UWB@RuShiftEval Measuring Semantic Difference as per-word Variation in Aligned Semantic Spaces

Ondřej Pražák, Pavel Přibáň, Stephen Taylor
{ondfa,pribanp,taylor}@kiv.zcu.cz

June 17, 2021

University of West Bohemia

Table of Contents

Task

- The task is to decide how similar the meanings of specified target words are between three different time periods.

Corpus	Sentences	Tokens
Pre-Soviet	5M	75M
Soviet	8M	97M
Post-Soviet	6M	85M

- **Enough to train good embeddings representation**
  - $\rightarrow$ Important for our approach
- A list of 99 target words
- In addition, RuSemShift gold data was available from previous work on comparable corpora.

Overview

- **completely unsupervised approach**
- Treat corpus C_1 and corpus C_2 as different languages L_1 and L_2
    - Train a separate semantic space for each corpus
- Normalize and zero-center
- Map the semantic spaces into a shared cross-lingual space

$$\hat{\mathrm{X}}^s = \mathrm{X}^s \mathrm{W}^{s \to t} \tag{1}$$

- Compute cosine similarity of mapped target words
    - $\to$ denotes the similarity between the time periods

The nickname for this approach – Procrustes's bed

Results

Of twelve named teams, we came in 6th.

We had 5 runs above BASELINE and 5 runs below.

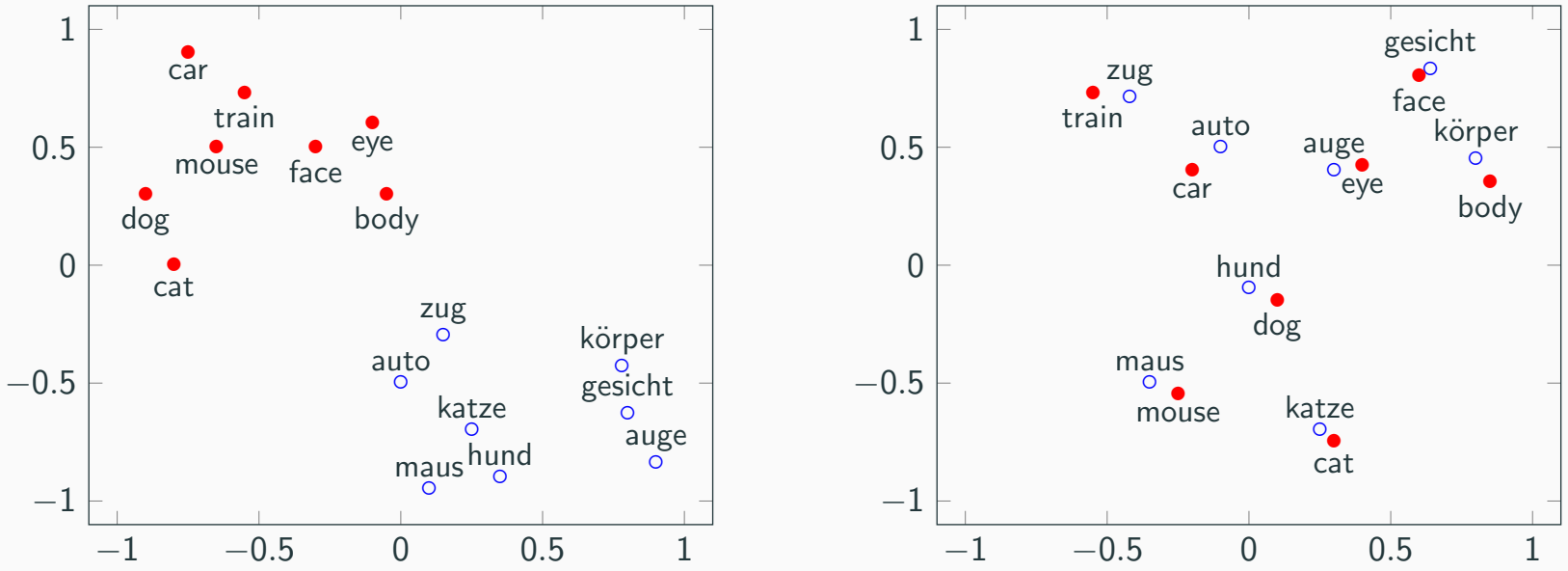
Rank	C_1/C_2	C_2/C_3	C_1/C_3	**Avg.**	**Description**
27	**.367**	**.354**	**.533**	**.417**	cca-150-OT
49	.277	.273	.464	.338	cca-100to220-agg
50	.239	.307	.450	.332	ort-200-OT
57	.220	.255	.446	.307	cca-100to500-agg
71	.096	.155	.317	.190	cca-300-pre-/post-trained

Table 1: Results of our selected submissions for all three corpora pairs.

Experiments within our approach

- Orthogonal vs CCA transformations
- What is a word?
- How wide an embedding vector?
- Variations of the translation dictionary
- Aggregating independent runs
- What we didn't do

Overview of transformations



(a) Before the projection.

(b) After the projection.

Figure 1: Sample visualisation of monolingual embeddings for English and German before and after their projections into a shared cross-lingual space.

.

Canonical Correlation Analysis

- CCA transforms spaces X^s and X^t into a third space X^o in which corresponding vectors from the two spaces are maximally correlated. $W^{s\to o}$ and $W^{t\to o}$ are built with the goal:

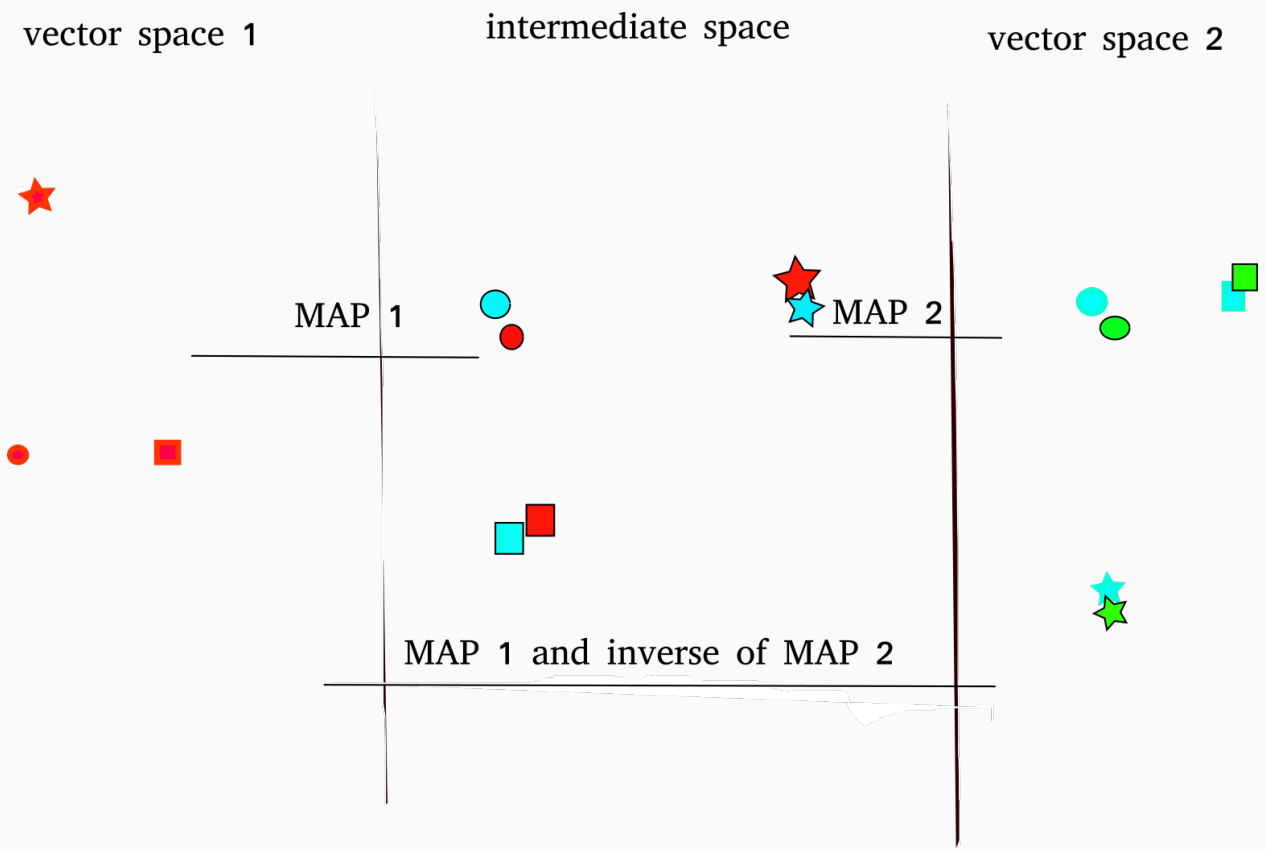
$$\begin{aligned}\underset{W^{s\to o},W^{t\to o}}{\arg\max} \sum_{i=1}^{n} \rho(x_i^s W^{s\to o}, x_i^t W^{t\to o}) = \\ \sum_{i=1}^{n} \frac{cov(x_i^s W^{s\to o}, x_i^t W^{t\to o})}{\sqrt{var(x_i^s W^{s\to o}) \times var(x_i^t W^{t\to o})}}\end{aligned} \tag{2}$$

- We use the two transforms to build one:

$$W^{s\to t} = W^{s\to o}(W^{t\to o})^{-1} \tag{3}$$

- It turns out there is closed form solution using SVD

Visual version of CCA-based transform



MAP 1 maps space 1 into the intermediate space, and MAP 2 maps space 2 into it. By combining MAP 1 and the inverse of MAP 2, we can map from space 1 to space 2.

Orthogonal Transformations

- Compute transformation matrix $W^{s \to t}$ under the **orthogonality** constraint: $WW^T = I$

$$\underset{W^{s \to t}}{\operatorname{argmin}} \sum_{i}^{|V|} (x_i^s W^{s \to t} - x_i^t)^2 \tag{4}$$

- Orthogonality preserves angles between vectors (words)
- Again, there is a closed-form solution using SVD

What is a word?

The target words on the evaluation list are all lemmas, which raises the questions:

- Do we ignore uses of the target lemma which are other wordforms, not the lemma itself?
  answer seems obviously no!
- Do we ignore the information provided by verb declension and noun case?
  Not nearly so clear!

We experimented with two strategies for words:

- Lemmatize all words. This was our 'default' strategy, and matches what we did in Diacr/It.
- Lemmatize only target words. We call this strategy 'Only Targets' and use it in our 'OT' runs.

How wide an embedding vector

We tried embedding vector widths from 50 to 500. All were trained for 5 epochs. Our pre-evaluation assessment was that widths of 100 to 200 worked best.

Post-evaluation experiments suggest that larger widths require more epochs of training.

Aggregating independent runs

Several of our submissions aggregated numerous runs with different parameters by averaging rank.

Rank seems more suitable for averaging than similarity, because randomness in training the embedding could result in a wider or narrower range of similarity scores, but the range of rank scores is fixed.

What we didn't do

We had settled on most of our hyper-parameters in previous workshops. We ran our systems against the RuSemShift data to check them for sanity, but not to refine them.

We made no effort to train against that data, either.

Conclusions

- Completely unsupervised
- Fast and easy to implement
    - No need to train a model

Thank you!

Questions?

{ondfa,pribanp,taylor}@kiv.zcu.cz

https://github.com/pauli31/SemEval2020-task1