Times Change

Tracking Latin Time Metaphors with Distributional Semantic Models

Krzysztof Nowak

Department of Medieval Latin
Institute of Polish Language
Polish Academy of Sciences

September 20 2017
Times Change
Tracking Latin Time Metaphors with Distributional Semantic Models
A Poor Historical Lexicographer’s Road to Distributional Semantic Models

Krzysztof Nowak

Department of Medieval Latin
Institute of Polish Language
Polish Academy of Sciences

September 20 2017
Motivation

- general: assessing interpretability of distributional semantics for everyday lexicography of an under-resourced language
- specific: tracking change in metaphorical patterns

Outline

1. What is distributional semantics ...
2. ... and why do we need it so badly?
3. Corpora
4. CMT and the study of metaphor
5. Lat. *tempus* 'time': clustering collocates
6. Conclusions
Method: Distributional Semantics

Procedure (Harispe 2017)
- choose your context (doc, par, sentence ..)
- weight raw frequencies (PMI...)
- reduce sparse matrix (SVD...)
- measure distance (cosine ...)

Distributional Hypothesis
"the degree of semantic similarity between two words (or other linguistic units) can be modelled ... as a function of the degree of overlap among their linguistic contexts" (Baroni and Lenci 2010)

<table>
<thead>
<tr>
<th></th>
<th>sacer 'saint'</th>
<th>lego 'to read'</th>
</tr>
</thead>
<tbody>
<tr>
<td>liber 'book'</td>
<td>73</td>
<td>422</td>
</tr>
<tr>
<td>uolumen 'volume'</td>
<td>234</td>
<td>158</td>
</tr>
<tr>
<td>bellum 'battle'</td>
<td>38</td>
<td>16</td>
</tr>
</tbody>
</table>

Table: Cooccurrence Matrix

Figure: Measuring Distance
More on DSMs

General Introductions
- general: Turney and Pantel 2010
- accessible: Widdows 2004

Survey
- Harispe et al. 2015 (cf. 2017)

DSMs in Linguistics
- Peirsman/Heylen et al. 2008-
Why this approach?

Poor resources
- a (potentially) huge corpus
- underserved with regard to the NLP tools
- no representative corpus at the moment

No sound evaluation framework
- no semantically annotated corpora
- no similar tools to the TOEFL dataset etc.
- no native speakers
- visual inspection relies on subjective assessment that the DSM was expected to eliminate
The Method

Workflow

1. frequency table

<table>
<thead>
<tr>
<th>Term</th>
<th>Property</th>
<th>Freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>doceo</td>
<td>iudex</td>
<td>4</td>
</tr>
<tr>
<td>periculum</td>
<td>repello</td>
<td>3</td>
</tr>
<tr>
<td>visus</td>
<td>forma</td>
<td>4</td>
</tr>
<tr>
<td>addo</td>
<td>membrum</td>
<td>2</td>
</tr>
</tbody>
</table>

2. R wordspace package by Stefan Evert:
   - weighting raw counts
   - matrix reduction

3. word similarity functions
   - pair.distances()
   - nearest.neighbours()

Corpora

Classical
- time: 2 BC - 6 AD
- size: 8M

Early Christian
- time: 1 AD - 6 AD
- size: 22M

EC & Medieval
- time: 1 AD - 13 AD
- size: 100M
The Problem: **TIME and its metaphors**

- more than stylistic device
- source domain → target domain
- abstracts regularly comprehended in more concrete terms

**Evans (2006, 2013) on time**
- metaphorical expression → lexical concept
- activation: formal and semantic properties
- ex.: *to loose one’s time* (mass noun, interest-related verb) → **Commodity Sense**

**Procedure**
1. wide collocational window
   - *vos tot milibus civium Romanorum*\(^5_L\) *uno*\(^4_L\) *nuntio*\(^3_L\) *atque*\(^2_L\) *uno*\(^1_L\) *tempore necatis*\(^1_R\)
   - *quo*\(^2_R\) *tandem*\(^3_R\) *animo*\(^4_R\) *esse*\(^5_R\) *debitis*
2. association strength
   - Dice: \[ \frac{2 \cdot f_{AB}}{f_A + f_B} \]
3. empirical collocations (Evert 2009) → clustering
Collocational stability (Classical corpus)

Attested in 5 periods
vernus ('springtime')

Attested in 4 periods
annus ('year')
brevis ('short')
certus ('certain')
dies ('day')
differo ('postpone')
longus ('long')
praesens ('present')
spatium ('space')

Attested in 3 periods
bellum ('war')
medius ('middle')
persona ('person')
venio ('to come')
Stable collocations (Classical Corpus)

**Time is Space**

*longus* 'long', *brevis* 'short', *medius* 'middle', *spatium* 'space'

**Syntagmatic similarity**

*certus* 'certain', *praesens* 'present'

**Paradigmatic similarity: hyponyms**

*annus* 'year', *dies* 'day'

**Moving Time?**

*venio* 'to come', *differo* 'to postpone'
All collocations (Classical Corpus)

nouns
verbs
adjectives
Nominal collocations (Classical Corpus)

**Time is Space**

*spatium 'space', locus 'place'*

Paradigmatic similarity: hyponyms and near synonyms

*annus 'year', dies 'day', nox 'night', aetas 'epoch'*...

**Time Reference**

*Caesar 'Cesar', consul 'consul' (ex. 'under consul X')*

**Legal uses**

*vis 'power', lex 'law', condicio 'condition', testamentum 'testament'*
Nominal collocations (Early Christian)

Time is Space

*spatium* 'space', *intervallum* (etym. 'between walls'), but also: *initium* 'beginning', *finis* 'end'

Paradigmatic similarity: hyponyms

*annus* 'year', *saeculum* 'century', *aetas* 'epoch'...

Time Reference

*resurrectio* 'resurrection', *persecutio* 'prosecution', *captivitas* 'captivity'...
Syntactic constraints: *temporis* 'of time'

**Time is Space**

*spatium* 'space', *intervallum* 'interval' (etym. 'between walls'), *angustia* 'narrowness', *brevitas* 'shortness'

**Time is Money**

*fructus* 'income', *aestimatio* 'evaluation', *usura* 'profit'
Semantic constraints: **FOOD terminology**

Sense Clusters of 50 CIBUS Collocations

**NOUNS AND ADJECTIVES**
- uentiar
- stomachus
- mensa
- sanguis
- uestis
- abstinence
- caro
- somnus
- solidus
- solidum
- pauper
- faculas
- uils
- delicatus
- dulcis
- laetus
- uentum
- spatialis
- corporalis
- communis
- huymord
- potes
- potus
- lac
- panis
- manna
- cibus
- uestio
- sumo
- indigeo
- capio
- utor
- Pan
- preparo
- distribuo
- ministro
- pereo
- abstineo
- leuino
- passo
- allo
- nutro
- reficio
- uesto
- eesco
- mundico
- coquo
- gusto

**VERBS**
Comparing diachronic "micro-thesauri"
Conclusions

- testing, testing, testing...
- building evaluation framework
- towards automatic comparison of vectors between corpora
- building tools
  - dictionary
  - testing framework

Our projects

eCorpus of ML
scriptores.pl/en/efontes

eDictionary of ML
scriptores.pl/elexicon


