Semantic Representations of Word Senses and Concepts

José Camacho Collados
Ignacio Iacobacci
Roberto Navigli

Mohammad Taher Pilehvar
Outline

- Foundations
- Sense representations
  - Introduction
  - Knowledge-based techniques
    - WordNet
    - Large knowledge resources
      - Wikipedia
      - BabelNet
      - FreeBase - WikiData

Coffee Break (30 mins)

- Unsupervised techniques
  - Advantages and limitations
- Applications
- Open problems and future work
Key points

• **What** do we want to represent?
• What does "**semantic representation**" mean?
• **Why** semantic representations?
• What **problems** affect mainstream representations?
• How to **address** these problems?
• What comes **next**?
What do we want to represent?

Linguistic items of different kinds:

- **Documents**: the Wikipedia page for "On the Internet, nobody knows you're a dog"

- **Sentences**: On the Internet, nobody knows you're a dog

- **Phrases**: on the Internet

- **Words**: dog

- **Senses**: 
What kinds of representation can we provide?

Vector representations (see Turney and Pantel, 2010 for a survey)
Vector space models

Words are represented as vectors

- Semantically similar words are close in the space
Term-document matrix

- Useful if you have a set collection of documents
- Rows are words, columns are documents
- For instance, Wikipedia documents and the terms occurring in it:

<table>
<thead>
<tr>
<th></th>
<th>dog</th>
<th>Internet</th>
<th>cartoon</th>
<th>cartoonist</th>
</tr>
</thead>
<tbody>
<tr>
<td>count</td>
<td>3</td>
<td>4</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>count</td>
<td>10</td>
<td>0</td>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>7</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Term-document matrix

- Useful if you have a set collection of documents
- Rows are words, columns are documents
- For instance, Wikipedia documents and the terms occurring in it:

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>dog</td>
<td>3</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Internet</td>
<td>4</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>7</td>
</tr>
<tr>
<td>cartoon</td>
<td>1</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>cartoonist</td>
<td>1</td>
<td>0</td>
<td>10</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Term-document matrix

- Useful if you have a set collection of documents
- Rows are words, columns are documents
- For instance, Wikipedia documents and the terms occurring in it:

<table>
<thead>
<tr>
<th>Term</th>
<th>3</th>
<th>10</th>
<th>0</th>
<th>0</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>dog</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Internet</td>
<td>4</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>7</td>
</tr>
<tr>
<td>cartoon</td>
<td>1</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>cartoonist</td>
<td>1</td>
<td>0</td>
<td>10</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

A word is defined by a **vector of counts over documents**
Generalization: the word-context matrix

A document might not be the best item for measuring word similarity.

What is the optimal granularity of "context" to measure the similarity between words?
  - n-gram, sentence, grammatical relations, paragraph, document, etc.

Distributional hypothesis (Harris, 1954): words occurring in similar context tend to have similar meanings.
A word is defined by a vector of counts over documents

Extract and count the cooccurrences in a corpus

| brown fox jumps over the lazy | dog | " . Changing the numbers will
| near his feet, is a sleeping pet | dog | . This effigy seems from the bearings
| feet is an animal, probably a | dog | , and the hands are joined in the
| again! The car is regarded as | dog | 's property, Sue gets her bottom
| only superstar who looks like a | dog | . Oh, and a final thought about
| you have to do to bring your pet | dog | , cat or ferret into ( or back
| burying a living child, a calf, a | dog | , goat, or lamb the lamb slain
| fly tipping / litter enforcement | dog | warden service dog fouling gypsies
| enforcement dog warden service | dog | fouling gypsies and travellers
| really coach a man like you would a | dog | ? Or is Katie about to learn that
Distributional semantics

Resulting in a cooccurrence vector, e.g.:

\[ \text{dog} = (10, 25, 3, 5, 7, 8, ..., 5) \]

Dog is expected to be more similar to other mammals than to, e.g., cartoonist

A word is defined by a vector of counts over contexts
What values should we use for non-zero components?

Beyond raw counts, we can calculate functions of term frequency, cooccurrence, frequency in topics, etc.

For term-document matrices, we can use TF-IDF:

$$TF - IDF(t, d) = \frac{f_{t,d}}{|d|} \times \log \frac{|D|}{|\{d: t \in d\}|}$$

or lexical specificity (less sensitive to document lengths)
What values should we use for non-zero components?

For term-context matrices, we can use

- Dice:
  \[
  Dice(w, w') = \frac{2c(w, w')}{c(w) + c(w')}
  \]

- Pointwise Mutual Information:
  \[
  PMI(w, w') = \log \frac{P(w, w')}{P(w)P(w')}
  \]

- Positive Pointwise Mutual Information:
  \[
  PPMI(w, w') = \begin{cases} 
  PMI(w, w') & \text{if } PMI(w, w') > 0 \\
  0 & \text{else}
  \end{cases}
  \]

  (However, biased towards infrequent words)
Comparing word representations

• Parametric
  – Cosine
  \[ \cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}} \]
  – Tanimoto similarity
  \[ f(A, B) = \frac{A \cdot B}{|A|^2 + |B|^2 - A \cdot B} \]
  – Kullback–Leibler (KL) divergence
  \[ D_{KL}(P \parallel Q) = \sum_{i} P(i) \log \frac{P(i)}{Q(i)} \]
  – Jensen–Shannon (JS) divergence
  \[ JSD(P \parallel Q) = \frac{1}{2} D(P \parallel M) + \frac{1}{2} D(Q \parallel M) \quad \text{where} \quad M = \frac{1}{2}(P + Q) \]

• Non-parametric
  – Rank-Biased Overlap
  \[ RBO(S_1, S_2) = (1 - p) \sum_{d=1}^{|H|} p^{d-1} \frac{|H_d|}{d} \]
  – Weighted Overlap
  \[ WO(v_1, v_2) = \frac{\sum_{q \in O} \left( \text{rank}(q, v_1) + \text{rank}(q, v_2) \right)^{-1}}{\sum_{i=1}^{|O|}(2i)^{-1}} \]
Small is good

• Vectors often have thousands to millions of dimensions
• The dimensionality of these vectors can be reduced in many different ways:
  – Random indexing
  – Non-negative matrix factorization
  – Singular Value Decomposition
  – Latent Dirichlet Allocation
  – Neural Network Embeddings
The word2vec architectures (Mikolov et al., 2013)
wevi: a tool for understanding word2vec [Rong, 2014]
Much work on vector representations of meaning

Bengio et al. (2003)

Collobert & Weston (2008)

<table>
<thead>
<tr>
<th>Probability and Ratio</th>
<th>$k = \text{solid}$</th>
<th>$k = \text{gas}$</th>
<th>$k = \text{water}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P(k</td>
<td>\text{ice})$</td>
<td>$1.0 \times 10^{-4}$</td>
<td>$6.6 \times 10^{-5}$</td>
</tr>
<tr>
<td>$P(k</td>
<td>\text{steam})$</td>
<td>$2.2 \times 10^{-6}$</td>
<td>$7.8 \times 10^{-4}$</td>
</tr>
<tr>
<td>$P(k</td>
<td>\text{ice})/P(k</td>
<td>\text{steam})$</td>
<td>$8.9$</td>
</tr>
</tbody>
</table>

Mikolov et al. (2013)

Pennington et al. (2014)
Why?

Embedded vector representations:

- are compact and fast to compute
- preserve important relational information between words (actually, meanings):

\[ \text{king} - \text{man} + \text{woman} \approx \text{queen} \]

- are geared towards general use (word2vec, GloVe)
- are a successful example of unsupervised learning
Applications for word representations

- Semantic similarity
- Word clustering
- Word Sense Induction
- Word Sense Disambiguation and Entity Linking
- Semantic role labeling
- Plagiarism detection
- Automated essay marking
- (Open) Information extraction
The dream: machine reading
The word level is not enough

Word representations alone are **not enough** to perform a number of tasks

- at the sentence, paragraph and document level
- at the sense level

Let's see for example what we could do with **semantic similarity**
Semantic Similarity at different levels

Sentence Level

Word Level

Sense Level
Semantic Similarity at different levels

Sentence Level

Word Level

Sense Level
Semantic Similarity at different levels

Word level

Applications

- Lexical simplification
  (Biran et al., 2011)

  *Locuacious* $\rightarrow$ *Talkative*

- Lexical substitution
  (McCarthy and Navigli, 2009)
Semantic Similarity at different levels

Sentence Level

Word Level

Sense Level
Semantic Similarity at different levels

Sentence level

➢ Applications

➢ Paraphrase recognition
  (Tsatsaronis et al., 2010)

➢ MT evaluation
  (Kauchak and Barzilay, 2006)

• Question Answering
  (Surdeanu et al., 2011)

• Textual Entailment
  (Dagan et al., 2006)

The worker was terminated

The boss fired him
Semantic Similarity at different levels

Sentence Level

Word Level

Sense Level

ACL Tutorial 2016: Semantic Representation of Word Senses and Concepts
Camacho-Collados, Iacobacci, Navigli, Pilehvar
Semantic Similarity at different levels

Sense level

Applications

• Coarsening sense inventories (Navigli, 2006; Snow et al., 2007)

• Semantic priming (Neely et al., 1989)

• Word Sense Disambiguation (Navigli, 2009)
Problem 1:
word representations cannot capture polysemy
Problem 1: word representations cannot capture polysemy
Problem 1: word representations cannot capture polysemy
Word representations and the triangular inequality

Example from Neelakantan et al (2014)

For distance $d$, $d(a, c) \leq d(a, b) + d(b, c)$.
Word representations and the triangular inequality

Example from Neelakantan et al (2014)

For distance $d$, $d(a, c) \leq d(a, b) + d(b, c)$.
Problem 2: word representations do not take advantage of existing semantic resources
Example: the sense inventory of "bank" in BabelNet

We want to create a separate representation for each sense of a given word.
Key goal: obtain sense representations
Key goal: obtain sense representations
Sense Representation
Techniques

Introduction
Two types of sense representation techniques

Linked to sense inventories

Knowledge-based

Not linked
Unsupervised (Multi prototype)
Unsupervised Sense Representations

Induce senses, then learn representations for the induced senses

Usually coupled with clustering
Unsupervised Sense Representations

Induce senses, then learn representations for the induced senses

... chose Zbigniew Brzezinski for the position of ...
... thus the symbol s position on his clothing was ...
... writes call options against the stock position ...
... offered a position with ...
... a position he would hold until his retirement in ...
... endanger their position as a cultural group...
... on the chart of the vessel s current position ...
... not in a position to help...

Image courtesy of Reisinger and Mooney (2010)
Unsupervised Sense Representations

Features:

- **Do not rely** on external *sense inventories*
- **Clustering** algorithms are generally used for distinguishing senses from each other
- Resulting sense representations are **not** linked to any inventory
Knowledge-based Sense Representations
Knowledge-based Sense Representations

Represent word senses as defined by sense inventories

**plant**

- *plant, works, industrial plant* (buildings for carrying on industrial labor)
- *plant, flora, plant life* ((botany) a living organism lacking the power of locomotion)
- *plant* (an actor situated in the audience whose acting is rehearsed but seems spontaneous to the audience)
- *plant* (something planted secretly for discovery by another)
Knowledge-based Sense Representations

Represent word senses as defined by sense inventories

Exploit various types of knowledge encoded in these resources:

sense definitions, synonymy, polysemy, semantic relations, structure, etc.
Knowledge-based Sense Representations

Features:

- Use **knowledge from lexical-semantic resources** for distinguishing senses from each other
- The resulting sense representations are **linked to the inventory**, hence useful for applications such as WSD
Knowledge-based Sense Representations

Represent word senses as defined by sense inventories

WordNet: the most commonly used

But also

Wikipedia
BabelNet
Freebase
WordNet

Main unit: synset (concept)

**synset**

**the middle of the day**
Noon, twelve noon, high noon, midday, noonday, noontide

**word sense**

**electronic device**
television, telly, television set, tv, tube, tv set, idiot box, boob tube, goggle box
WordNet semantic relations

- a protective covering that is part of a plant hood, cap
- (botany) a living organism lacking the power of locomotion plant, flora, plant life
- a living thing that has (or can develop) the ability to act or function independently organism, being
- the branch of biology that studies plants botany
- any of a variety of plants grown indoors for decorative purposes houseplant
WordNet

WordNet Search - 3.1
- WordNet home page - Glossary - Help

Word to search for: plant  
Search WordNet

Display Options: (Select option to change)  
Change

Key: "S." = Show Synset (semantic) relations, "W." = Show Word (lexical) relations
Display options for sense: (gloss) "an example sentence"

Noun

- S: (n) plant, works, industrial plant (buildings for carrying on industrial labor) "they built a large plant to manufacture automobiles"
- S: (n) plant, flora, plant life ((botany) a living organism lacking the power of locomotion)
- S: (n) plant (an actor situated in the audience whose acting is rehearsed but seems spontaneous to the audience)
- S: (n) plant (something planted secretly for discovery by another) "the police used a plant to trick the thieves"; "he claimed that the evidence against him was a plant"

Verb

- S: (v) plant, set (put or set (seeds, seedlings, or plants) into the ground) "Let’s plant flowers in the garden"
- S: (v) implant, engraft, embed, imbed, plant (fix or set securely or deeply) "He planted a knee in the back of his opponent"; "The dentist implanted a tooth in the gum"
- S: (v) establish, found, plant, constitute, institute (set up or lay the groundwork for) "establish a new department"
- S: (v) plant (place into a river) "plant fish"
- S: (v) plant (place something or someone in a certain position in order to secretly observe or deceive) "Plant a spy in Moscow"; "plant bugs in the dissident’s apartment"
- S: (v) plant, implant (put firmly in the mind) "Plant a thought in the students’ minds"

Link to online browser
Knowledge-based Sense Representations


M. T. Pilehvar and N. Collier, De-Conflated Semantic Representations (EMNLP 2016)

Chen et al (2014)

A Unified Model for Word Sense Representation and Disambiguation

Basic idea: word sense representation and Word Sense Disambiguation can benefit from each other

Joint word sense representation and disambiguation
Chen et al (2014)

1- Use a sense definition to initialize its representation

plant, flora, plant life ((botany) a living organism lacking the power of locomotion)

Ssense representation + word embeddings
Chen et al (2014)

1- Use a sense definition to initialize its representation
2- Automatically disambiguate large amounts of text

They proposed simple disambiguation techniques based on the obtained initial sense representations and used these disambiguation techniques to disambiguate large amounts of texts
Chen et al (2014)

Disambiguation Technique

To disambiguate a content word (plant):

\textit{water is absorbed by roots of a plant from the soil}

- Obtain the sentence representation (by averaging word embeddings)
- Pick the sense of \textit{plant} which has the highest cosine similarity to the sentence vector
Chen et al (2014)

1- Use a sense definition to initialize its representation
2- Automatically disambiguate large amounts of text
3- Modify the objective of Skip-gram to learn sense representations
Chen et al (2014)

Experiments and evaluation

Word similarity measurement

The most commonly used benchmark for the evaluation of sense representation techniques
Chen et al (2014)

Experiments and evaluation

Word similarity measurement

The most common measures for the evaluation of semantic relatedness are:

- RG-65
- MC-30
- TOEFL
- MEN
- WordSim-353
- SimLex-999
- SCWS
- ....and many more

Also, the SemEval-2017 task on Multilingual and Cross-lingual Word Similarity
Chen et al (2014)

Experiments and evaluation

Word similarity measurement

But:

- How **vector** representations are used to measure **semantic similarity**?
- How **sense** representations are used for measuring **word similarity**?
Vector Comparison
Cosine Similarity

The most commonly used measure for the similarity of vector space model (sense) representations

\[ \text{sim}(A, B) = \cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|} \]
How are sense representations used for word similarity?

Usually, four techniques are used (Reisinger and Mooney, 2010):

1- MaxSim

2- AvgSim

3- MaxSimC

4- AvgSimC
How are sense representations used for word similarity?

1- **MaxSim**: pick the similarity between the most similar senses across two words

\[
\text{MaxSim}(w, w') \overset{\text{def}}{=} \max_{1 \leq j \leq K, 1 \leq k \leq K'} d(\pi_k(w), \pi_j(w'))
\]
How are sense representations used for word similarity?

2- **AvgSim**: average the similarities between senses across two words

\[
\text{AvgSim}(w, w') \overset{\text{def}}{=} \frac{1}{KK'} \sum_{j=1}^{K} \sum_{k=1}^{K'} d(\pi_k(w), \pi_j(w'))
\]
How are sense representations used for word similarity?

For some datasets, words are provided with contexts, e.g., Stanford Contextual Word Similarity (SCWS)

**plant**

In a thermal power plant heat energy is converted to electric power.

**tree**

Almost 400 billion trees grow in the Amazon rainforest.
How are sense representations used for word similarity?

3- **MaxSimC**: the similarity between the “most appropriate” senses of the two words

\[
\text{MaxSimC}(w, w') \overset{\text{def}}{=} d(\hat{\pi}(w), \hat{\pi}(w'))
\]

The most appropriate sense of the word w given the context
How are sense representations used for word similarity?

4- **AvgSimC**: average of pairwise similarities weighted by their appropriateness in context

\[
\text{AvgSimC}(w, w') \overset{\text{def}}{=} \frac{1}{KK'} \sum_{j=1}^{K} \sum_{k=1}^{K'} d_{c,w,k}^{d_{c',w',j}} d(\pi_k(w), \pi_j(w'))
\]

In a thermal power plant 1 heat energy is converted to electric power.

Almost 400 billion growing in the Amazon rainforest.

The likelihood of a sense in the context
Chen et al (2014)

Results on the SCWS dataset:

<table>
<thead>
<tr>
<th>Model</th>
<th>$\rho \times 100$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Our Model-S</td>
<td>64.2</td>
</tr>
<tr>
<td>Our Model-M</td>
<td>68.9</td>
</tr>
</tbody>
</table>

Sense representations usually improve over word representations on word similarity benchmarks.
Chen et al (2014)

Limitations:

- Content words in definitions are not always enough for accurately pinpointing the semantics of a word sense
- The disambiguation technique is far from optimal which introduces noise to the representation procedure
Rothe and Schütze (2015)

AutoExtend: Extending Word Embeddings to Embeddings for Synsets and Senses

*the middle of the day*
Noon, twelve noon, high noon, midday, noonday, noontide
Rothe and Schütze (2015)

AutoExtend: Extending Word Embeddings to Embeddings for Synsets and Senses

Leverages WordNet properties (constraints) for learning sense representations

polysemy and synonymy
Rothe and Schütze (2015)

Two basic premises:

1- A word is the sum of its senses

e.g., embedding of plant is the sum of embeddings of plant(organism), plant(industry), etc.

2- A synset is the sum of its senses

e.g., embedding of this synset is:
plant (organism) + flora (organism) + plant_life (organism)
Rothe and Schütze (2015)

An autoencoder framework for learning

Illustration from Rothe and Schütze (2015)

Words  Senses  Synsets  Senses  Words
Rothe and Schütze (2015)

Word similarity experiments

Stanford Contextual Word Similarity

<table>
<thead>
<tr>
<th></th>
<th>AvgSim</th>
<th>AvgSimC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>62.8,\uparrow</td>
<td>65.7,\uparrow</td>
</tr>
<tr>
<td>2</td>
<td>–</td>
<td>65.4,\uparrow</td>
</tr>
<tr>
<td>3</td>
<td>67.2</td>
<td>69.3</td>
</tr>
<tr>
<td>4</td>
<td>66.2,\uparrow</td>
<td>68.9</td>
</tr>
<tr>
<td>5</td>
<td>66.6,\uparrow</td>
<td>66.6,\uparrow</td>
</tr>
<tr>
<td>6</td>
<td>62.6,\uparrow</td>
<td>63.7,\uparrow</td>
</tr>
<tr>
<td>7</td>
<td>68.9</td>
<td>69.8</td>
</tr>
</tbody>
</table>
Johansson and Nieto Piña (2015)

Embedding a Semantic Network in a Word Space
(NAACL 2015, short)

Learns sense embeddings in the same semantic space as (pre-trained) word embeddings

Applied to Swedish data:

SALDO semantic network
The distances between neighbours to be minimized, while satisfying the mix constraint for each lemma:

*a word vector is a convex combination of its senses vectors*
Johansson and Nieto Piña (2015)

Evaluation on classifying frames in FrameNet

<table>
<thead>
<tr>
<th>Frame</th>
<th>P</th>
<th>R</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANIMALS</td>
<td>0.741</td>
<td>0.643</td>
<td>0.689</td>
</tr>
<tr>
<td>FOOD</td>
<td>0.684</td>
<td>0.679</td>
<td>0.682</td>
</tr>
<tr>
<td>PEOPLE_BY_VOCATION</td>
<td>0.595</td>
<td>0.651</td>
<td>0.622</td>
</tr>
<tr>
<td>ORIGIN</td>
<td>0.789</td>
<td>0.691</td>
<td>0.737</td>
</tr>
<tr>
<td>PEOPLE_BY_ORIGIN</td>
<td>0.693</td>
<td>0.481</td>
<td>0.568</td>
</tr>
<tr>
<td>Overall</td>
<td>0.569</td>
<td>0.292</td>
<td>0.386</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Frame</th>
<th>P</th>
<th>R</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANIMALS</td>
<td>0.826</td>
<td>0.663</td>
<td>0.736</td>
</tr>
<tr>
<td>FOOD</td>
<td>0.726</td>
<td>0.743</td>
<td>0.735</td>
</tr>
<tr>
<td>PEOPLE_BY_VOCATION</td>
<td>0.605</td>
<td>0.637</td>
<td>0.621</td>
</tr>
<tr>
<td>ORIGIN</td>
<td>0.813</td>
<td>0.684</td>
<td>0.742</td>
</tr>
<tr>
<td>PEOPLE_BY_ORIGIN</td>
<td>0.756</td>
<td>0.508</td>
<td>0.608</td>
</tr>
<tr>
<td>Overall</td>
<td>0.667</td>
<td>0.332</td>
<td>0.443</td>
</tr>
</tbody>
</table>

(a) Using lemma embeddings.

(b) Using sense embeddings.
Retrofitting (Faruqui et al., NAACL 2015)


Distributional approaches usually rely only on the statistics derived from text corpora. They usually ignore all the valuable information encoded in knowledge resources.
Retrofitting (Faruqui et al., NAACL 2015)

WordNet

The Paraphrase Database

Benefit from synonymy and other semantic relationships in resources

plant, flora, plant life

houseplant

originator
creator
draftsman
proposer
co-author
authorship

writer ≈
Retrofitting (Faruqui et al., NAACL 2015)

Make these vectors more similar to each other
Jauhar et al. (NAACL 2015)

Ontologically Grounded Multi-sense Representation Learning for Semantic Vector Space Models (S. K. Jauhar, C. Dyer and E. Hovy)

Two techniques for learning sense-specific embeddings that are linked to WordNet: Retro and EM
\[ C(V) = \arg \min_V \sum_{i-j} \alpha \|\hat{u}_i - v_{ij}\|^2 + \sum_{ij-i'j'} \beta_r \|v_{ij} - v_{i'j'}\|^2 \]
EM: Extends the skip-gram model to learn ontologically-grounded sense vectors

\[ C(\theta) = \arg \max_{\theta} \sum_{(w_i, c_i) \in D} \log \left( \sum_{s_{ij}} p(c_i | s_{ij}; \theta) \right) \times \\
p(s_{ij} | w_i; \theta) - \gamma \sum_{ij-i'j'} \beta_r \| v_{ij} - v_{i'j'} \|^2 \]

Ontological prior
De-Conflated Semantic Representations

Approaches so far

M. T. Pilehvar and N. Collier
(EMNLP 2016)
De-Conflated Semantic Representations

Uses Personalized PageRank algorithm to exploit WordNet for sense specific information

\[ \tilde{v}^{(t)} = (1 - \alpha) M \tilde{v}^{(t-1)} + \alpha \tilde{v}^{(0)} \]

**Digit**

<table>
<thead>
<tr>
<th>#</th>
<th>Sense biasing words</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>dactyl, finger, toe, thumb, pollex, body_part, nail, minimus, tarsier, webbed, extremity, appendage</td>
</tr>
<tr>
<td>2</td>
<td>figure, cardinal_number, cardinal, integer, whole_number, numeration_system, number_system, system_of_numeration, large_integer, constituent, element, digital</td>
</tr>
</tbody>
</table>
De-Conflated Semantic Representations

M. T. Pilehvar and N. Collier (EMNLP 2016)
De-Conflated Semantic Representations

- Learns a representation $v_{s_i}^*$ for a sense $s_i$ that is:

  - Close to its **lemma** embedding
  - Close to a weighted average of embeddings of its sense biasing words

\[
\arg\min_{v^*} \alpha d(v_{s_i}^* , v_{s_i}) + \sum_{b_{ij} \in B_i} \delta_{ij} d(v_{s_i}^* , v_{b_{ij}})
\]
De-Conflated Semantic Representations

appendage

finger

toe

nail

thumb

wrist

bone

limb

lobe

heel

hip

ankle

foot
## De-Conflated Semantic Representations

### Evaluation: Word Similarity

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Approach</th>
<th>Score</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$r$</td>
<td>$\rho$</td>
</tr>
<tr>
<td>MEN-3K</td>
<td>Iacobacci et al. (2015)</td>
<td>–</td>
<td>80.5</td>
</tr>
<tr>
<td></td>
<td>DECONF</td>
<td>78.0</td>
<td>78.6</td>
</tr>
<tr>
<td></td>
<td>Faruqui et al. (2015)</td>
<td>–</td>
<td>75.9</td>
</tr>
<tr>
<td></td>
<td>Initial word vectors</td>
<td>72.3</td>
<td>73.2</td>
</tr>
<tr>
<td>RG-65</td>
<td>Iacobacci et al. (2015)</td>
<td>–</td>
<td>87.1</td>
</tr>
<tr>
<td></td>
<td>Faruqui et al. (2015)</td>
<td>–</td>
<td>84.2</td>
</tr>
<tr>
<td></td>
<td>Initial word vectors</td>
<td>77.2</td>
<td>76.1</td>
</tr>
<tr>
<td>YP-130</td>
<td>Iacobacci et al. (2015)</td>
<td>–</td>
<td>63.9</td>
</tr>
<tr>
<td></td>
<td>Initial word vectors</td>
<td>58.0</td>
<td>55.9</td>
</tr>
<tr>
<td>SL-999</td>
<td>Goikoetxea et al. (2015)</td>
<td>–</td>
<td>55.2</td>
</tr>
<tr>
<td></td>
<td>Initial word vectors</td>
<td>45.4</td>
<td>44.2</td>
</tr>
</tbody>
</table>
De-Conflated Semantic Representations

Evaluation: **Word to Sense Similarity** (SemEval-2014 task on Cross-Level Semantic Similarity)

Word similarity:  
- plant  farm

Word to sense similarity:  
- plant#2  farm

<table>
<thead>
<tr>
<th>System</th>
<th>MaxSim</th>
<th></th>
<th>AvgSim</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>r</td>
<td>ρ</td>
<td>r</td>
<td>ρ</td>
</tr>
<tr>
<td>DECONF*</td>
<td>36.4</td>
<td>37.6</td>
<td>36.8</td>
<td>38.8</td>
</tr>
<tr>
<td>Rothe and Schütze (2015)*</td>
<td>34.0</td>
<td>33.8</td>
<td>34.1</td>
<td>33.6</td>
</tr>
<tr>
<td>Iacobacci et al. (2015)*</td>
<td>19.1</td>
<td>21.5</td>
<td>21.3</td>
<td>24.2</td>
</tr>
<tr>
<td>Chen et al. (2014)*</td>
<td>17.7</td>
<td>18.0</td>
<td>17.2</td>
<td>16.8</td>
</tr>
<tr>
<td>DECONF</td>
<td>35.5</td>
<td>36.4</td>
<td>36.2</td>
<td>38.0</td>
</tr>
<tr>
<td>Iacobacci et al. (2015)</td>
<td>19.0</td>
<td>21.5</td>
<td>20.9</td>
<td>23.2</td>
</tr>
</tbody>
</table>
Align, Disambiguate, and Walk (ADW)

Align, Disambiguate and Walk: A Unified Approach for Measuring Semantic Similarity (Pilehvar, Jurgens and Navigli, ACL 2013)

From Senses to Texts: An All-in-one Graph-based Approach for Measuring Semantic Similarity (Pilehvar and Navigli, 2015, Artificial Intelligence, 2015)

- Purely based on the knowledge derived from WordNet (no corpus statistics)
- Human-interpretable sense representations (all sense representations covered so far were non-interpretable)
ADW: Semantic Signature
Human-interpretable dimensions

plant (living organism)
ADW: Personalized PageRank
Representing multiple senses

\[ \text{woman}^1_n, \text{fry}^2_v, \text{food}^1_n \]
ADW: Personalized PageRank
Representing multiple senses

$\text{woman}_n^1$, $\text{fry}_v^2$, $\text{food}_n^1$
ACL Tutorial 2016: Semantic Representation of Word Senses and Concepts
Camacho-Collados, Iacobacci, Navigli, Pilehvar
These weights form a semantic signature
Vector Comparison
Weighted Overlap

\[ WO(v_1, v_2) = \frac{\sum_{q \in O} \left( rank(q, v_1) + rank(q, v_2) \right)^{-1}}{\sum_{i=1}^{\left| O \right|} (2i)^{-1}} \]
ADW

Alignment-based disambiguation

a simple technique for using sense representations for measuring semantic similarity of word, phrase or sentence pairs.
ADW

Online demo: http://lcl.uniroma1.it/adw/
ADW
Advantages and limitation

+ Interpretable dimensions
+ Unified representation for all lexical levels: senses, words, phrases and sentences
+ Uses only WordNet as its knowledge resource
+ Rich and highly accurate representations: state-of-the-art performance on multiple NLP tasks and datasets

- Limited coverage (that of WordNet)
  > Solution: use large-scale lexical resources
Large knowledge resources
Large knowledge resources

Wikipedia

BabelNet

FreeBase
Wikipedia

The Free Encyclopedia
Wikipedia

High coverage of named entities and specialized concepts from different domains
A car is a wheeled, self-powered motor vehicle used for transportation. Most definitions of the term specify that cars are designed to run primarily on roads, to have seating for one to eight people, to typically have four wheels, and to be constructed principally for the transport of people rather than goods.\textsuperscript{[3][4]} The year 1886 is regarded as the birth year of the modern car. In that year, German inventor Karl Benz built the Benz Patent-Motorwagen. Cars did not become widely available until the early 20th century. One of the first cars that was accessible to the masses was the 1908 Model T, an American car manufactured by the Ford Motor Company.
Wikipedia hyperlinks

A car is a wheeled, self-powered motor vehicle used for transportation. Most definitions of the term specify that cars are designed to run primarily on roads, to have seating for one to eight people, to typically have four wheels, and to be constructed principally for the transport of people rather than goods.\textsuperscript{[3]}[4] The year 1886 is regarded as the birth year of the modern car. In that year, German inventor Karl Benz built the Benz Patent-Motorwagen. Cars did not become widely available until the early 20th century. One of the first cars that was accessible to the masses was the 1908 Model T, an American car manufactured by the Ford Motor Company.
Wikipedia

~4.3M * Wikipedia pages

>71M * hyperlinks

977M * lemmas

[Based on the slides of Raganato and Delli Bovi (2016)]
Wikipedia

~4.3M * Wikipages

>71M * hyperlinks

977M * lemmas

The Free Encyclopedia

Constantly updating and growing!

270+ active languages!

*English dump 11/2014

[Based on the slides of Raganato and Delli Bovi (2016)]
Wikipedia

- ~4.3M * Wikipages
- >71M hyperlinks
- 977M lemmas

*Based on the slides of Raganato and Delli Bovi (2016)*
Wikipedia as a sense-annotated corpus

~4.3M * concepts or entities

>71M sense annotations

Named Entity Disambiguation (Wikification)

Semantic similarity

Information Extraction

Taxonomies, ontologies and semantic networks

*[English dump 11/2014]*

[Based on the slides of Raganato and Delli Bovi (2016)]
Wikipedia as a semantic network

~4.3M * concept nodes

>71M semantic connections

Named Entity Disambiguation (Wikification)

Semantic similarity

Information Extraction

Taxonomies, ontologies and semantic networks

*English dump 11/2014

[Based on the slides of Raganato and Delli Bovi (2016)]
Semantic Representations exploiting Wikipedia

- **SSA** (Hassan and Mihalcea, AAAI 2011)
- **SaSA** (Wu and Giles, AAAI 2015)
SSA: Salient Semantic Analysis

(Hassan and Mihalcea, AAAI 2011)

It exploits Wikipedia as a sense-annotated corpus using its hyperlinks

It increases the number of links by exploiting the one sense per page heuristic.
SSA: Salient Semantic Analysis

(Hassan and Mihalcea, AAAI 2011)

It exploits Wikipedia as a sense-annotated corpus using its hyperlinks.

It increases the number of links by exploiting the one sense per page heuristic.

This property and other structural properties of Wikipedia have been exploited in Raganato et al. (IJCAI 2016) to build a large sense-annotated corpus.
SSA: Salient Semantic Analysis

(Hassan and Mihalcea, AAAI 2011)

For a given word, it constructs an explicit vector where dimensions are co-occurring Wikipedia pages (weights correspond to normalized frequencies).

Strong results in word, sentence and document relatedness.
SaSA: Sense-aware Semantic Analysis

(Wu and Giles, AAAI 2015)

To be explained in the next section of “Unsupervised sense representations”!
BabelNet

(Navigli and Ponzetto, AIJ 2012)

Thanks to an automatic mapping algorithm, it merges Wikipedia and WordNet, among other resources (Wiktionary, OmegaWiki, WikiData, VerbNet, FrameNet)
BabelNet as a very large semantic network (13.8M synsets and 380M relations)
BabelNet

(Navigli and Ponzetto, AIJ 2012)

Other features:

- **Multilinguality**: 270+ languages
- Integration of **encyclopedic** (named entities) and **lexicographic knowledge** (concepts)
- Synsets associated with **images, domains, definitions, examples**, etc.
BabelNet

Nome

**jaguar, panther, Felis onca**

A large spotted feline of tropical America similar to the leopard; in some classifications considered a member of the genus Felis

ID: 00033987n | Concept

**Jaguar Cars, Jaguar**

Jaguar Cars is a brand of Jaguar Land Rover, a British multinational car manufacturer headquartered in Whitley, Coventry, England, owned by Tata Motors since 2008.

ID: 00038731n | Entity

**Atari Jaguar, Jaguar (video game console)**

The Atari Jaguar is a home video game console that was released by Atari Corporation in 1993.

ID: 02142312n | Entity

**Mac OS X v10.2, Jaguar (macos)**

Mac OS X version 10.2 Jaguar is the third major release of Mac OS X, Apple's desktop and server operating system.
BabelNet

Concept

Entity

jaguar, panther, Felis onca
A large spotted feline of tropical America similar to the leopard; in some classifications considered a member of the genus Felis.

Jaguar Cars, Jaguar
Jaguar Cars is a brand of Jaguar Land Rover, a British multinational car manufacturer headquartered in Whitley, Coventry, England, owned by Tata Motors since 2008.

Atari Jaguar, Jaguar (video game console)
The Atari Jaguar is a home video game console that was released by Atari Corporation in 1993.

Mac OS X v10.2, Jaguar (macos)
Mac OS X version 10.2 Jaguar is the third major release of Mac OS X, Apple’s desktop and server operating system.
BabelNet

It follows the same structure of WordNet: **synsets** are the main units

jaguar, panther, Felis onca

A large spotted feline of tropical America similar to the leopard; in some classifications considered a member of the genus Felis

ID: 00033987n | Concetto
BabelNet

In this case, synsets are multilingual

Nome

**jaguar, panther, Felis onca**

A large spotted feline of tropical America similar to the leopard; in some classifications considered a member of the genus Felis

ID: 00033987n | Concetto
Knowledge-based sense representations exploiting Wikipedia and BabelNet

- **NASARI** (Camacho-Collados et al.; NAACL and ACL 2015, AIJ 2016)

- **SensEmbed** (Iacobacci et al. ACL 2015)
NASARI: Integrating Explicit Knowledge and Corpus Statistics for a Multilingual Representation of Concepts and Entities
(Camacho-Collados et al., AIJ 2016)

Goal

Build vector representations for multilingual BabelNet synsets.

How?

It exploits Wikipedia semantic network and the WordNet taxonomy to construct a subcorpus contextual information for any given BabelNet synset.
NASARI: Integrating Explicit Knowledge and Corpus Statistics for a Multilingual Representation of Concepts and Entities

(Camacho-Collados et al., AIJ 2016)

Process of obtaining contextual information for a BabelNet synset exploiting BabelNet taxonomy and Wikipedia as a semantic network.
NASARI: Integrating Explicit Knowledge and Corpus Statistics for a Multilingual Representation of Concepts and Entities
(Camacho-Collados et al., AIJ 2016)

Three types of vector representations:

- **Lexical** (dimensions are words)

- **Unified** (dimensions are multilingual BabelNet synsets)

- **Embedded** (latent dimensions)
NASARI: Integrating Explicit Knowledge and Corpus Statistics for a Multilingual Representation of Concepts and Entities
(Camacho-Collados et al., AIJ 2016)

Three types of vector representations:

- **Lexical** (dimensions are words): Dimensions are weighted via *lexical specificity* (statistical measure based on the hypergeometric distribution)

- **Unified** (dimensions are multilingual BabelNet synsets): This representation uses a *hypynym-based clustering technique* and can be used in *cross-lingual* applications

- **Embedded** (latent dimensions)
NASARI: Integrating Explicit Knowledge and Corpus Statistics for a Multilingual Representation of Concepts and Entities
(Camacho-Collados et al., AIJ 2016)

Three types of vector representations:

- **Lexical** (dimensions are words): Dimensions are weighted via *lexical specificity* (statistical measure based on the hypergeometric distribution)

- **Unified** (dimensions are multilingual BabelNet synsets): This representation uses a *hyponym-based clustering technique* and can be used in *cross-lingual* applications

- **Embedded** (latent dimensions)
NASARI: Integrating Explicit Knowledge and Corpus Statistics for a Multilingual Representation of Concepts and Entities

(Camacho-Collados et al., AIJ 2016)
NASARI: Integrating Explicit Knowledge and Corpus Statistics for a Multilingual Representation of Concepts and Entities

(Camacho-Collados et al., AIJ 2016)

Three types of vector representations:

- **Lexical** (dimensions are words)
- **Unified** (dimensions are multilingual BabelNet synsets)
- **Embedded**: Low-dimensional vectors (latent) exploiting word embeddings obtained from text corpora. This representation is obtained by plugging word embeddings on the lexical vector representations.
NASARI: Integrating Explicit Knowledge and Corpus Statistics for a Multilingual Representation of Concepts and Entities

(Camacho-Collados et al., AIJ 2016)

Three types of vector representations:

- **Lexical** (dimensions are words)
- **Unified** (dimensions are multilingual BabelNet synsets)

- **Embedded**: Low-dimensional vectors (latent) exploiting *word embeddings* obtained from text corpora. This representation is obtained by plugging word embeddings on the lexical vector representations.

Word and synset embeddings share the same vector space!
NASARI: Integrating Explicit Knowledge and Corpus Statistics for a Multilingual Representation of Concepts and Entities
(Camacho-Collados et al., AIJ 2016)

High coverage of concepts and named entities in several languages (covers all Wikipedia pages).

Useful for multilingual and cross-lingual semantic similarity, Sense Clustering, Domain Labeling and Word Sense Disambiguation.
NASARI: Integrating Explicit Knowledge and Corpus Statistics for a Multilingual Representation of Concepts and Entities

(Camacho-Collados et al., AIJ 2016)

<table>
<thead>
<tr>
<th>English</th>
<th>r</th>
<th>ρ</th>
<th>French</th>
<th>r</th>
<th>ρ</th>
<th>German</th>
<th>r</th>
<th>ρ</th>
<th>Spanish</th>
<th>r</th>
<th>ρ</th>
</tr>
</thead>
<tbody>
<tr>
<td>NASARI</td>
<td>0.81</td>
<td>0.78</td>
<td>NASARI</td>
<td>0.82</td>
<td>0.73</td>
<td>NASARI</td>
<td>0.69</td>
<td>0.65</td>
<td>NASARI</td>
<td>0.85</td>
<td>0.79</td>
</tr>
<tr>
<td>NASARI_lexical</td>
<td>0.80</td>
<td>0.78</td>
<td>NASARI_lexical</td>
<td>0.80</td>
<td>0.70</td>
<td>NASARI_lexical</td>
<td>0.69</td>
<td>0.67</td>
<td>NASARI_lexical</td>
<td>0.85</td>
<td>0.79</td>
</tr>
<tr>
<td>NASARI_unified</td>
<td>0.80</td>
<td>0.76</td>
<td>NASARI_unified</td>
<td>0.82</td>
<td>0.76</td>
<td>NASARI_unified</td>
<td>0.71</td>
<td>0.68</td>
<td>NASARI_unified</td>
<td>0.82</td>
<td>0.77</td>
</tr>
<tr>
<td>NASARI_embed</td>
<td>0.82</td>
<td>0.80</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>SOC-PMI</td>
<td>0.61</td>
<td>–</td>
<td>SOC-PMI</td>
<td>0.19</td>
<td>–</td>
<td>SOC-PMI</td>
<td>0.27</td>
<td>–</td>
<td>SOC-PMI</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>PMI</td>
<td>0.41</td>
<td>–</td>
<td>PMI</td>
<td>0.34</td>
<td>–</td>
<td>PMI</td>
<td>0.40</td>
<td>–</td>
<td>PMI</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>LSA-Wiki</td>
<td>0.65</td>
<td>0.69</td>
<td>LSA-Wiki</td>
<td>0.57</td>
<td>0.52</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Wiki-wup</td>
<td>0.59</td>
<td>–</td>
<td>Wiki-wup</td>
<td>0.65</td>
<td>–</td>
<td>Wiki-wup</td>
<td>0.65</td>
<td>–</td>
<td>Wiki-wup</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Word2Vec</td>
<td>–</td>
<td>0.73</td>
<td>Word2Vec</td>
<td>–</td>
<td>0.47</td>
<td>Word2Vec</td>
<td>–</td>
<td>0.53</td>
<td>Best-Word2Vec</td>
<td>0.80</td>
<td>0.80</td>
</tr>
<tr>
<td>Retrofitting</td>
<td>–</td>
<td>0.77</td>
<td>Retrofitting</td>
<td>–</td>
<td>0.61</td>
<td>Retrofitting</td>
<td>–</td>
<td>0.60</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>NASARI_poly-embed</td>
<td>0.74</td>
<td>0.77</td>
<td>NASARI_poly-embed</td>
<td>0.60</td>
<td>0.69</td>
<td>NASARI_poly-embed</td>
<td>0.46</td>
<td>0.52</td>
<td>NASARI_poly-embed</td>
<td>0.68</td>
<td>0.74</td>
</tr>
<tr>
<td>Polyglot-embed</td>
<td>0.51</td>
<td>0.55</td>
<td>Polyglot-embed</td>
<td>0.38</td>
<td>0.35</td>
<td>Polyglot-embed</td>
<td>0.18</td>
<td>0.15</td>
<td>Polyglot-embed</td>
<td>0.51</td>
<td>0.56</td>
</tr>
<tr>
<td>IAA</td>
<td>0.85</td>
<td>–</td>
<td>IAA</td>
<td>–</td>
<td>–</td>
<td>IAA</td>
<td>0.81</td>
<td>–</td>
<td>IAA</td>
<td>0.83</td>
<td>–</td>
</tr>
</tbody>
</table>

Multilingual Word Similarity
NASARI: Integrating Explicit Knowledge and Corpus Statistics for a Multilingual Representation of Concepts and Entities
(Camacho-Collados et al., AIJ 2016)

<table>
<thead>
<tr>
<th>Measure</th>
<th>EN-FR</th>
<th>EN-DE</th>
<th>EN-ES</th>
<th>FR-DE</th>
<th>FR-ES</th>
<th>DE-ES</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$r$</td>
<td>$\rho$</td>
<td>$r$</td>
<td>$\rho$</td>
<td>$r$</td>
<td>$\rho$</td>
<td>$r$</td>
</tr>
<tr>
<td><strong>NASARI\textsubscript{unified}</strong></td>
<td>0.84</td>
<td>0.79</td>
<td>0.79</td>
<td>0.84</td>
<td>0.75</td>
<td>0.70</td>
<td>0.86</td>
</tr>
<tr>
<td><strong>CL-MSR-2.0</strong></td>
<td>0.30</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><strong>NASARI\textsubscript{pivot}</strong></td>
<td>0.79</td>
<td>0.69</td>
<td>0.78</td>
<td>0.76</td>
<td>0.80</td>
<td>0.74</td>
<td>0.79</td>
</tr>
<tr>
<td><strong>ADW\textsubscript{pivot}</strong></td>
<td>0.80</td>
<td>0.82</td>
<td>0.73</td>
<td>0.82</td>
<td>0.72</td>
<td>0.77</td>
<td>0.81</td>
</tr>
<tr>
<td><strong>Word2Vec\textsubscript{pivot}</strong></td>
<td>0.77</td>
<td>0.82</td>
<td>0.70</td>
<td>0.73</td>
<td>0.76</td>
<td>0.80</td>
<td>0.75</td>
</tr>
<tr>
<td><strong>Best-Word2Vec\textsubscript{pivot}</strong></td>
<td>0.75</td>
<td>0.84</td>
<td>0.69</td>
<td>0.76</td>
<td>0.75</td>
<td>0.82</td>
<td>0.74</td>
</tr>
<tr>
<td><strong>Best-PMI-SVD\textsubscript{pivot}</strong></td>
<td>0.76</td>
<td>0.76</td>
<td>0.72</td>
<td>0.74</td>
<td>0.77</td>
<td>0.77</td>
<td>0.65</td>
</tr>
</tbody>
</table>

Cross-lingual Word Similarity
SensEmbed (Iacobacci et al., ACL 2015)

It leverages BabelNet and Word2Vec to build sense embeddings. Two steps:

• First, it uses Babelfy (Moro et al., TACL 2014), a multilingual joint disambiguation and entity linking system, to disambiguate a corpus.
Babelfy (Moro et al. TACL 2014)

Disambiguation and Entity Linking

Napoléon Bonaparte was a French military and political leader during the French Revolution.

Napoléon Bonaparte
French general who became emperor of the French (1789-1821)

French
Of or pertaining to France or the people of France

military
Of or relating to the study of the principles of warfare

political leader
A person active in party politics

political
Involving or characteristic of politics or parties or politicians- Daniel…

leader
A person who rules or guides or inspires others

French Revolution
The revolution in France against the Bourbons; 1789-1799

Revolution
The overthrow of a government by those who are governed

French
Of or pertaining to France or the people of France
SensEmbed (Iacobacci et al., ACL 2015)

It leverages BabelNet and Word2Vec to build sense embeddings. Two steps:

• First, it uses Babelfy (Moro et al., TACL 2014), a multilingual joint disambiguation and entity linking system, to disambiguate a corpus.
• Then, it uses Word2Vec to learn sense embeddings from the sense-annotated corpus.
SensEmbed (Iacobacci et al., ACL 2015)

SensEmbed construction

...survey on the relationship between the banks and our industry, in preparation for a forthcoming forum.
...and it stands on the right bank of the Drava River, bounded by the river to the north...
...If you have dividend or receive bank or building society interest on which tax has been paid,
...workplaces and unions. Corporations, banks and trusts controlled a great deal and, although machines...
...The critical decision for the banks will come if their own adviser sticks to his view of the costs.
countryside of high hedges and tall earth banks with trees on top. The heavily wooded area was criss-crossed...
SensEmbed (Iacobacci et al., ACL 2015)

SensEmbed construction

...survey on the relationship between the banks and our industry, in preparation for a forthcoming forum.
...and it stands on the right bank of the Drava River, bounded by the river to the north...
...If you have dividend or receive bank or building society interest on which tax has been paid,
...workplaces and unions. Corporations, banks and trusts controlled a great deal and, although machines...
...the critical decision for the banks will come if their own adviser sticks to his view of the costs.
countryside of high hedges and tall earth banks with trees on top. The heavily wooded area was criss-crossed...
SensEmbed (Iacobacci et al., ACL 2015)

**SensEmbed** construction

---

...survey on the relationship between the banks and our industry, in preparation for a forthcoming forum.

...and it stands on the right bank of the Drava River, bounded by the river to the north...

...If you have dividend or receive bank or building society interest on which tax has been paid,

...workplaces and unions. Corporations,

...The critical decision for the countrysides of high hedges and tall earth banks will come if their own adviser sticks to his view of the costs.

banks with trees on top. The heavily wooded area was criss-crossed...

---

...bank...
SensEmbed (Iacobacci et al., ACL 2015)

It leverages the BabelNet semantic network and the sense embeddings for word and relational similarity, tasks in which SensEmbed proves to be very competitive.
## SensEmbed (Iacobacci et al., ACL 2015)

<table>
<thead>
<tr>
<th>Measure</th>
<th>RG-65</th>
<th>WS-Sim</th>
<th>WS-Rel</th>
<th>YP-130</th>
<th>MEN</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pilehvar et al. (2013)</td>
<td>0.868</td>
<td>0.677</td>
<td>0.457</td>
<td>0.710</td>
<td>0.690</td>
<td>0.677</td>
</tr>
<tr>
<td>Zesch et al. (2008)</td>
<td>0.820</td>
<td>—</td>
<td>—</td>
<td>0.710</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Collobert and Weston (2008)</td>
<td>0.480</td>
<td>0.610</td>
<td>0.380</td>
<td>—</td>
<td>0.570</td>
<td>—</td>
</tr>
<tr>
<td>Word2vec (Baroni et al., 2014)</td>
<td>0.840</td>
<td>0.800</td>
<td>0.700</td>
<td>—</td>
<td>0.800</td>
<td>—</td>
</tr>
<tr>
<td>GloVe</td>
<td>0.769</td>
<td>0.666</td>
<td>0.559</td>
<td>0.577</td>
<td>0.763</td>
<td>0.737</td>
</tr>
<tr>
<td>ESA</td>
<td>0.749</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>PMI-SVD</td>
<td>0.738</td>
<td>0.659</td>
<td>0.523</td>
<td>0.337</td>
<td>0.726</td>
<td>0.695</td>
</tr>
<tr>
<td>Word2vec</td>
<td>0.732</td>
<td>0.707</td>
<td>0.476</td>
<td>0.343</td>
<td>0.665</td>
<td>0.644</td>
</tr>
<tr>
<td><strong>SensEmbed_closest</strong></td>
<td><strong>0.894</strong></td>
<td><strong>0.756</strong></td>
<td><strong>0.645</strong></td>
<td><strong>0.734</strong></td>
<td><strong>0.779</strong></td>
<td><strong>0.769</strong></td>
</tr>
<tr>
<td><strong>SensEmbed_weighted</strong></td>
<td><strong>0.871</strong></td>
<td><strong>0.812</strong></td>
<td><strong>0.703</strong></td>
<td><strong>0.639</strong></td>
<td><strong>0.805</strong></td>
<td><strong>0.794</strong></td>
</tr>
</tbody>
</table>

---

**Word Similarity (Spearman correlation)**
SensEmbed (Iacobacci et al., ACL 2015)

<table>
<thead>
<tr>
<th>Measure</th>
<th>MaxDiff</th>
<th>Spearman</th>
</tr>
</thead>
<tbody>
<tr>
<td>Com</td>
<td>45.2</td>
<td>0.353</td>
</tr>
<tr>
<td>PairDirection</td>
<td>45.2</td>
<td>—</td>
</tr>
<tr>
<td>RNN-1600</td>
<td>41.8</td>
<td>0.275</td>
</tr>
<tr>
<td>UTD-LDA</td>
<td>—</td>
<td>0.334</td>
</tr>
<tr>
<td>UTD-NB</td>
<td>39.4</td>
<td>0.229</td>
</tr>
<tr>
<td>UTD-SVM</td>
<td>34.7</td>
<td>0.116</td>
</tr>
<tr>
<td>PMI baseline</td>
<td>33.9</td>
<td>0.112</td>
</tr>
<tr>
<td>Word2vec</td>
<td>43.2</td>
<td>0.288</td>
</tr>
<tr>
<td><strong>SensEmbed_{closest}</strong></td>
<td><strong>45.9</strong></td>
<td><strong>0.358</strong></td>
</tr>
</tbody>
</table>

Relational Similarity
SensEmbed (Iacobacci et al., ACL 2015)

It has also shown its effectiveness in Taxonomy Learning (Espinosa-Anke et al. AAAI, 2016) and Open Information Extraction (Delli Bovi et al., EMNLP 2015) tasks.

We will see more about this on the “Applications” section!
FreeBase

FreeBase was a large collaborative knowledge base.

It was finally shut down on May 2016, but the data was transferred to WikiData.

It is the core of the Google Knowledge Graph.
WikiData

WikiData is a large collaborative knowledge base (18M items).

It is based on Wikipedia and it provides a large set of relations (including a large taxonomy) among item. It exploits Wikipedia infoboxes.

Example: Madrid capital of Spain
WikiData

Spain (Q29)

country in southwestern Europe

Kingdom of Spain | ES | España

In more languages

<table>
<thead>
<tr>
<th>Language</th>
<th>Label</th>
<th>Description</th>
<th>Also known as</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>Spain</td>
<td>country in southwestern Europe</td>
<td>Kingdom of Spain, ES, España</td>
</tr>
<tr>
<td>Spanish</td>
<td>España</td>
<td>país de Europa</td>
<td>Reino de España</td>
</tr>
<tr>
<td>Italian</td>
<td>Spagna</td>
<td>Stato dell'Europa sud-occidentale, membro dell'Unione europea</td>
<td>Regno di Spagna</td>
</tr>
<tr>
<td>French</td>
<td>Espagne</td>
<td>pays d'Europe</td>
<td>Royaume d'Espagne</td>
</tr>
</tbody>
</table>

More languages

Statements

| capital | Madrid |

1 reference
WikiData
TransE: Translating Embeddings for Modeling Multi-relational Data

(Bordes et al., NIPS 2013)

Idea: Learn representations in a vector space not only for entities but also for relations.
TransE: Translating Embeddings for Modeling Multi-relational Data

Relations become Translations

(Bordes et al., NIPS 2013)
Given a training set of triples \((h,l,t)\), TransE learns embeddings for entities and their relationships by minimizing the following loss function:

\[
\mathcal{L} = \sum_{(h,l,t)\in S} \sum_{(h',l,t')\in S'} \left[ \gamma + d(h + l, t) - d(h' + l, t') \right]_+
\]

(Bordes et al., NIPS 2013)
TransE: Translating Embeddings for Modeling Multi-relational Data

(Bordes et al., NIPS 2013)

It has proved its effectiveness in learning relations in two different lexical resources: WordNet and FreeBase.
TransE: Translating Embeddings for Modeling Multi-relational Data

(Bordes et al., NIPS 2013)

New works based on the original TransE:

- **pTransE**: Joint embedding of words and entities (Wang et al., EMNLP 2014)

- **TransH**: Improving the relation mapping (Wang et al., AAAI 2014)

- **TransR**: Learning embeddings of entities and relations in separate spaces (Lin et al., AAAI 2015)

- **TransD**: Dynamic mapping for each entity-relation pair in separated spaces (Ji et al., ACL 2015)
Unsupervised sense representations
Multi-prototype Representations

Why Unsupervised?

Why do we need them?

What for?
Unsupervised Learning

“Given a set of observations [...] the goal is to directly infer the properties of this probability density without the help of a supervisor or teacher providing correct answers or degree-of-error for each observation.”

Hastie, Friedman, Tibshirani, 2001
Unsupervised Learning

Most commonly-used techniques:

- **Clustering** or data segmentation has the goal of **grouping** a collection of objects into subsets or “clusters,” such that those within each cluster are more **closely related**

- **Principal components** are a sequence of projections of features which are **mutually uncorrelated** and ordered in variance
Distributional Hypothesis

“words that occur in the same contexts tend to have similar meanings”
Harris, 1954

“a word is characterized by the company it keeps”
Firth, 1957
Unsupervised Word Sense Disambiguation

It aims to divide “the occurrences of a word into a number of classes by determining for any two occurrences whether they belong to the same sense or not”

Schütze 1998
Unsupervised Word Sense Disambiguation

Main approaches

- Based on Clustering
- Joint training of multiple prototypes
- Exploiting bilingual corpora
Cluster-based sense representations
Cluster-based sense representations

- They are generally split in two steps:
  - Discrimination of senses
  - Single/Multiple prototype training

- They have a bounded (fixed) amount of prototypes

- Generally clustering considers no overlaps between clusters
It presents a vector-space model that represents a word’s meaning by a set of distinct “sense specific” vectors.

The set of vectors for a word is determined by clustering the contexts in which a word appears.

Explicit feature vectors based on unigrams
... chose Zbigniew Brzezinski for the position of ... 
... thus the symbol's position 
on his clothing was ... 
... writes call options against 
the stock position ... 
... offered a position with ... 
... a position he would hold 
until his retirement in ... 
... endanger their position as 
a cultural group... 
... on the chart of the vessel's 
current position ... 
... not in a position to help...

(cluster#1) location importance bombing
(cluster#2) post appointme nt, role, job
(cluster#3) intensity, winds, hour, gust
(cluster#4) lineman, tackle, role, scorer

(collect contexts) (cluster) (similarity)
It measures similarity between two words, $w$ and $w'$, by calculating the minimum distance in terms of cosine similarity between $w$ and $w'$ sense vectors:

$$\text{MaxSim}(w, w') \overset{\text{def}}{=} \max_{1 \leq j \leq K, 1 \leq k \leq K} d(\pi_k(w), \pi_j(w'))$$
It presents a model that unlike Reisinger and Mooney, where only local context (i.e., co-occurrences) is used, leverages also global context (i.e. document topics) for learning multiple prototype vectors.
Improving Word Representations via Global Context and Multiple Word Prototypes

Huang et al., ACL 2012
Senses are represented with latent features in a 50-dimensional embedding space.

The representations are clustered via fixed-size context windows in order to discriminate the single-prototype representation into its different meanings.
## Improving Word Representations via Global Context and Multiple Word Prototypes

Huang et al., ACL 2012

<table>
<thead>
<tr>
<th>Center Word</th>
<th>Nearest Neighbors</th>
</tr>
</thead>
<tbody>
<tr>
<td>bank_1</td>
<td>corporation, insurance, company</td>
</tr>
<tr>
<td>bank_2</td>
<td>shore, coast, direction</td>
</tr>
<tr>
<td>star_1</td>
<td>movie, film, radio</td>
</tr>
<tr>
<td>star_2</td>
<td>galaxy, planet, moon</td>
</tr>
<tr>
<td>cell_1</td>
<td>telephone, smart, phone</td>
</tr>
<tr>
<td>cell_2</td>
<td>pathology, molecular, physiology</td>
</tr>
<tr>
<td>left_1</td>
<td>close, leave, live</td>
</tr>
<tr>
<td>left_2</td>
<td>top, round, right</td>
</tr>
</tbody>
</table>
It also includes a new dataset for measuring Multi-Prototype representations that has become the *de facto* evaluation for sense-based representations: Stanford Contextual Word Similarity or *SCWS*.
Improving Word Representations via Global Context and Multiple Word Prototypes

Huang et al., ACL 2012

<table>
<thead>
<tr>
<th>Model</th>
<th>$\rho \times 100$</th>
</tr>
</thead>
<tbody>
<tr>
<td>C&amp;W-S</td>
<td>57.0</td>
</tr>
<tr>
<td>Our Model-S</td>
<td>58.6</td>
</tr>
<tr>
<td>Our Model-M AvgSim</td>
<td>62.8</td>
</tr>
<tr>
<td>Our Model-M AvgSimC</td>
<td>65.7</td>
</tr>
<tr>
<td>$tfidf-S$</td>
<td>26.3</td>
</tr>
<tr>
<td>Pruned $tfidf-S$</td>
<td>62.5</td>
</tr>
<tr>
<td>Pruned $tfidf-M$ AvgSim</td>
<td>60.4</td>
</tr>
<tr>
<td>Pruned $tfidf-M$ AvgSimC</td>
<td>60.5</td>
</tr>
</tbody>
</table>

Reisinger and Mooney, 2010
It provides “sense-specific” prototypes of a word by clustering Wikipedia pages based on both local (i.e. co-occurrences) and global contexts (i.e. links and categories) of the word in Wikipedia.

Each dimension of the vector space is a Wikipedia concept or article where a word appears or co-occurs with.

Wu and Giles, AAAI 2015

Wu and Giles, AAAI 2015

<table>
<thead>
<tr>
<th>Model</th>
<th>$\rho$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ESA</td>
<td>0.518</td>
</tr>
<tr>
<td>SSA</td>
<td>0.509</td>
</tr>
<tr>
<td>Pruned tfidf-M</td>
<td>0.605</td>
</tr>
<tr>
<td>Huang et al. 2012</td>
<td>0.657</td>
</tr>
<tr>
<td>$SaSA_1$</td>
<td>0.662</td>
</tr>
<tr>
<td>$SaSA_K$</td>
<td>0.664</td>
</tr>
</tbody>
</table>

Reisinger and Mooney, 2010
K-Embeddings: Learning Conceptual Embeddings for Words using Context

Vu and Parker, NAACL 2016

It proposes an extension of word embedding as an iterative algorithm.

It has latent representations based on the chosen word embeddings model.
K-Embeddings: Learning Conceptual Embeddings for Words using Context

Vu and Parker, NAACL 2016

It clusters the context embeddings and uses those clusters as sense annotations for training sense embeddings.

The resulting annotation could be used as input to refine the clusters (iterative).
K-Embeddings: Learning Conceptual Embeddings for Words using Context

Vu and Parker, NAACL 2016
K-Embeddings: Learning Conceptual Embeddings for Words using Context

Vu and Parker, NAACL 2016

The convergence of the number of prototypes

<table>
<thead>
<tr>
<th>$K$</th>
<th>total embeddings</th>
<th>vocabulary size</th>
<th>ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1,965,139</td>
<td>1,965,139</td>
<td>1.00</td>
</tr>
<tr>
<td>5</td>
<td>2,807,016</td>
<td>1,443,061</td>
<td>1.95</td>
</tr>
<tr>
<td>10</td>
<td>2,740,351</td>
<td>1,474,704</td>
<td>1.86</td>
</tr>
<tr>
<td>15</td>
<td>3,229,945</td>
<td>1,374,055</td>
<td>2.35</td>
</tr>
<tr>
<td>20</td>
<td>3,236,882</td>
<td>1,410,521</td>
<td>2.29</td>
</tr>
<tr>
<td>25</td>
<td>3,382,722</td>
<td>1,383,162</td>
<td>2.45</td>
</tr>
<tr>
<td>30</td>
<td>3,404,150</td>
<td>1,418,027</td>
<td>2.40</td>
</tr>
</tbody>
</table>
K-Embeddings: Learning Conceptual Embeddings for Words using Context

Vu and Parker, NAACL 2016

Accuracy on Microsoft Research Syntactic Analogies Dataset (Mikolov et al., 2013)
Joint training of sense representations
Joint training of sense representations

- The training is done in a single step
- No assumption on sense overlap (unlike cluster-based techniques)
- No assumption in the number of prototypes
- Allows to have a shared space of words and senses as an emergent behavior of the model
Efficient Non-parametric Estimation of Multiple Embeddings per Word in Vector Space
Neelakantan et al., EMNLP 2014

• An extension of Skip-gram model

• It allows to learn multiple embeddings per word type with no assumptions about the number of senses per word type.

• Improve the computational expense of the two-step (cluster-based) process.
Efficient Non-parametric Estimation of Multiple Embeddings per Word in Vector Space

Neelakantan et al., EMNLP 2014

MSSG: Multi-Sense Skip-gram
Efficient Non-parametric Estimation of Multiple Embeddings per Word in Vector Space
Neelakantan et al., EMNLP 2014

NP-MSSG: Multi-Sense Skip-gram

Similar to MSSG but instead of choosing across the k possible sense vectors, if the Context Cluster Center is not similar enough (given a threshold) a new cluster is created.
Efficient Non-parametric Estimation of Multiple Embeddings per Word in Vector Space

Neelakantan et al., EMNLP 2014

<table>
<thead>
<tr>
<th>Model</th>
<th>avgSimC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pruned TF-IDF</td>
<td>60.5</td>
</tr>
<tr>
<td>Huang et al-50d</td>
<td>65.7</td>
</tr>
<tr>
<td>MSSG-50d</td>
<td>66.9</td>
</tr>
<tr>
<td>MSSG-300d</td>
<td>69.3</td>
</tr>
<tr>
<td>NP-MSSG-50d</td>
<td>66.1</td>
</tr>
<tr>
<td>NP-MSSG-300d</td>
<td>69.1</td>
</tr>
</tbody>
</table>
A Probabilistic Model for Learning Multi-Prototype Word Embeddings

A Mixture model

Tian et al., COLING 2014
A Probabilistic Model for Learning Multi-Prototype Word Embeddings

The main idea is to combine

- **Skip-Gram Model**
  - Provides less parameters
  - Only needs local context

- **Mixture Model**
  - Provides a probabilistic framework
  - Avoid additional clustering efforts

Tian et al., COLING 2014
A Probabilistic Model for Learning Multi-Prototype Word Embeddings

Skip-gram Model

\[
P(w_0|w_I) = \frac{\exp(V_{w_I}^T U_{w_0})}{\sum_{w \in W} \exp(V_{w_I}^T U_{w})}
\]

Tian et al., COLING 2014
A Probabilistic Model for Learning Multi-Prototype Word Embeddings

Tian et al., COLING 2014

Multi-Prototype Skip-gram Model

\[
p(w_o|w_i) = \sum_{i=1}^{N_{w_i}} P(w_o|h_i = i, w_i) P(h_i = i|w_i)
\]

\[
= \sum_{i=1}^{N_{w_i}} \frac{\exp(U^T_{w_o} V_{w_i,i})}{\sum_{w \in W} \exp(U^T_{w} V_{w_i,i})} P(h_i = i|w_i),
\]

- Suppose \(N_{apple} = 2\)
  - \(h_{apple} = 1\): ‘apple’ is a fruit
  - \(h_{apple} = 2\): ‘apple’ is a company
- Denote
  \[\psi_i = P(tree|h_{apple} = i, apple)\]
A Probabilistic Model for Learning Multi-Prototype Word Embeddings

<table>
<thead>
<tr>
<th>Model</th>
<th>$\rho \times 100$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word2Vec</td>
<td>61.7</td>
</tr>
<tr>
<td>EHMModel</td>
<td>65.7</td>
</tr>
<tr>
<td>Model_M</td>
<td>63.6</td>
</tr>
<tr>
<td>Model_W</td>
<td>65.4</td>
</tr>
</tbody>
</table>

Tian et al., COLING 2014

Huang et al., 2012
A Probabilistic Model for Learning Multi-Prototype Word Embeddings

Tian et al., COLING 2014

<table>
<thead>
<tr>
<th>Model</th>
<th>EHModel</th>
<th>Our Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>#parameters</td>
<td>$d_n\text{words} + d_n\text{embeddings} + (d_n\text{window} + 1)h_l + (2d + 1)h_g$</td>
<td>$d_n\text{words} + d_n\text{embeddings}$</td>
</tr>
</tbody>
</table>
Topical Word Embeddings

Liu et al., AAAI 2015

It proposes a multi-prototype word embeddings model with interpretable dimensions.

The dimensions are topics, rather than words, obtained by latent Dirichlet allocation (LDA).

It is based on the assumption that words will have different embeddings under different topics.
Topical Word Embeddings

Three models

- **TWE-1.** Each topic is treated as an extra word. Embeddings of words and topics are learned separately. The topical embeddings are build with both contributions.

- **TWE-2.** Each word-topic pair is considered as a pseudo word, and learn topical word embeddings directly.

- **TWE-3.** Words and topics are separate but learned jointly. The embedding of each word-topic pair is the concatenation of both word and topic embeddings.

Liu et al., AAAI 2015
Topical Word Embeddings

Liu et at., AAAI 2015
Topical Word Embeddings

<table>
<thead>
<tr>
<th>Model</th>
<th>(\rho \times 100)</th>
<th>(\text{AvgSimC})</th>
<th>(\text{MaxSimC})</th>
</tr>
</thead>
<tbody>
<tr>
<td>C&amp;W</td>
<td>57.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TFIDF</td>
<td>26.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pruned TFIDF</td>
<td>62.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LDA-S</td>
<td>56.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LDA-C</td>
<td>50.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Skip-Gram</td>
<td>65.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pruned TFIDF-M</td>
<td></td>
<td>60.5</td>
<td>60.4</td>
</tr>
<tr>
<td>Tian</td>
<td>65.4</td>
<td></td>
<td>63.6</td>
</tr>
<tr>
<td>Huang</td>
<td>65.3</td>
<td></td>
<td>58.6</td>
</tr>
<tr>
<td>TWE-1</td>
<td>68.1</td>
<td></td>
<td>67.3</td>
</tr>
<tr>
<td>TWE-2</td>
<td>67.9</td>
<td></td>
<td>63.6</td>
</tr>
<tr>
<td>TWE-3</td>
<td>67.1</td>
<td></td>
<td>65.5</td>
</tr>
</tbody>
</table>

Liu et al., AAAI 2015
Do Multi-Sense Embeddings Improve Natural Language Understanding?

Li and Jurafsky, EMNLP 2015

It criticizes multi-prototype models by questioning if there is clear evidence how these models improve single-prototype approaches on real NLU tasks.

It introduces a multisense embeddings model based on Chinese Restaurant Processes
Do Multi-Sense Embeddings Improve Natural Language Understanding?

Li and Jurafsky, EMNLP 2015

Chinese restaurant process

A restaurant where a new customer finds table and is likely to choose those tables which are more populated.
Do Multi-Sense Embeddings Improve Natural Language Understanding?

Li and Jurafsky, EMNLP 2015

Idea:

A word is associated with a new sense vector just when evidence in the context suggests that it is sufficiently different from its early senses.
Do Multi-Sense Embeddings Improve Natural Language Understanding?

Li and Jurafsky, EMNLP 2015

<table>
<thead>
<tr>
<th>Model</th>
<th>SCWS Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>SkipGram</td>
<td>66.4</td>
</tr>
<tr>
<td>SG+Greedy</td>
<td>69.1</td>
</tr>
<tr>
<td>SG+Expect</td>
<td>69.7</td>
</tr>
<tr>
<td>Chen</td>
<td>68.4</td>
</tr>
<tr>
<td>Neelakantan</td>
<td>69.3</td>
</tr>
</tbody>
</table>
Sense-based representations by exploiting bilingual resources
Sense-based representations by exploiting bilingual resources

“The other major potential source of sense-tagged data comes from parallel aligned bilingual corpora. Here, translation distinctions can provide a practical correlate to sense distinctions, as when instances of the English word”

Resnik & Yarowsky, 1997
Learning Sense-specific Word Embeddings By Exploiting Bilingual Resources

 Guo et al., COLING 2014

It proposes a method for learning sense-specific word embeddings by using bilingual parallel data.

It is supported by a language model based on neural networks

“same word in the source language with different senses […] has different translations in the foreign language”
Learning Sense-specific Word Embeddings By Exploiting Bilingual Resources

Guo et al., COLING 2014

Idea:

The words in the **source language** are **tagged** with their **translation** in the **foreign language**

The translations are **clustered**, exhibiting **different senses in different clusters**

The **sense-annotated data** is used to learn **sense-specific word embeddings**
Learning Sense-specific Word Embeddings By Exploiting Bilingual Resources

Guo et al., COLING 2014
Learning Sense-specific Word Embeddings By Exploiting Bilingual Resources

Guo et al., COLING 2014

<table>
<thead>
<tr>
<th>System</th>
<th>MaxSim</th>
<th></th>
<th>AvgSim</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\rho \times 100$</td>
<td>$\tau \times 100$</td>
<td>$\rho \times 100$</td>
<td>$\tau \times 100$</td>
</tr>
<tr>
<td>Ours</td>
<td>55.4</td>
<td>40.9</td>
<td>49.3</td>
<td>35.2</td>
</tr>
<tr>
<td>SingleEmb</td>
<td>42.8</td>
<td>30.6</td>
<td>42.8</td>
<td>30.6</td>
</tr>
<tr>
<td>Multi-prototype</td>
<td>40.7</td>
<td>29.1</td>
<td>38.3</td>
<td>27.4</td>
</tr>
</tbody>
</table>

Bilingual Learning of Multi-sense Embeddings with Discrete Autoencoders

Šuster et al., NAACL 2016

Uses second-language embeddings as a supervisory signal in learning multisense representations in the first language.

“Polysemy in one language can be at least partially resolved by looking at the translation of the word and its context in another language.”
Bilingual Learning of Multi-sense Embeddings with Discrete Autoencoders

Šuster et al., NAACL 2016

It is designed as an autoencoder: a feed forward neural network model that learns to mimic its input layer in the output layer.

Two parts:

• An encoding part which assigns a sense to a pivot word given the word and the context in both languages

• A reconstruction (decoding) part recovering context words based on the pivot word and its sense
Bilingual Learning of Multi-sense Embeddings with Discrete Autoencoders

Šuster et al., NAACL 2016
## Bilingual Learning of Multi-sense Embeddings with Discrete Autoencoders

Šuster et al., NAACL 2016

<table>
<thead>
<tr>
<th>Model (300-dim.)</th>
<th>SCWS</th>
</tr>
</thead>
<tbody>
<tr>
<td>SG</td>
<td>65.0</td>
</tr>
<tr>
<td><strong>MU</strong></td>
<td>66.7</td>
</tr>
<tr>
<td><strong>BiMU</strong></td>
<td>69.0</td>
</tr>
<tr>
<td>Chen et al. (2014)</td>
<td>68.4</td>
</tr>
<tr>
<td>Neelakantan et al. (2014)</td>
<td>69.3</td>
</tr>
<tr>
<td>Li and Jurafsky (2015)</td>
<td>69.7</td>
</tr>
</tbody>
</table>

Omitting the Bilingual corpora
Advantages and limitations of both types of sense representations
Advantages and Limitations

Knowledge-based sense representations

Advantages

- The learned sense representations are **linked to sense inventories**
  - This in turn might enable **multilinguality** (see BabelNet)
  - May exploit extra-information available in the underlying resource
- The number of senses for each word varies and is decided by expert lexicographers
Advantages and Limitations

Knowledge-based sense representations

Disadvantages

- Require sense inventories
  - Might not be available/complete in some languages
- Given that sense inventories are fixed, to cover emerging senses the inventory needs to be updated before we can create a vector representation
Advantages and Limitations

Unsupervised sense representations

Advantages

- **Fully unsupervised** (no need for external knowledge resources) and allows to have an entire end-to-end approach

- Can be adapted to specific corpora and domains
Advantages and Limitations

Unsupervised sense representations

Disadvantages

- The learned sense representations are not linked to any sense inventory
- Usually assume the number of senses to be fixed for all words
- The representations are generally not fine grained and difficult to evaluate.
- Rare words and less frequent meanings are not represented properly
Applications
Applications

• Semantic Similarity \textit{(used in other applications)}
• Word Sense Disambiguation / Entity Linking
• Link Prediction
• Ontology learning
• Information Extraction
• Sense Clustering
• Alignment of Lexical Resources
Sense-based Semantic Similarity

Based on the semantic similarity between senses.

Two main measures:

• **Cosine similarity** for low-dimensional vectors
• **Weighted Overlap** for sparse high-dimensional vectors (usually interpretable)
Sense-based Semantic Similarity: Words

Different sense-based measures as explained in the previous section.

Sense-based similarity performs on par or better than word-based approaches.
How to compose vectors for sentence/document representation?

Averaging word vectors is the most common approach

Drawbacks:

- **Word order** is not taken into account (new neural network approaches take word order into account, e.g. LSTMs)
- **Syntax** is not taken into account
- **Ambiguity** is not taken into account
How to model sentences and documents using sense representations?

There are some interesting compositionality ideas and approaches to test the use of sense representations to model sentences and documents: e.g. ADW or Li and Jurafsky (2015).

However, sense-based representation of sentences and documents remains an open problem (same applies to word-based).
Word Sense Disambiguation

Two ways to use sense representations for WSD:

• Integrated as a feature in a supervised disambiguation system (Rothe and Schütze, ACL 2015)

• Knowledge-based disambiguation (Camacho-Collados et al., ACL 2015)
Integration of sense representations in a supervised WSD system

(Rothe and Schütze, ACL 2015)

IMS (Zhong and Ng, ACL 2010 demo) is a state-of-the-art supervised disambiguation system. It is a SVM classifier which uses features based on the surrounding words of the target word (local context).

Idea: Use word and sense embeddings of the surrounding words and add it as a new feature.
Integration of sense representations in a supervised WSD system

(Rothe and Schütze, ACL 2015)

<table>
<thead>
<tr>
<th>IMS feature sets</th>
<th>Senseval-2</th>
<th>Senseval-3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 POS</td>
<td>53.6</td>
<td>58.0</td>
</tr>
<tr>
<td>2 surrounding word</td>
<td>57.6</td>
<td>65.3</td>
</tr>
<tr>
<td>3 local collocation</td>
<td>58.7</td>
<td>64.7</td>
</tr>
<tr>
<td>4 $S_{naive}$-product</td>
<td>56.5</td>
<td>62.2</td>
</tr>
<tr>
<td>5 S-cosine</td>
<td>55.5</td>
<td>60.5</td>
</tr>
<tr>
<td>6 S-product</td>
<td>58.3</td>
<td>64.3</td>
</tr>
<tr>
<td>7 S-raw</td>
<td>56.8</td>
<td>63.1</td>
</tr>
<tr>
<td>8 MFS</td>
<td>47.6†</td>
<td>55.2†</td>
</tr>
<tr>
<td>9 Rank 1 system</td>
<td>64.2†</td>
<td>72.9</td>
</tr>
<tr>
<td>10 Rank 2 system</td>
<td>63.8†</td>
<td>72.6</td>
</tr>
<tr>
<td>11 IMS</td>
<td>65.2‡</td>
<td>72.3‡</td>
</tr>
<tr>
<td>12 IMS + $S_{naive}$-prod.</td>
<td>62.6†</td>
<td>69.4†</td>
</tr>
<tr>
<td>13 IMS + S-cosine</td>
<td>65.1†</td>
<td>72.4†</td>
</tr>
<tr>
<td>14 IMS + S-product</td>
<td>66.5</td>
<td>73.6</td>
</tr>
<tr>
<td>15 IMS + S-raw</td>
<td>62.1†</td>
<td>66.8†</td>
</tr>
<tr>
<td>16 IMS + S_{optimized}-prod.</td>
<td>66.6</td>
<td>73.6</td>
</tr>
</tbody>
</table>

WSD using WordNet as sense inventory (lexical sample)
Knowledge-based Word Sense Disambiguation

(Camacho-Collados et al., AIJ 2016)

Basic idea

Select the sense which is semantically closer to the semantic representation of the whole document

(global context).

\[
\hat{d}(s) = \arg\max_{d \in D} WO(\text{NASARI}_{\text{lex}}(s), \vec{v}_{\text{lex}}(d))
\]
Knowledge-based Word Sense Disambiguation
(Camacho-Collados et al., AIJ 2016)

*Kobe, which is one of Japan's largest cities, [...]"
Knowledge-based Word Sense Disambiguation

(Camacho-Collados et al., AIJ 2016)

*Kobe, which is one of Japan's largest cities,* [...]
Knowledge-based Word Sense Disambiguation
(Camacho-Collados et al., AIJ 2016)

*Kobe, which is one of Japan's largest cities, [...]"
## Knowledge-based Word Sense Disambiguation

(Camacho-Collados et al., AIJ 2016)

<table>
<thead>
<tr>
<th>System</th>
<th>English</th>
<th>French</th>
<th>Italian</th>
<th>German</th>
<th>Spanish</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Nasari</strong></td>
<td>86.3</td>
<td>76.2</td>
<td>83.7</td>
<td>83.2</td>
<td>82.9</td>
<td>82.5</td>
</tr>
<tr>
<td>Muffin</td>
<td>84.5</td>
<td>71.4</td>
<td>81.9</td>
<td>83.1</td>
<td>85.1</td>
<td>81.2</td>
</tr>
<tr>
<td>Babelry</td>
<td>87.4</td>
<td>71.6</td>
<td>84.3</td>
<td>81.6</td>
<td>83.8</td>
<td>81.7</td>
</tr>
<tr>
<td>UMCC-DLSI</td>
<td>54.8</td>
<td>60.5</td>
<td>58.3</td>
<td>61.0</td>
<td>58.1</td>
<td>58.5</td>
</tr>
<tr>
<td>MFS</td>
<td>80.2</td>
<td>74.9</td>
<td>82.2</td>
<td>83.0</td>
<td>82.1</td>
<td>79.3</td>
</tr>
</tbody>
</table>

Multilingual WSD using Wikipedia as sense inventory (all-words)
Knowledge-based Word Sense Disambiguation

(Camacho-Collados et al., AIJ 2016)

<table>
<thead>
<tr>
<th>System</th>
<th>SemEval-2013</th>
<th>SemEval-2007</th>
</tr>
</thead>
<tbody>
<tr>
<td>NASARI</td>
<td>66.7</td>
<td>66.7</td>
</tr>
<tr>
<td>NASARI+IMS</td>
<td>67.0</td>
<td>68.5</td>
</tr>
<tr>
<td>MUFFIN</td>
<td>66.0</td>
<td>66.0</td>
</tr>
<tr>
<td>Babelfy</td>
<td>65.9</td>
<td>62.7</td>
</tr>
<tr>
<td>UKB</td>
<td>61.3</td>
<td>56.0</td>
</tr>
<tr>
<td>UMCC-DLSI</td>
<td>64.7</td>
<td>–</td>
</tr>
<tr>
<td>Multi-Objective</td>
<td>72.8</td>
<td>66.0</td>
</tr>
<tr>
<td>IMS</td>
<td>65.3</td>
<td>67.3</td>
</tr>
<tr>
<td>MFS</td>
<td>63.2</td>
<td>65.8</td>
</tr>
</tbody>
</table>

WSD using WordNet as sense inventory (All-Words)
# Knowledge-based Word Sense Disambiguation

(Camacho-Collados et al., AIJ 2016)

<table>
<thead>
<tr>
<th>System</th>
<th>SemEval-2013</th>
<th>SemEval-2007</th>
</tr>
</thead>
<tbody>
<tr>
<td>NASARI</td>
<td>66.7</td>
<td>66.7</td>
</tr>
<tr>
<td>NASARI+IMS</td>
<td>67.0</td>
<td>68.5</td>
</tr>
<tr>
<td>MUFFIN</td>
<td>66.0</td>
<td>66.0</td>
</tr>
<tr>
<td>Babelfy</td>
<td>65.9</td>
<td>62.7</td>
</tr>
<tr>
<td>UKB</td>
<td>61.3</td>
<td>56.0</td>
</tr>
<tr>
<td>UMCC-DLSI</td>
<td>64.7</td>
<td>–</td>
</tr>
<tr>
<td>Multi-Objective</td>
<td>72.8</td>
<td>66.0</td>
</tr>
<tr>
<td>IMS</td>
<td>65.3</td>
<td>67.3</td>
</tr>
<tr>
<td>MFS</td>
<td>63.2</td>
<td>65.8</td>
</tr>
</tbody>
</table>

WSD using WordNet as sense inventory (All-Words)
Word Sense Disambiguation

Open problem

Integration of **knowledge-based** (exploiting global contexts) and **supervised** (exploiting local contexts) systems to overcome the **knowledge-acquisition bottleneck**.
Link Prediction

Bordes et al. (NIPS 2013)

Add automatically relations between entities in a knowledge base.

How?

Embedding entities and relationships together
-\rightarrow TransE
### Link Prediction

Bordes et al. (NIPS 2013)

#### Link Prediction results in WordNet and FreeBase

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Metric</th>
<th>WN</th>
<th>FB15k</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Eval. setting</td>
<td>Mean Rank</td>
<td>Hits@10 (%)</td>
</tr>
<tr>
<td>Unstructured</td>
<td></td>
<td>315</td>
<td>304</td>
</tr>
<tr>
<td>RESCAL</td>
<td></td>
<td>1,180</td>
<td>1,163</td>
</tr>
<tr>
<td>SE</td>
<td></td>
<td>1,011</td>
<td>985</td>
</tr>
<tr>
<td>SME(Linear)</td>
<td></td>
<td>545</td>
<td>533</td>
</tr>
<tr>
<td>SME(Bilinear)</td>
<td></td>
<td>526</td>
<td>509</td>
</tr>
<tr>
<td>LFM</td>
<td></td>
<td>469</td>
<td>456</td>
</tr>
<tr>
<td>TransE</td>
<td></td>
<td>263</td>
<td>251</td>
</tr>
</tbody>
</table>
Taxonomy Learning
Espinosa-Anke et al. (AAAI 2016)

Global approach which exploits a large semantic network to **extend, taxonomize and semantify** domain terminologies.

How are sense representations used?
Taxonomy Learning

Espinosa-Anke et al. (AAAI 2016)

It uses sense representations to disambiguate and provide semantic coherence for taxonomies

![Taxonomy Diagram]

- Fruit
  - Apple
  - Orange
  - Lemon
Taxonomy Learning

Espinosa-Anke et al. (AAAI 2016)

It uses sense representations to disambiguate and provide semantic coherence for taxonomies.
Taxonomy Learning
Espinosa-Anke et al. (AAAI 2016)

It uses sense representations to disambiguate and provide semantic coherence for taxonomies

Fruit

Apple
Orange
Lemon

Apple

Semantically incoherent according to the sense vectors (SensEmbed)

Apple

Similar to the root (fruit) according to cosine similarity
Semantically coherent
Open Information Extraction
Delli Bovi et al. (EMNLP 2015)

Idea

Integrate the output of different Open Information Extraction systems into a single unified and fully disambiguated knowledge repository.
Similarly to the taxonomy learning approach, it uses sense representations to **disambiguate** and give a semantic coherence to the extracted relations.
Sense Clustering

• Current sense inventories suffer from the high granularity of their sense inventories.
• A meaningful clustering of senses would help boost the performance on downstream applications

(Hovy et al., AIJ 2013)

• Examples:
  - Street (with sidewalks or without sidewalks) in WordNet
  - Parameter (computer programming) - Parameter in Wikipedia
Sense Clustering

Basic approach

Using a clustering algorithm based on the semantic similarity between sense vectors
Sense Clustering

- ADW (Pilehvar et al. ACL 2013) for WordNet

- NASARI (Camacho-Collados et al. AIJ 2016) for Wikipedia
Sense Clustering
(Pilehvar et al., ACL 2013)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$R_{Cos}$</td>
<td>0.406</td>
<td>0.522</td>
<td>0.450</td>
<td>0.465</td>
<td>0.484</td>
<td>0.441</td>
<td>0.485</td>
</tr>
<tr>
<td>$R_{WO}$</td>
<td>0.421</td>
<td>0.544</td>
<td>0.483</td>
<td>0.482</td>
<td>0.531</td>
<td>0.470</td>
<td>0.503</td>
</tr>
<tr>
<td>$R_{Jac}$</td>
<td>0.418</td>
<td>0.531</td>
<td>0.478</td>
<td>0.473</td>
<td>0.501</td>
<td>0.465</td>
<td>0.493</td>
</tr>
<tr>
<td>SVM</td>
<td>0.370</td>
<td>0.455</td>
<td>NA</td>
<td>NA</td>
<td>0.473</td>
<td>0.423</td>
<td>0.432</td>
</tr>
<tr>
<td>ODE</td>
<td>0.218</td>
<td>0.396</td>
<td>NA</td>
<td>NA</td>
<td>0.371</td>
<td>0.331</td>
<td>0.288</td>
</tr>
</tbody>
</table>

Clustering of WordNet senses (F-Measure)
## Sense Clustering

(Camacho-Collados et al., AIJ 2016)

<table>
<thead>
<tr>
<th>Measure</th>
<th>System type</th>
<th>500-pair</th>
<th></th>
<th>SemEval</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Acc.</td>
<td>F1</td>
<td>Acc.</td>
</tr>
<tr>
<td>NASARI</td>
<td>unsupervised</td>
<td>83.8</td>
<td>70.5</td>
<td>87.4</td>
</tr>
<tr>
<td>NASARI\textsubscript{lexical}</td>
<td>unsupervised</td>
<td>81.6</td>
<td>65.4</td>
<td>85.7</td>
</tr>
<tr>
<td>NASARI\textsubscript{unified}</td>
<td>unsupervised</td>
<td>82.6</td>
<td>69.5</td>
<td>87.2</td>
</tr>
<tr>
<td>NASARI\textsubscript{embed}</td>
<td>unsupervised</td>
<td>81.2</td>
<td>65.9</td>
<td>86.3</td>
</tr>
<tr>
<td>SVM-monolingual</td>
<td>supervised</td>
<td>77.4</td>
<td>-</td>
<td>83.5</td>
</tr>
<tr>
<td>SVM-multilingual</td>
<td>supervised</td>
<td>84.4</td>
<td>-</td>
<td>85.5</td>
</tr>
<tr>
<td>Baseline\textsubscript{no-cluster}</td>
<td>-</td>
<td>71.4</td>
<td>0.0</td>
<td>82.5</td>
</tr>
<tr>
<td>Baseline\textsubscript{cluster}</td>
<td>-</td>
<td>28.6</td>
<td>44.5</td>
<td>17.5</td>
</tr>
</tbody>
</table>

### Clustering of Wikipedia pages
Alignment of Lexical Resources

Pilehvar and Navigli (ACL 2014)
Alignment of Lexical Resources

Idea: Ontologization of lexical resources to build a graph (semantic network) for each resource

Definition page for windmill

1. A **machine** which translates **linear motion** of **wind** to **rotational motion** by means of **adjustable vanes** called **sails**.
Alignment of Lexical Resources
Pilehvar and Navigli (ACL 2014)

Once the graph for each resource is constructed, PageRank is used to build a sense representation (i.e. semantic signature) for each concept.

Finally, sense representations with a very high degree of similarity are aligned.
Open Problems and Future Work
Open Problems and Future Work

1. Improve evaluation
   - Move from word similarity gold standards to end-to-end applications
     - Integration in Natural Language Understanding tasks (Li and Jurafsky, EMNLP 2015)
     - SemEval task? see e.g. WSD & Induction within an end user application @ SemEval 2013
Open Problems and Future Work

2. Make semantic representations more meaningful
   - unsupervised representations are hard to inspect (clustering is hard to evaluate)
   - but also knowledge-based approaches have issues:
     • e.g. top-10 closest vectors to the military sense of “company” in AutoExtend
Open Problems and Future Work

3. Interpretability

- The reason why things work or do not work is not obvious
  - E.g. avgSimC and maxSimC are based on implicit disambiguation that improves word similarity, but is not proven to disambiguate well
  - Many approaches are tuned to the task
- Embeddings are difficult to interpret and debug
Open Problems and Future Work

4. Link the representations to rich semantic resources like WikiData and BabelNet
   – Enabling applications that can readily take advantage of huge amounts of multilinguality and information about concepts and entities
   – Improving the representation of low-frequency/isolated meanings
Open Problems and Future Work

5. Scaling semantic representations to sentences and documents
   – Sensitivity to word order
   – Combine vectors into syntactic-semantic structures
   – Requires disambiguation, semantic parsing, etc.
   – Compositionality
Open Problems and Future Work

6. Addressing multilinguality
   – a key trend in today’s NLP research
   • We are already able to perform POS tagging and dependency parsing in dozens of languages
     – Also mixing up languages
Open Problems and Future Work

• We can perform Word Sense Disambiguation and Entity Linking in hundreds of languages
  – Babelfy (Moro et al. 2014)
  – but with only a few sense vector representations
• Now: it is crucial that sense and concept representations are language-independent
• Enabling comparisons across languages
• Also useful in semantic parsing
Open Problems and Future Work

• Representations are most of the time evaluated in English
  – single words only
• It is important to evaluate sense representations in other languages and across languages
  – Check out the SemEval 2017 Task 2: multilingual and cross-lingual semantic word similarity (multilwords, entities, domain-specific, slang, etc.)
Open Problems and Future Work

7. Integrate sense representations into Neural Machine Translation
   - Previous results in the 2000s working on semantically-enhanced SMT are not very encouraging
   - However, many options have not been considered
Thank you!

Questions please!