Inverted Indexing For Cross-Lingual NLP

Polina Stadnikova

Saarland University

18th January 2018

Outline

Want to obtain cross-lingual word representations?

What has been done before:

Cross-lingual learning:

- *Annotation projection*: using manually annotated word alignments to project from the source to the target
- *Delexicalized transfer*: remove lexical features from monolingual data, retain reliable PoS-tags for the target

Want to obtain cross-lingual word representations?

The linguistic ressources problem:

- unevenly distributed
- how to transfer from the target to the source?
- Wikipedia articles in English and Greenlandic: 5549690 vs. 1643
- State of the art: cross-lingual representations with English as source language

Want to obtain cross-lingual word representations?

Why do we need a new approach?

We do not want to depend on:

- Neural networks training
- Parallel data availability

And we also want to:

- Keep lexical features to make the *truly* cross-lingual transfer

Approach in a nutshell

- Make clusters of Wikipedia articles linking to the same concept
- Count occurrences of the words in clusters
- Simultaneously train models with lexical features on different source languages
- Test models on different tasks (PoS-tagging, parsing, etc)

Distributional representations

The problem with word representations:

- High dimensionality, sparseness, no fine-grained representation of relatedness

Vector representations

- *Count-based*: represent words by their co-occurrences; raw or weighted co-occurrence matrices
- *Prediction-based*: represent words in in the middle layer of NN; learn to predict words from the context, or vice versa

Monolingual representations

Count-based

- Co-occurrence information
- Binary matrices, raw counts, or point-wise mutual information
- Dimensionality reduction: SVD

Prediction-based

- 3 layers architecture: input, word representations, output
- Skip-gram model: input - target word, output - context
- CBOW model: input - context, output - target word

Bilingual representations

Klementiev et al. (2012): learn word embeddings from target and source languages

Method

- Use parallel texts with word alignments
- Minimize the loss between the target model and the source model
- Modifiable interaction matrix which enforces aligned words to have similar representations

Bilingual representations

Chandar et al. (2014): bag-of-words representations

Method

- Do not use word alignments
- Auto-encoder architecture
- Input layer $\rightarrow$ source bag-of-words vectors, output layer $\rightarrow$ target bag-of-words vectors
- Try to reconstruct the input at the output layer, passing representations through a middle layer
- Dimensionality reduction is provided by middle layers

Bilingual representations

What about count-based bilingual representations?

Inverted indexing

Basic idea:

Words *glasses*[en], *Brille*[de], *gafas*[es] occur in the Wikipedia article about Harry Potter
They should have the same representations

Why Wikipedia?

A lot of articles in different languages on the same topic → linked to the same Wikipedia concept

Inverted indexing

Method: represent words by Wikipedia concepts they are used to describe

- For a set of languages (German, English, French, Spanish, and Swedish), find a common subset of Wikipedia concepts
- Describe each concept by a set of term occurring in the articles
- Create a concept - to - term set matrix
- Describe each word by a row in the inverted indexing of the matrix

Inverted indexing has been used for text categorization, cross-lingual relatedness measure

Settings

Baseline embeddings

- From **Klementiev et al. (2012)** and **Chandar et al. (2014)**
- Perform the nearest cross-language neighbors test in some representations
- **Chandar** and **Inverted** contain less noise

Parameters

- Fixed dimensionality in SVD: $\delta \in \{40, 80, 160\}$
- Scaling factor: $\sigma \in \{1.0, 0.1, 0.01, 0.001\}$

Tasks

- Document classification, PoS-tagging, dependency parsing, word alignments

Data sets

	TRAIN		TEST		TOKEN COVERAGE		
lang	data points	tokens	data points	tokens	KLEMENTIEV	CHANDAR	INVERTED
			RCV – DOCUMENT CLASSIFICATION				
en	10000	–	–	–	0.314	0.314	0.779
de	–	–	4998	–	0.132	0.132	0.347
			AMAZON – DOCUMENT CLASSIFICATION				
en	6000	–	–	–	0.314	0.314	0.779
de	–	–	6000	–	0.132	0.132	0.347
			GOOGLE UNIVERSAL TREEBANKS – POS TAGGING & DEPENDENCY PARSING				
en	39.8k	950k	2.4k	56.7k	–	–	–
de	2.2k	30.4k	1.0k	16.3k	0.886	0.884	0.587
es	3.3k	94k	0.3k	8.3k	0.916	0.916	0.528
fr	3.3k	74.9k	0.3k	6.9k	0.888	0.888	0.540
sv	4.4k	66.6k	1.2k	20.3k	n/a	n/a	0.679
			CONLL 07 – DEPENDENCY PARSING				
en	18.6	447k	–	–	–	–	–
es	–	–	206	5.7k	0.841	0.841	0.455
de	–	–	357	5.7k	0.616	0.612	0.294
sv	–	–	389	5.7k	n/a	n/a	0.561
			EUROPARL – WORD ALIGNMENT				
en	–	–	100	–	0.370	0.370	0.370
es	–	–	100	–	0.533	0.533	0.533

Document classification

- Represent each document by the average of the word representations occurring both in documents and in embeddings
- No scaling, 40 dimensions
- Ignore stopwords
- No effect of OOV words

Dataset	KLEMENTIEV	CHANDAR	INVERTED
AMAZON	0.32	0.36	**0.49**
RCV	0.75	**0.90**	0.55

PoS-tagging

- Tags from the Google Universal Treebanks
- Scaled word representations
- Delexicalized PoS tagger with the inverted word representations

		de	es	fr	sv	av-sv
EN→TARGET						
EMBEDS	K12	80.20	73.16	47.69	-	67.02
	C14	74.85	83.03	48.24	-	68.71
INVERTED	SVD	**81.18**	82.12	49.68	78.72	70.99
MULTI-SOURCE→TARGET						
INVERTED	SVD	80.10	**84.69**	49.68	78.72	70.66

$\sigma = 0.01, \delta = 160, i = 10$

Dependency parsing

- Google Universal Treebanks, CoNLL treebanks for German, Spanish, Swedish
- **Delex** baseline: learns without lexical features, iterates over the data (single-source and multi-source setup), parameter set on the Spanish development data
- OOV words: mean vector for words with a specific PoS

		UAS		
		de	es	sv
EN→TARGET				
DELEX	-	44.78	47.07	56.75
DELEX-XIAO	-	46.24	52.05	57.79
EMBEDS	K12	44.77	47.31	-
	C14	44.32	47.56	
INVERTED	-	45.01	47.45	56.15
XIAO	-	49.54	55.72	61.88

CoNLL, unlabeled, $\sigma = 0.005$, $\delta = 20$, $i = 3$

Dependency parsing

		UAS de	UAS es	UAS fr	UAS sv	LAS de	LAS es	LAS fr	LAS sv
EN→TARGET									
DELEX	-	56.26	62.11	64.30	66.61	48.24	53.01	54.98	56.93
EMBEDS	K12	56.47	61.92	61.51	-	48.26	52.88	51.76	-
	C14	56.19	61.97	62.95	-	48.11	52.97	53.90	-
INVERTED	-	56.18	61.71	63.81	66.54	48.82	53.04	54.81	57.18
MULTI-SOURCE→TARGET									
DELEX	-	**56.80**	63.21	66.00	**67.49**	**49.32**	54.77	56.53	**57.86**
INVERTED	-	56.56	**64.03**	**66.22**	67.32	48.82	**55.03**	**56.79**	57.70

Google Universal Treebanks, unlabeled and labeled, $\sigma = 0.005$, $\delta = 20$, $i = 3$

Word alignments

- English-Spanish data with possible and certain alignments
- For each word representation, align every aligned word in the gold standard to its nearest neighbor

	KLEMENTIEV	CHANDAR	INVERTED
EN-ES (S+P)	0.20	0.24	**0.25**
ES-EN (S+P)	0.35	0.32	**0.41**
EN-ES (S)	0.20	**0.25**	**0.25**
ES-EN (S)	0.38	0.39	**0.53**

Precision, S = sure (certain), P = possible

Results

- **Document classification**
  For RCV: *Klementiev* and *Chandar* developed their methods on this data
  For Amazon: *Inverted* outperforms

Results

- **Document classification**
  For RCV: *Klementiev* and *Chandar* developed their methods on this data
  For Amazon: *Inverted* outperforms
- **PoS-tagging**
  Best results with *Inverted*
  No general gain from multiple source languages

Results

- **Document classification**
  For RCV: *Klementiev* and *Chandar* developed their methods on this data
  For Amazon: *Inverted* outperforms
- **PoS-tagging**
  Best results with *Inverted*
  No general gain from multiple source languages
- **Dependency parsing**
  No significant improvements
  Klementiev and *Chandar* hurt performance, *Inverted* improves on *some* languages

Results

- **Document classification**
  For RCV: *Klementiev* and *Chandar* developed their methods on this data
  For Amazon: *Inverted* outperforms
- **PoS-tagging**
  Best results with *Inverted*
  No general gain from multiple source languages
- **Dependency parsing**
  No significant improvements
  Klementiev and *Chandar* hurt performance, *Inverted* improves on *some* languages
- **Word alignments**
  Consistent improvements with *Inverted*

Let's wrap up

We have seen the first *count-based* approach that enables multi-source learning using cross-lingual word representations

Let's wrap up

This approach...

...does not require training neural networks

Let's wrap up

This approach...

...does not require training neural networks

...does not depend on the parallel data between source and target

Let's wrap up

This approach...

...does not require training neural networks

...does not depend on the parallel data between source and target

...enables obtaining *truly* cross-lingual word representations

Let's wrap up

This approach...

...does not require training neural networks

...does not depend on the parallel data between source and target

...enables obtaining *truly* cross-lingual word representations

...is computationally efficient and almost parameter-free

Let's wrap up

This approach...

...does not require training neural networks

...does not depend on the parallel data between source and target

...enables obtaining *truly* cross-lingual word representations

...is computationally efficient and almost parameter-free

...but, nevertheless, parameter-sensitive

Let's wrap up

This approach...

...does not require training neural networks

...does not depend on the parallel data between source and target

...enables obtaining *truly* cross-lingual word representations

...is computationally efficient and almost parameter-free

...but, nevertheless, parameter-sensitive

...outperforms two state-of-the-art approaches in 14 of 17 datasets in 4 tasks

References

- Anders Søgaard, Zeljko Agic, Hector Martinez Alonso, Barbara Plank, and Bernd Bohnet (2015), *Inverted indexing for cross-lingual NLP*. In *ACL*, Vol. 1, pages 1713-1722.
- Alexandre Klementiev, Ivan Titov, and Binod Bhattarai (2012), *Inducing crosslingual distributed representations of words.* In *COLING*.
- Sarath Chandar, Stanislas Lauly, Hugo Larochelle, Mitesh Khapra, Balaraman Ravindran, Vikas C Raykar, and Amrita Saha (2014), *An autoencoder approach to learning bilingual word representations*. In *NIPS*.

Thanks for your attention!

Questions?

Backup slides

Bilingual representations

Xiao and Guo (2014): learn from bilingual dictionaries

Method

- Use unambiguous source-target pairs from Wiktionary
- Force translations to have the same representations

Gouws and Søgaard (2015): a simple approach to learn prediction-based representations

Method

- Collect source-target pivot pairs of words
- Randomly replace pivot words with their equivalents from other languages

Baseline embeddings

	KLEMENTIEV	CHANDAR	INVERTED
es			
coche ('car', NOUN)	approximately beyond upgrading	car bicycle cars	driving car cars
expressed ('expressed', VERB)	1.61 55.8 month-to-month	reiterates reiterating confirming	exists defining example
teléfono ('phone', NOUN)	alexandra davison creditor	phone telephone e-mail	phones phone telecommunication
árbol ('tree', NOUN)	tree market-oriented assassinate	tree bread wooden	tree trees grows
escribió ('wrote', VERB)	wrote alleges testified	wrote paul palace	wrote inspired inspiration
amarillo ('yellow', ADJ)	yellow louisiana 1911	crane grabs outfit	colors yellow oohs
de			
auto ('car', NOUN)			car cars camaro
ausgedrückt ('expressed', VERB)			adjective decimal imperative
fr			
voiture ('car', NOUN)			mercedes-benz cars quickest
exprimé ('expressed', VERB)			simultaneously instead possible
téléphone ('phone', NOUN)			phone create allowing
arbre ('tree', NOUN)			tree trees grows
écrit ('wrote', VERB)			published writers books
jaune ('yellow', ADJ)			classification yellow stages
sv			
bil ('car', NOUN)			cars car automobiles
uttryckte ('expressed', VERB)			rejected threatening unacceptable
telefon ('phone', NOUN)			telephones telephone share
träd ('tree', NOUN)			trees tree trunks
skrev ('wrote', VERB)			death wrote biography
gul ('yellow', ADJ)			greenish bluish colored

Three nearest neighbors in the English training data for words from the Spanish test data