Rome:Italy
	capital-word	Apia:Samoa :: Cairo-Egypt
	currency	Brazil:real :: japan:yen
	city-in-state	Houstan:Texas :: Detroit:Michigan
	family	boy:girl :: brother:sister
Syntactic	opposite	ethical:unethical :: aware:unaware
	comparative	bad:worse :: big:bigger
	superlative	bright:brightest :: low:lowest
	present-participle	code:coding :: dance:dancing
	nation-adjectiv	Australia:Australian :: India:Indian
	past-tense	flying:flew :: hitting:hit
	plural-nouns	dollar:dollars :: bird:birds
	plural-verbs	speak:speaks :: find:finds

Dataset covers mostly morpho-syntactic relations !

➢ *It is still unknown to what extent predict vector spaces encode deep semantic relatedness*

→ We present a systematic exploration of morpho-syntactic and semantic relatedness in **English** and the morphologically richer language **German**.

Analogy and Semantic Relatedness Tasks

- The **Google semantic/syntactic** (analogies)
  → Constructed German counterpart through manual translation by 3 human judges
- The **paradigmatic semantic relation** (analogies)
  → The dataset was adapted by Lenci & Benotto for English and by Scheible & Schulte im Walde for German. Examples:

Antonym-Adj	psychological :	physical ::	maximum :	minimum
Antonym-NN	biblical :	secular ::	deaf :	hearing
Antonym-Verb	split :	join ::	sum :	subtract
Hyperonym-NN	groove :	dance ::	maze :	puzzle
Synonym-NN	skyline :	horizon ::	rumor :	gossip
Antonym-Adj	faul :	fleißig ::	traurig :	heiter
Antonym-NN	Ausnahmefall :	Regelfall ::	Deutlichkeit :	Undeutlichkeit
Antonym-Verb	lockern ::	festigen :	ärgern :	freuen
Hyperonym-NN	Stuhl :	Möbel ::	Bibel :	Buch
Synonym-NN	Pflanze :	Gewächs ::	Zeit :	Dauer

- The **general semantic** dataset
  → Correlation tasks : RG/Gur65, WordSim353/Schm280
  → Synonym task : TOEFL

NEW Datasets available!
- German version of Google analogies
- New paradigmatic relation analogies
- New rated Schm280 correlation task (German WordSim353)

All datasets are accessible at:
http://www.ims.uni-stuttgart.de/data/analogies/

Corpora

English
ENCOW14$_{\text{subset}}$
7.9 billion tokens

German
DECOW12
7.9 billion tokens

Preprocessing

- Shuffled sentences
- POS-Tagged + lemmatized
- Lowercased all words
- Basic filtering (punctuation, URL...)

Setup

- 400 Dimensions
- SubSampling $P = 10^{-5}$
- Sym. Window of size 2
- 15 Neg. Samples

{CBOW (predict), SKIP (predict), PPMI-SVD (count)} × {Word forms, Lemmatized forms} × {German, English}

TreeTagger

Results

Recall at ten! still very low

	Google-Sem (Acc)			Google-Syn (Acc)			Sem-Gen (ρ)			Sem-Para (R@10)			TOEFL (Acc)		
	CBOW	SKIP	BOW	CBOW	SKIP	BOW	CBOW	SKIP	BOW	CBOW	SKIP	BOW	CBOW	SKIP	BOW
EN W	68.8	**71.8**	39.5	**81.9**	80.5	57.9	**77.9**	77.8	77.8	**19.3**	16.4	15.6	**96.2**	**96.2**	72.2
EN L	68.3	**71.8**	40.3	47.1	**47.4**	29.3	**80.5**	78.6	66.4	18.4	**15.9**	15.8	**90.0**	87.5	66.2
DE W	42.4	**45.9**	27.3	**48.4**	47.1	31.0	**75.6**	73.3	58.9	14.7	14.4	**14.8**	**69.0**	68.3	54.4
DE L	43.5	**45.9**	28.9	**31.8**	31.5	23.7	73.3	**75.7**	64.7	**15.1**	13.8	14.9	**69.4**	68.5	55.8