Predictability of Distributional Semantics in Derivational Word Formation

Sebastian Padó*, Aurélie Herbelot+, Max Kisselew*,
Jan Šnajder†

*

\+

†

TakeLab

Overview

1. Introduction

2. Analyzing Models of Morphological Derivation

Processes of Word Formation

- **Composition**: *file + name → filename*
- **Inflection:** *make → make+s, computer → computer+s*
- **Morphological derivation ...**
  - can mean attaching an affix to a base word (e. g. *drive* + *ER* → *driver*)
  - can be more complex, involving stem alternation, deletion of previous affixes, circumfixation
  - can take place both within parts of speech and across parts of speech
  - is very productive process in many languages, notably Slavic languages

Introduction

Compositional models of distributional semantics (CDSMs)

- are generally applied to *compositionally compute phrase meaning* (Baroni and Zamparelli, 2010; Coecke et al., 2010)
- have been applied to model word formation processes like composition and (morphological) derivation (Lazaridou et al., 2013)
- Goal: Predict vector for the derived word from vector of base and vector of affix

Introduction

Modeling Derivation through Compositional Distributional Semantics Models (CDSMs):

$$\overrightarrow{\mathit{derived}} = \overrightarrow{\mathit{base}} + \overrightarrow{\mathit{affix}}$$

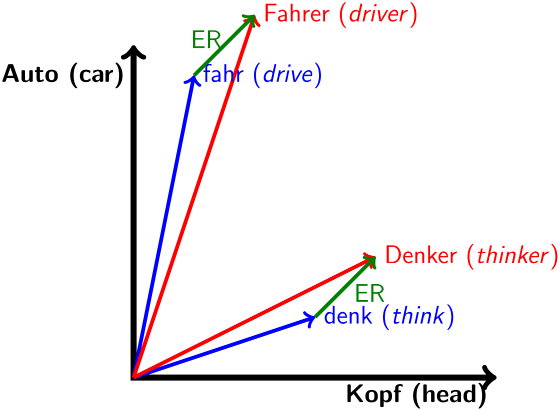
Examples:

$$\overrightarrow{\mathit{Fahrer}} = \overrightarrow{\mathit{fahr}} + \overrightarrow{\mathit{ER}}$$

driver *drive* *ER*

$$\overrightarrow{\mathit{Denker}} = \overrightarrow{\mathit{denk}} + \overrightarrow{\mathit{ER}}$$

thinker *think* *ER*



Introduction

Challenges:

- Morphological derivation is often irregular
  $\Rightarrow$ Meaning changes not completely predictable
  (Plank, 1981; Laca, 2001; Plag, 2003; Dressler, 2005)
- Practical concerns, e.g. different frequencies of base and derived word
- No clear picture about factors that affect CDSMs performance in modeling of derivation (Lazaridou et al., 2013)
- Very uneven performance of CDSMs across words and word pairs (Kisselew et al., 2015)

Our contribution:

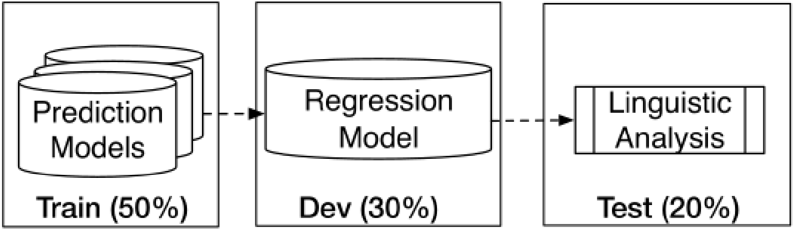
$\Rightarrow$ We investigate linguistic factors that govern the success or failure of CDSMs to predict distributional vectors for derived words

Overview

1. Introduction

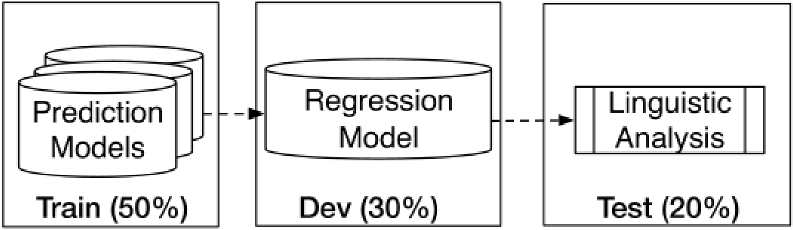
2. Analyzing Models of Morphological Derivation

Overall workflow



- Step 1: Train CDSMs on Train set; run CDSMs on Dev and Test sets
- Step 2: Learn regression model on CDSM performance numbers from Dev set
- Step 3: Test regression model on CDSM performance numbers from Test set

Step 1



- Step 1: Train CDSMs on Train set; run CDSMs on Dev and Test sets
- Step 2: Learn regression model on CDSM performance numbers from Dev set
- Step 3: Test regression model on CDSM performance numbers from Test set

Data: Derivational word pairs

Extracted from DErivBase (Zeller et al. 2013). Examples:

POS + ID	Pattern	Sample word pair
A → N 16	*+ität*	produktiv → Produktivität
		(productive → productivity)
N → A 26	*-ung +end*	Einigung → einigend
		(agreement → agreeing)
V → N 09	*(null)*	aufatmen → Aufatmen
		(to breathe → sigh of relief)

- 74 patterns (49 cross-POS patterns)
- 30,757 word pairs
- Median per pattern: 194.5 word pairs
- Min. 83, max. 3028 word pairs

Data: Vector space

- CBOW vectors (Mikolov et al., 2013), 300 dimensions, context window: ± 2
- Corpus: SdeWaC (Faaß and Eckart, 2013)

CDSMs

Employed CDSMs:

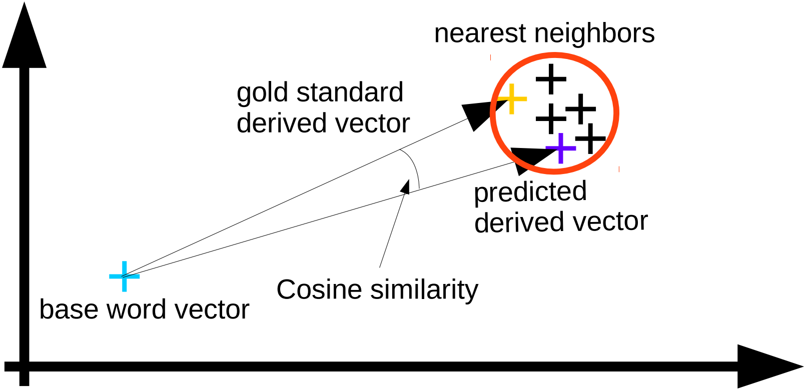
- Simple additive model: $\overrightarrow{deriv} = \overrightarrow{base} + \overrightarrow{affix}$
- Weighted additive model: $\overrightarrow{deriv} = \alpha\overrightarrow{base} + \beta\overrightarrow{affix}$
- Simple multiplicative model: $\overrightarrow{deriv} = \overrightarrow{base} \odot \overrightarrow{affix}$
- Lexical function model: $\overrightarrow{deriv} = A\overrightarrow{base}$

Baseline:

- Baseline: $\overrightarrow{deriv} = \overrightarrow{base}$

Evaluation Measure

How well does the predicted vector align with the corpus-observed vector?



Evaluation Measure

Reciprocal rank (RR): 1 divided by the position of the predicted vector in the similarity-ranked list of the observed vector's neighbors

Example:

Base word	*vernünftig*	*harmonisch*	*absichtlich*
Correct derived word	*unvernünftig*	*unharmonisch*	*unabsichtlich*
Nearest neighbor 1	**unvernünftig**	*wohlausgewogen*	**unabsichtlich**
Nearest neighbor 2	*akzeptabel*	*spannungsvoll*	*wissentlich*
Nearest neighbor 3	*rational*	*stimmig*	*vorsätzlich*
Nearest neighbor 4	*sinnvoll*	**unharmonisch**	*falsch*
RR	$\frac{1}{1}$	$\frac{1}{4}$	$\frac{1}{1}$
Aggregate RRs into Mean Reciprocal Ranks (MRRs)	$\frac{\frac{1}{1}+\frac{1}{4}+\frac{1}{1}}{3} = \frac{2.25}{3} = 0.75$		

CDSM Models - Results

Results for individual CDSM prediction models on test set

	Baseline	Simple Add	Weighted Add	Mult	LexFun
Mean Reciprocal Rank	0.271	0.309	**0.316**	0.272	0.150

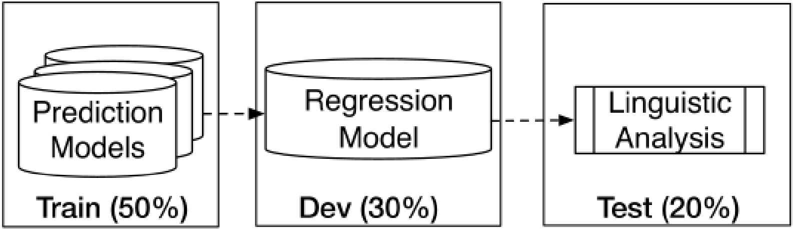
CDSM Models - Results by Pattern

Performance is highly variable across patterns and words pairs

Examples:

POS + ID	Pattern	Sample word pair	RR
V $\rightarrow$ V 01	*-en +eln*	zucken $\rightarrow$ zuckeln (twitch $\rightarrow$ saunter)	0.03
A $\rightarrow$ N 10	*-(a|e)nt +(a|e)nz*	präsent $\rightarrow$ Präsenz (present $\rightarrow$ presence)	0.69

Step 2



- Step 1: Train CDSMs on Train set; run CDSMs on Dev and Test sets
- Step 2: Learn regression model on CDSM performance numbers from Dev set
- Step 3: Test regression model on CDSM performance numbers from Test set

Regression Model

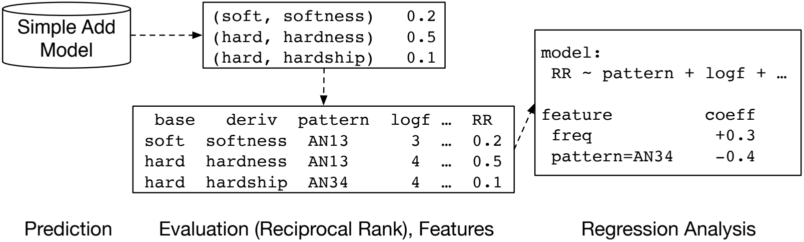
Task: Predict the performance of the CDSM models (measured as RR) at the word pair level using a regression model

Three classes of predictors:

Predictor class	**Description**
Base word level	lemma frequency number of WordNet synsets productivity of the base word etc.
Prediction level	similarity of the derived vector to its nearest neighbors similarity between base vector and derived vector etc.
Pattern level	Identity of the pattern

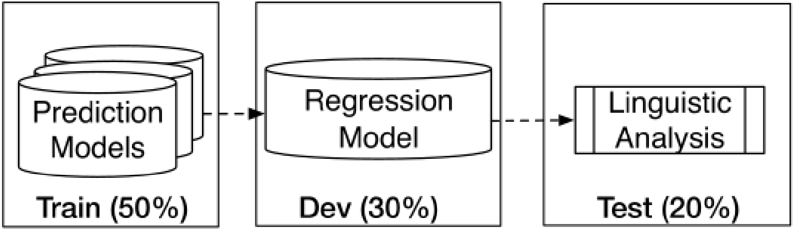
Analysis toy example

Toy example for a single CDSM prediction model (simple additive):



1. Run the CDSM model on unseen data
2. Evaluate its reciprocal ranks at the word pair level
3. Compute features from the same data
4. Learn regression model: Yields **coefficients** for features indicating their impact on CDSM performance

Step 3



- Step 1: Train CDSMs on Train set; run CDSMs on Dev and Test sets
- Step 2: Learn regression model on CDSM performance numbers from Dev set
- Step 3: Test regression model on CDSM performance numbers from Test set

Linguistic Analysis - Experiment

- **Research question:** Which properties of the base word and the pattern make the prediction easy or difficult?
- **Estimate** the following linear regression model to predict RR on a test set (use pattern-level and base-level features):

  ```
  RR ~  pattern + base_productivity + base_typicality
  + base_polysemy + base_freq
  ```

Linguistic Analysis - Results

Coefficients, significances, and effect sizes for the predictors (negative coefficients indicate poorer CDSM performance):

Predictor	Estimate	LMG score
`pattern`	N/A	87.2%
`base_productivity`	−0.13***	7.6%
`base_freq`	0.21***	4.1%
`base_polysemy`	−0.03**	0.8%
`base_typicality`	0.04***	0.2%

Linguistic Analysis - Results

Coefficients, significances, and effect sizes for the predictors (negative coefficients indicate poorer CDSM performance):

Predictor	Estimate	LMG score
pattern	N/A	87.2%
base_productivity	−0.13***	7.6%
base_freq	0.21***	4.1%
base_polysemy	−0.03**	0.8%
base_typicality	0.04***	0.2%

- pattern (the derivation pattern) accounts for a large percentage of the variance.

Linguistic Analysis - Results

Coefficients, significances, and effect sizes for the predictors (negative coefficients indicate poorer CDSM performance):

Predictor	Estimate	LMG score
pattern	N/A	87.2%
base_productivity	−0.13***	7.6%
base_freq	0.21***	4.1%
base_polysemy	−0.03**	0.8%
base_typicality	0.04***	0.2%

- More productive bases are more difficult to predict.

Linguistic Analysis - Results

Coefficients, significances, and effect sizes for the predictors (negative coefficients indicate poorer CDSM performance):

Predictor	Estimate	LMG score
pattern	N/A	87.2%
base_productivity	−0.13***	7.6%
base_freq	0.21***	4.1%
base_polysemy	−0.03**	0.8%
base_typicality	0.04***	0.2%

- More frequent bases are easier to predict.

Linguistic Analysis - Results

Coefficients, significances, and effect sizes for the predictors (negative coefficients indicate poorer CDSM performance):

Predictor	Estimate	LMG score
pattern	N/A	87.2%
base_productivity	−0.13***	7.6%
base_freq	0.21***	4.1%
base_polysemy	−0.03**	0.8%
base_typicality	0.04***	0.2%

- Polysemy (number of WordNet senses) and typicality of the base word play very small roles – they show expected effects but these hardly matter.

Analysis by pattern – Results

1) Cross-POS derivations: 🙂

Reason: Cross-POS derivations often syntactically motivated – context remains similar.

For example:

- *-ung* nominalization pattern:
  verarbeiten → *Verarbeitung* / *(to) process* → *processing*

Analysis by pattern – Results

2) Derivation patterns that are semantically regular: 🙂

Reason: Patterns that are semantically irregular/ambiguous are hard to learn.

For example:

- Noun $\rightarrow$ verb derivation patterns generate verbs from nouns that are only loosely semantically related
  (*Zweig* $\rightarrow$ *abzweigen* / *(tree) branch* $\rightarrow$ *branch off*)

Analysis by pattern – Results

3) Patterns with a change in argument structure: 🙁

Reason: Arguments incorporated through derivation drop out of the context of the derived word.

For example:

- agentive/instrumental nominalization pattern +*er* (*fahren* → *Fahrer* / *drive* → *driver*)

Ensemble Prediction - Experiments

- If we have different models, can we combine them to obtain better prediction?
- Follow-up study: Select one vector from among the predictions of multiple CDSMs (ensemble prediction)
- Two models:
  1. Oracle model:
     Compares all prediction models and picks the one with the highest RR
  2. Ensemble model:
     Predicts the CDSMs' expected performances at the word pair level using a linear regression model

Ensemble Prediction - Results

Model	MRR
Oracle model	0.362
Ensemble model	0.321
Weighted Add (best individual model)	0.316

- Small improvement by oracle model
  $\Rightarrow$ Reason: almost all models highly correlated with one another

Conclusions

- First analysis of CDSMs on derivational phenomena that is both detailed and broad-coverage
- Three main factors for bad performance of CDSMs:
  1. modifications of argument structure
  2. semantic irregularity
  3. within-POS derivations
- Our dataset with derivationally related word pairs and CDSM performance predictors is available at: `http://www.ims.uni-stuttgart.de/data/derivsem`

The End

Thank you!

Any questions?