Lexical Substitution for Evaluating Compositional Distributional Models

M. Buljan\textsuperscript{1}  S. Padó\textsuperscript{1}  J. Šnajder\textsuperscript{2}

\textsuperscript{1}IMS
University of Stuttgart

\textsuperscript{2}TakeLab
University of Zagreb

February 2, 2018
Distributional Semantics

A brief summary

Distributional Hypothesis

You shall know a word by the company it keeps (J.R. Firth)

- Semantic similarity $\rightarrow$ distributional similarity
- Algebraic representation
  - co-occurrence statistics
  - linear algebra, vector and tensor representation
Distributional Semantics
Vector space model

- Co-occurrence matrix
- Weighting
- Dimensionality reduction

|     | blue | good | cute |
|-----|------|------|------|
| car | 5    | 3    | 1    |
| dog | 0    | 7    | 4    |
| cat | 0    | 1    | 9    |
Distributional Semantics
Moving beyond individual words

Bag-of-words approach

Problem:
▶ Longer phrases and sentences?
  ▶ Higher-order tensor representation – computational issues
▶ Modelling syntactic relations?
  ▶ BoW insufficient to model dependencies
Compositional DS Models

Modelling phrase meaning; e.g. "fluffy cat sees big dog"

Phrase – function of constituents

CDSMs

- Algebraic vector operations
  \[
  \text{fluffy} + \text{cat} + \text{sees} + \text{big} + \text{dog}
  \]
  - Issue: "cat sees dog" vs. "dog sees cat"

- Higher-order tensor representation
  \[
  \text{sees}^{□}(\text{fluffy cat}_S, \text{big dog}_O)
  \]
  - Issue: computational complexity and data sparsity
CDSMs

Additive and Multiplicative model
\[ p = u \odot v \]; component-wise addition/multiplication

Practical Lexical Function (PLF)
Predicate: function over argument

- adjective and noun: \( A \odot \cdot \vec{N} \)
  \[ \text{fluffy} \odot \cdot \text{cat} \]
- transitive verb and subject/object: \( V_S \odot \cdot \vec{S} + V_O \odot \cdot \vec{O} + \vec{V} \)
  \[ \text{sees}_S \odot \cdot \text{cat} + \text{sees}_O + \text{sees}_O \odot \cdot \text{dog} \]
- matrix representation of semantic roles
  – simpler to train, compute
CDSMs

Training the PLF

Composition: \( V_S \cdot \overrightarrow{AN} + V_O \cdot \overrightarrow{AN} + \overrightarrow{V} = \overrightarrow{ANVA\bar{N}} \)

- Learning values of function matrices
  - Corpus-observed bigrams (\( \overrightarrow{AN}, \overrightarrow{VN_S}, \overrightarrow{VN_O} \))
  - Regression learning; \( V_S \cdot \overrightarrow{N} = \overrightarrow{VN_S} \)
  - Function matrices: \( A, V_S, V_O \)

PLF: train- vs. test-time discrepancy

- \( \text{PLF}_{\text{Paperno}}: \overrightarrow{V} \cdot \overrightarrow{AN} + \overrightarrow{V} = \overrightarrow{ANV} \)
- \( \text{PLF}_{\text{Gupta}}: \overrightarrow{V} \cdot \overrightarrow{AN} = \overrightarrow{ANV} \)
Experimental Setup

So far, CDSMs tested on bi- and trigram similarity tasks

RQs:
- Performance with more complex phrases? → natural text
- What is n-gram similarity, anyway? → real-world task
- How do CDSMs compare? → algebraic vs. functional
- Test-vs-train discrepancy? → composition variants
Experimental Setup
LexSub dataset

- ColnCo corpus – manually annotated substitutes
- Extraction of ANVAN sentences/ clauses
  \[(\text{Adjective} + \text{Noun} + \text{Verb} + \text{Adjective} + \text{Noun})\]
- 165 phrases → 732 substitution targets

| target       | substitute\textsubscript{1} | substitute\textsubscript{2} | confounder\textsubscript{1} | confounder\textsubscript{2} |
|--------------|-------------------------------|-------------------------------|-----------------------------|-----------------------------|
| construction | construction | arm | build | large | airfield |
| construction | construction | branch | build | large | airfield |
| construction | construction | part | build | large | airfield |
| construction | construction | back | build | large | airfield |
| construction | construction | hand | build | large | airfield |
Experimental Setup

Tested Models

Phrase similarity; 4-way ranking task; MAP

Baselines

▷ Random baseline
▷ Lemma-level similarity

Algebraic models

▷ Additive
▷ Multiplicative

PLF

▷ PLF_{\text{Paperno}}
▷ PLF_{\text{Gupta}}

Comparison

▷ Word-embeddings lexical substitution model; context2vec
### Evaluation Results

|     | $\text{BL}_{\text{Rnd}}$ | $\text{BL}_{\text{Lem}}$ | Add | Mult | PLF$_P$ | PLF$_G$ | C2V$_{\text{Phr}}$ | C2V$_{\text{Sent}}$ |
|-----|--------------------------|---------------------------|-----|------|---------|---------|----------------|----------------|
| overall | .680 | .599 | .656 | .669 | .681 | **.706** | **.702** | **.731** |
| Anvan | .680 | .680 | .716 | .715 | .730 | .727 | .694 | .707 |
| aNvan | .680 | .575 | .652 | .633 | .695 | .688 | .708 | .744 |
| anVan | .680 | .537 | .618 | .670 | **.536** | **.680** | .697 | .723 |
| anvAn | .680 | .625 | .668 | .668 | .721 | .715 | .690 | .710 |
| anvaN | .680 | .580 | .633 | .666 | .725 | .723 | .723 | .772 |

- Lemma similarity worst overall – importance of context
- Only PLF beats baseline – BoW vs. semantic roles
- Trouble with verbs – saliency or composition?
- PLF$_{\text{Gupta}}$ > PLF$_{\text{Paperno}}$ – composition and training concerns
Ongoing Work
Expanding the PLF

- PLF: more train- vs. test-time discrepancy
  - training: $V \triangleleft \cdot \overrightarrow{N} = \overrightarrow{NV}$
  - composition: $V \triangleleft \cdot \overrightarrow{AN} = \overrightarrow{ANV}$

- Heuristic composition
  - $\overrightarrow{A} + (V \triangleleft \overrightarrow{N}) + \overrightarrow{V}$
Ongoing Work

Evaluation Results

|              | NVN          | A + (NV) + (VN) | (AN) + (NV) + (VN) |
|--------------|--------------|-----------------|-------------------|
|              | Add  | Mult | PLFₚ | PLF₇ | PLFₚ | PLF₇ | PLFₚ,AVA | PLFₚ,V | PLFₗ    |
| overall      | .618 | .643 | .694 | .730 | .690 | .709 | .725      | .718    | .742    |
| Anvan        | −     | −    | −    | −    | .683 | .686 | −         | −       | −       |
| aNvan        | .613 | .617 | .705 | .681 | .721 | .714 | .724      | .730    | .712    |
| anVan        | .616 | .653 | .639 | .778 | .650 | .755 | .767      | .645    | .774    |
| anvAn        | −     | −    | −    | −    | .641 | .647 | −         | −       | −       |
| anvaN        | .625 | .658 | .734 | .734 | .744 | .735 | .753      | .748    | .743    |

- performance improves with heuristic composition
- the trouble with verbs – overpowering predicate vectors
Summary

- Simpler vs. more complex compositional models
- Performance comparable to state-of-the-art on LexSub
- More context → better disambiguation

Next steps
- Adjusting the composition
- Modelling more functions – adverbs, quantifiers
- Expanding the dataset – more natural syntax
- PhD tie-in: DS + (non-)compositional MWEs
