

Monolingual and Cross-Lingual Information Retrieval Models Based on (Bilingual) Word Embeddings

Ivan Vulić and Marie-Francine Moens

KU Leuven
Department of Computer Science

KU LEUVEN

ivan.vulic@cs.kuleuven.be

marie.francine.moens@cs.kuleuven.be

SIGIR 2015, Santiago de Chile; August 11, 2015

I. Learning Bilingual Word Embeddings (BWEs)

- Dense word representations, word embeddings, bilingual word embeddings
- Monolingual and bilingual embedding spaces
- Multilingual text data → why document-aligned data?
- New BWE learning model: BWESG → learning monolingual and bilingual embedding spaces

II. BWEs in IR

- Semantically-aware representations in the ad-hoc retrieval process?
- From word representations to query and document representations
- Monolingual embeddings → monolingual retrieval; Bilingual embeddings → cross-lingual retrieval
- The same conceptual model of retrieval for MoIR and CLIR with bilingual embeddings spaces!
- Results and discussion

Part I: Learning BWEs

Key idea

Distributional hypothesis → words with similar meanings are likely to appear in similar contexts

[Harris, Word 1954]



shouts:

“Meaning as use!”



calmly states:

“You shall know a word by the company it keeps.”

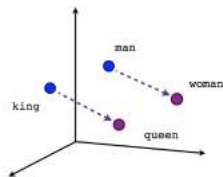
Dense representations → real-valued low-dimensional vectors
(seen already? LSI?)

Word embedding induction

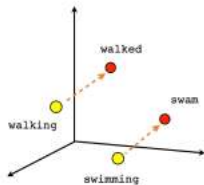
→ learn word-level features which generalize well across tasks and languages

→ **bilingual word embeddings** (this talk)

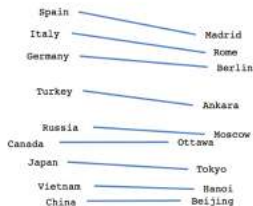
Word embeddings capture interesting and universal features:



Male-Female

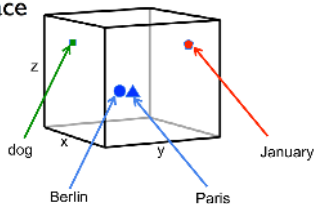


Verb tense



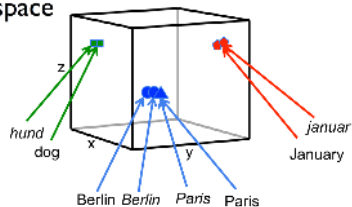
Country-Capital

3D embedding space



Monolingual

3D embedding space



vs.

Bilingual

[Image courtesy of Stephan Gouws]

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

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

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

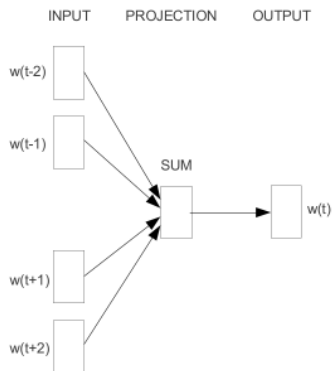
$$\text{vec}(w_2^T) = [f_1^2, f_2^2, \dots, f_{dim}^2]$$

Language-independent word representations in the same shared semantic (or *embedding*) space!

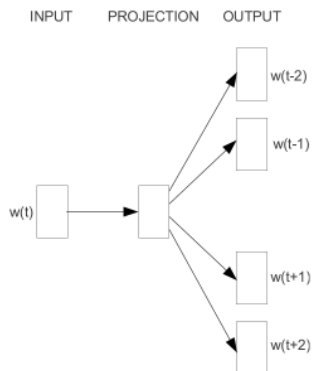
Word representation \rightarrow **A dense real-valued dim -dimensional vector**, these dimensions are no longer interpretable (unlike with other semantic representations).

Skip-gram with negative sampling (SGNS)

[Mikolov et al.: NIPS 2013]



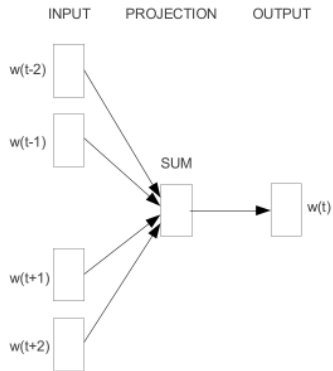
CBOW



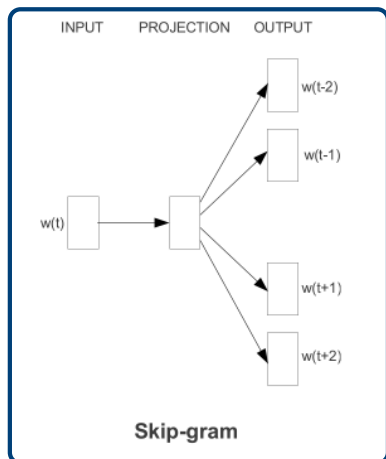
Skip-gram

Skip-gram with negative sampling (SGNS)

[Mikolov et al.; NIPS 2013]



CBOW



Skip-gram

Skip-gram with negative sampling (SGNS)

[Mikolov et al.; NIPS 2013]

Learning from the set D of $(word, context)$ pairs observed in a corpus:
 $(w, v) = (w(t), w(t \pm i)); i = 1, \dots, cs; cs = \text{context window size}$

SG learns to predict the context of the pivot word

John saw a **cute gray** **huhblub** **running in** the field.

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

$\text{vec}(\text{huhblub}) = [-0.23, 0.44, -0.76, 0.33, 0.19, \dots]$

Negative sampling = learning using both positive (“observed”) examples (set D), and negative (“unobserved”) examples (set D')

SGNS is actually doing something very similar to the older approaches → factorizing the traditional word-context matrix!

[Levy et al., NIPS 2014, TACL 2015]

More research focused on learning monolingual WEs:

- Full-fledged neural-net approaches [Bengio et al., JMLR 2003; Collobert and Weston, ICML 2008]
- Other factorization methods (e.g., Hellinger PCA) [Lebret and Collobert, EACL 2014]
- GloVe [Pennington et al., EMNLP 2014]
- ...

Probability for one word-context pair (w, v) :

$$P(D = 1|w, v, \theta) = \frac{1}{1 + \exp(-\vec{w} \cdot \vec{v}_c)}$$

General objective:

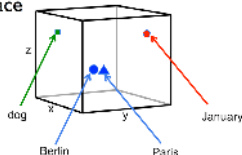
$$J = \arg \max_{\theta} \sum_{(w,v) \in D} \log \frac{1}{1 + \exp(-\vec{w} \cdot \vec{v}_c)}$$

General objective with negative sampling:

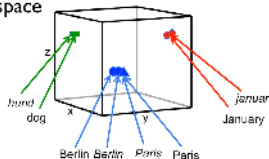
$$J = \arg \max_{\theta} \sum_{(w,v) \in D} \log \frac{1}{1 + \exp(-\vec{w} \cdot \vec{v}_c)} + \sum_{(w,v') \in D'} \log \frac{1}{1 + \exp(\vec{w} \cdot \vec{v}'_c)}$$

Generalizing the WE learning in bilingual settings using the similar principles...

3D embedding space



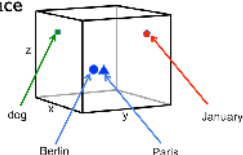
3D embedding space



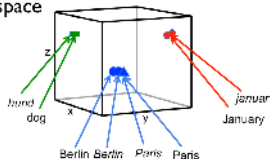
1. Align pretrained monolingual embedding spaces (**offline**) using *dictionaries* [Mikolov et al., arXiv 2013; Lazaridou et al., ACL 2015]
2. Jointly learn and align embeddings (**online**) using *parallel-only data* [Hermann and Blunsom, ACL 2014; Chandar et al., NIPS 2014]
3. Jointly learn and align embeddings (**online**) using *mono and parallel data* [Gouws et al., ICML 2015; Soyer et al., ICLR 2015, Shi et al., ACL 2015]

Generalizing the WE learning in bilingual settings using the similar principles...

3D embedding space



3D embedding space



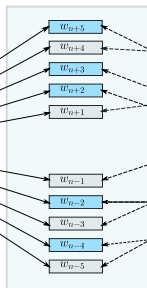
1. Align pretrained monolingual embedding spaces (**offline**) using *dictionaries* [Mikolov et al., arXiv 2013; Lazaridou et al., ACL 2015]
2. Jointly learn and align embeddings (**online**) using *parallel-only data* [Hermann and Blunsom, ACL 2014; Chandar et al., NIPS 2014]
3. Jointly learn and align embeddings (**online**) using *mono and parallel data* [Gouws et al., ICML 2015; Soyer et al., ICLR 2015, Shi et al., ACL 2015]
4. Can we do it without readily available dictionaries and parallel data? → Using document-aligned data (e.g., Wikipedia) [our model: BWESG]

Input: Pivot word representation

$$\sum_i P(w_{n\pm i}|w_n)$$



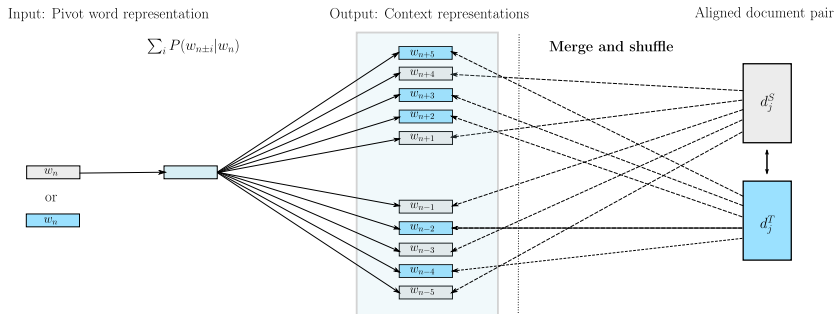
Output: Context representations



Merge and shuffle

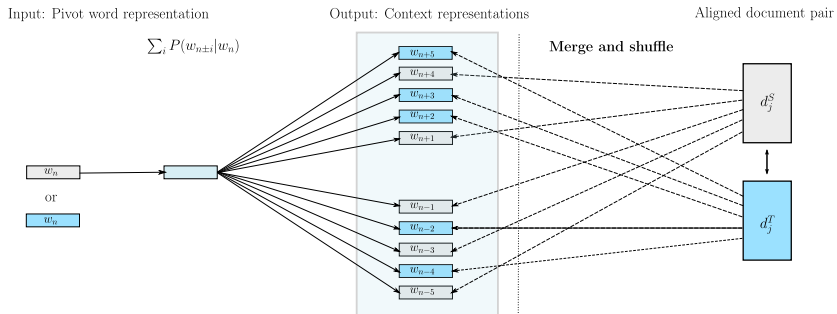
Aligned document pair





→ **Merge & Shuffle:** Training a SGNS (or any other monolingual model!) on shuffled “pseudo-bilingual” documents →

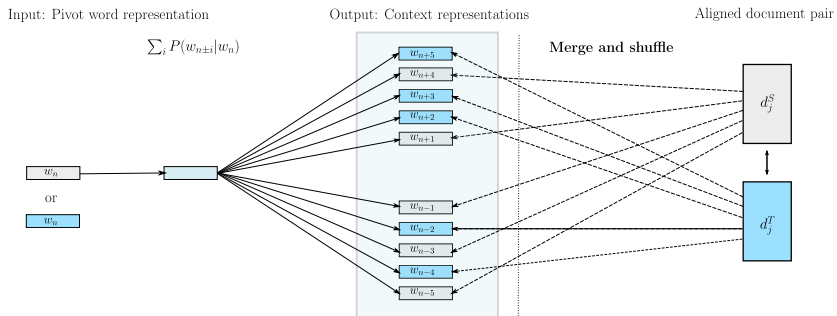
→ Our model: **BWESG**



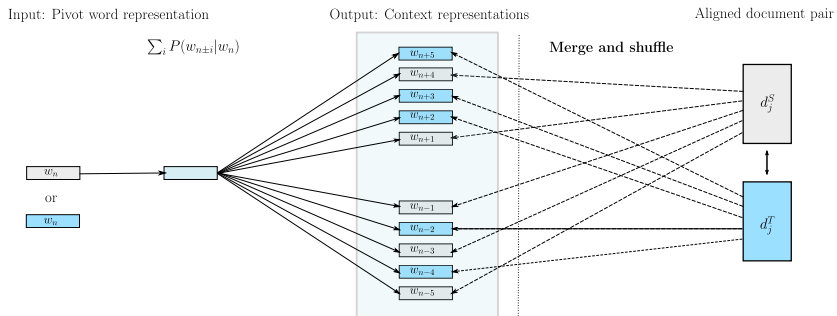
→ **Merge & Shuffle:** Training a SGNS (or any other monolingual model!) on shuffled “pseudo-bilingual” documents →

→ Our model: **BWESG**

→ **1. dumb shuffling: random (this work);** **2. slightly more intelligent: length ratio-based (after this work);** **3. even more intelligent: future work**



→ shuffling ensures bilingual (instead of monolingual) contexts → learning a bilingual embedding space jointly (**online**)



→ shuffling ensures bilingual (instead of monolingual) contexts → learning a bilingual embedding space jointly (**online**)

→ **No longer a local model:** Window size controls the number of **document-level** positive samples

Spanish-English (ES-EN)			Italian-English (IT-EN)			Dutch-English (NL-EN)		
(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
reina	reina	reina	madre	madre	madre	schilder	schilder	schilder
(Spanish)	(English)	(Combined)	(Italian)	(English)	(Combined)	(Dutch)	(English)	(Combined)
rey	queen(+)	queen(+)	padre	mother(+)	mother(+)	kunstschilder	painter(+)	painter(+)
trono	heir	rey	moglie	father	padre	schilderij	painting	kunstschilder
monarca	throne	trono	sorella	sister	moglie	kunstenaar	portrait	painting
heredero	king	heir	figlia	wife	father	olieverf	artist	schilderij
matrimonio	royal	throne	figlio	daughter	sorella	olieverfschilder	art	kunstenaar
hijo	reign	monarca	fratello	son	figlia	schilderen	impressionist	portrait
reino	succession	heredero	casa	friend	figlio	frans	cubism	olieverf
reinado	princess	king	amico	childhood	sister	nederlands	art	olieverfschilderij
regencia	marriage	matrimonio	marito	family	fratello	componist	poet	schilderen
duque	prince	royal	donna	cousin	wife	beeldhouwer	drawing	artist

$$\overrightarrow{\text{reina}} - \overrightarrow{\text{woman}} + \overrightarrow{\text{man}} \approx \overrightarrow{\text{rey}}$$

$$\overrightarrow{\text{queen}} - \overrightarrow{\text{mujer}} + \overrightarrow{\text{hombre}} \approx \overrightarrow{\text{king}}$$

$$\overrightarrow{\text{reina}} - \overrightarrow{\text{mujer}} + \overrightarrow{\text{hombre}} \approx \overrightarrow{\text{rey}}$$

Useful in bilingual lexicon extraction!

- A novel model for learning bilingual word embeddings (BWEs) from non-parallel document-aligned data
- A simple framework for constructing query and document embeddings
- A unified framework for MoIR and CLIR based on (bilingual) word embeddings

Part II: BWEs in IR

We learn **word embeddings**:

$vec(\text{huhblub}) = [-0.23, 0.44, -0.76, 0.33, 0.19, \dots]$

$vec(\text{fluffy}) = [0.31, 0.02, -0.11, -0.28, 0.52, \dots]$

→ How to build document and query embeddings?

$vec(\text{huhblup is fluffy}) = ??$

Adapting the framework from *compositional distributional semantics*: [Mitchell and Lapata, ACL 2008; Socher et al., EMNLP 2011; Milajevs et al., EMNLP 2014] and many more...

A generic composition with a **bag-of-words** assumption:

$(d = \{w_1, w_2, \dots, w_{|N_d|}\})$

$$\vec{d} = \vec{w}_1 \star \vec{w}_2 \star \dots \star \vec{w}_{|N_d|}$$

\star = compositional vector operator (addition, multiplication, tensor product,...)

A general framework → in this work the simple and effective **additive composition**:

[Mitchell and Lapata, ACL 2008]

$$\vec{d} = \vec{w}_1 + \vec{w}_2 + \dots + \vec{w}_{|N_d|}$$

The *dim*-dimensional **document embedding** in the same bilingual word embedding space:

$$\vec{d} = [f_{d,1}, \dots, f_{d,k}, \dots, f_{d,dim}]$$

→ the **ADD-BASIC** composition model

A slightly more intelligent idea → weighting the summands using their **self information** computed in the target collection:

$$si_w = -\ln \frac{freq(w, \mathcal{DC})}{|N_{\mathcal{DC}}|}$$

$freq(w, \mathcal{DC})$ = frequency of w in the collection

A SI-weighted sum:

$$\vec{d} = si_{w_1} \cdot \vec{w}_1 + si_{w_2} \cdot \vec{w}_2 + \dots + si_{w_{|N_d|}} \cdot \vec{w}_{|N_d|}$$

→ the **ADD-SI** composition model

- The same principles with **queries**
- Using only ADD-BASIC

$$\vec{Q} = \vec{q}_1 + \vec{q}_2 + \dots + \vec{q}_m$$

The *dim*-dimensional **query embedding** in the same bilingual word embedding space:

$$\vec{Q} = [f_{Q,1}, \dots, f_{Q,k}, \dots, f_{Q,dim}]$$

- 1 **Induce** a bilingual word embedding space using any BWE induction model \rightarrow *in this work: BWESG*
- 2 Given is a target document collection $\mathcal{DC} = \{d'_1, \dots, d'_{N'}\}$.
Compute dim -dimensional document embeddings \vec{d}' for each $d' \in \mathcal{DC}$ using the dim -dimensional WEs from the set \mathcal{BWE} obtained in the previous step and a semantic composition model (ADD-BASIC or ADD-SI something anything else).
- 3 After the query $Q = \{q_1, \dots, q_m\}$ is issued in language L_S , **compute** a dim -dimensional query embedding using the ADD-BASIC composition model.

- 4 For each $d' \in \mathcal{DC}$, compute the semantic similarity score $sim(d', Q)$ which quantifies each document's relevance to the query Q :

$$sim(d', Q) = SF(d', Q) = \frac{\vec{d}' \cdot \vec{Q}}{|\vec{d}'| \cdot |\vec{Q}|}$$

- 5 Rank all documents from \mathcal{DC} according to their similarity scores from the previous step.

WE-VS: WE-based MoIR and CLIR models (using ADD-BASIC)

Part IIb: Experiments

- Stochastic gradient descent with a default global learning rate 0.025
- Other default word2vec parameters: subsampling rate $1e - 4$, negative sampling with 25 negative samples, 15 epochs
- 10 random corpora shuffles, although **we advocate the use of a more intelligent shuffling procedure** (developed after the paper was released)
- $d = 100 - 800$ in steps of 100
- $cs = 10 - 100$ in steps of 10

[English|Dutch] → [English|Dutch] retrieval

Exactly the same setup as in: [Vulić et al., Information Retrieval 2013, ECIR 2013]

Training data Europarl 6,206 documents (parallel)
 Wikipedia 7,612 documents (comparable)

Vocabulary size English 76,555 words
 Dutch 71,168 words

→ Stop words removed

→ We exploit document-level alignments as the only bilingual signal (even for Europarl)

[English|Dutch] → [English|Dutch] retrieval (using CLEF 2001-2003 campaigns)

Monolingual				
<i>Direction</i>	<i>DC</i>	<i># Docs</i>	<i>Query Set</i>	<i># Queries</i>
EN→EN 2001	LAT	110,861	EN'01: 41-90	47
EN→EN 2002	LAT	110,861	EN'02: 91-140	42
EN→EN 2003	LAT+GH	166,753	EN'03: 141-200	53
NL→NL 2001	NC+AD	190,604	NL'01: 41-90	50
NL→NL 2002	NC+AD	190,604	NL'02: 91-140	50
NL→NL 2003	NC+AD	190,604	NL'03: 141-200	56

[English|Dutch] → [English|Dutch] retrieval (using CLEF 2001-2003 campaigns)

Cross-lingual				
<i>Direction</i>	<i>DC</i>	# Docs	Query Set	# Queries
NL→EN 2001	LAT	110,861	NL'01: 41-90	47
NL→EN 2002	LAT	110,861	NL'01: 91-140	42
NL→EN 2003	LAT+GH	166,753	NL'03: 141-200	53
EN→NL 2001	NC+AD	190,604	EN'01: 41-90	50
EN→NL 2002	NC+AD	190,604	EN'02: 91-140	50
EN→NL 2003	NC+AD	190,604	EN'03: 141-200	56

→ Queries extracted from the *title + description* fields

→ Stop words removed → Measuring MAP

Single models:

1. **WE-VS**: Our WE-based retrieval model
2. **LM-UNI**: Unigram query likelihood language model with standard Dirichlet smoothing
3. **LDA-IR**: Semantically-aware (Bi)LDA-based QL model
[Wei and Croft, SIGIR 2006; Vulić et al, IR 2013]

A detailed description of all the models along with their parameter setup in the paper!

Combined models:

1. **LM-UNI+LDA-IR**: A linear combination of the two single models:

[Wei and Croft, SIGIR 2006; Vulić et al, IR 2013]

$$P(q_i|d) = \lambda P_{lda}(q_i|d) + (1 - \lambda) P_{lm}(q_i|d)$$

2. **LM-UNI+WE-VS**: A linear combination of LM-UNI and WE-VS (to directly compare the “quality of semantic awareness” in the retrieval process)

x. **GT+LM+LDA** (only for CLIR): Translating a query using *Google Translate*, and then employing LM-UNI+LDA-IR on the translated query

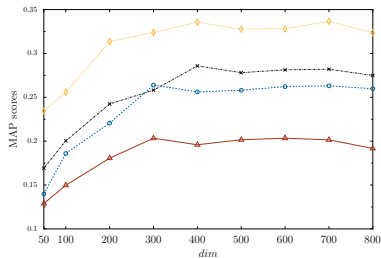
Again, a detailed description of all the models along with their parameter setup in the paper!

Model	EN→EN			NL→NL		
	2001	2002	2003	2001	2002	2003
LM-UNI	.381	.360	.359	.256	.323	.357
LDA-IR <i>dim:300; cs:60</i>	.279	.216	.241	.131	.143	.130
WE-VS <i>dim:600; cs:60</i>	.324x	.258x	.257y	.203x	.237x	.224x
WE-VS	.329x	.281x	.262y	.204x	.262x	.231x
LM+LDA <i>dim:300; cs:60</i>	.399	.360	.379	.260	.326	.357
LM+WE ($\lambda=0.3$)	.412y	.381x	.401y	.271x	.349x	.372x
LM+WE ($\lambda=0.5$)	.429x	.394x	.407x	.279x	.370x	.382x
LM+WE ($\lambda=0.7$) <i>dim:600; cs:60</i>	.451x	.392y	.389	.270	.364x	.373y
LM+WE ($\lambda=0.3$)	.419y	.382x	.403y	.274x	.350x	.373x
LM+WE ($\lambda=0.5$)	.436x	.391x	.408x	.282x	.371x	.383x
LM+WE ($\lambda=0.7$)	.430x	.392y	.381	.268	.367x	.374y

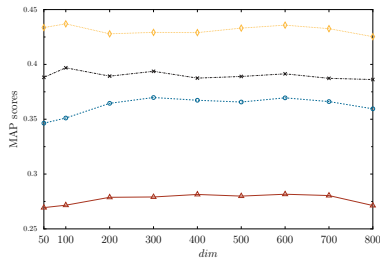
Model	EN→EN			NL→NL		
	2001	2002	2003	2001	2002	2003
LM-UNI	.381	.360	.359	.256	.323	.357
LDA-IR <i>dim:300; cs:60</i>	.279	.216	.241	.131	.143	.130
WE-VS <i>dim:600; cs:60</i>	.324x	.258x	.257y	.203x	.237x	.224x
WE-VS	.329x	.281x	.262y	.204x	.262x	.231x
LM+LDA <i>dim:300; cs:60</i>	.399	.360	.379	.260	.326	.357
LM+WE ($\lambda=0.3$)	.412y	.381x	.401y	.271x	.349x	.372x
LM+WE ($\lambda=0.5$)	.429x	.394x	.407x	.279x	.370x	.382x
LM+WE ($\lambda=0.7$) <i>dim:600; cs:60</i>	.451x	.392y	.389	.270	.364x	.373y
LM+WE ($\lambda=0.3$)	.419y	.382x	.403y	.274x	.350x	.373x
LM+WE ($\lambda=0.5$)	.436x	.391x	.408x	.282x	.371x	.383x
LM+WE ($\lambda=0.7$)	.430x	.392y	.381	.268	.367x	.374y

Model	EN→EN			NL→NL		
	2001	2002	2003	2001	2002	2003
LM-UNI	.381	.360	.359	.256	.323	.357
LDA-IR <i>dim:300; cs:60</i>	.279	.216	.241	.131	.143	.130
WE-VS <i>dim:600; cs:60</i>	.324x	.258x	.257y	.203x	.237x	.224x
WE-VS	.329x	.281x	.262y	.204x	.262x	.231x
LM+LDA <i>dim:300; cs:60</i>	.399	.360	.379	.260	.326	.357
LM+WE ($\lambda=0.3$)	.412y	.381x	.401y	.271x	.349x	.372x
LM+WE ($\lambda=0.5$)	.429x	.394x	.407x	.279x	.370x	.382x
LM+WE ($\lambda=0.7$) <i>dim:600; cs:60</i>	.451x	.392y	.389	.270	.364x	.373y
LM+WE ($\lambda=0.3$)	.419y	.382x	.403y	.274x	.350x	.373x
LM+WE ($\lambda=0.5$)	.436x	.391x	.408x	.282x	.371x	.383x
LM+WE ($\lambda=0.7$)	.430x	.392y	.381	.268	.367x	.374y

Testing the influence of dimensionality...

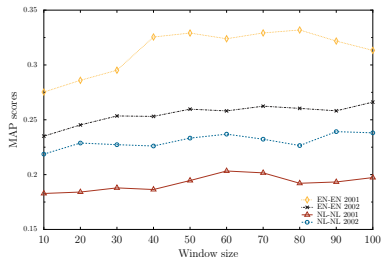


(a) WE-VS, $cs = 60$

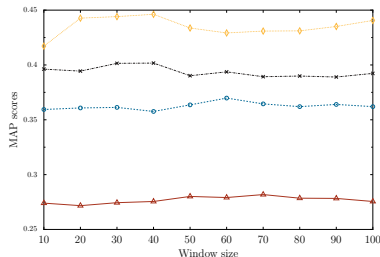


(b) LM-UNI+WE-VS, $cs = 60$

..and window size... (controlling the data dropout)



(c) WE-VS, $dim = 300$



(d) LM-UNI+WE-VS, $dim = 300$

Model	NL→EN			EN→NL		
	2001	2002	2003	2001	2002	2003
LM-UNI	.094	.108	.092	.078	.125	.112
LDA-IR	.197	.139	.123	.145	.137	.171
<i>dim:300; cs:60</i>						
WE-VS	.187	.204x	.120	.174	.185y	.157
<i>dim:600; cs:60</i>						
WE-VS	.222y	.230x	.127	.178y	.219x	.181
LM+LDA	.267	.225	.199	.225	.268	.278
GT+LM+LDA	.307	.275	.248	.230	.240	.244
<i>dim:300; cs:60</i>						
LM+WE ($\lambda=0.3$)	.189	.273	.197	.101	.159	.150
LM+WE ($\lambda=0.5$)	.218	.283y	.220	.113	.184	.167
LM+WE ($\lambda=0.7$)	.255	.307x	.219	.180	.209	.208
<i>dim:600; cs:60</i>						
LM+WE ($\lambda=0.3$)	.205	.281y	.198	.107	.167	.154
LM+WE ($\lambda=0.5$)	.236	.299x	.215	.123	.203	.183
LM+WE ($\lambda=0.7$)	.286	.317x	.222	.190	.249	.225

Model	NL→EN			EN→NL		
	2001	2002	2003	2001	2002	2003
LM-UNI	.094	.108	.092	.078	.125	.112
LDA-IR	.197	.139	.123	.145	.137	.171
<i>dim:300; cs:60</i>						
WE-VS	.187	.204x	.120	.174	.185y	.157
<i>dim:600; cs:60</i>						
WE-VS	.222y	.230x	.127	.178y	.219x	.181
LM+LDA	.267	.225	.199	.225	.268	.278
GT+LM+LDA	.307	.275	.248	.230	.240	.244
<i>dim:300; cs:60</i>						
LM+WE ($\lambda=0.3$)	.189	.273	.197	.101	.159	.150
LM+WE ($\lambda=0.5$)	.218	.283y	.220	.113	.184	.167
LM+WE ($\lambda=0.7$)	.255	.307x	.219	.180	.209	.208
<i>dim:600; cs:60</i>						
LM+WE ($\lambda=0.3$)	.205	.281y	.198	.107	.167	.154
LM+WE ($\lambda=0.5$)	.236	.299x	.215	.123	.203	.183
LM+WE ($\lambda=0.7$)	.286	.317x	.222	.190	.249	.225

Model	NL→EN			EN→NL		
	2001	2002	2003	2001	2002	2003
LM-UNI	.094	.108	.092	.078	.125	.112
LDA-IR	.197	.139	.123	.145	.137	.171
<i>dim:300; cs:60</i>						
WE-VS	.187	.204x	.120	.174	.185y	.157
<i>dim:600; cs:60</i>						
WE-VS	.222y	.230x	.127	.178y	.219x	.181
LM+LDA	.267	.225	.199	.225	.268	.278
GT+LM+LDA	.307	.275	.248	.230	.240	.244
<i>dim:300; cs:60</i>						
LM+WE ($\lambda=0.3$)	.189	.273	.197	.101	.159	.150
LM+WE ($\lambda=0.5$)	.218	.283y	.220	.113	.184	.167
LM+WE ($\lambda=0.7$)	.255	.307x	.219	.180	.209	.208
<i>dim:600; cs:60</i>						
LM+WE ($\lambda=0.3$)	.205	.281y	.198	.107	.167	.154
LM+WE ($\lambda=0.5$)	.236	.299x	.215	.123	.203	.183
LM+WE ($\lambda=0.7$)	.286	.317x	.222	.190	.249	.225

Model	NL→EN			EN→NL		
	2001	2002	2003	2001	2002	2003
LM-UNI	.094	.108	.092	.078	.125	.112
LDA-IR <i>dim:300; cs:60</i>	.197	.139	.123	.145	.137	.171
WE-VS <i>dim:600; cs:60</i>	.187	.204x	.120	.174	.185y	.157
WE-VS	.222y	.230x	.127	.178y	.219x	.181
LM+LDA	.267	.225	.199	.225	.268	.278
GT+LM+LDA <i>dim:300; cs:60</i>	.307	.275	.248	.230	.240	.244
LM+WE ($\lambda=0.3$)	.189	.273	.197	.101	.159	.150
LM+WE ($\lambda=0.5$)	.218	.283y	.220	.113	.184	.167
LM+WE ($\lambda=0.7$) <i>dim:600; cs:60</i>	.255	.307x	.219	.180	.209	.208
LM+WE ($\lambda=0.3$)	.205	.281y	.198	.107	.167	.154
LM+WE ($\lambda=0.5$)	.236	.299x	.215	.123	.203	.183
LM+WE ($\lambda=0.7$)	.286	.317x	.222	.190	.249	.225

Model	NL→EN			EN→NL		
	2001	2002	2003	2001	2002	2003
LM+LDA	.267	.225	.199	.225	.268	.278
GT+LM+LDA	.307	.275	.248	.230	.240	.244
<i>dim:600; cs:60</i>						
LM+WE ($\lambda=0.3$)	.205	.281y	.198	.107	.167	.154
LM+WE ($\lambda=0.5$)	.236	.299x	.215	.123	.203	.183
LM+WE ($\lambda=0.7$)	.286	.317x	.222	.190	.249	.225
<i>dim:600; cs:60</i>						
LM+LDA+WE ($\lambda=0.3$)	.277	.263	.210	.229	.288	.283
LM+LDA+WE ($\lambda=0.5$)	.281y	.281y	.214	.240	.297y	.290
LM+LDA+WE ($\lambda=0.7$)	.302x	.302x	.227	.244y	.311x	.302y

Composition	Monolingual					
	EN→EN			NL→NL		
	2001	2002	2003	2001	2002	2003
ADD-BASIC (300-60)	.324	.258	.257	.203	.237	.224
ADD-SI (300-60)	.338	.278y	.255	.212	.253y	.227
ADD-BASIC (600-60)	.329	.281	.262	.204	.262	.231
ADD-SI (600-60)	.344y	.301y	.263	.215	.275y	.234

Composition	Cross-lingual					
	NL→EN			EN→NL		
	2001	2002	2003	2001	2002	2003
ADD-BASIC (300-60)	.187	.204	.120	.174	.185	.157
ADD-SI (300-60)	.216 _x	.213 _y	.122	.189 _y	.208 _x	.161
ADD-BASIC (600-60)	.221	.230	.127	.178	.219	.181
ADD-SI (600-60)	.237 _y	.233	.130	.189	.229 _x	.184

- A novel model for learning bilingual word embeddings (BWEs) from non-parallel document-aligned data
- A simple framework for constructing query and document embeddings
- A unified framework for MoIR and CLIR based on (bilingual) word embeddings

The proposed framework is very general:

Designing other shuffling procedures for BWESG

Building new BWE induction models for the same multilingual data type (remove the need for pseudo-bilingual documents?)

Experimenting with other monolingual WE induction models for BWESG besides SGNS

Investigating other BWE induction models (different bilingual signals) in the same (CL)IR pipeline

Investigating more elaborate composition models to construct document and query embeddings (what about syntax?)

Testing true *paragraph and phrase embeddings* in the same (CL)IR pipeline

[Le and Mikolov, ICML 2014; Soyer et al., ICLR 2015]

Other (more distant) language pairs, other queries+test collections

Combining the semantic BWE-based knowledge with other IR modeling paradigms (besides the ones mentioned here)

