Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

FACULTY
OF MATH, COMPUTER SCIENCE
AND NATURAL SCIENCES

Language Technology

STEFFEN REMUS, CHRIS BIEMANN

# RETROFITTING WORD REPRESENTATIONS FOR UNSUPERVISED SENSE AWARE WORD SIMILARITIES

HTTPS://UHH-LT.GITHUB.IO/SENSEASIM

## WORD SIMILARITY

**Task:**
- Compute the **similarity or relatedness** between two arbitrary words

**Common Methodology:**
1. use vector representation $\mathbf{v}_u$ and $\mathbf{v}_w$ of words $u$ and $w$
2. compute cosine similarity

$$\text{sim}(w, u) = \cos(\mathbf{v}_w, \mathbf{v}_u)$$

## ISSUE

### MEANING IS LOST

- Standard word embeddings project **one word to exactly one vector**
- Different senses of words are neglected
  → **Minor senses are underrepresented**
  → **Major senses are amplified**
  → **Source corpus / embedding defines the quality of the similarities**

## IDEA

### RETROFIT WORD VECTORS TO VECTORS OF SENSES USING SENSE INVENTORIES

- A **sense inventory** $\mathbf{S}_w$ is a **collection of synsets** $S_w^k \in \mathbf{S}_w$ for a particular word $w$
- Where the word $w$ itself is not present any of its synsets: $S_w^k \backslash w$
- Example: $\mathbf{S}_{\text{iron}}$

{zinc,calcium,magnesium,potassium,...},
{steel, metal, copper, aluminum, titanium, ...},
{shirt, steam, jeans, clothes, wool, coat, ...},
{putter, club, green, wood, driver, ...},
{pink, gold, black, white, orange, ...}

Nutrition
Atomic Elements
Clothes
Golf Sports
Color

## SENSE INVENTORY

- **JoBimText*** unsupervised sense inventory
  - via Web API `http://jobimtext.org`
- **Sense induction** by clustering → in the paper!

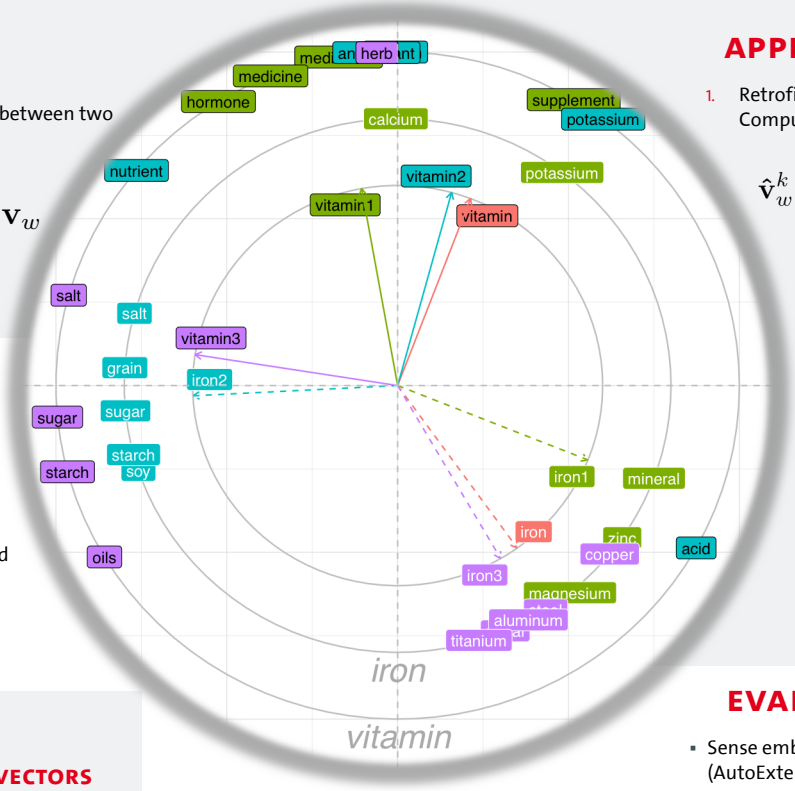


Figure:
Visualization is based on terms on the unit circle. Inner circle: sense vectors and original word vectors. Outer circles: Sense inventory terms of the words.

## APPROACH

1. Retrofit existing word embeddings:
   Compute vector representation of word senses

$$\hat{\mathbf{v}}_w^k = \lambda \mathbf{v}_w + (1-\lambda)\frac{1}{m}\sum_{u \in \text{top}_m(S_w^k)} \mathbf{v}_u$$

2. Compute similarity as argmax
   (we tested 3 options)

$$\text{sim}(w, u) =$$
$$\arg\max_k \cos(\hat{\mathbf{v}}_w^k, \mathbf{v}_u)$$
$$\arg\max_l \cos(\mathbf{v}_w, \hat{\mathbf{v}}_u^l)$$
$$\arg\max_{k,\,l} \cos(\hat{\mathbf{v}}_w^k, \hat{\mathbf{v}}_u^l)$$

## EVALUATION

- Sense embeddings with sense induction (AutoExtend, AdaGram)
- Several word embeddings and retrofitted word embeddings: SGNS (Word2Vec), Glove, SymPat (symmetric patterns), LSA, ParaGram (from PPDB)
- Datasets: SimLex999, MEN, SimVerb, WordSim353

| | AUTOEXTEND | ADAGRAM | SGNS | SGNS-S | GLOVE | GLOVE-S | SYMPAT | SYMPAT-S | LSABOW | LSABOW-S | LSAHAL | LSAHAL-S | PARAGRAMSL | PARAGRAMSL-S | PARAGRAMWS | PARAGRAMWS-S |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| SIMLEX999 | 0.45 | 0.29 | 0.44 | 0.46 | 0.37 | **0.41** | 0.54 | 0.55 | 0.30 | **0.39** | 0.27 | **0.38** | 0.68 | 0.64 | 0.66 | 0.64 |
| MEN | 0.72 | 0.67 | 0.77 | 0.78 | 0.73 | **0.77** | 0.53 | **0.68** | 0.67 | **0.70** | 0.71 | **0.74** | 0.77 | **0.80** | 0.80 | 0.81 |
| SIMVERB | 0.43 | 0.27 | 0.36 | **0.39** | 0.23 | **0.30** | 0.37 | **0.45** | 0.15 | **0.22** | 0.19 | **0.28** | 0.53 | 0.53 | 0.51 | 0.50 |
| WORDSIM353 | 0.58 | 0.61 | 0.70 | 0.69 | 0.61 | **0.65** | 0.47 | **0.62** | 0.67 | 0.66 | 0.59 | **0.63** | 0.72 | 0.73 | 0.77 | 0.75 |
| SIMLEX999-N | 0.44 | 0.33 | 0.45 | **0.50** | 0.39 | **0.47** | 0.48 | **0.55** | 0.32 | **0.46** | 0.34 | **0.44** | 0.68 | 0.66 | 0.64 | 0.64 |
| MEN-N | 0.72 | 0.68 | 0.77 | 0.79 | 0.76 | **0.80** | 0.57 | **0.74** | 0.71 | 0.73 | 0.73 | **0.76** | 0.78 | **0.81** | 0.80 | 0.82 |

Table:
Spearman correlation scores. Sense-aware similarities are marked with '-S'. Underlined values: the winning system with a slight margin (< 0.03). Bold face values: winning system with a larger margin.
Lower part: evaluates only the noun pair parts of the datasets (indicated by '-N').

LANGUAGE TECHNOLOGY GROUP
UNIVERSITY OF HAMBURG
uhh-lt.github.io/senseasim/

Scan me

LREC 2018, MIYAZAKI, JAPAN