Cross-Lingual Syntactically Informed Distributed Word Representations

Ivan Vulić
University of Cambridge

EACL 2017; Valencia; April 6, 2017

iv250@cam.ac.uk
Motivation (High-Level)

The NLP community has developed useful features for several tasks but finding features that are...

1. **task-invariant** (POS tagging, SRL, NER, parsing, ...)
   
   (monolingual word embeddings)

2. **language-invariant** (English, Dutch, Chinese, Spanish, ...)
   
   (cross-lingual word embeddings → this talk)

...is non-trivial and time-consuming (20+ years of feature engineering...)
The NLP community has developed useful features for several tasks but finding features that are...

1. **task-invariant** (POS tagging, SRL, NER, parsing, ...)
   (monolingual word embeddings)

2. **language-invariant** (English, Dutch, Chinese, Spanish, ...)
   (cross-lingual word embeddings → this talk)

...is non-trivial and time-consuming (20+ years of feature engineering...)

**Learn word-level features which generalise across tasks and languages**
Motivation (Low-Level)

Inject **syntactic** information into **cross-lingual** word embeddings

- Similar structures in English and Italian
- **Universal Dependencies**: syntactic contexts in multiple languages
Learning from Context

Skip-gram with negative sampling (SGNS)
[Mikolov et al.; NIPS 2013]

Learning from the set $D$ of \textit{(word, context)} pairs observed in a corpus:
$(w, v) = (w_t, w_{t \pm c}); i = 1, ..., c; c =$ context window size

SG learns to predict the \textbf{context} of each pivot word.

John saw a \textit{cute gray huhblub running in} the field.

$D = (\text{huhblub, cute}), (\text{huhblub, gray}), (\text{huhblub, running}), (\text{huhblub, in})$

$vec(\text{huhblub}) = [-0.23, 0.44, -0.76, 0.33, 0.19, ...]$
Learning from Context

**Representation model** → Skip-gram with negative sampling (SGNS)

SGNS may be trained with *arbitrary contexts*
[Levy and Goldberg, ACL 2014]

**Context is crucial**
Different context types result in different SGNS vectors.

[Schwartz et al, NAACL 2016; Melamud et al, NAACL 2016]

**Some standard context types:**
1. (Ordinary) bag-of-words **(BOW)**
2. Positional **(POSIT)**
3. Dependency-based: Basic **(DEPS-NAIVE)**
4. Dependency-based: with prepositional arc collapsing **(DEPS-ARC)**
4. (Universal) Dependency-based: with prepositional arc collapsing

\{(\text{discovers, scientist\_nsubj}), (\text{discovers, stars\_dobj}), (\text{discovers, telescope\_nmod}), (\text{stars, discovers\_dobj-1}), (\text{scientist, australian\_amod}), (\text{discovers, telescope\_prep\_with}), (\text{telescope, discovers\_prep\_with-1})\}, ...
Cross-Lingual Word Embeddings

Representation of a word $w_1^S \in V^S$:

$$vec(w_1^S) = [f_1^1, f_2^1, \ldots, f_{\text{dim}}^1]$$

Exactly the same representation for $w_2^T \in V^T$:

$$vec(w_2^T) = [f_1^2, f_2^2, \ldots, f_{\text{dim}}^2]$$

Language-independent word representations in the same shared semantic (or embedding) space!
Cross-Lingual Word Embeddings

Q1 → How to align semantic spaces in two different languages?

Q2 → Which bilingual signals are used for the alignment?

See also:
[Upadhyay et al., ACL 2016; Vulić and Korhonen, ACL 2016]
Exploiting Syntax and Translation Pairs

Using translation dictionaries, e.g., [en_stars, it_stelle], [en_scientist, it_scienzato]

Extracting context pairs from hybrid “cross-lingual” trees
Exploiting Syntax and Translation Pairs

**Online training** with monolingual and cross-lingual dependency-based contexts

(a) T1

(australian scientist) discovers (stars) with telescope prep:with

(b) T2

(Scienziato australiano) scopre (stelle) con telescopio

(c) T3

(australian scienzato) discovers (stars) with telescope prep:with

(d) T4

(Scientist australiano) scopre (stelle) con telescopio

(e) T5

(australian scientist) discovers (stelle) with telescope prep:with

(f) T6

(Scienziato australiano) scopre (stars) con telescopio
Online training with monolingual and cross-lingual dependency-based contexts

(discovers, scientist_nsubj)
(stars, discovers_dobj⁻¹)
(scienzato, australiano_amod)
(scopre, stelle_dobj)

(scientist, australiano_amod)
(australiano, scientist_amod⁻¹)
(stars, scopre_dobj⁻¹)
(discovers, scienzato_nsubj)

Training word2vecf SGNS on these (word, context) pairs
Experimental Setup

Language pairs

Results reported with two language pairs: **IT-EN, DE-EN**. Experiments conducted with more language pairs (SV-EN, FR-EN, NL-EN).

Translation dictionaries

1. BNC-Lemma+GT
2. dict.cc

Training Data and Setup

→ SGNS model; Data: Wikipedias in EN, IT, DE
→ Universal Dependencies v1.4
→ SOTA UPOS tagger [Martins et al., ACL 2013]
→ SOTA dependency parser [Bohnet, COLING 2010]

[Vulić and Korhonen, ACL 2016]
Baselines

Cross-lingual embeddings relying on exactly the same supervision signal: translation dictionaries

[Mikolov et al., arXiv 2013], [Lazaridou et al., ACL 2015], ...

word2vec SGNS trained with three context types:

1. BOW ($win = 2$)
2. Positional ($win = 2$)
3. Monolingual DEPS (exactly the same signal used as with our model)

Online vs offline: These models train monolingual SGNS offline and learn a mapping function
### Task I: (Monolingual) Word Similarity

Results on **multilingual SimLex-999**

[Leviant and Reichart, arXiv 2015]

<table>
<thead>
<tr>
<th>Model</th>
<th>IT All</th>
<th>IT Verbs</th>
<th>DE All</th>
<th>DE Verbs</th>
<th>EN (with IT) All</th>
<th>EN (with IT) Verbs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mono-sgns</td>
<td>0.235</td>
<td>0.318</td>
<td>0.305</td>
<td>0.259</td>
<td>0.331</td>
<td>0.281</td>
</tr>
<tr>
<td>off-bow2</td>
<td>0.254</td>
<td>0.317</td>
<td>0.306</td>
<td>0.263</td>
<td>0.328</td>
<td>0.279</td>
</tr>
<tr>
<td>off-posit2</td>
<td>0.227</td>
<td>0.323</td>
<td>0.283</td>
<td>0.194</td>
<td>0.336</td>
<td>0.316</td>
</tr>
<tr>
<td>off-deps</td>
<td>0.199</td>
<td>0.308</td>
<td>0.258</td>
<td>0.214</td>
<td>0.334</td>
<td>0.311</td>
</tr>
<tr>
<td>CL-DepEmb</td>
<td><strong>0.287</strong></td>
<td><strong>0.358</strong></td>
<td><strong>0.306</strong></td>
<td><strong>0.319</strong></td>
<td><strong>0.356</strong></td>
<td><strong>0.308</strong></td>
</tr>
</tbody>
</table>
Task II: (Bilingual) Lexicon Induction

Results on three BLI datasets:
1. Translations of SimLex words (IT-EN and DE-EN)
2. IT-EN test set [Vulić and Moens, EMNLP 2013]
3. DE-EN test set [Upadhyay et al., ACL 2016]

<table>
<thead>
<tr>
<th>Model</th>
<th>IT-EN</th>
<th>DE-EN</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SL-TRANS</td>
<td>VULIC1k</td>
</tr>
<tr>
<td>off-bow2</td>
<td>0.328 [0.457]</td>
<td>0.405</td>
</tr>
<tr>
<td>off-posit2</td>
<td>0.219 [0.242]</td>
<td>0.272</td>
</tr>
<tr>
<td>off-deps</td>
<td>0.169 [0.065]</td>
<td>0.271</td>
</tr>
<tr>
<td>CL-DepEmb</td>
<td><strong>0.541 [0.597]</strong></td>
<td><strong>0.532</strong></td>
</tr>
</tbody>
</table>

BLI results (Top 1 scores). For SL-Trans we also report results on the verb translation subtask (numbers in square brackets).
More Results: Highlights (Not Really)

- Improvements with CL-DepEmb on verb similarity; tested on SimVerb-3500

  → DE SimLex-999, adjectives: 0.585, best baseline: 0.417
  → IT SimLex-999, adjectives: 0.334, best baseline: 0.266

  → DE SimLex-999, verbs: 0.319, best baseline: 0.263
  → IT SimLex-999, verbs: 0.358, best baseline: 0.323
These preliminary experiments show that injecting syntactic information into cross-lingual tasks helps semantic tasks which stress similarity...

- Porting this idea to more (typologically diverse) languages
- More accurate dependency parsers? Selection of (reliable) translation pairs?
- More sophisticated approaches to constructing hybrid “cross-lingual” trees
- Other semantic tasks: cross-lingual lexical entailment, lexical substitution?
Questions?