Studying Laws of Semantic Divergence across Languages using Cognate Sets

Ana-Sabina Uban, Alina Maria Ciobanu, Liviu P. Dinu
Human Language Technologies Research Center (http://nlp.unibuc.ro)
Faculty of Mathematics and Computer Science, University of Bucharest

Motivation

- Languages are continuously changing, and words shift their meanings for various reasons.
- Semantic divergence in related languages is a key concern of historical linguistics.
- Laws of semantic change have been studied only monolingually using diachronic texts.

Measuring the Semantic Distance between Cognates

1. Obtain word embeddings for each of the two languages.
2. Obtain a shared embedding space, common to the two languages, using an algorithm that finds the optimal linear transformation between the two spaces, minimizing the distance between a few seed word pairs with the same meaning.
3. Compute semantic distances for each pair of cognates using a vectorial distance on their corresponding vectors in the shared embedding space.

Our Approach

- We propose a method for measuring semantic divergence **cross-lingually**, based on cognate pairs and cross-lingual word embeddings.
- We develop an algorithm for detecting and correcting false friends, based on the idea of **deceptive** cognate pairs.
- We study the correlation between properties of words (polysemy and frequency) and the degree of their semantic change across languages.

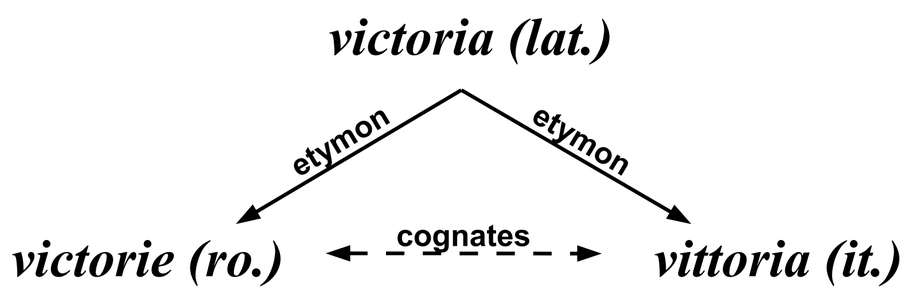
Detecting and Correcting False Friends

Given the cognates pair (c_1, c_2) where c_1 is a word in $lang_1$ and c_2 is a word in $lang_2$:
Find the word w_2 in $lang_2$ such that for any w_i in $lang_2$, $distance(c_2, w_2) < distance(c_2, w_i)$
if $w_2 \neq c_2$ **then**
    (c_1, c_2) is a pair of false friends
    Degree of falseness $= distance(c_1, w_2) - distance(c_1, c_2)$
return w_2 as potential correction
end if

Data

- We use a list 3,218 complete cognate sets in Romanian, French, Italian, Spanish and Portuguese.
- A subset of 305 cognate sets include English.

Example:



Properties of Cross-lingual Semantic Change

- Correlation between falseness and frequency rank

	ES	PT	IT	FR	EN
ES	-	-23.4	-31.5	-39.8	-20.9
PT	-42.0	-	-37.7	-34.2	-31.4
IT	-29.5	-28.5	-	-33.9	-36.2
FR	-25.9	-16.3	-23.3	-	-31.9
EN	-27.7	-39.3	-39.7	-39.2	-

- Correlation between falseness and polysemy

	ES	PT	IT	FR	EN
ES	-	56.2	47.3	26.5	12.1
PT	20.2	-	34.5	28.8	4.2
IT	18.6	15.0	-	6.2	2.1
FR	14.2	26.0	16.4	-	-5.4
EN	-9.1	-11.2	-16.5	-14.0	-

Properties of Cross-lingual Semantic Change

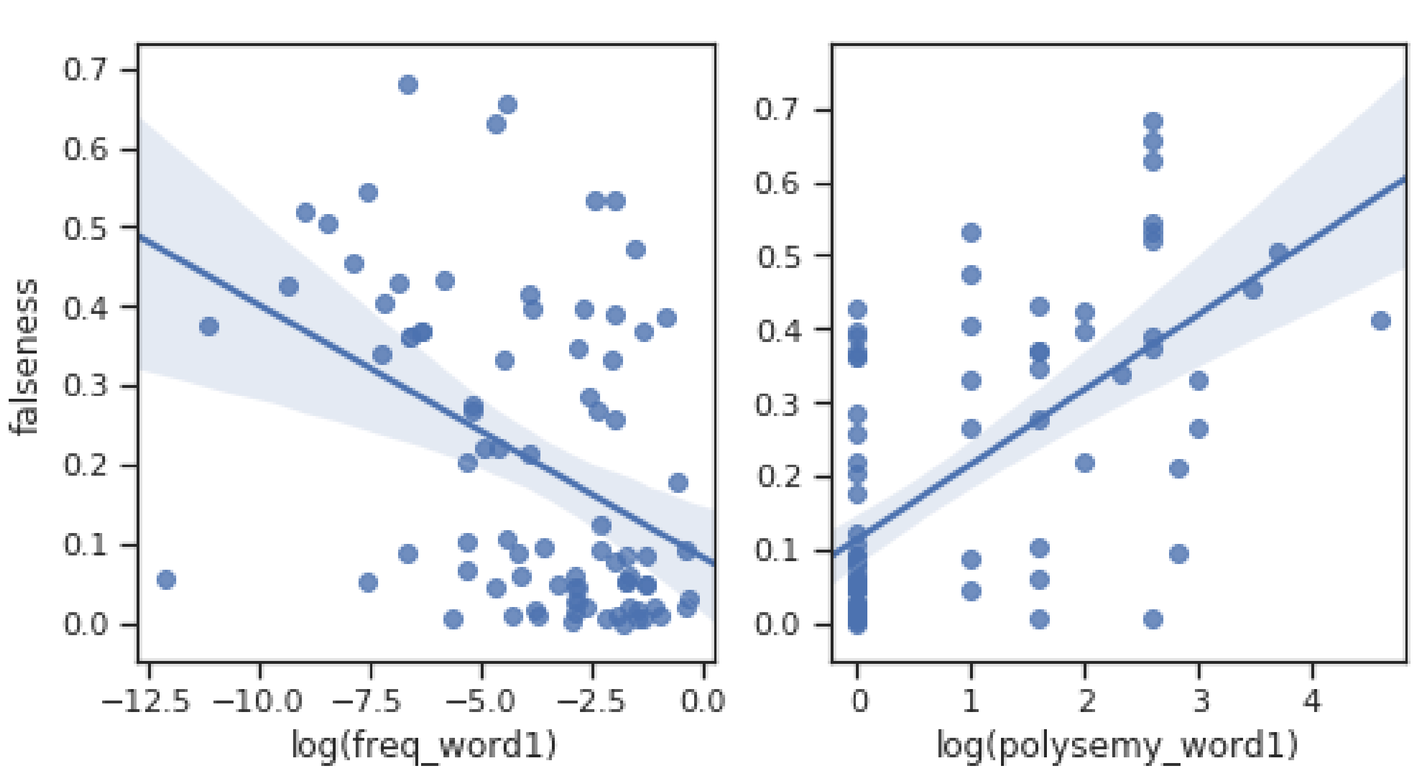


Figure: Correlation between falseness and frequency/polysemy for ES-PT

Examples

word	deceptive cognate	correction	falseness
long $_{FR}$	luengo $_{ES}$	largo $_{ES}$	0.50
face $_{FR}$	faz $_{ES}$	cara $_{ES}$	0.39
change $_{FR}$	caer $_{ES}$	cambia $_{ES}$	0.46
stnga $_{RO}$	stanco $_{IT}$	destra $_{IT}$	0.52
tnr $_{RO}$	tenero $_{IT}$	giovane $_{IT}$	0.41
inim $_{RO}$	anima $_{IT}$	cuore $_{IT}$	0.13
amic $_{RO}$	amico $_{IT}$	amichetto $_{IT}$	0.04

Conclusions

- We proposed a method to compute the semantic change of words across languages using cognate pairs.
- We studied how cross-lingual semantic change relates to word properties (polysemy, frequency).