Cross-Lingual Word Representations: Induction and Evaluation

Ivan Vulić, Anders Søgaard, Manaal Faruqui

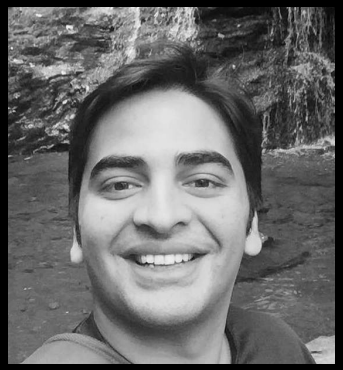

EMNLP 2017 Tutorial

Copenhagen; September 8, 2017

Before We Start...

Ivan

Anders



Manaal

Tutorial slides are available here:

http://people.ds.cam.ac.uk/iv250/tutorial/xlingrep-tutorial.pdf

#emnlp2017 #XlingWordRep

Motivation

We want to understand and model the meaning of...

Source: dreamstime.com

...without manual/human input and without perfect MT

Motivation

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 $\rightarrow$ this tutorial)

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

Motivation

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)

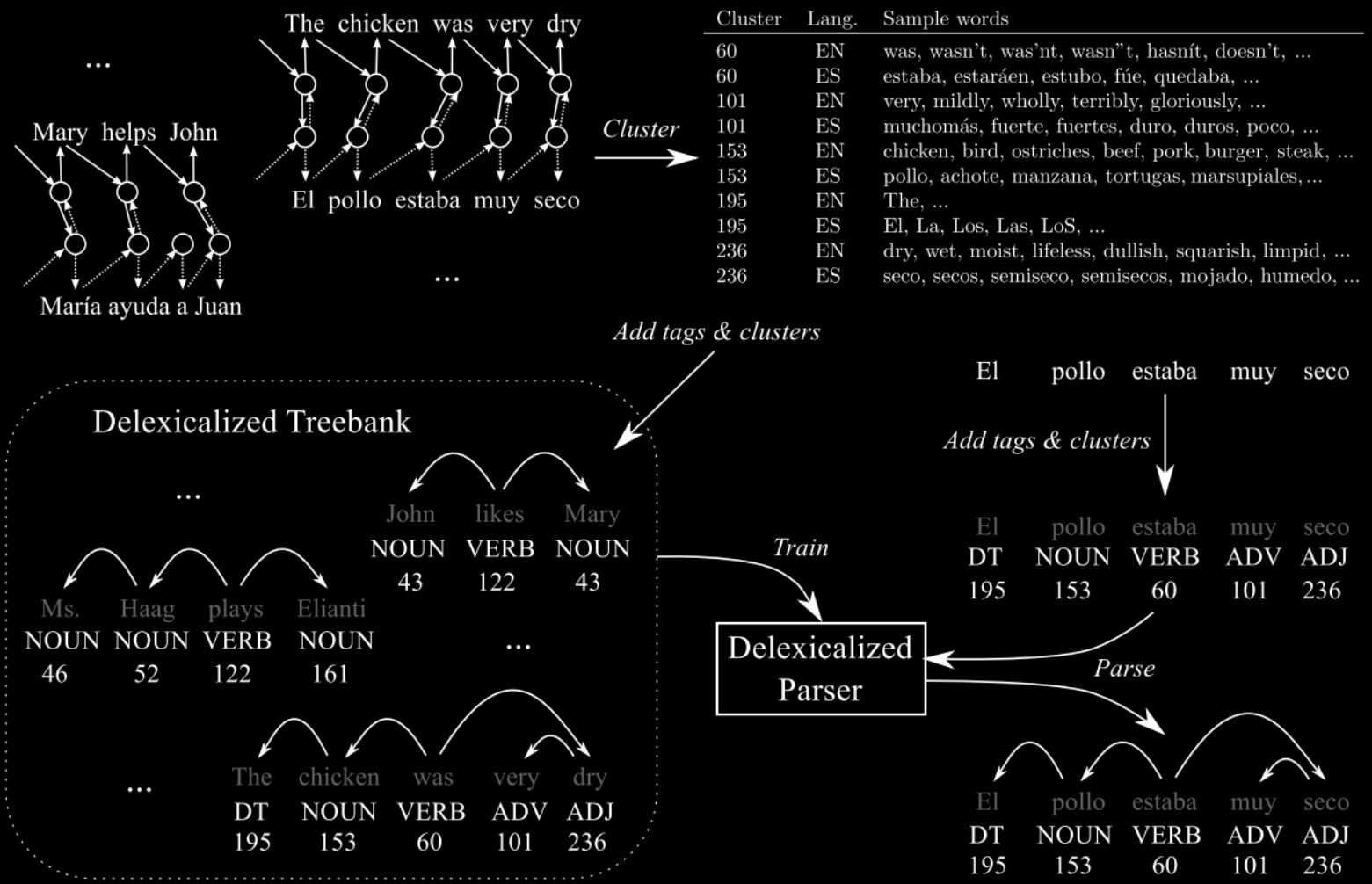
2. **language-invariant** (English, Dutch, Chinese, Spanish, ...)

(cross-lingual word embeddings $\rightarrow$ this tutorial)

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

Learn word-level features which generalise across tasks and languages

The World Existed B.E. (Before Embeddings)

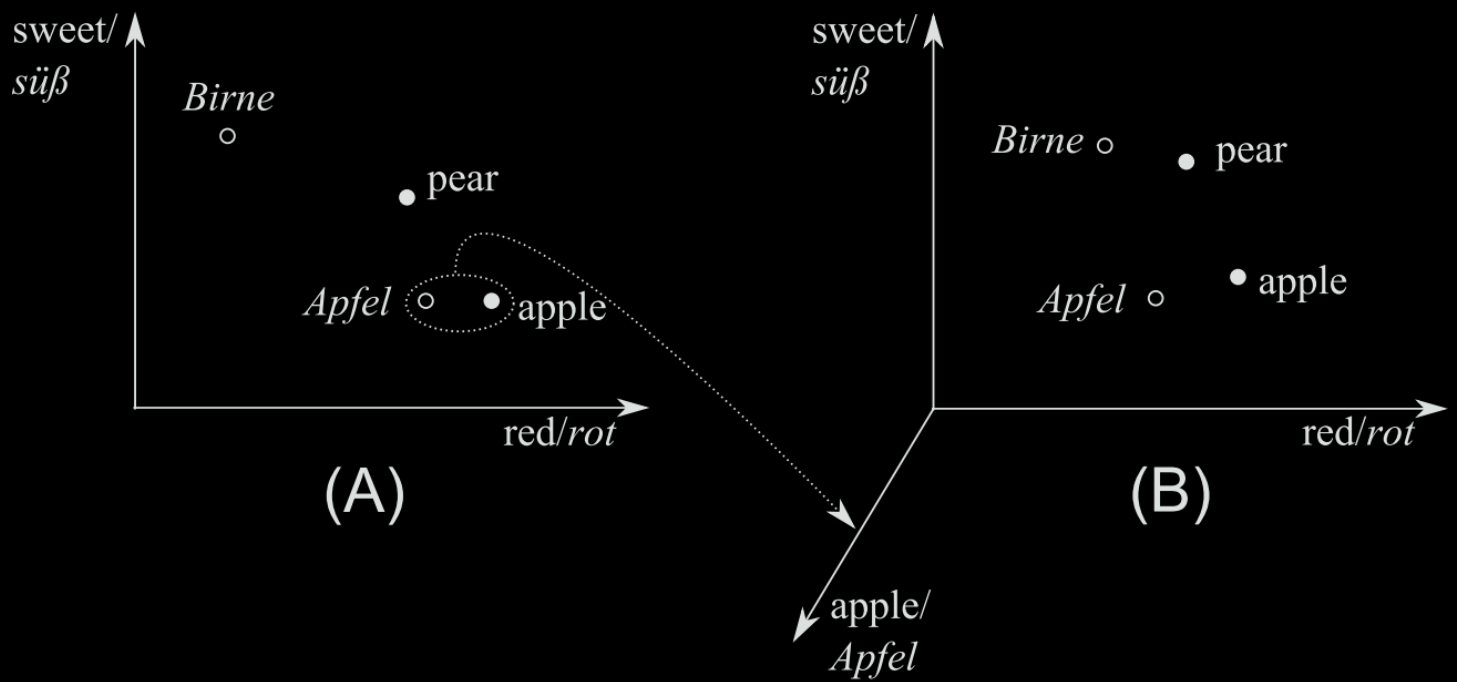


Cluster	Lang.	Sample words
60	EN	was, wasn't, was'nt, wasn"t, hasnít, doesn't, ...
60	ES	estaba, estaráen, estubo, fúe, quedaba, ...
101	EN	very, mildly, wholly, terribly, gloriously, ...
101	ES	muchomás, fuerte, fuertes, duro, duros, poco, ...
153	EN	chicken, bird, ostriches, beef, pork, burger, steak, ...
153	ES	pollo, achote, manzana, tortugas, marsupiales, ...
195	EN	The, ...
195	ES	El, La, Los, Las, LoS, ...
236	EN	dry, wet, moist, lifeless, dullish, squarish, limpid, ...
236	ES	seco, secos, semiseco, semisecos, mojado, humedo, ...

B.E. Example 1: Cross-lingual (Brown) clusters

[Tackstrom et al., NAACL 2012, Faruqui and Dyer, ACL 2013]

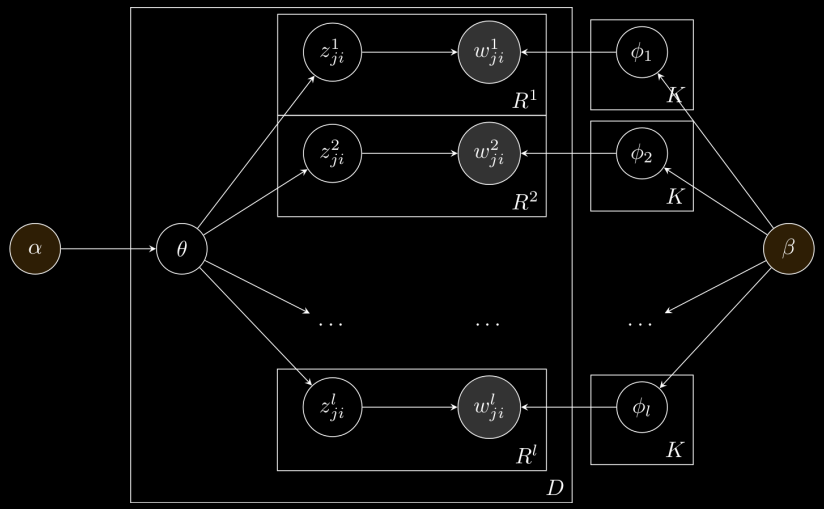
The World Existed B.E.

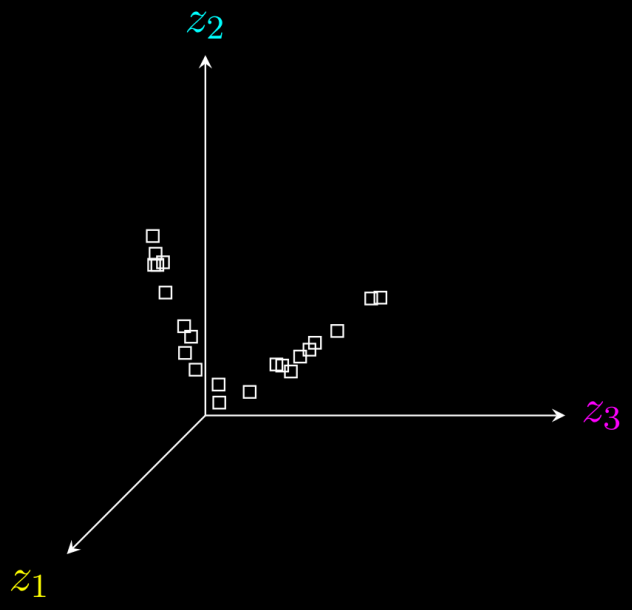


B.E. Example 2: Traditional "count-based" cross-lingual spaces

[Gaussier et al., ACL 2004; Peirsman and Padó, NAACL 2010]

The World Existed B.E.



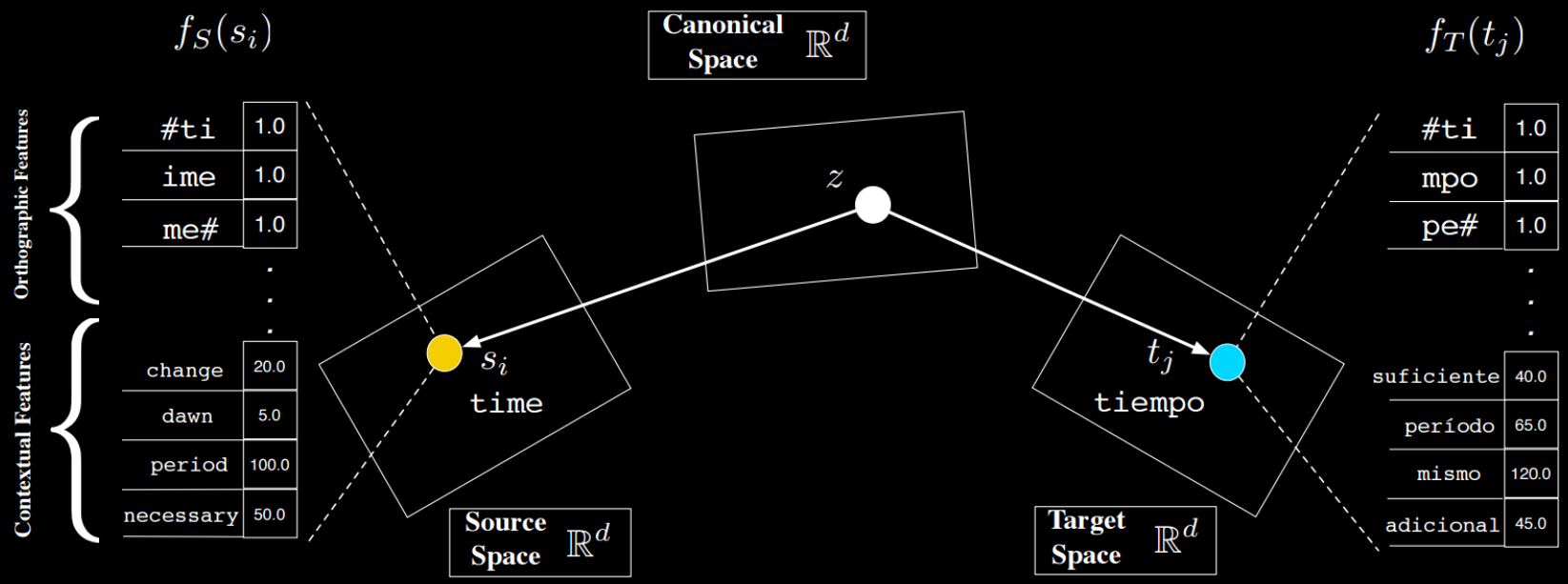


	soccer	fútbol
z_1	0.54	0.55
z_2	0.02	0.03
z_3	0.22	0.23

B.E. Example 3: Cross-lingual topical spaces

[Mimno et al., EMNLP 2009; Vulić et al., ACL 2011]

The World Existed B.E.



B.E. Example 4: Cross-lingual CCA-based vector spaces based on distributional and orthographic information

[Haghighi et al., ACL 2008]

Motivation Revisited

Cross-lingual word embeddings:

- simple: quick and efficient to train
- state-of-the-art and omnipresent
- lightweight and cheap

"Cross-lingual word embedding models are to MT what (monolingual) word embedding models are to language modeling."

by Sebastian Ruder

Are they?

They support cross-lingual NLP: enabling cross-lingual modeling and cross-lingual transfer

Tutorial Goals

- Link cross-lingual word embeddings to prior work in cross-lingual NLP
- Provide a systematic typology of cross-lingual word embedding models
- Analyze the importance of data requirements vs. algorithm choices
- Demonstrate that many current architectures are in fact very similar in their modeling assumptions and formulations
- Stress the importance of cross-lingual word embeddings in cross-lingual downstream tasks and applications

Tutorial Overview

Introduction *(15 mins)*

- Word embeddings and cross-lingual word embeddings
- Bilingual/multilingual data requirements:
  motivating the model typology

Part I: Learning from Word-Level Supervision *(45 mins)*

- Word-level supervision: word alignments and dictionaries
- Mapping-based approaches, extension and variations
- Approaches based on pseudo-bilingual corpora
- Joint online models
- From bilingual to multilingual spaces

Tutorial Overview

Part II: Learning from Sentence-Level Supervision *(30 mins)*

- Sentence-level supervision: sentence alignments without word alignment, sentence IDs
- Overview of standard approaches: bilingual autoencoders, BilBOWA, BiCVM, matrix factorization models
- Similarities to word-level model formulations
- From bilingual to multilingual spaces

Checkpoint: Coffee break

Tutorial Overview

Part III: Learning from Document-Level Supervision *(20 mins)*

- Document alignments vs sentence alignments: analogies and differences
- Inverse indexing and approaches based on pseudo-bilingual corpora
- Similarities to word-level and sentence-level model formulations

Part IV: Learning from Other Sources *(15 mins)*

- Other modalities and data sources (e.g., image captions, geospatial data, eye-tracking data)
- Multi-modal bilingual representations

Tutorial Overview

Part V: Evaluation and Application *(40 mins)*

- Intrinsic vs. extrinsic evaluation, current evaluation challenges
- Standard tasks: bilingual lexicon extraction, cross-lingual document classification
- Other possible (new and emerging) evaluation protocols
- Cross-lingual knowledge transfer (semantic vs. syntactic information)
- Applications in cross-lingual NLP and IR
- Multilingual embeddings: supporting massively multilingual NLP?

Outro: Useful Software, Future Challenges, Concluding Remarks *(15 mins)*

Word Embeddings

Representation of each word $w \in V$:

$$vec(w) = [f_1, f_2, \ldots, f_{dim}]$$

Word representations in the same semantic (or *embedding*) space!

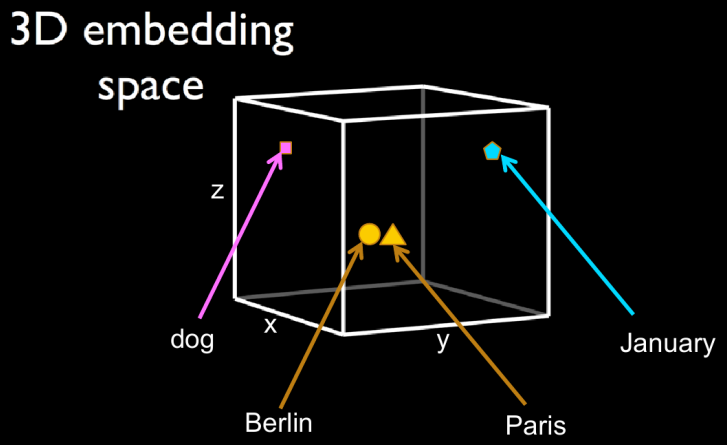


Image courtesy of [Gouws et al., ICML 2015]

Cross-Lingual Word Embeddings (CLWEs)

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

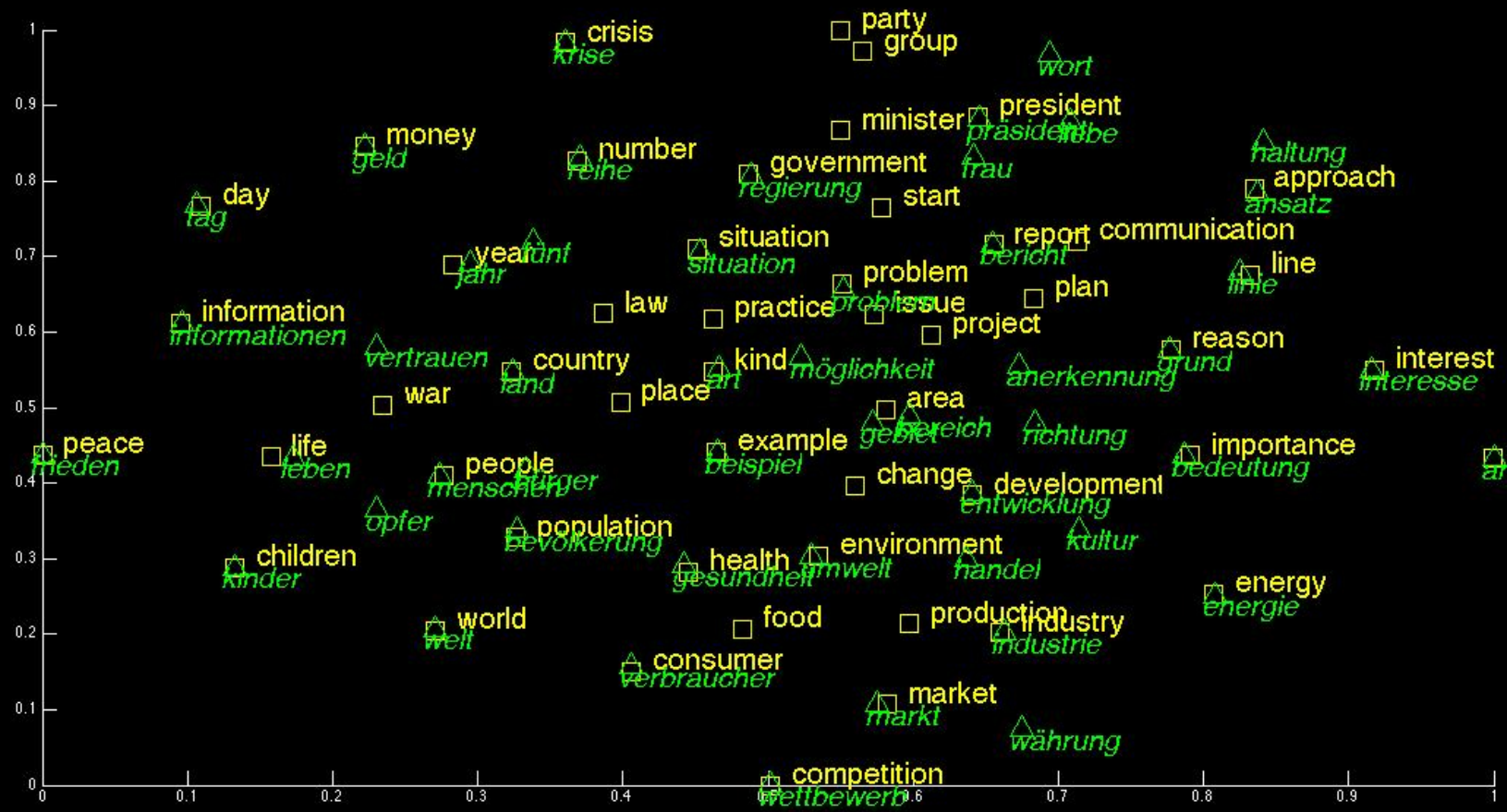
$$vec(w_1^S) = [f_1^1, f_2^1, \ldots, f_{dim}^1]$$

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

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

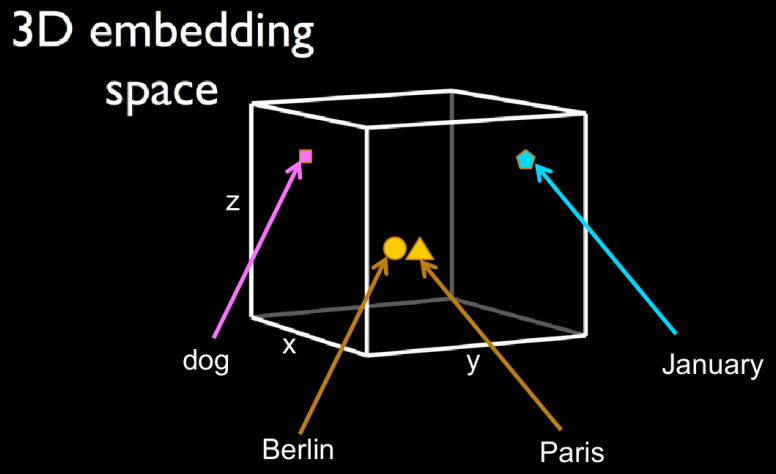
Language-independent word representations in the same shared semantic (or *embedding*) space!

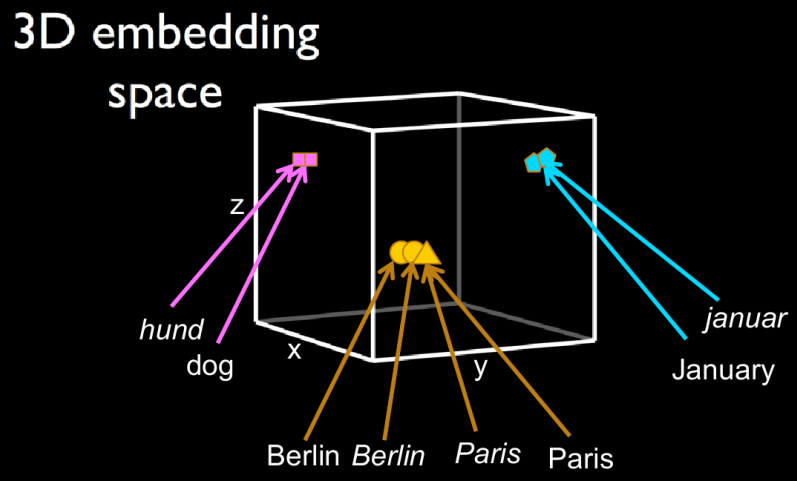
Cross-Lingual Word Embeddings



[Luong et al., NAACL 2015]

Cross-Lingual Word Embeddings





Monolingual vs. **Cross-lingual**

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

Q2 → Data Requirements: Which **bilingual signals** are used for the alignment?

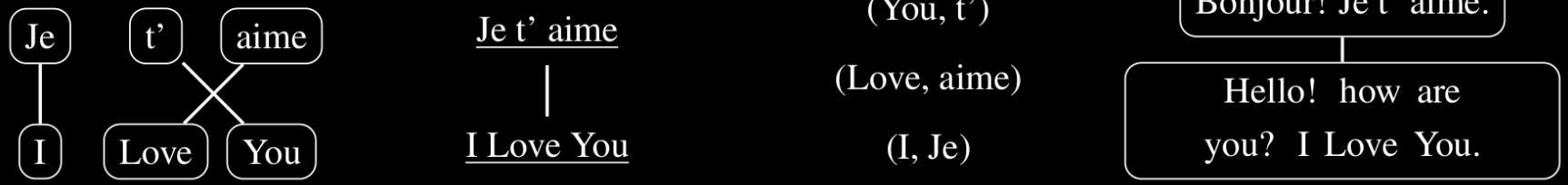
Data Requirements: Bilingual Signals

	Parallel	Comparable
Word	Dictionaries	Images
Sentence	Translations	Captions
Document	-	Wikipedia

Nature and alignment level of data sources required by CLWE models

Data requirements are more important for the final CLWE model performance than the actual underlying architecture
[Levy et al., EACL 2017]

Data Requirements: Bilingual Signals



[Upadhyay et al., ACL 2016]

Illustration of different data requirements and bilingual signals

1. Word-level: word alignments and translation dictionaries (in **Part I**)
2. Sentence-level: sentence alignments (in **Part II**)
3. Document-level: Wikipedia/news articles (in **Part III**)
4. Other: visual alignment: image captions, eye-tracking data (in **Part IV**)

Welcome to the CLWE Jungle I

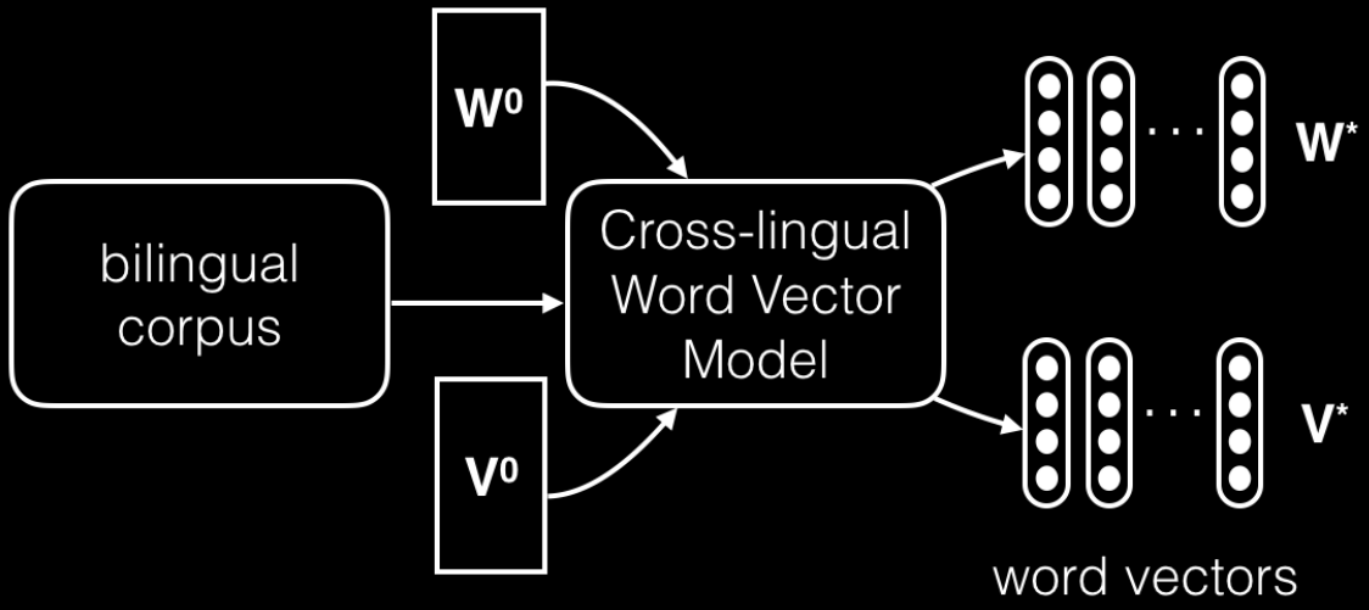
	Parallel	Comparable
Word	Mikolov et al. (2013)	Kiela et al. (2015)
	Faruqui & Dyer (2014)	Vulić et al. (2016)
	Xing et al. (2015)	Vulić et al. (2017)
	Dinu et al. (2015)	Zhang et al. (2017)
	Lazaridou et al. (2015)	Hauer et al. (2017)
	Zhang et al. (2016)	
	Zou et al. (2013)	
	Ammar et al. (2016)	
	Artexte et al. (2016)	
	Xiao & Guo (2014)	
	Gouws and Søgaard (2015)	
	Duong et al. (2016)	
	Gardner et al. (2015)	
	Smith et al. (2017)	
	Mrkšić et al. (2017)	
	Artetxe et al. (2017)	

Welcome to the CLWE Jungle II

	Parallel	Comparable
Sentence alignments	Guo et al. (2015) Vyas and Carpuat (2016) Hermann and Blunsom (2013) Lauly et al. (2013) Kočiský et al. (2014) Chandar et al. (2014) Pham et al. (2015) Klementiev et al. (2012) Soyer et al. (2015) Luong et al. (2015) Gouws et al. (2015) Coulmance et al. (2015) Shi et al. (2015) Rajendran et al. (2016) Levy et al. (2017)	Calixto et al. (2017) Gella et al. (2017)
Document	Hermann and Blunsom (2014)	Vulić & Korhonen (2016) Vulić and Moens (2016) Søgaard et al. (2015) Mogadala and Rettinger (2016)

General (Simplified) Methodology

[Upadhyay et al., ACL 2016]

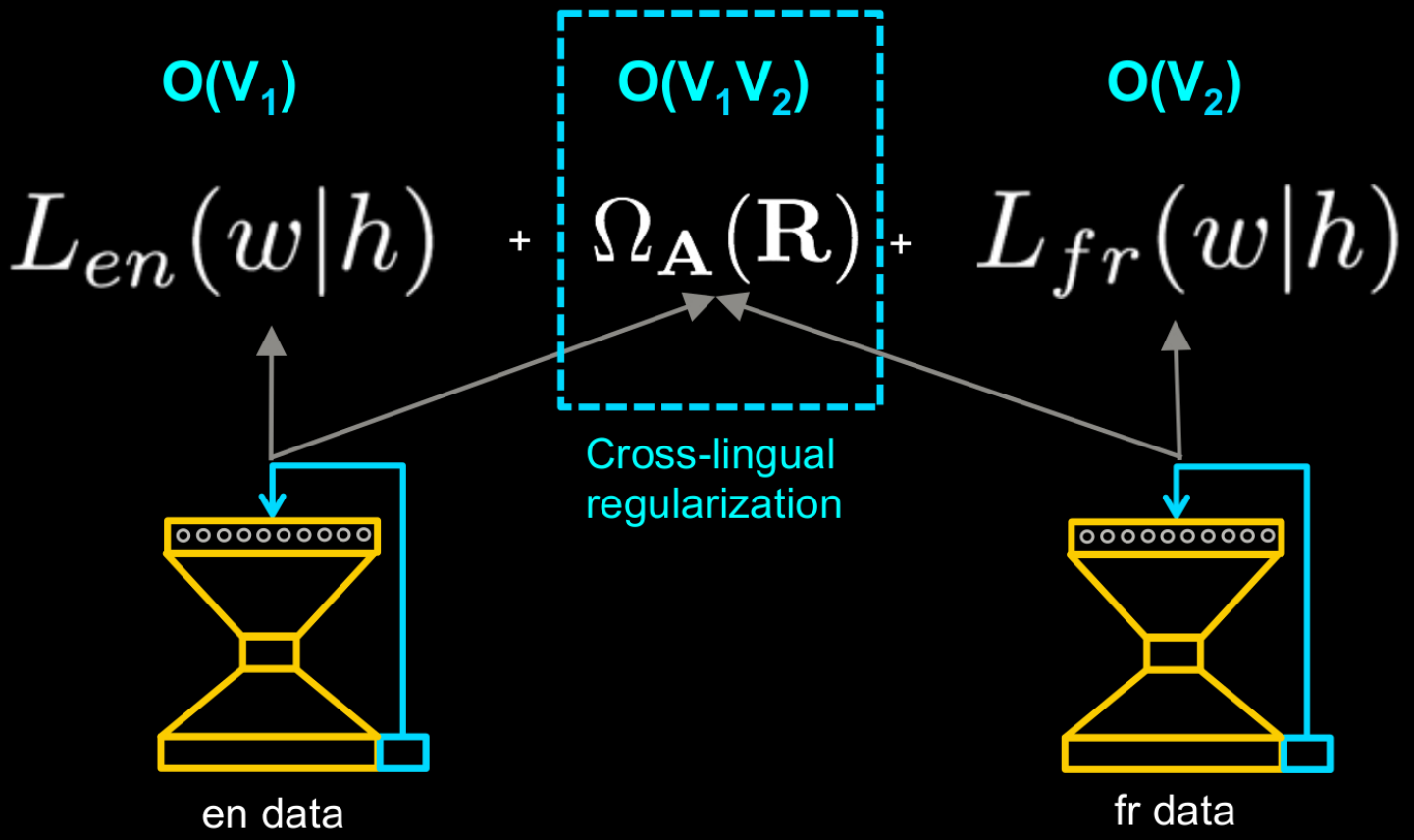


1. Initialize $\mathbf{W} \leftarrow \mathbf{W}^0, \mathbf{V} \leftarrow \mathbf{V}^0$
2. $(\mathbf{W}^f, \mathbf{V}^f) \leftarrow \arg\min \alpha\mathcal{L}^1(\mathbf{W}) + \beta\mathcal{L}^2(\mathbf{V}) + \Omega(W, V)$

General (Simplified) Methodology

(Bilingual) data sources are more important than the chosen algorithm

Most algorithms are formulated as: $\mathcal{L}^1(\mathbf{W}) + \mathcal{L}^2(\mathbf{V}) + \Omega(W, V)$



Part I: Learning from Word-Level Signal

Learning from Context

Most CLWE models are multilingual extensions of standard mono WE models

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

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

SG learns to predict the **context** of each pivot word.

John saw a cute gray huhblub running in the field.

$D =$ (huhblub, cute), (huhblub, gray), (huhblub, running), (huhblub, in)
$vec(huhblub) = [-0.23, 0.44, -0.76, 0.33, 0.19, \ldots]$

Three Clusters of *Word-Level Signal* Models

1. Mapping-based or offline approaches: train monolingual word representations independently and then learn a transformation matrix from one language to the other

2. Pseudo-bilingual approaches: train a monolingual WE model on automatically constructed corpora containing words from both languages

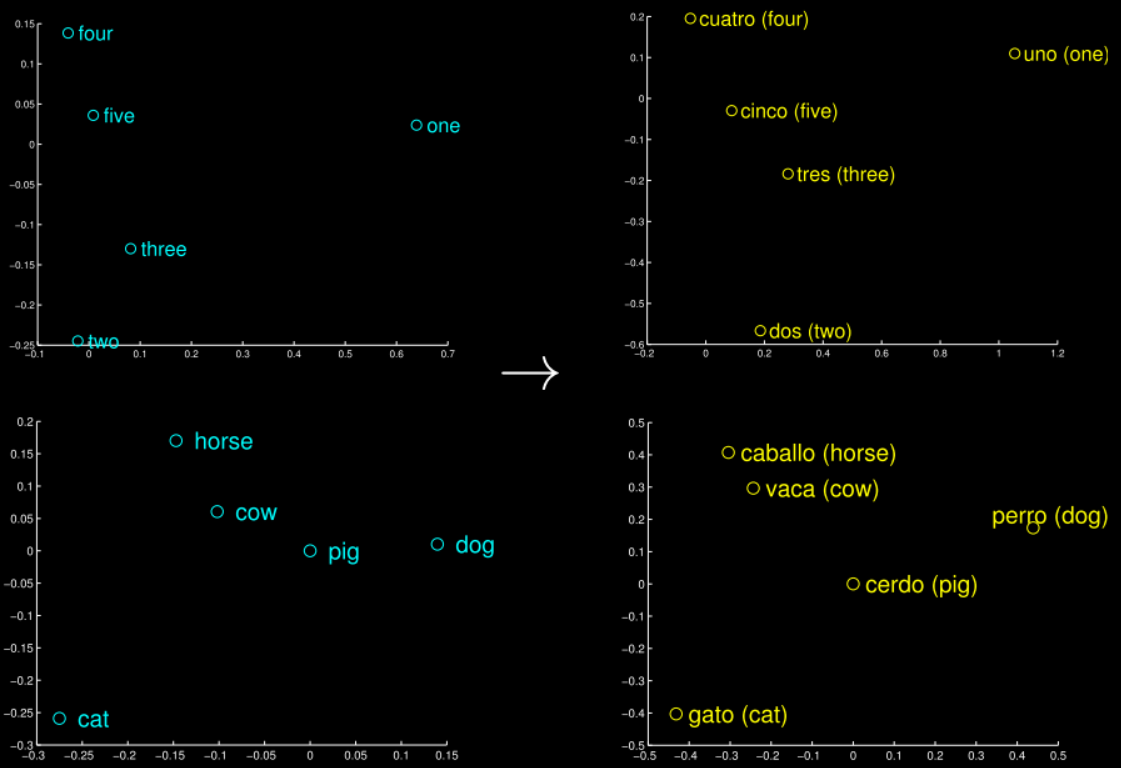
3. Joint online approaches: take parallel text as input and minimize the source and target language losses jointly with the regularization term

These approaches are equivalent, *modulo* optimization strategies

Basic Mapping Approach

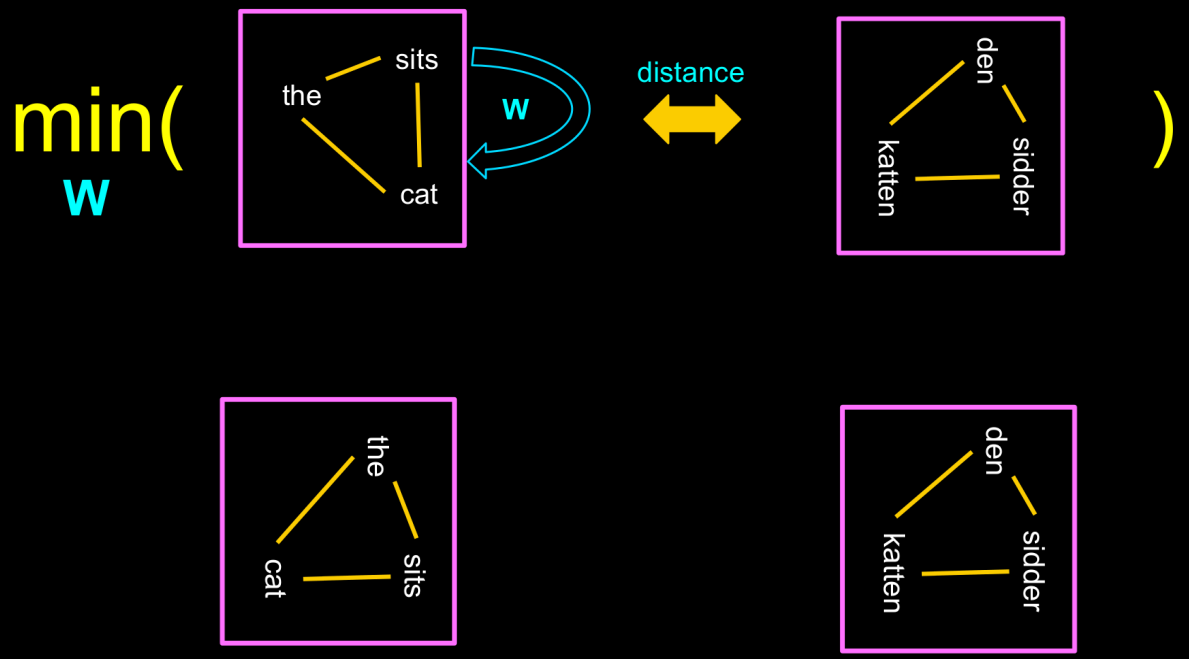
The geometric relations that hold between words are similar across languages: e.g., numbers and animals in English show a similar geometric constellation as their Spanish counterparts

[Mikolov et al., arXiv 2013]



Basic Mapping Approach

[Mikolov et al., arXiv 2013]



Learn to transform the pre-trained source language embeddings into a space where the distance between a word and its translation pair is minimised

Basic Mapping Approach

Based on given word translation pairs (x_i^s, x_i^t), $i = 1, \ldots, N$

Mean-squared error between the Euclidean distances (of the actual and transformed vector)

$$\Omega_{\text{MSE}} = \sum_{i=1}^{n} \|\mathbf{W}\mathbf{x}_i^s - \mathbf{x}_i^t\|^2$$

$$\Omega_{\text{MSE}} = \|\mathbf{W}\mathbf{X}^s - \mathbf{X}^t\|_F^2$$

Standard (re)formulation:

$$J = \mathcal{L}_{\text{SGNS}}^1 + \mathcal{L}_{\text{SGNS}}^2 + \Omega_{\text{MSE}}$$

Mapping Approach: Variations and Extensions

One tip: Matrix $\mathbf{W}$ may be computed more efficiently by taking the Moore-Penrose pseudoinverse:

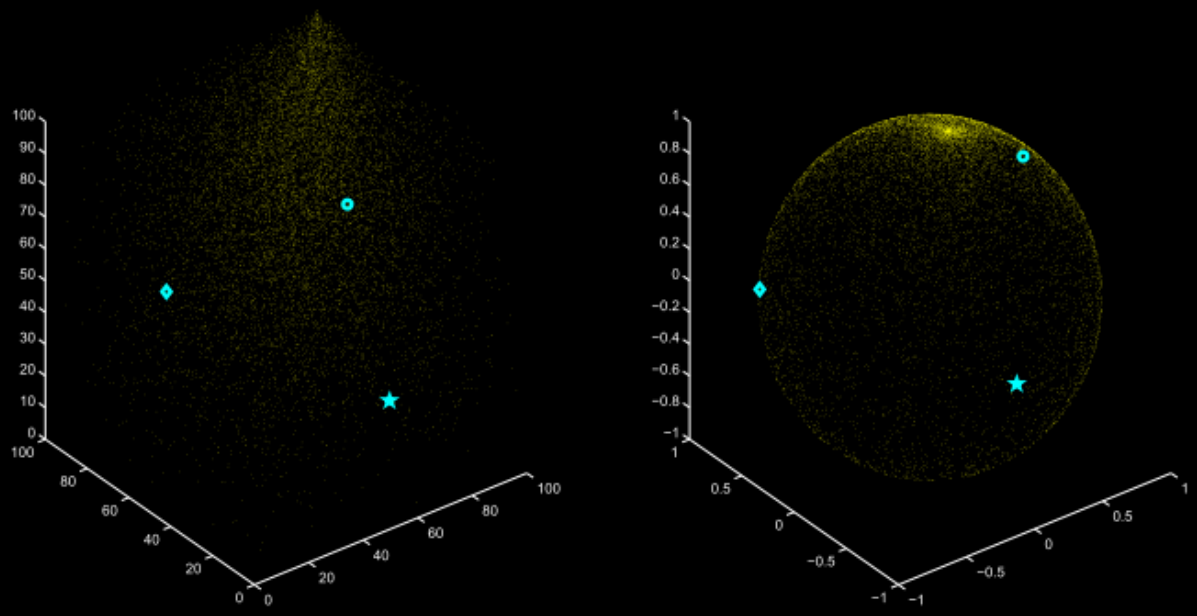
$$\mathbf{X}^{+} = (\mathbf{X}^{s\top}\mathbf{X}^{s})^{-1}\mathbf{X}^{S\top} \text{ as } \mathbf{W} = \mathbf{X}^{+}\mathbf{X}^{t}$$

[Artetxe et al., EMNLP 2016; Glavaš et al., EMNLP 2017]

Mapping Approach: Variations and Extensions

[Xing et al., NAACL 2015]: there is a mismatch between the initial objective function, the distance measure, and the transformation objective

Solution: 1. Normalization of word vectors to unit length + 2. Replacing MSE with cosine similarity for learning the mapping



$$\Omega_{\cos} = \max_{W} \sum_{i=1}^{n} cos(W x_i^s, x_i^t)$$

Mapping Approach: Variations and Extensions

[Zhang et al., NAACL 2016; Artetxe et al., EMNLP 2016; Smith et al., ICLR 2017]: analytical solution to the transformation problem

Unit length normalization and orthogonality imposed

Another requirement [Artetxe et al., EMNLP 2016]:

Two randomly selected words are generally not expected to be similar $\rightarrow$ the cosine of their embeddings in any dimension as well as their cosine similarity should be zero $\rightarrow$ adding centering matrix

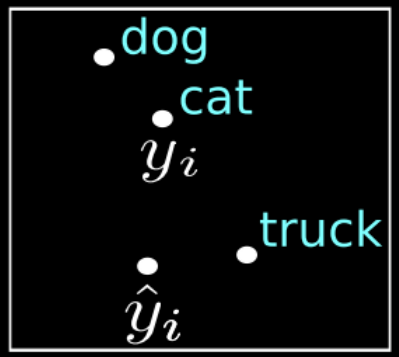
$$\Omega_{\text{centered}} = \max_{W} \sum_{i=1}^{n} cov(Wx_i^s, x_i^t)$$

Mapping Approach: Variations and Extensions

[Lazaridou et al., ACL 2015]:
Max-margin hinge loss (MMHL) with intruders instead of MSE

This reduces hubness and improves similarity computations

$$\sum_{j \neq i}^{k} \max\{0, \delta_{sim} + cos(\hat{y}_i, y_i) - cos(\hat{y}_i, y_j)\}$$

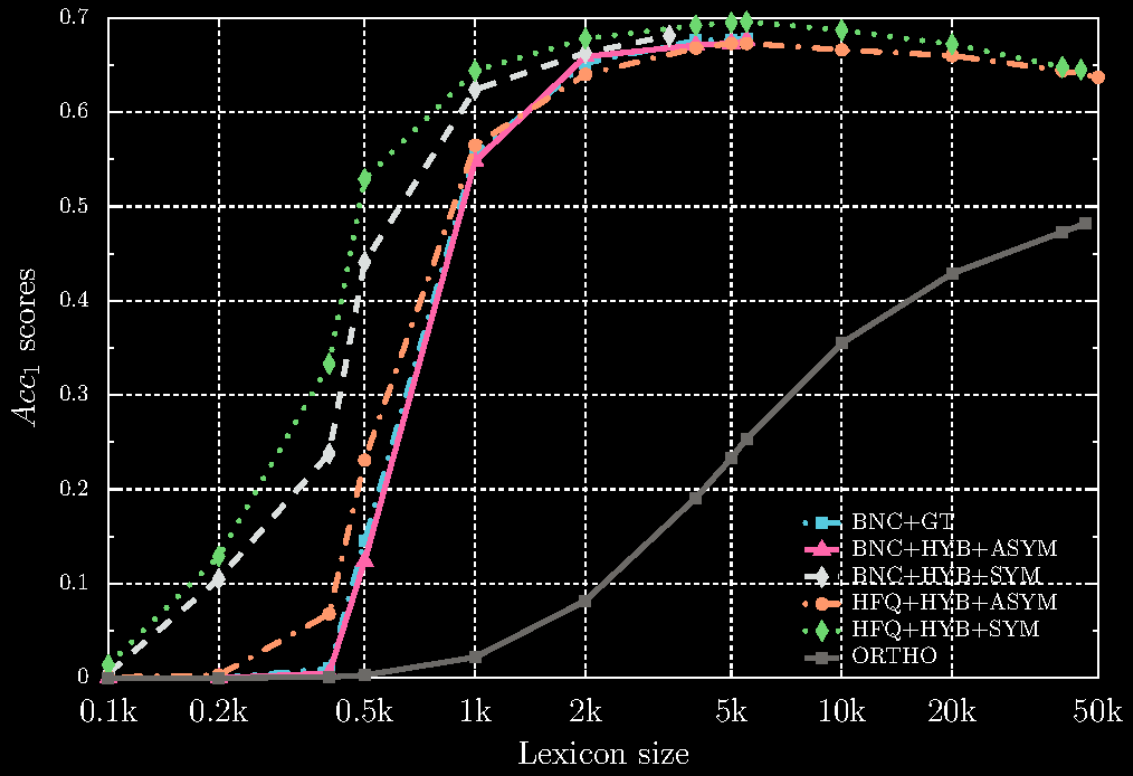


$$J = \mathcal{L}^1_{\text{CBOW}} + \mathcal{L}^2_{\text{CBOW}} + \Omega_{\text{MMHL}}$$

Mapping Approach: Variations and Extensions

[Vulić and Korhonen, ACL 2016]: Analysis of seed lexicons

Dictionary size and translation reliability are important factors

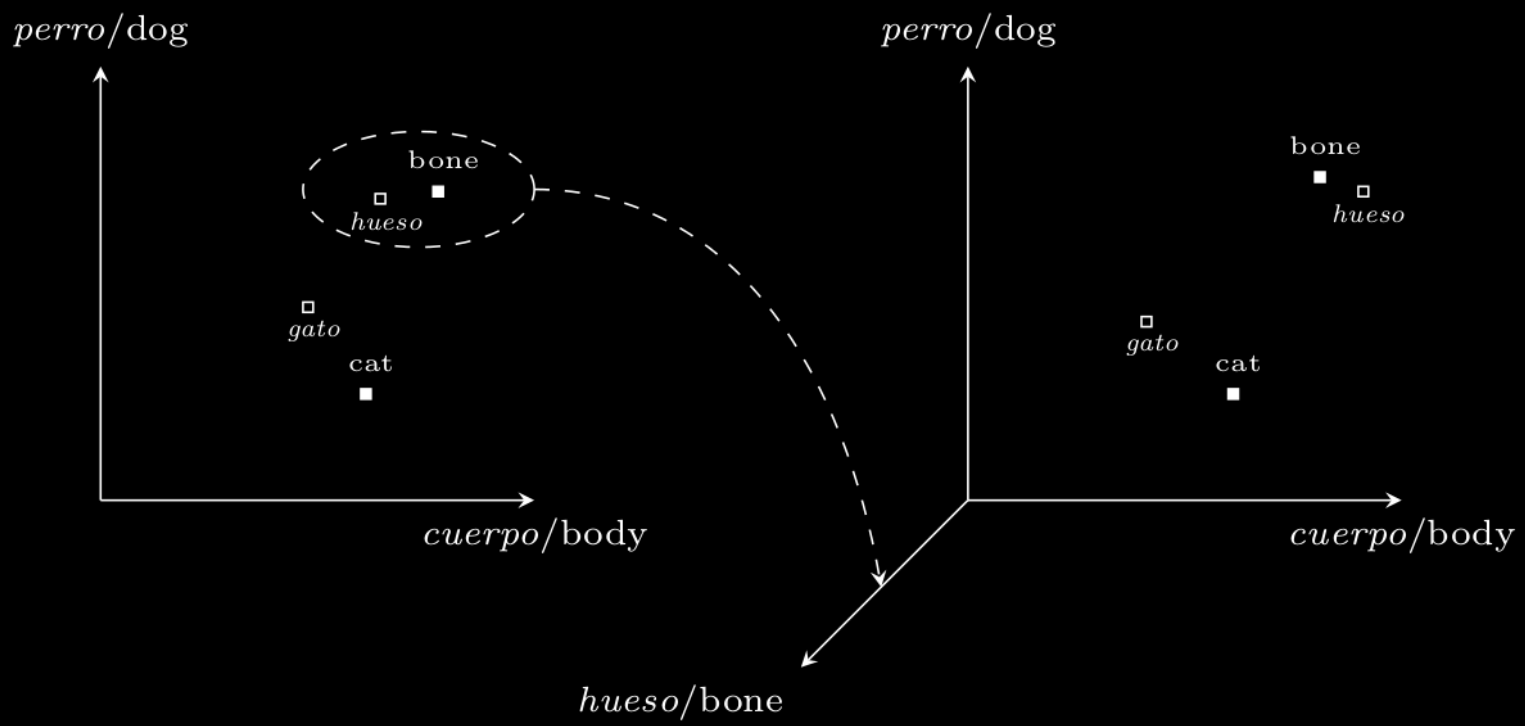


Hybrid model: it learns reliable translation pairs from non-parallel data and then seeds the transformation matrix learning

Bootstrapping from Small Seed Dictionaries

Minimising the requirements: the bootstrapping idea dates back to "pre-embedding" times

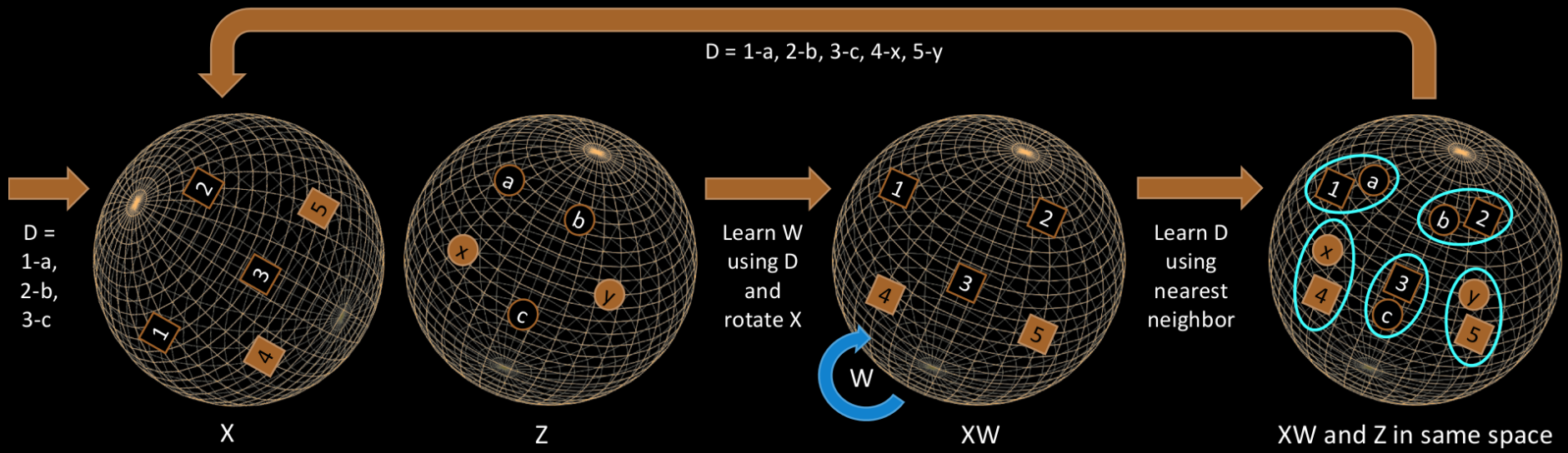
[Peirsman and Padó, NAACL 2010; Vulić and Moens; EMNLP 2013]



Application to truly low-resource cross-lingual settings?

Bootstrapping from Small Seed Dictionaries

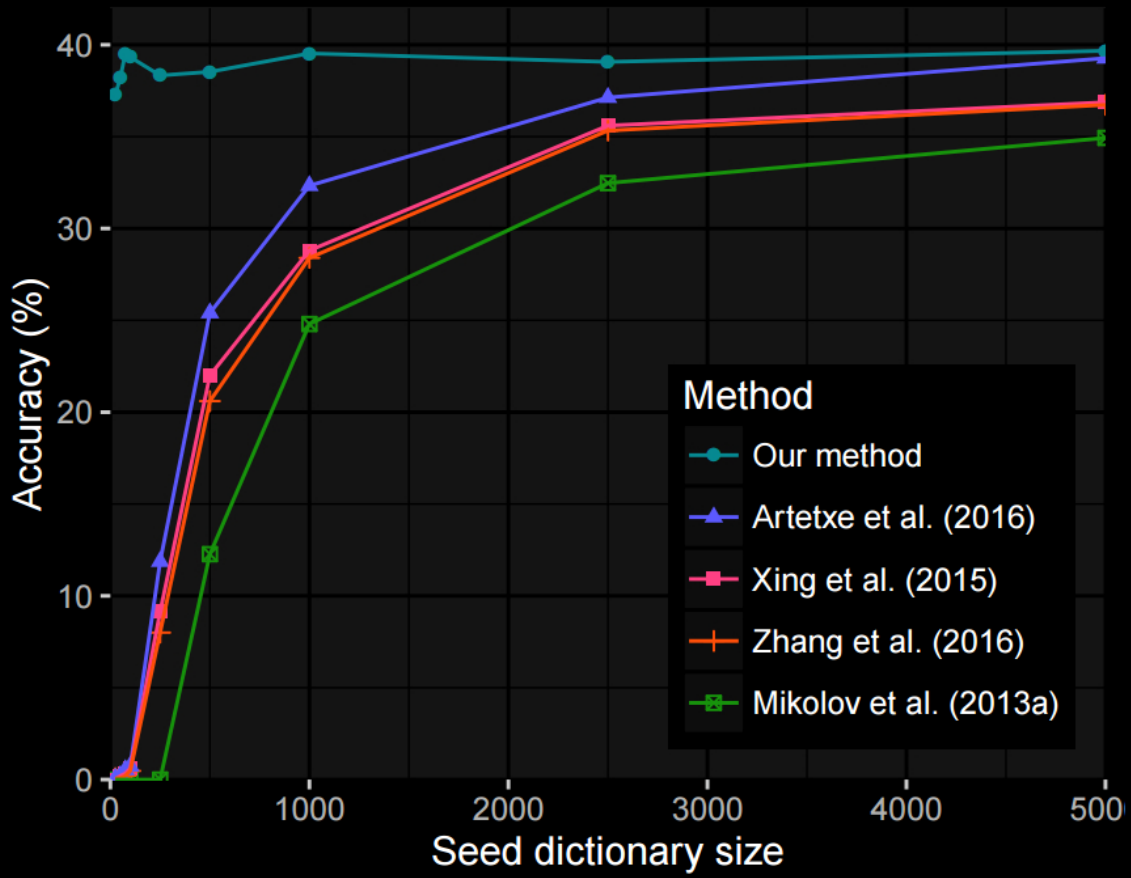
The latest instance of the bootstrapping idea
[Artetxe et al., ACL 2017]



Self-learning iterative framework: starting from cognates or numerals

Bootstrapping from Small Seed Dictionaries

Lexicon induction results from [Artetxe et al., ACL 2017]



Bootstrapping helps with really small seed dictionaries

A performance plateau with simple mapping approaches has been reached?

Mapping Approach: Bilingual to Multilingual

Fixing one space (English) and learning transformations for all monolingual word embeddings in other languages

This constructs a multilingual space from input monolingual spaces

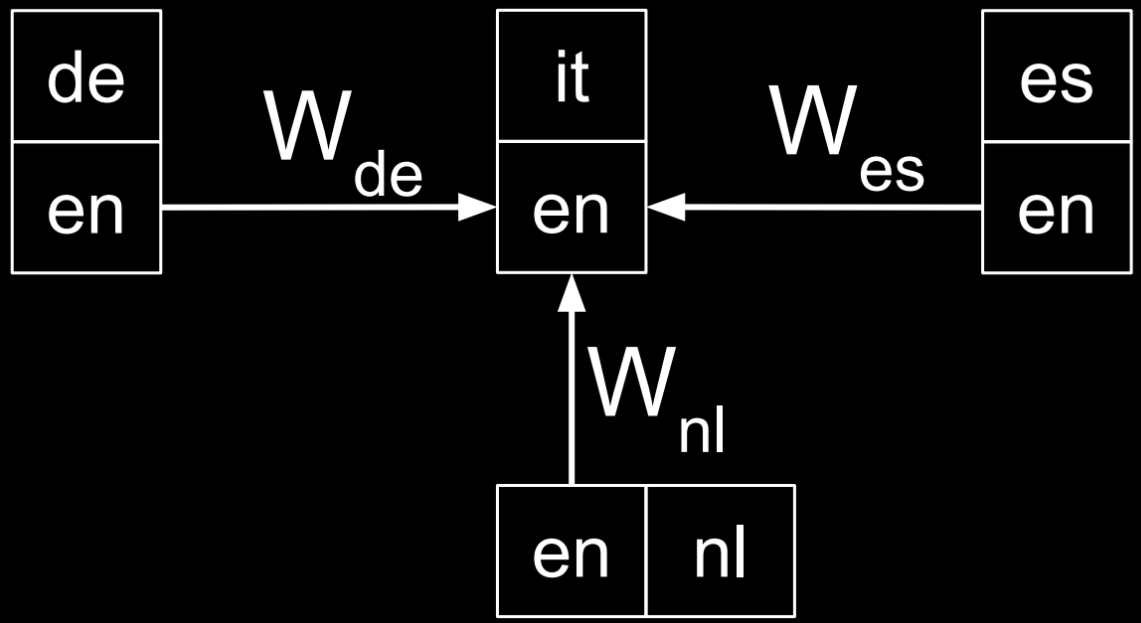
[Smith et al., ICLR 2017]:

A unified multilingual space for 89 languages using fastText vectors and 5k Google Translate pairs for each language pair

Mapping Approach: Bilingual to Multilingual

Multiple transformations with pivoting

This constructs a multilingual space from input bilingual spaces



[Duong et al., EACL 2017: Multilingual training of crosslingual word embeddings]

Mapping Approach: Bilingual to Multilingual

chico$_{es}$ - bruder$_{de}$ + sorella$_{it}$ (boy - brother + sister)	ehemann$_{de}$ - padre$_{es}$ + madre$_{it}$ (husband - father + mother)	principe$_{it}$ - junge$_{de}$ + meisje$_{nl}$ (prince - boy + girl)
chica$_{es}$ (girl)	**echtgenote$_{nl}$** (wife)	**principessa$_{it}$** (princess)
ragazza$_{it}$ (girl)	**moglie$_{it}$** (wife)	**princess$_{en}$**
meisje$_{nl}$ (girl)	her$_{en}$	**princesa$_{es}$** (princess)
girl$_{en}$	marito$_{it}$ (husband)	príncipe$_{es}$ (prince)
mädchen$_{de}$ (girl)	haar$_{nl}$ (her)	**prinzessin$_{de}$** (princess)

Multilingual semantic space is more expressive than separate bilingual spaces?

[Duong et al., EACL 2017]

Pseudo-Bilingual Training

[Gouws and Søgaard, NAACL 2015]

Barista: Bilingual Adaptive Reshuffling with Individual Stochastic Alternatives

1. Build a pseudo-bilingual corpus (or corrupt the original training data) using a set of predefined equivalences (e.g., POS tags, word translation pairs)

2. Train a monolingual WE model on the corrupted/shuffled corpus

Original: we build the house

POS: we build la voiture / they run la house

Translations: we construire the maison / nous build la house

Pseudo-Bilingual Training

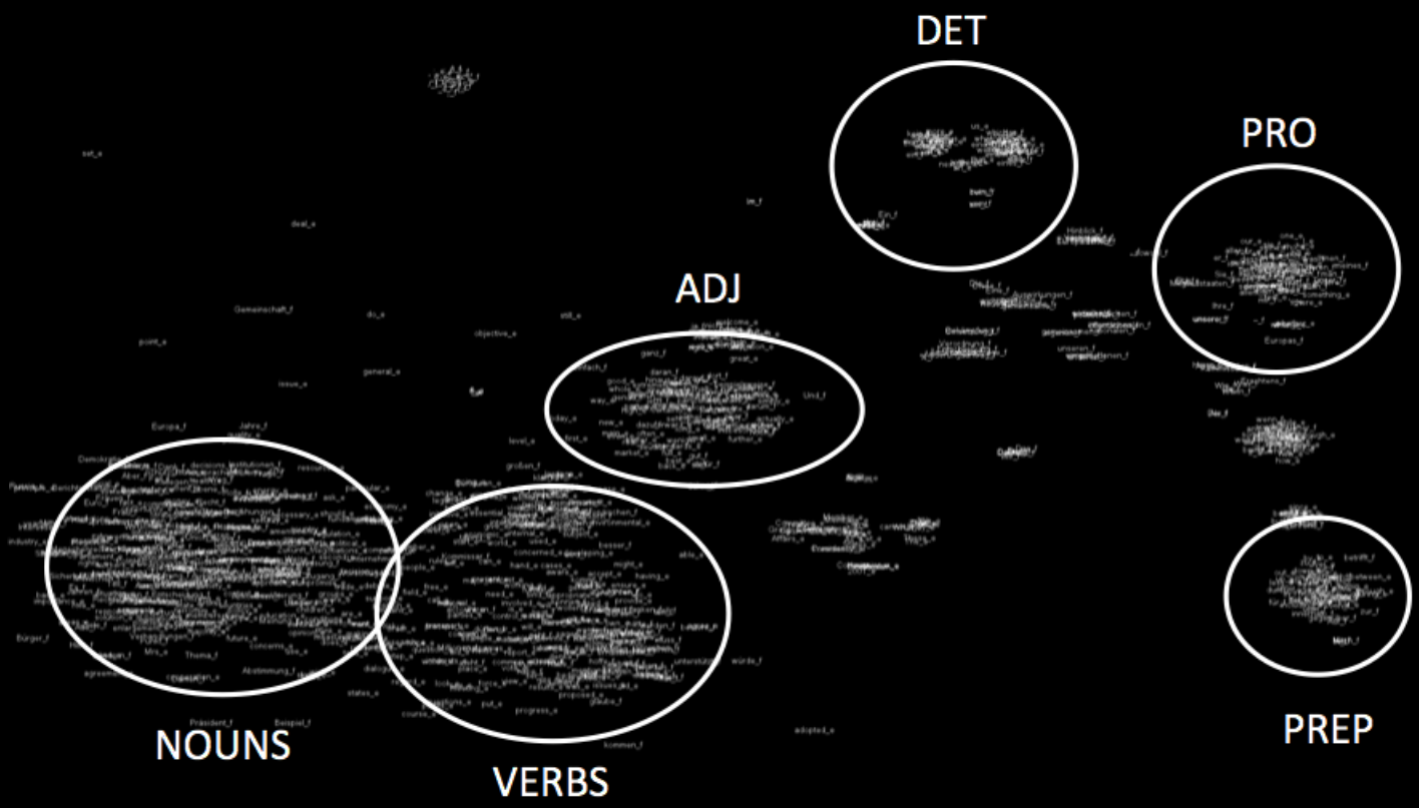
[Gouws and Søgaard, NAACL 2015]

Barista: Bilingual Adaptive Reshuffling with Individual Stochastic Alternatives

1. **Input:** Source corpus C_s, target corpus C_t, dictionary D
2. Concatenate C_s and C_t and then shuffle the concatenated corpus
3. Pass over the corpus, for each word w: if $\{w'|(w,w') \vee (w',v) \in D\}$ is non-empty and of cardinality k, replace w with w' with probability $1/2k$; keep w otherwise
4. Train a standard WE model on the pseudo-bilingual corpus C'

Pseudo-Bilingual Training

The chosen equivalence class dictates the shape of the output space

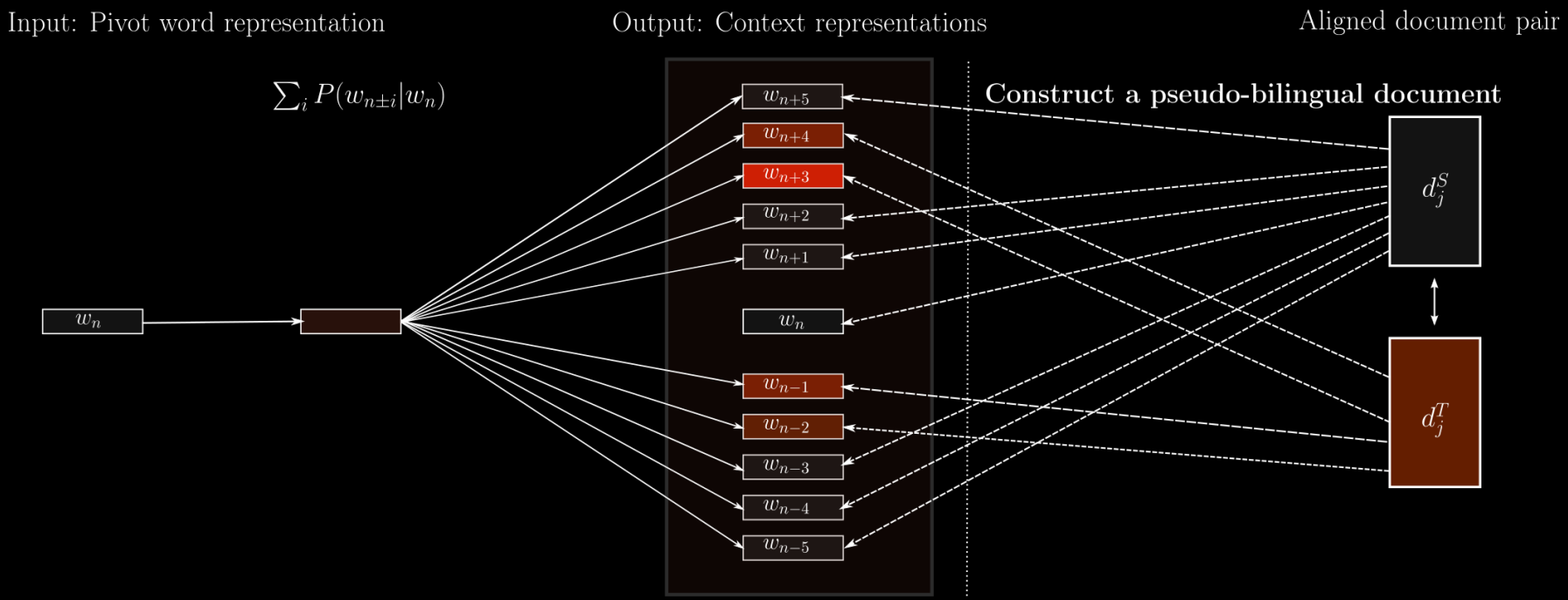


Cross-lingual embedding space with POS equivalences

Pseudo-Bilingual Training

Back to the future...

A similar idea was developed with sentence and document alignments



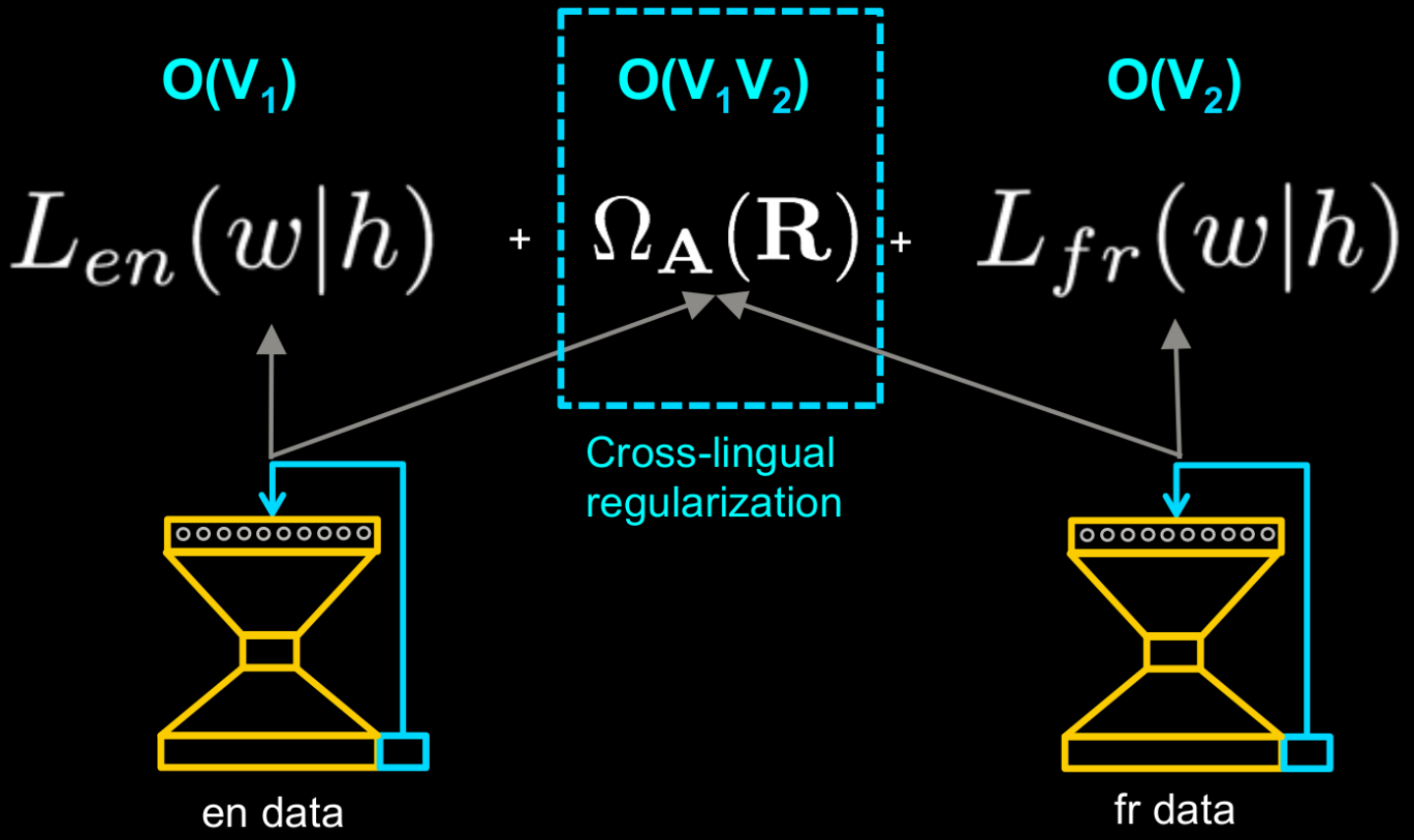
Shuffling aligned sentences/documents

Online Monolingual and Cross-Lingual Training

Combining monolingual and cross-lingual objectives in joint training

Monolingual → using monolingual data and preserving monolingual relations

Cross-lingual → tying the two spaces together



Online Monolingual and Cross-Lingual Training

[Zou et al., EMNLP 2013]: using normalized word alignment counts estimated from parallel texts stored in two matrices: $\mathbf{A}_{L_1 \rightarrow L_2}$ and $\mathbf{A}_{L_2 \rightarrow L_1}$

Monolingual objectives: global context with noise contrastive estimation

Cross-lingual objective

$$J_{L_1 \rightarrow L_2} = ||\mathbf{X}^t - \mathbf{A}_{L_1 \rightarrow L_2}\mathbf{X}^s||^2$$
$$J_{L_2 \rightarrow L_1} = ||\mathbf{X}^s - \mathbf{A}_{L_2 \rightarrow L_1}\mathbf{X}^t||^2$$
$$\Omega_{\text{WA}} = J_{L_1 \rightarrow L_2} + J_{L_2 \rightarrow L_1}$$

Again, standard formulation:

$$J = \mathcal{L}^1_{\text{NCE}} + \mathcal{L}^2_{\text{NCE}} + \Omega_{\text{WA}}$$

Online Monolingual and Cross-Lingual Training

[Klementiev et al., COLING 2012]: using word alignment counts plus neural language models on both sides

Monolingual objectives: two full-fledged neural language models

$$\mathcal{L} = -\log P(w_i \mid w_{i-C+1:i-1}) \tag{1}$$

Cross-lingual objective: regularization based on word alignments

$$J_{L_1 \rightarrow L_2} = \sum_{i=1}^{|V|^s} \frac{1}{2} \mathbf{x}_i^{s\top} (\mathbf{A}^{s \rightarrow t} \otimes \mathbf{I}) \mathbf{x}_i^s \tag{2}$$

$J_{L_2 \rightarrow L_1}$ computed in an analogous fashion

$$\Omega_{\text{WA}} = J_{L_1 \rightarrow L_2} + J_{L_2 \rightarrow L_1}$$

Again, standard formulation:

$$J = \mathcal{L}^1_{\text{NLM}} + \mathcal{L}^2_{\text{NLM}} + \Omega_{\text{WA}}$$

Online Monolingual and Cross-Lingual Training: Bilingual to Multilingual

Trivial extension to multilingual training

Combining N monolingual objectives with $\binom{N}{2}$ cross-lingual objectives

$$J = \mathcal{L}^1_{\text{NLM}} + \mathcal{L}^2_{\text{NLM}} + \mathcal{L}^3_{\text{NLM}} + \Omega_{\text{WA},L_1,L_2} + \Omega_{\text{WA},L_1,L_3} + \Omega_{\text{WA},L_2,L_3}$$

In practice:

Using full NLMs makes the models inefficient: using monolingual SGNS objectives instead

Online Monolingual and Cross-Lingual Training: BOW SGNS with Word-Level Information

Bilingual Skip-Gram (BiSkip) proposed by Luong et al., NAACL 2015

Very similar to the original mapping approach, but formulated as online monolingual and cross-lingual training on pseudo-bilingual corpora

$$J = \mathcal{L}^1_{\text{SGNS}} + \mathcal{L}^2_{\text{SGNS}} + \Omega_{L_1,L_2}$$

$$\Omega_{L_1,L_2} = SGNS_{L_1 \to L_2} + SGNS_{L_2 \to L_1}$$

Online Monolingual and Cross-Lingual Training: Bilingual to Multilingual

Multilingual Skip-Gram (MultiSkip) by Ammar et al., NAACL 2016

Simple extension (again) $\rightarrow$ summing up bilingual objectives for all available parallel corpora

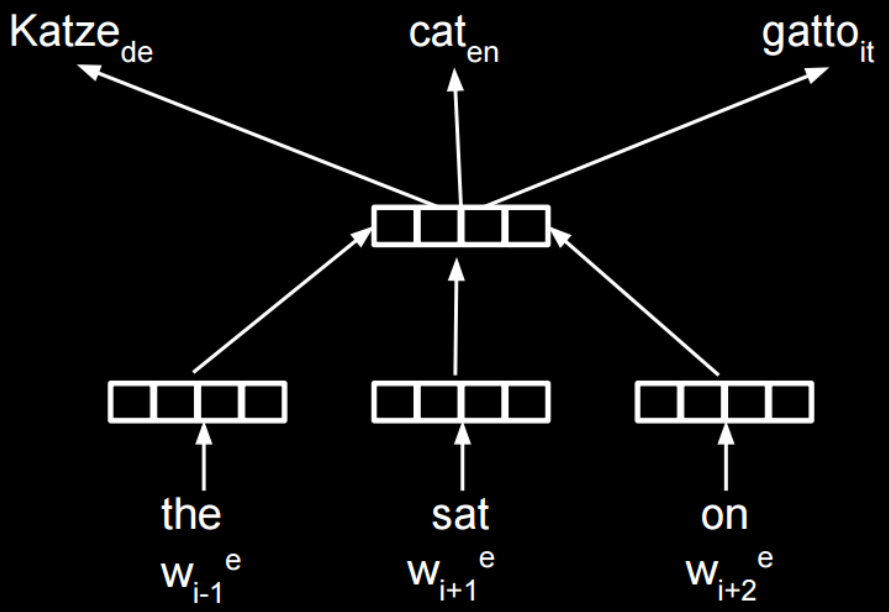
$$J = \mathcal{L}^1_{\text{SGNS}} + \mathcal{L}^2_{\text{SGNS}} + \mathcal{L}^3_{\text{SGNS}} + \Omega_{L_1,L_2} + \Omega_{L_2,L_3} + \Omega_{L_1,L_3}$$

Online Monolingual and Cross-Lingual Training: Bilingual to Multilingual

Using bilingual dictionaries in all language pairs or word alignments (per language pair)

1. Select the closest translation (if more are possible)

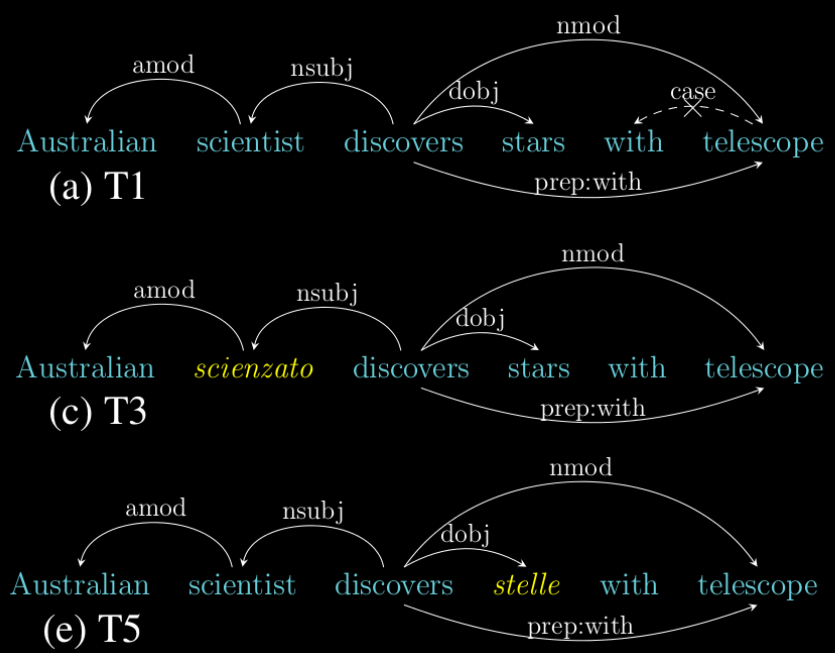
2. (See the figure below)



Online Monolingual and Cross-Lingual Training: SGNS with Dependency-Based Contexts

[Vulić, EACL 2017]

Extracting context pairs from hybrid "cross-lingual" trees



Online Monolingual and Cross-Lingual Training

Extracting monolingual and cross-lingual dependency-based contexts

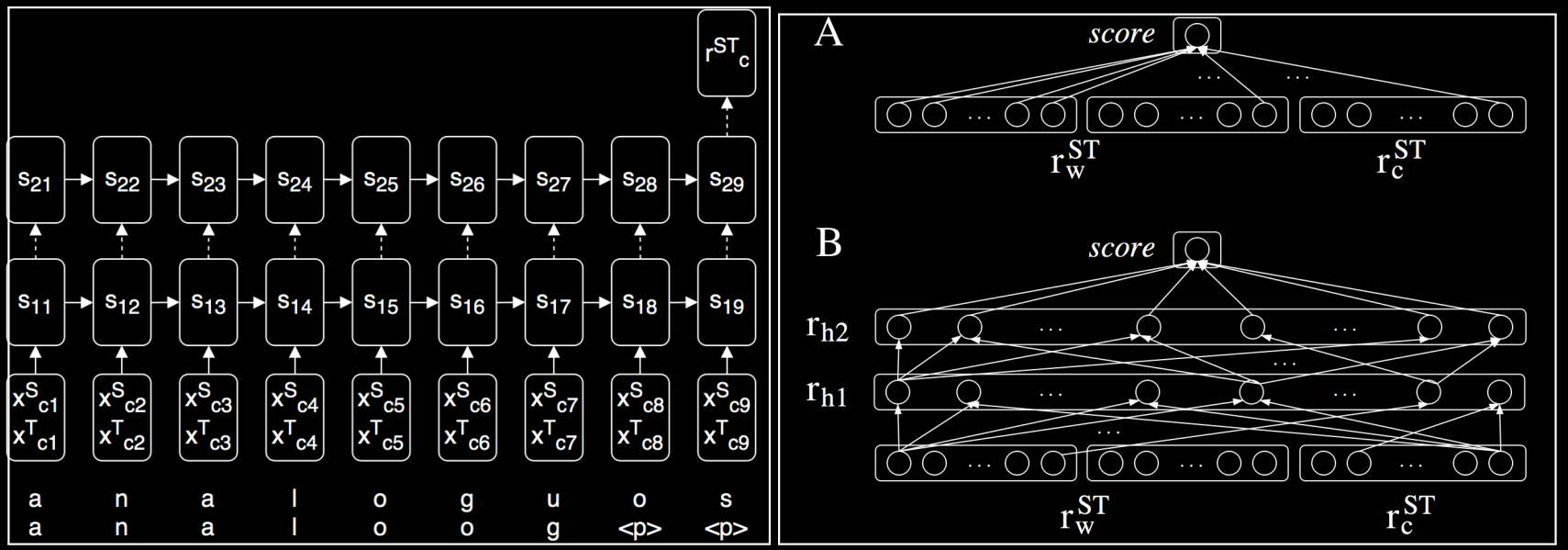
(discovers, scientist_nsubj)
(stars, discovers_dobj^{-}1)
(scienzato, australiano_amod)
(scopre, stelle_dobj)

(scientist, australiano_amod)
(australiano, scientist_amod^{-}1)
(stars, scopre_dobj^{-}1)
(discovers, scienzato_nsubj)

Training `word2vecf` SGNS on these $(word, context)$ pairs

Combining Contextual and Orthographic Info

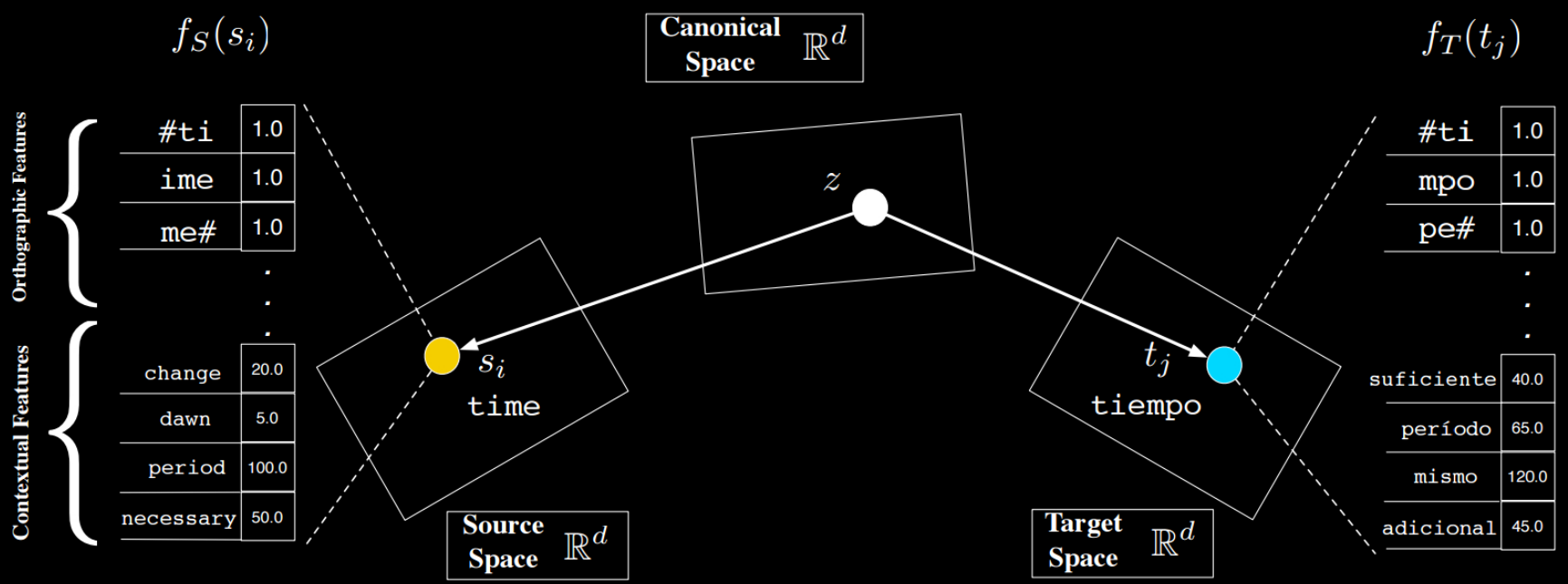
Metric learning using translation pairs with contextual and orthographic info [Heyman et al., EACL 2017]



Combining Contextual and Orthographic Info

A historical note: learning word representations relying on distributional and ortographic evidence is not a brand new idea

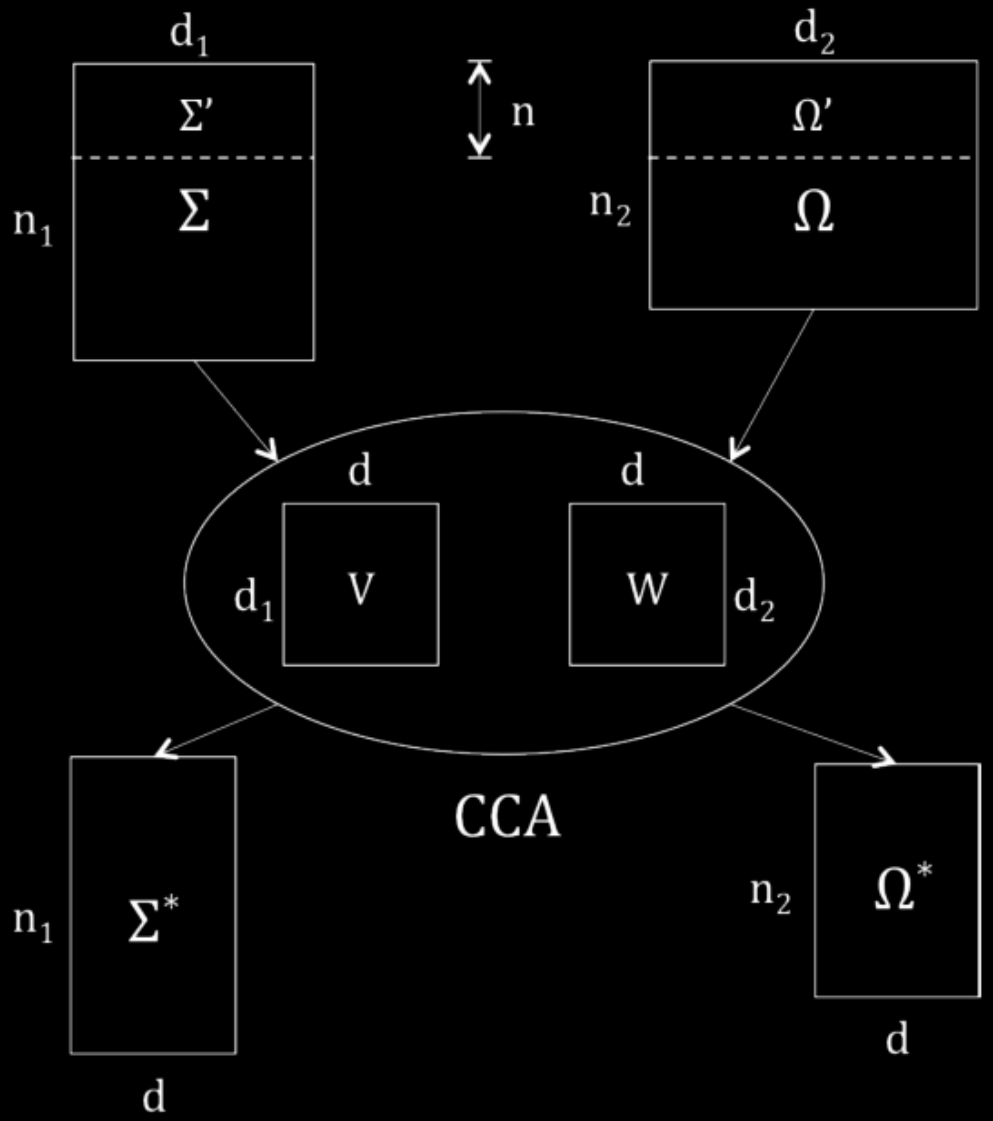
We just have more powerful algorithms and more data today



[Haghighi et al., ACL 2008]

CCA-Based Approaches

[Faruqui and Dyer, EACL 2014]



CCA-Based Approaches

Cross-lingual projection using Canonical Correlation Analysis

The objective may (again) be formulated as:

$$J = \mathcal{L}^1_{\text{LSA}} + \mathcal{L}^2_{\text{LSA}} + \Omega_{\text{CCA}}$$

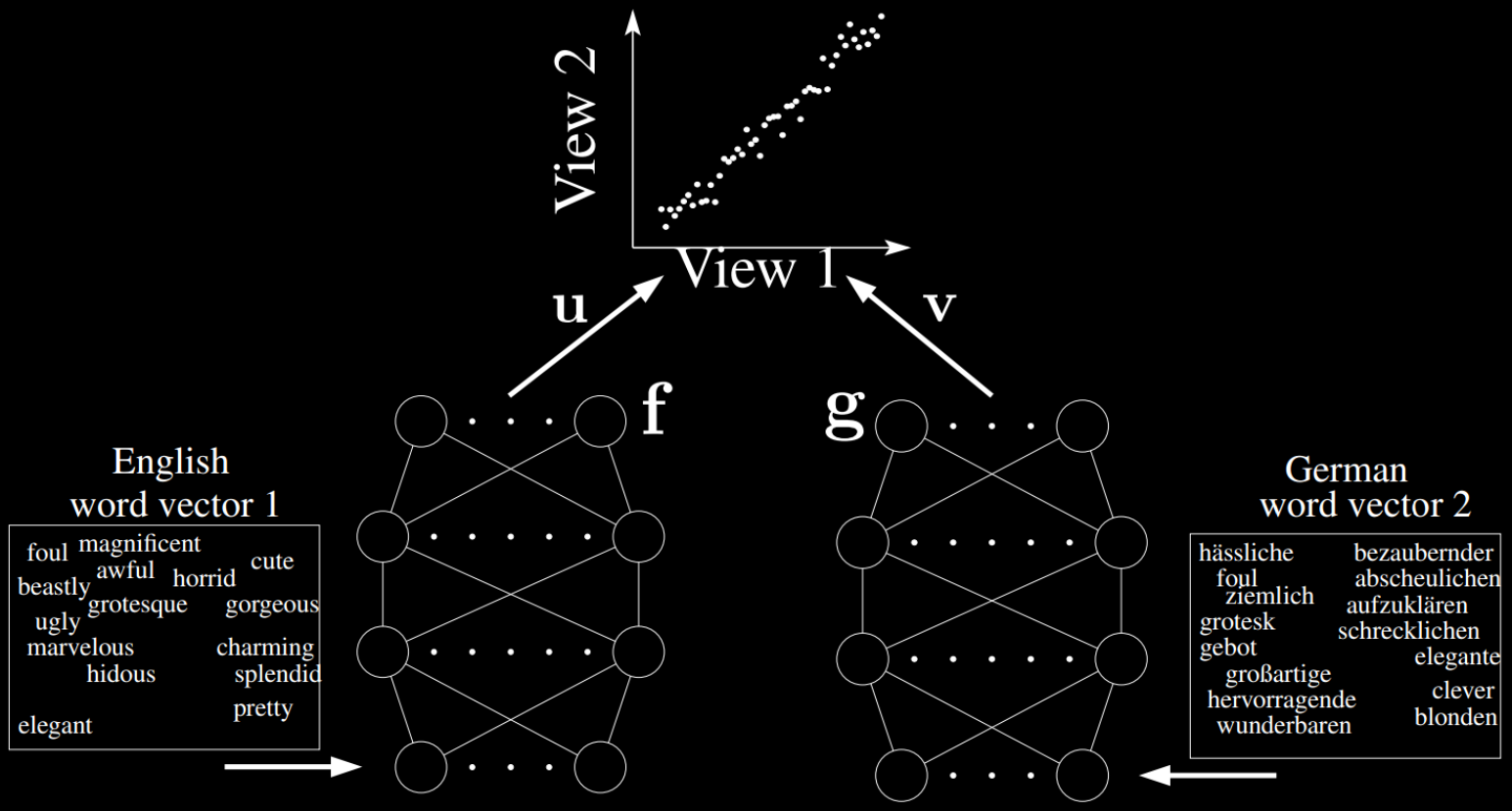
$$\text{where } \Omega_{\text{CCA}} = -\sum_{i=1}^{n} \rho(\mathbf{x}_i^s \mathbf{v}_i, \mathbf{x}_i^t \mathbf{w}_i)$$

$$\text{where } \rho(\mathbf{x}_i^{s'}, \mathbf{x}_i^{t'}) = \frac{\text{cov}(\mathbf{x}_i^{s'}, \mathbf{z}_i')}{\sqrt{\text{cov}(\mathbf{x}_i^{s'2})\text{cov}(\mathbf{x}_i^{t'2})}}$$

$$\text{where } \mathbf{x}_i^{s'} = \mathbf{x}_i^s \mathbf{v}_i \quad \mathbf{x}_i^{t'} = \mathbf{x}_i^t \mathbf{w}_i$$

where v_i and w_i are two projection vectors

CCA-Based Approaches: Extensions



Extension of the basic approach: replacing CCA with Deep CCA

[Lu et al., NAACL 2015]

CCA-Based Approaches: Extensions

CCA-based approaches with pivoting: other language or visual modality

[Rajendran et al., NAACL 2016; Gella et al., EMNLP 2017]

Multi-view and generalized CCA: bilingual $\rightarrow$ multilingual

[Rastogi et al., NAACL 2015, Ammar et al., arXiv 2016]

Cross-Lingual Word Representations via Semantic Specialisation

Combining resource-based and distributional information within a semantic specialisation framework

State-of-the-art: Paragram and Attract-Repel
[Wieting et al., TACL 2015; Mrkšić et al., TACL 2017]

If S is the set of synonymous word pairs, the procedure iterates over mini-batches of such constraints $\mathcal{B}_S$, optimising the following cost function:

$$\begin{aligned} S(\mathcal{B}_S) = \sum_{(x_l, x_r) \in \mathcal{B}_S} & \left(ReLU\left(\delta_{sim} + \mathbf{x}_l\mathbf{t}_l - \mathbf{x}_l\mathbf{x}_r\right)\right. \\ + & \left. ReLU\left(\delta_{sim} + \mathbf{x}_r\mathbf{t}_r - \mathbf{x}_l\mathbf{x}_r\right)\right) \end{aligned}$$

where δ_{sim} is the similarity margin and $\mathbf{t}_l$ and $\mathbf{t}_r$ are **negative examples** for the given word pair (x_l, x_r).

CLWEs via Semantic Specialisation

Cross-lingual supervision: again word-level "linguistic constraints" (e.g., synonyms)

Constraint	Relation	Source
(en_sweet, it_dolce)	SYN	BabelNet
(en_work, fr_travail)	SYN	BabelNet
(en_school, de_schule)	SYN	BabelNet
(fr_montagne, de_gebirge)	SYN	BabelNet
(sh_gradonačelnik, en_mayor)	SYN	BabelNet
(nl_vrouw, it_donna)	SYN	BabelNet
(en_sour, it_dolce)	ANT	BabelNet
(en_asleep, fr_éveillé)	ANT	BabelNet
(en_cheap, de_teuer)	ANT	BabelNet
(de_langsam, es_rápido)	ANT	BabelNet
(sh_obeshrabriti, en_encourage)	ANT	BabelNet
(fr_jour, nl_nacht)	ANT	BabelNet

CLWEs via Semantic Specialisation

The **Attract** term

Negative Examples for each Synonymy Pair

For each synonymy pair $(\mathbf{x}_l, \mathbf{x}_r)$, the negative example pair $(\mathbf{t}_l, \mathbf{t}_r)$ is chosen from the remaining in-batch vectors so that $\mathbf{t}_l$ is the one closest (cosine similarity) to $\mathbf{x}_l$ and $\mathbf{t}_r$ is closest to $\mathbf{x}_r$.

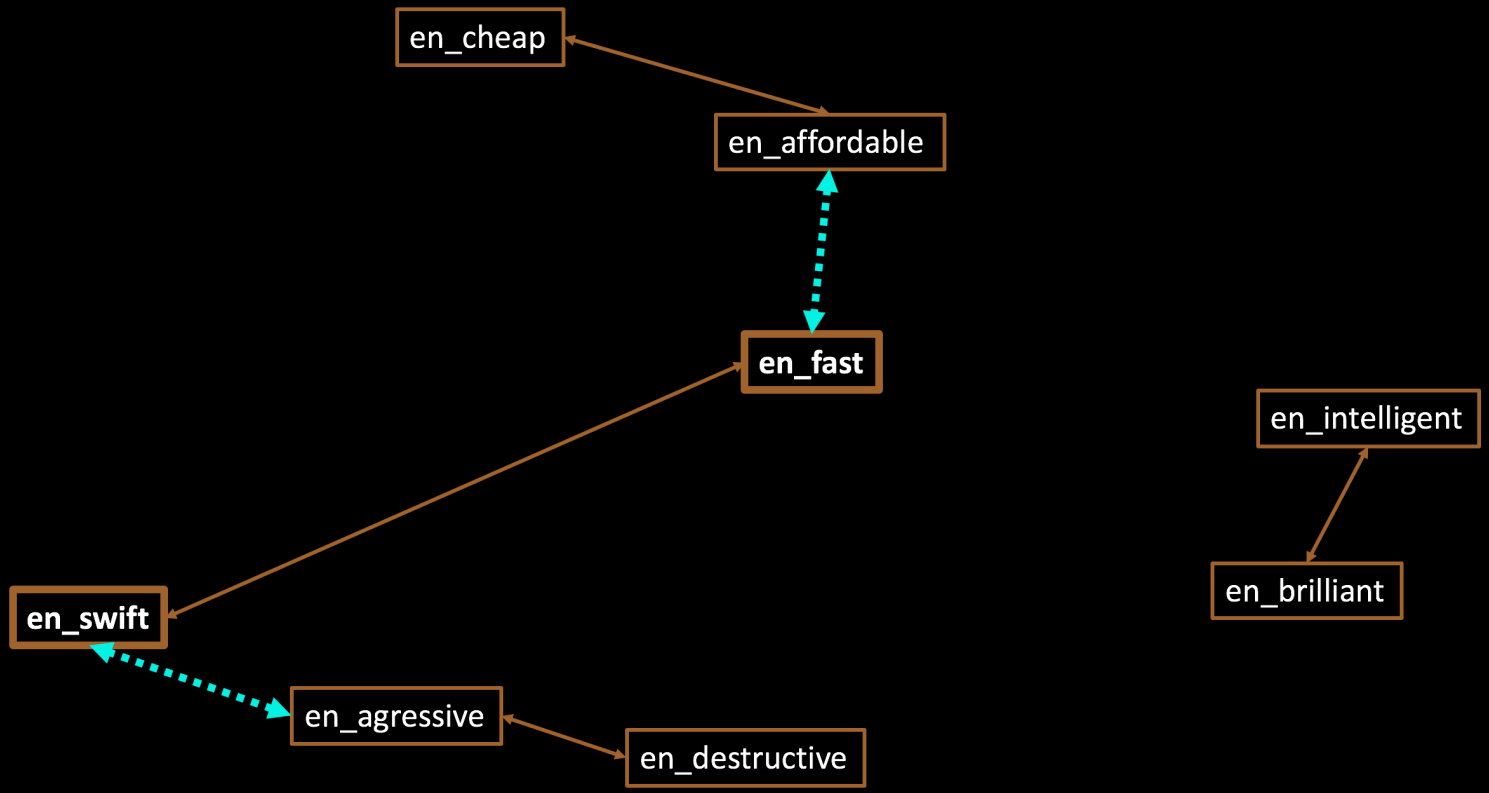
$$S(\mathcal{B}_S) = \sum_{(x_l, x_r) \in \mathcal{B}_S} \left(ReLU\left(\delta_{sim} + \mathbf{x}_l\mathbf{t}_l - \mathbf{x}_l\mathbf{x}_r\right)\right. + \left. ReLU\left(\delta_{sim} + \mathbf{x}_r\mathbf{t}_r - \mathbf{x}_l\mathbf{x}_r\right)\right)$$

The two negative examples are used to force synonymous pairs to be closer to each other than to their respective negative examples (i.e. to any of the remaining words in the current mini-batch).

CLWEs via Semantic Specialisation

Negative Examples for each Synonymy Pair

For each synonymy pair $(\mathbf{x}_l, \mathbf{x}_r)$, the negative example pair $(\mathbf{t}_l, \mathbf{t}_r)$ is chosen from the remaining in-batch vectors so that $\mathbf{t}_l$ is the one closest (cosine similarity) to $\mathbf{x}_l$ and $\mathbf{t}_r$ is closest to $\mathbf{x}_r$.



CLWEs via Semantic Specialisation

The Regularization term

The second term tries to retain the beneficial semantic content embedded in the initial vector space.

L2 Regularization

$$R(V) = \sum_{x_i \in V} \lambda_{reg} \left\| \widehat{\mathbf{x}}_i - \mathbf{x}_i \right\|_2$$

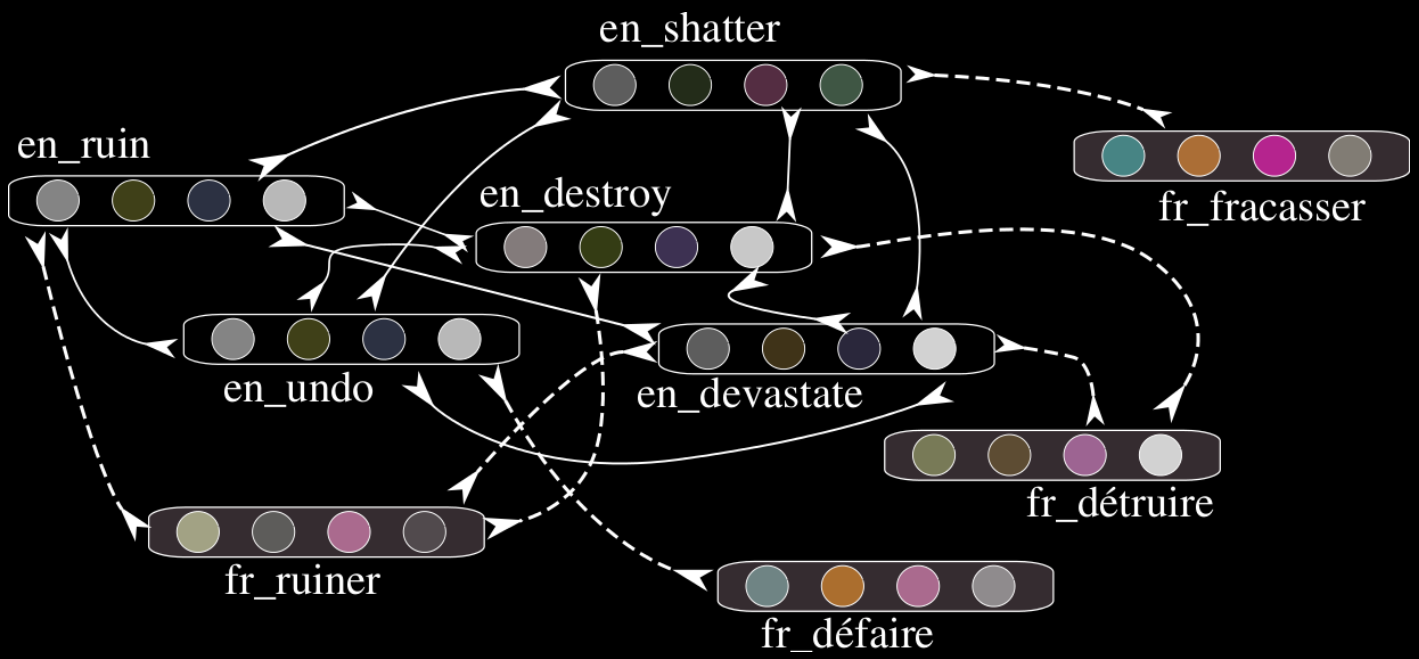
This term preserves semantic relations learned by distributional models that do not contradict the injected similarity constraints.

CLWEs via Semantic Specialisation

$$J = \mathcal{L}^1_{\text{Mono}} + \mathcal{L}^2_{\text{Mono}} + \Omega_{\text{Spec}}$$

Cross-lingual connections may be used for direct transfer of semantic resources (e.g., VerbNet, FrameNet)

[Vulić et al., EMNLP 2017]



CLWEs via Semantic Specialisation

Trivially Multilingual

Cross-lingual word links between all languages are used to bring the word vector spaces of various languages into a single unified vector space.

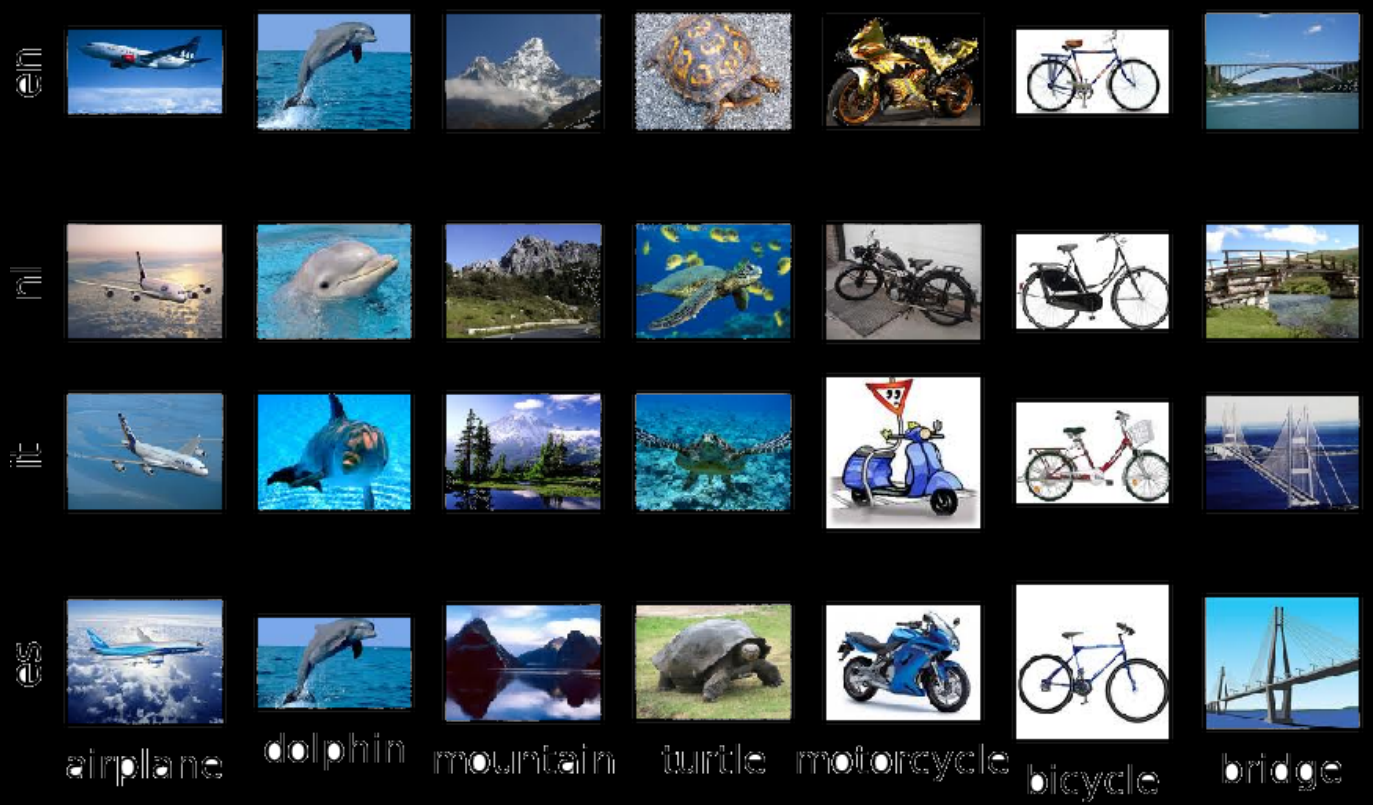
[Mrkšić et al., TACL 2017]

	en_carpet			**en_woman**	
Slavic+EN	**Germanic**	**Romance+EN**	**Slavic + EN**	**Germanic**	**Romance+EN**
en_rug	de_teppichboden	en_rug	ru_женщина	de_frauen	fr_femme
bg_килим	nl_tapijten	it_moquette	bg_жените	sv_kvinnliga	en_womanish
ru_ковролин	en_rug	it_tappeti	sh_žena	sv_kvinna	es_mujer
bg_килими	de_teppich	pt_tapete	en_womanish	sv_kvinnor	pt_mulher
pl_dywany	en_carpeting	es_moqueta	bg_жена	de_weib	es_fémina
bg_мокет	de_teppiche	it_tappetino	pl_kobieta	en_womanish	en_womens
pl_dywanów	sv_mattor	en_carpeting	sh_treba	sv_kvinno	pt_feminina
sh_tepih	sv_matta	pt_carpete	bg_жени	de_frauenzimmer	pt_femininas
pl_wykładziny	en_carpets	pt_tapetes	en_womens	sv_honkön	es_femina
ru_ковер	nl_tapijt	fr_moquette	pl_kobiet	sv_kvinnan	fr_femelle
ru_коврик	nl_kleedje	en_carpets	sh_žene	nl_vrouw	pt_fêmea
sh_ćilim	nl_vloerbedekking	es_alfombra	pl_niewiasta	de_madam	fr_femmes
en_carpeting	de_brücke	es_alfombras	sh_žensko	sv_kvinnligt	it_donne
pl_dywan	de_matta	fr_tapis	sh_ženke	sv_gumman	es_mujeres
ru_ковров	nl_matta	pt_tapeçaria	pl_samica	sv_female	pt_fêmeas
en_carpets	en_mat	it_zerbino	ru_самка	sv_gumma	es_hembras

Comparable Word-Level Signal: Images

Bilingual space need not be linguistic in nature

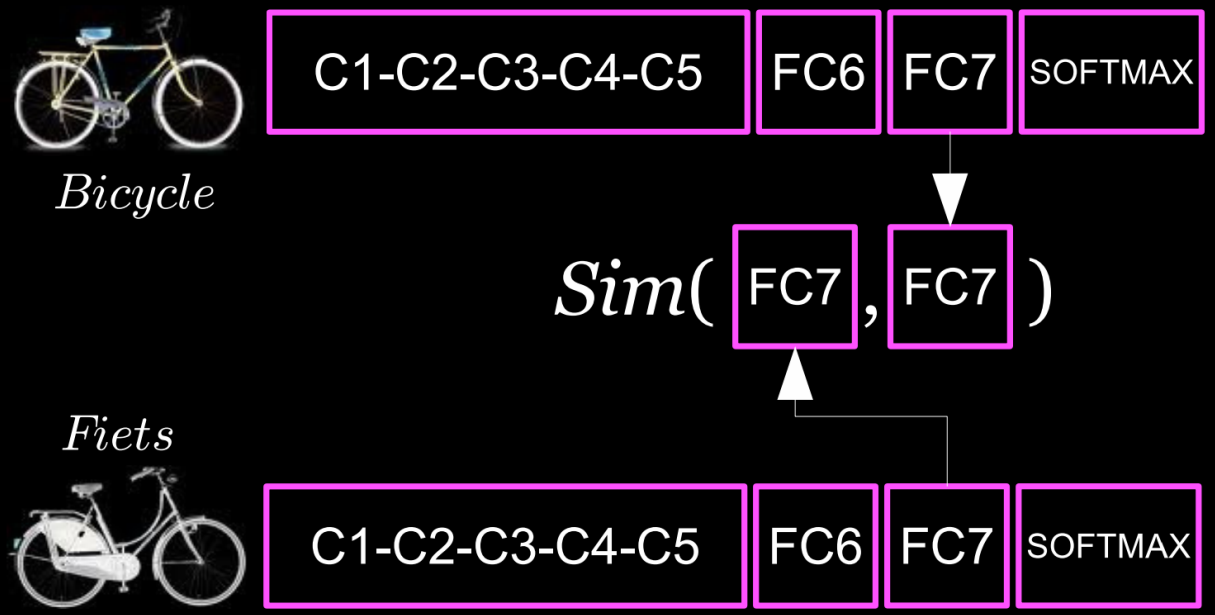
Visual and multi-modal cross-lingual word representations



[Kiela et al., EMNLP 2015; Vulić et al., ACL 2016]

Comparable Word-Level Signal: Images

Visual features: transferred CNN features from the ImageNet task



Recently: Multilingual image captioning data: multi-modal representations beyond word-level

[Specia et al., WMT 2016; Elliott et al., VL 2016; Gella et al., EMNLP 2017]

Parts II-IV

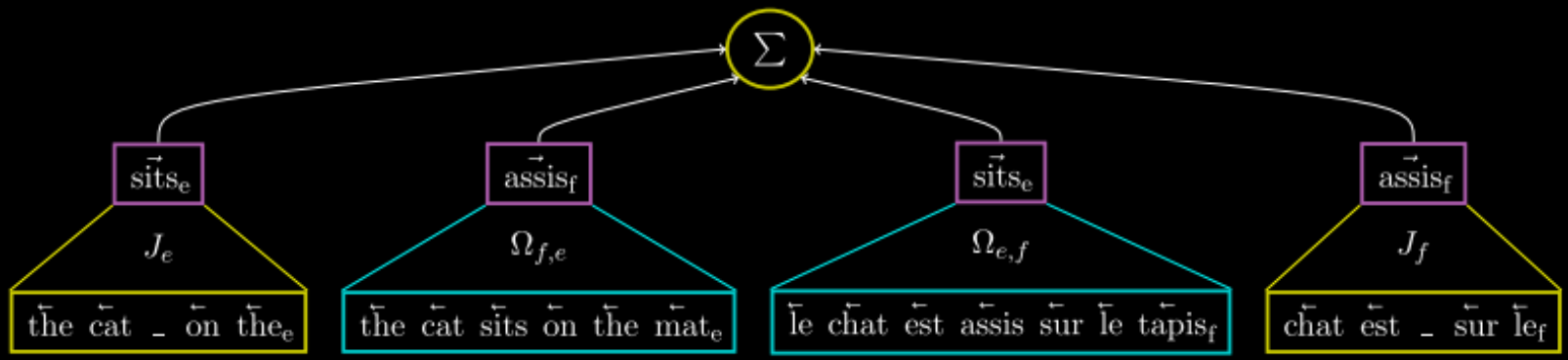
Anders Søgaard
http://cst.dk/anders/

A cold open...

Sentence-Level (Alignment) Signal

Trans-gram model: very similar to BiSkip, but it relies on full sentential contexts

[Coulmance et al., EMNLP 2015]



$$J = \mathcal{L}^1_{\text{SGNS}} + \mathcal{L}^2_{\text{SGNS}} + \Omega_{L_1,L_2}$$

$$\Omega_{L_1,L_2} = SGNS_{L_1 \to L_2} + SGNS_{L_2 \to L_1}$$

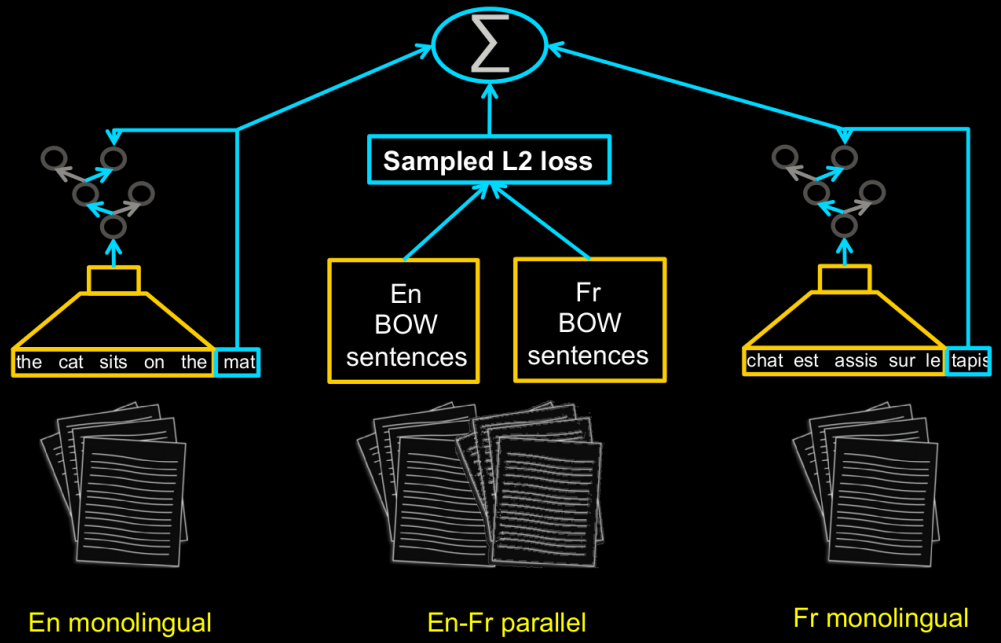
Deja vu, right?

Sentence-Level (Alignment) Signal

BilBOWA model: very similar to online models, but using sentence-level cross-lingual information

[Gouws et al., ICML 2017]

Deja vu, right?



$$J = \mathcal{L}^1_{\text{SGNS}} + \mathcal{L}^2_{\text{SGNS}} + \Omega_{BilBOWA}$$

$$\Omega_{BilBOWA} = \| \frac{1}{m} \sum_{w^s_i \in sent^s}^{m} \mathbf{x}^s_i - \frac{1}{n} \sum_{w^t_j \in sent^t}^{n} \mathbf{x}^t_j \|^2$$

Outline of my part

- Recap
- Part II: Learning from sentences
- Part III: Learning from documents
- Part IV: Learning from extra-linguistic context

Recap, so far..

Cross-lingual word embeddings

- Learned from aligned words, sentences, or dictionaries
- Alignments can either be of parallel or comparable units
- Most, if not all, approaches minimize a sum of *n* losses (for *n* languages) + a regularizer

Cross-lingual word embeddings

- Many approaches optimize for the same objectives:
  - If you sample contexts through word alignments (Duong et al., 2015), you add a penalty $\Omega_{\sum |w_t - w_s|^2}$
  - If you create a mixed corpus using a translation function (Gouws and Søgaard, 2015), you add penalty $\Omega_{\sum |w_t - w_s|^2}$
  - If you hard code your translation (Xiao and Guo, 2014), you add a strong penalty $\Omega_{\sum |w_t - w_s|^2}$

Cross-lingual word embeddings

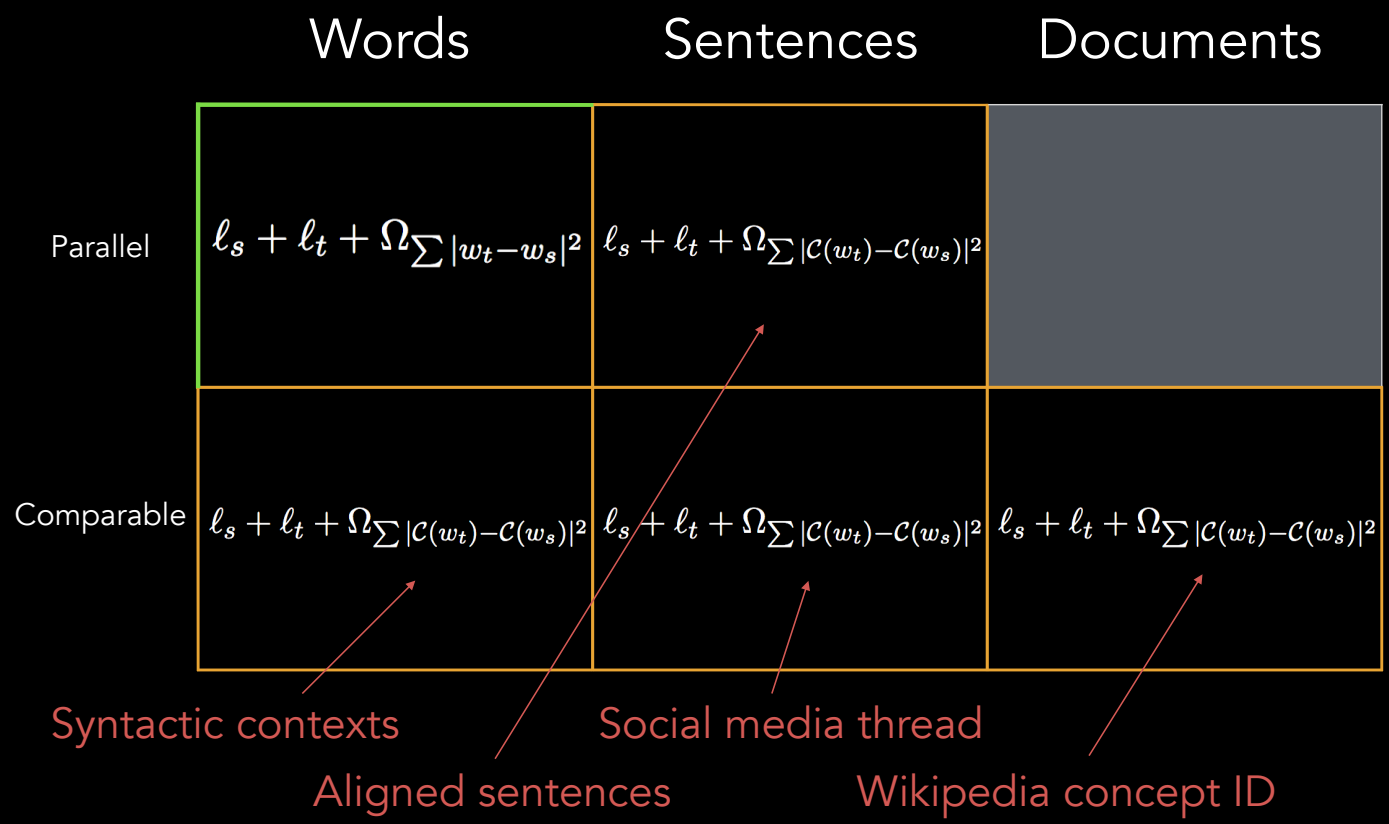
	Words	Sentences	Documents
Parallel	$\ell_s + \ell_t + \Omega_{\sum \lvert w_t - w_s \rvert^2}$	$\ell_s + \ell_t + \Omega_{\sum \lvert \mathcal{C}(w_t) - \mathcal{C}(w_s) \rvert^2}$	
Comparable	$\ell_s + \ell_t + \Omega_{\sum \lvert \mathcal{C}(w_t) - \mathcal{C}(w_s) \rvert^2}$	$\ell_s + \ell_t + \Omega_{\sum \lvert \mathcal{C}(w_t) - \mathcal{C}(w_s) \rvert^2}$	$\ell_s + \ell_t + \Omega_{\sum \lvert \mathcal{C}(w_t) - \mathcal{C}(w_s) \rvert^2}$

Shout-out to Upadhyay et al. (2016)

Cross-lingual word embeddings

	Words	Sentences	Documents
Parallel	$\ell_s + \ell_t + \Omega_{\sum \lvert w_t - w_s \rvert^2}$	$\ell_s + \ell_t + \Omega_{\sum \lvert \mathcal{C}(w_t) - \mathcal{C}(w_s) \rvert^2}$	
Comparable	$\ell_s + \ell_t + \Omega_{\sum \lvert \mathcal{C}(w_t) - \mathcal{C}(w_s) \rvert^2}$	$\ell_s + \ell_t + \Omega_{\sum \lvert \mathcal{C}(w_t) - \mathcal{C}(w_s) \rvert^2}$	$\ell_s + \ell_t + \Omega_{\sum \lvert \mathcal{C}(w_t) - \mathcal{C}(w_s) \rvert^2}$

Cross-lingual word embeddings



Real-world assumptions

á-jŭà

á-jŭà	**wŭá**	**á-vĭò**
CLASS-dog	kill	CLASS-goat

‘the dog killed the goat’

4+ million speakers

65-70% Internet

500b USD GDP

150 universities

á-jǔà	**wǔá**	**á-vǐò**
CLASS-dog	**kill**	**CLASS-goat**

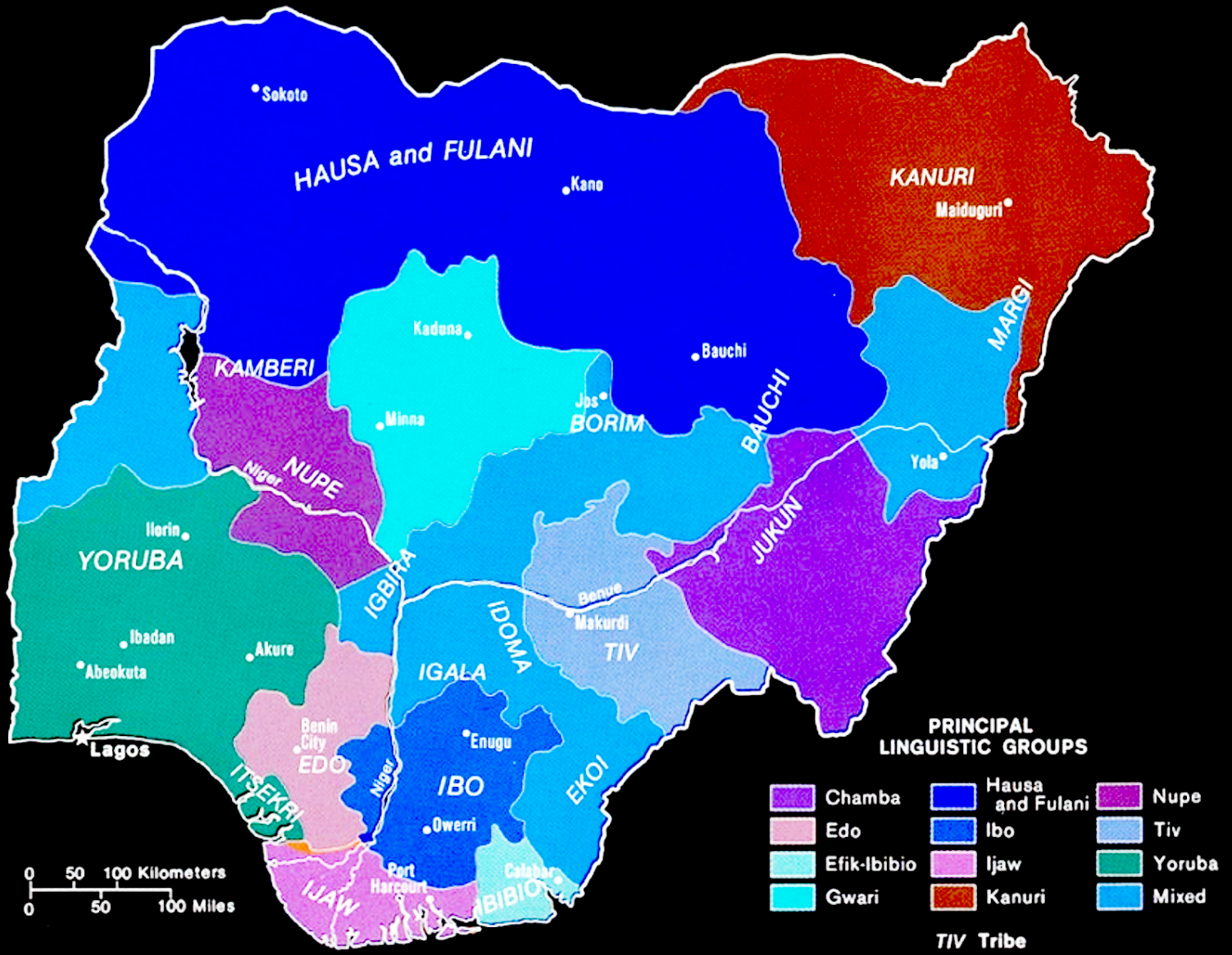
‘the dog killed the goat’

Google Translate

Wiktionary

Europarl

POS taggers


Sokoto
HAUSA and FULANI
Kano
KANURI
Maiduguri
Kaduna
Bauchi
KAMBERI
MARGI
Jos
BORIM
BAUCHI
Minna
Yola
Niger
NUPE
Ilorin
YORUBA
JUKUN
IGBIRA
IDOMA
Benue
Makurdi
TIV
Ibadan
Akure
Abeokuta
IGALA
Benin City
EDO
Enugu
Lagos
ITSEKRI
Niger
IBO
EKOI
Owerri
Port Harcourt
Calabar
IJAW
IBIBIO
0 50 100 Kilometers
0 50 100 Miles
PRINCIPAL LINGUISTIC GROUPS
Chamba
Edo
Efik-Ibibio
Gwari
Hausa and Fulani
Ibo
Ijaw
Kanuri
Nupe
Tiv
Yoruba
Mixed
TIV Tribe

Supervision

- Online parallel text: the Bible, the Watchtower, a few declarations
- OCR printed dictionaries?
- Comparable text through reference points: Wiktionary conceptIDs (*not* for Tiv, though), real-life events, hashtags, URLs, linguistic annotation, images, …

Cross-lingual word embeddings

	Words	Sentences	Documents
Parallel	$\ell_s + \ell_t + \Omega_{\sum \lvert w_t - w_s \rvert^2}$	Bible	
Comparable	#hashtags, images (search, co-clicks), time stamps	Captions	Wikipedia

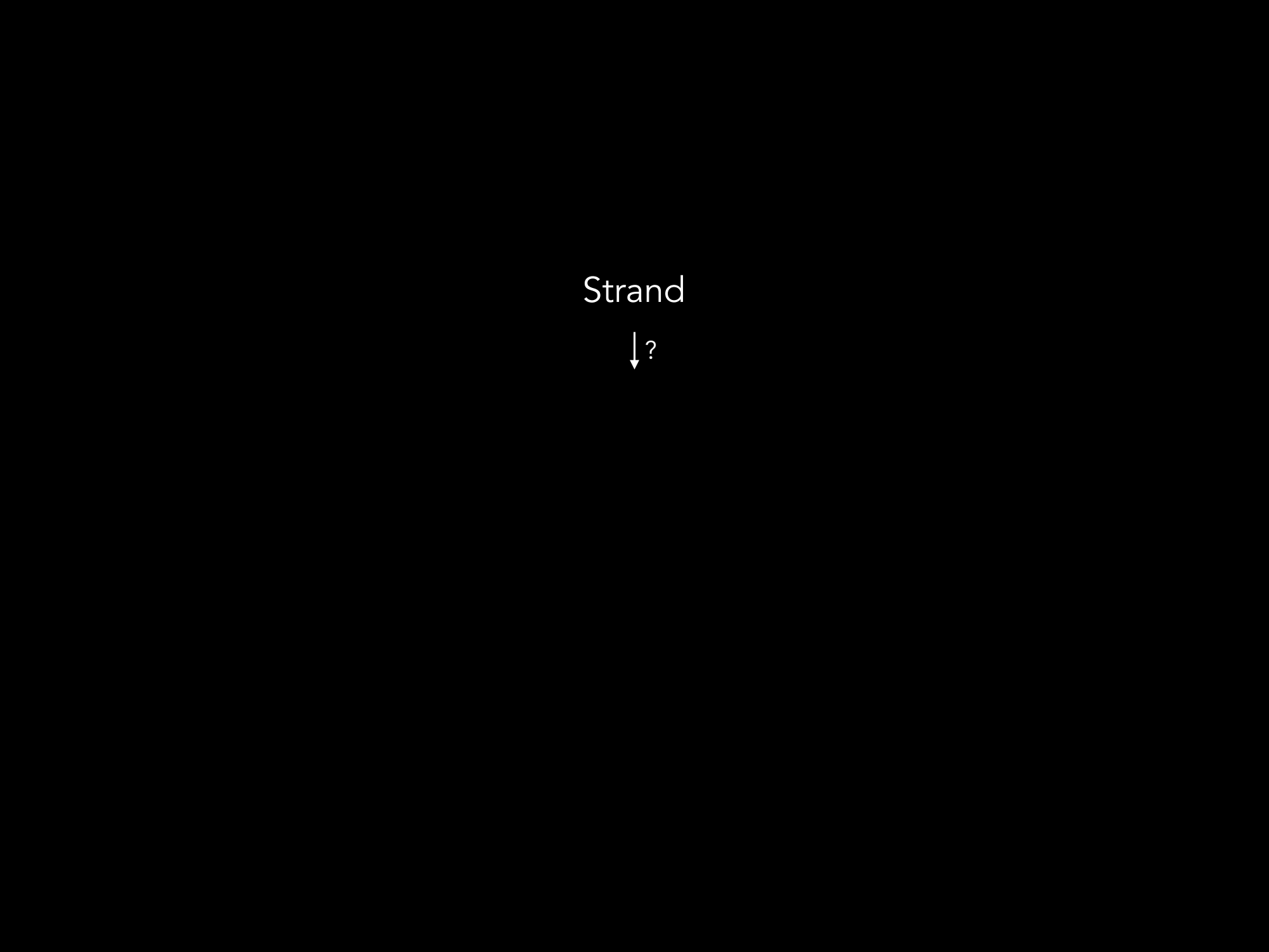
Parallel sentences

	Words	Sentences	Documents
Parallel	$\ell_s + \ell_t + \Omega_{\sum \lvert w_t - w_s \rvert^2}$	Bible	
Comparable	#hashtags, images (search, co-clicks), time stamps	Captions	Wikipedia

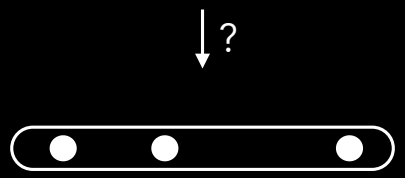
Feature spaces

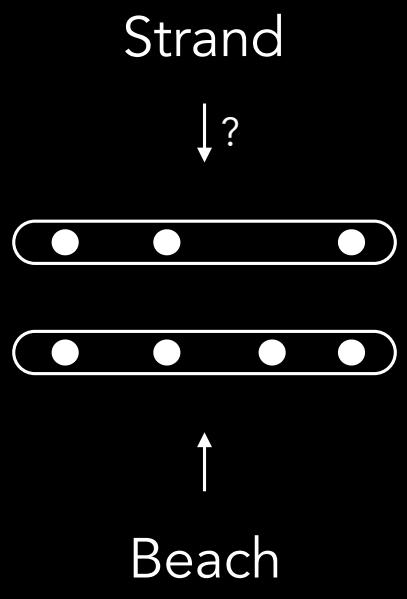
Sentence alignments

Is Valencia on the coast?	Ist Valencia an der Küste?
So you will go swimming?	Also wirst du schwimmen gehen?
If you meet a shark, eat it.	Wenn Sie einen Hai treffen, essen Sie es.
Don’t step on corals.	Treten Sie nicht auf Korallen.
Name a Spanish hip hop band.	Nennen Sie eine spanische Hip-Hop-Band.


Strand
?

Strand



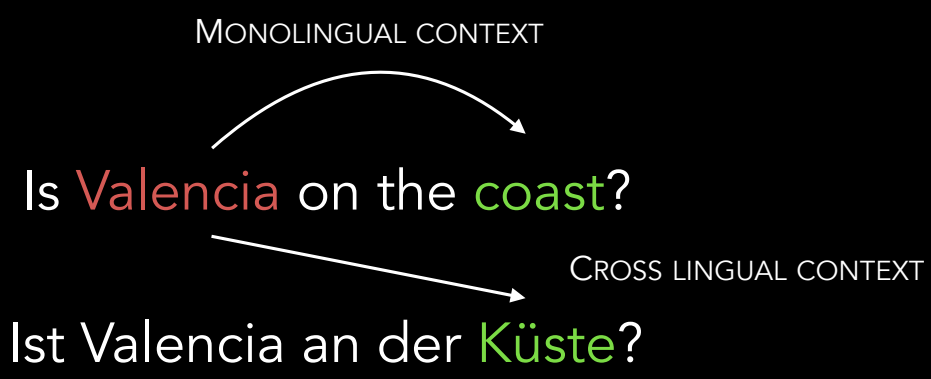

Strand
?
Beach

Is Valencia on the coast?

Ist Valencia an der Küste?

Is Valencia on the coast?

Ist Valencia an der Küste?


MONOLINGUAL CONTEXT
Is Valencia on the coast?
CROSS LINGUAL CONTEXT
Ist Valencia an der Küste?


MONOLINGUAL CONTEXT
Is Valencia on the coast?
SENTENCE ID
CROSS LINGUAL CONTEXT
Ist Valencia an der Küste?

Four algorithms

	Monolingual	Cross-lingual	Sentence ID
BilBOWA (Gouws et al.)	Yes	Yes	
BWE (Vulic and Moens)	Yes	Yes	
BA (Chandar et al.)	Yes	Yes	Yes
Inverted (Søgaard et al.)			Yes

BilBOWA (Gouws et al.)

$$\ell_s + \ell_t + \Omega_{\sum |\mathcal{C}(w_t) - \mathcal{C}(w_s)|^2}$$

- CBOW source language losses
- Equivalence class is weighted by mean of
  - P(w in S | v in S)
  - P(v in S | w in S)

BWE (Vulic and Moens)

$$\ell_s + \ell_t + \Omega_{\sum |\mathcal{C}(w_t) - \mathcal{C}(w_s)|^2}$$

- SNGS source language losses
- Equiv class is weighted by (weighted) mean of
  - P(w in S | v in S)
  - P(v in S | w in S)

BWE (Vulic and Moens)

$$\ell_s + \ell_t + \Omega_{\sum |\mathcal{C}(w_t) - \mathcal{C}(w_s)|^2}$$

- SNGS source language losses
- Equiv class is weighted by (weighted) mean of
  - P(w in S | v in S)
  - P(v in S | w in S)

Very similar to BilBOWA,
but performs much better
(Levy et al., 2017)

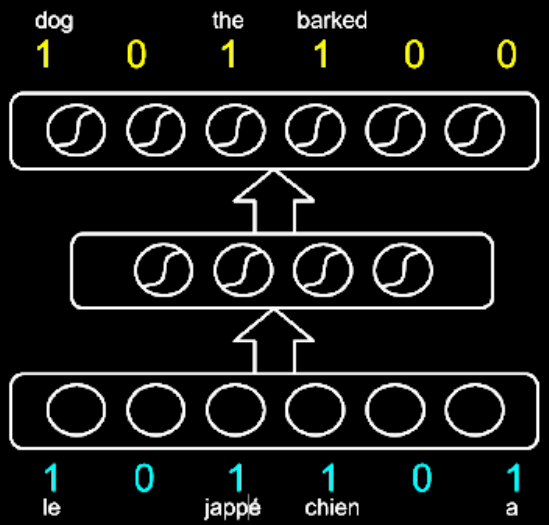
BWE (Vulic and Moens)

$$\ell_s + \ell_t + \Omega_{\sum |\mathcal{C}(w_t) - \mathcal{C}(w_s)|^2}$$

- SNGS source language losses
- Equiv class is weighted by (weighted) mean of
  - P(w in S | v in S)
  - P(v in S | w in S)

Very similar to BilBOWA, but performs much better (Levy et al., 2017)

BA (Chandar et al.)



- Superior to much of previous work (Levy et al., 2017)
- Why?

Inverted (Søgaard et al.)

A dog runs

A dog barks

A cat runs

ID	Term	Document
1	dog	1, 2
2	runs	1, 3
3	barks	2
4	cat	3

Inverted (Søgaard et al.)

Ein Hund läuft	A dog runs
Ein Hund bellt	A dog barks
Eine Katze läuft	A cat runs

ID	Term	Document
1	dog	1, 2
2	runs	1, 3
3	barks	2
4	cat	3
5	hund	1, 2
6	läuft	1, 3
7	bellt	2
8	katze	3

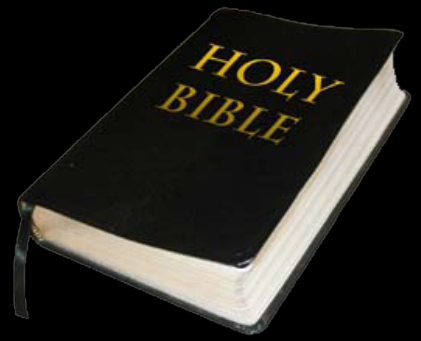
Inverted (Søgaard et al.)

SVD ←

ID	Term	Document
1	dog	1, 2
2	runs	1, 3
3	barks	2
4	cat	3
5	hund	1, 2
6	läuft	1, 3
7	bellt	2
8	katze	3

Data and tasks

SENTENCE ID = VERSE ID



SENTENCE ID = VERSE ID

Word alignment

Strand
↓?

Disappointed the venue is not by the beach

SENTENCE ID = VERSE ID

Word alignment

Strand
↓?

Disappointed the venue is not by the beach

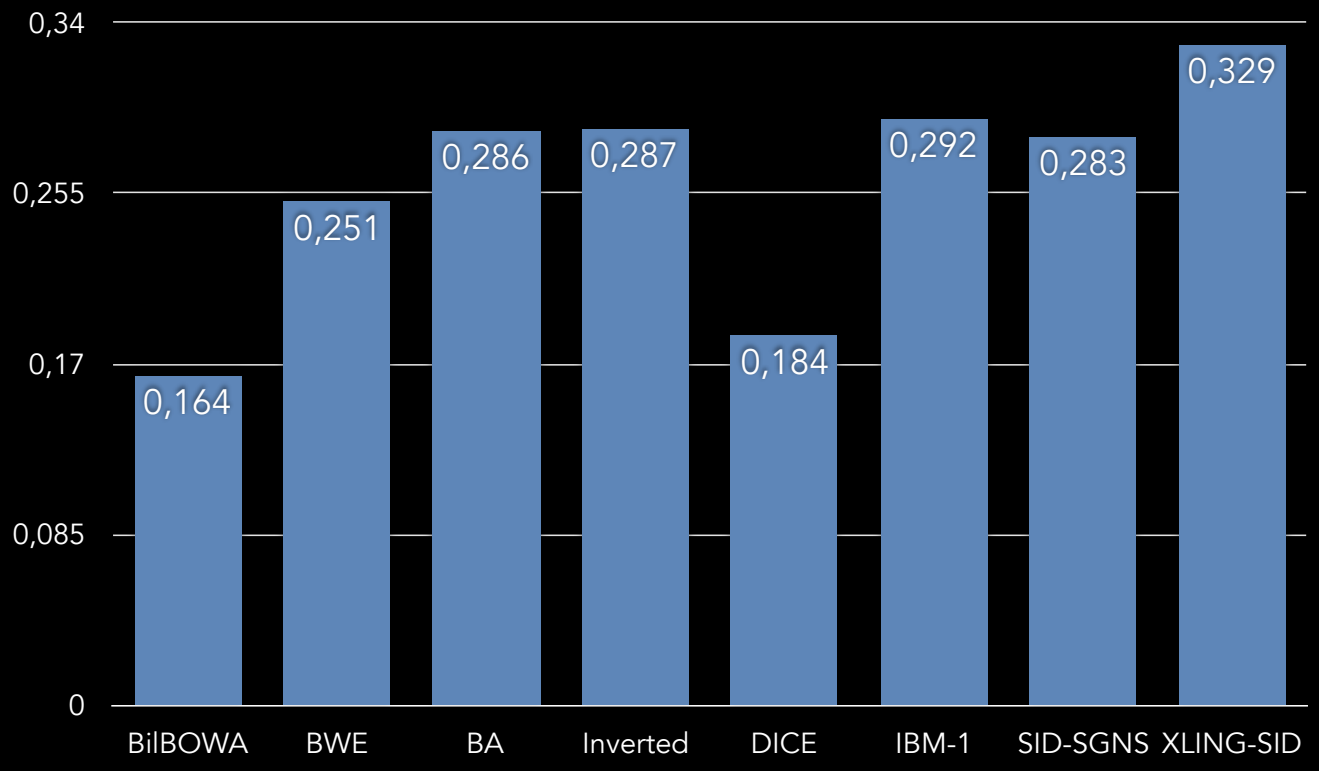
Dictionary induction

Strand
↓?

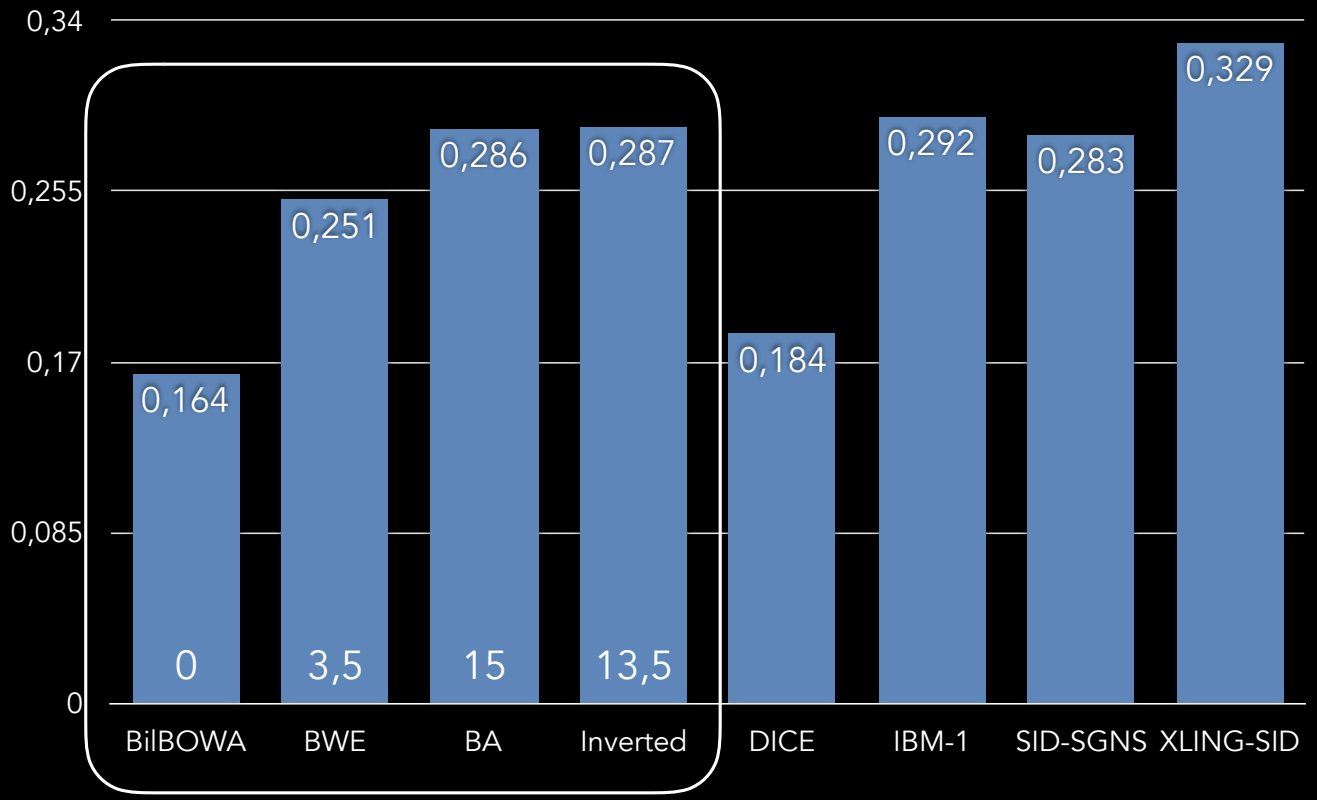
shore, Kantian, yo-yo, beach, pokemon, dried

Graca	
Hansards	
Lambert	
Mihalcea	
Holmqvist	
Cakmak	
Wiktionary	

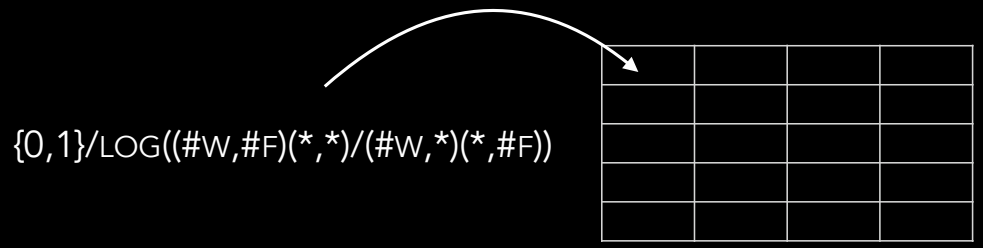

WORD ALIGNMENT
DICTIONARY INDUCTION
Graca
Hansards
Lambert
Mihalcea
Holmqvist
Cakmak
Wiktionary


0,34
0,255
0,17
0,085
0
0,164
0,251
0,286
0,287
0,184
0,292
0,283
0,329
BilBOWA
BWE
BA
Inverted
DICE
IBM-1
SID-SGNS
XLING-SID

State of the art



Generalizing over inverted indexing



Counting

Step 1

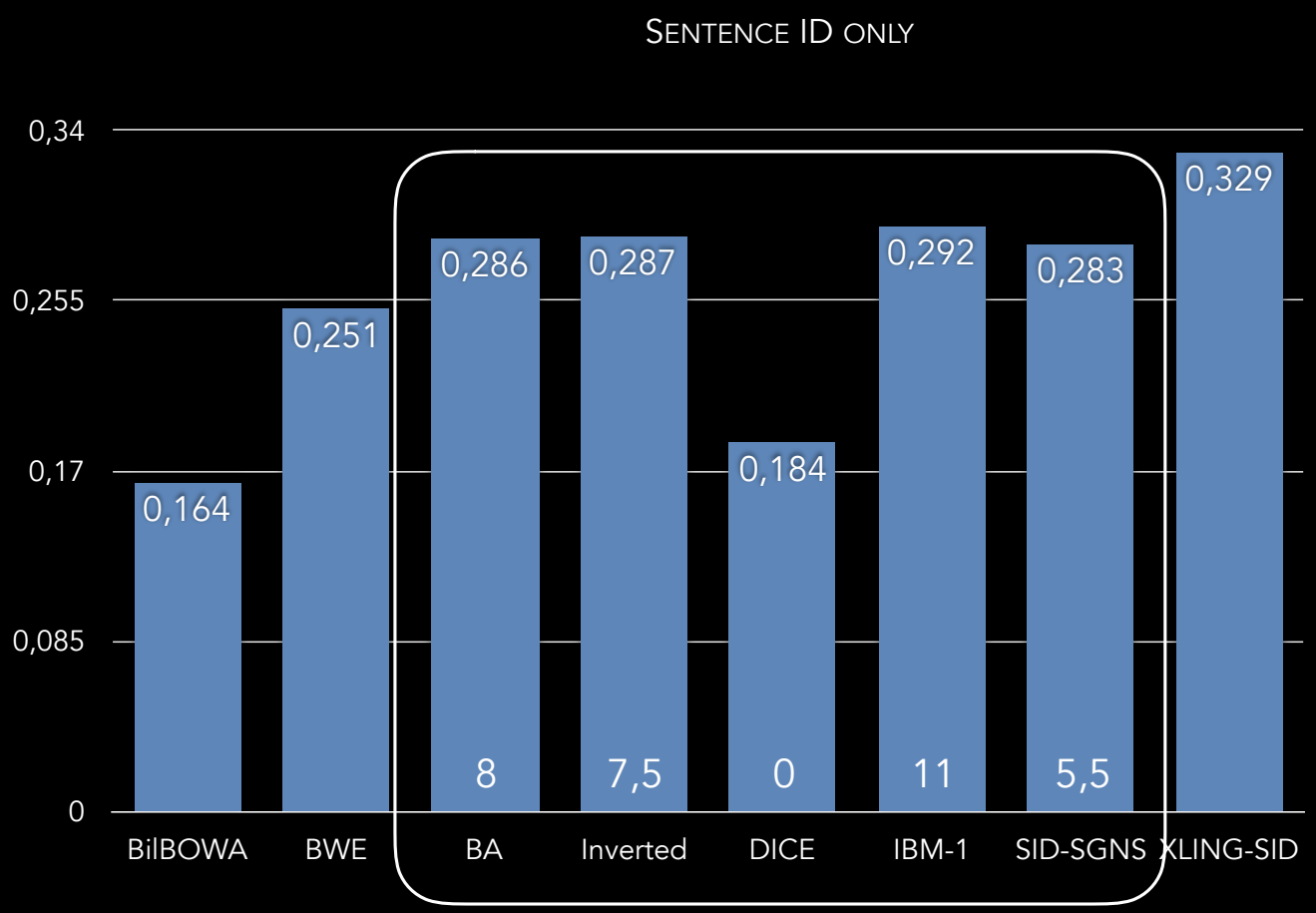
Dimensionality reduction

Step 2

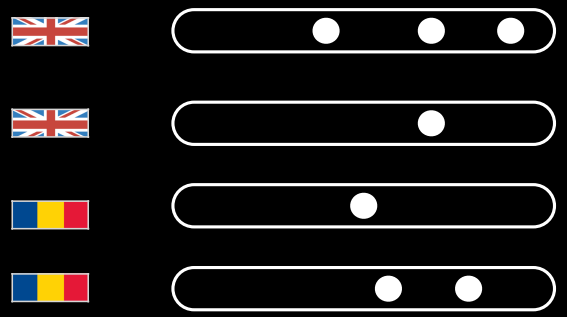
There **is** no step 3

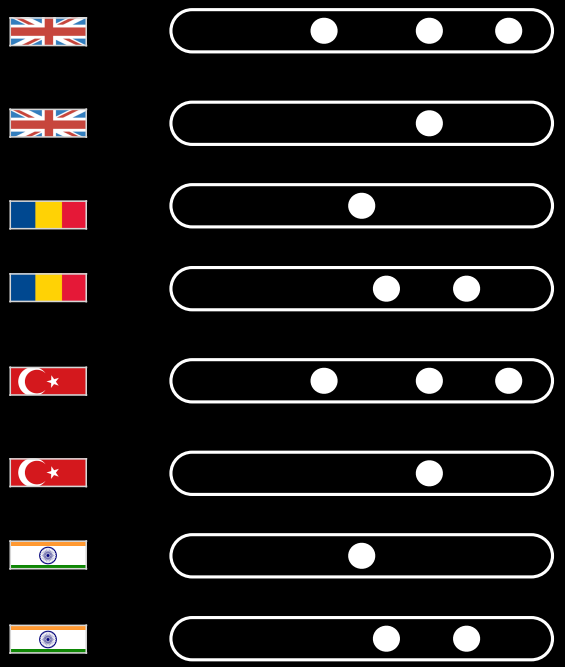
Step 3

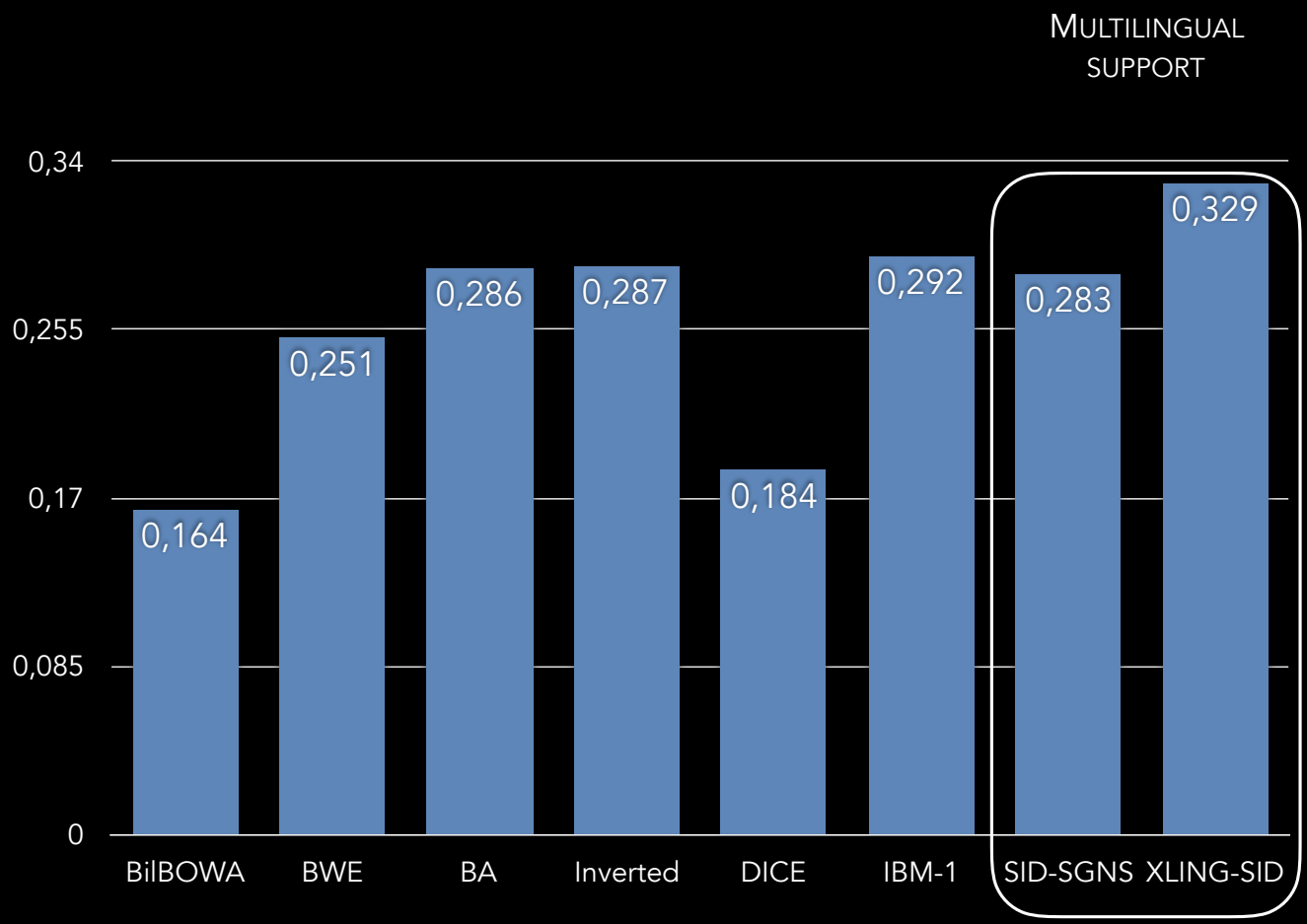
Dice	L1-normalized binary	None
Inverted	L2-normalized binary	SVD
BA	Binary	Auto encoder
This$_{0.1}$	Positive PMI	SVD
This	Binary	SGNS


SENTENCE ID ONLY
0,34
0,255
0,17
0,085
0
0,164
0,251
0,286
0,287
0,184
0,292
0,283
0,329
8
7,5
0
11
5,5
BilBOWA
BWE
BA
Inverted
DICE
IBM-1
SID-SGNS
XLING-SID

Multilingual support






MULTILINGUAL SUPPORT
0,34
0,255
0,17
0,085
0
0,164
0,251
0,286
0,287
0,184
0,292
0,283
0,329
BilBOWA
BWE
BA
Inverted
DICE
IBM-1
SID-SGNS
XLING-SID

Other work on multi-lingual support

- Ammar et al. (2016): *Massively Multilingual Word Embeddings*
- Duong et al. (2017): *Multilingual Training of Crosslingual Word Embeddings*
- Smith et al. (2017): *Offline Bilingual Word Vectors, Orthogonal Transformations and the Inverted Softmax*

Comparable sentences

$\ell_s + \ell_t + \Omega_{\sum \lvert w_t - w_s \rvert^2}$	Bible	
#hashtags, images (search, co-clicks), time stamps	Captions	Wikipedia

Comparable sentences

- Image captioning

The brown dog is running after the black dog.

Ein brauner Hund und ein schwarzer Hund.

Comparable sentences

- Image captioning
- Tweets with similar time stamps and hashtags
- ...

August 7
Diebesbanden machen NRW-Krankenhäuser unsicher [url] via @rponline Dank offener Grenzen und #Merkel s Neubürger
@JunckerEU needs to listen to Bill Gates and control European borders. #refugeecrisis #Merkel #btw17

Comparable documents

$\ell_s + \ell_t + \Omega_{\sum \lvert w_t - w_s \rvert^2}$	Bible	
#hashtags, images (search, co-clicks), time stamps	Captions	Wikipedia

Comparable documents

- Previously discussed: BWEs, Inverted Indexing
- Multilingual topic modeling
- Document classification?

Comparable words

$\ell_s + \ell_t + \Omega_{\sum \lvert w_t - w_s \rvert^2}$	Bible	
#hashtags, images (search, co-clicks), time stamps	Captions	Wikipedia

Comparable words

- Ortography
- Images
- Hashtags, names, and URLs
- Co-clicks, timestamps, and demographics
- Wikipedia ConceptIDs?

Comparable words

- Ortography
- Images
- Hashtags, names, and URLs
- Co-clicks, timestamps, and demographics
- Gaze data, fMRI
- Wikipedia ConceptIDs?

Images

dog

red

run

hund

rød

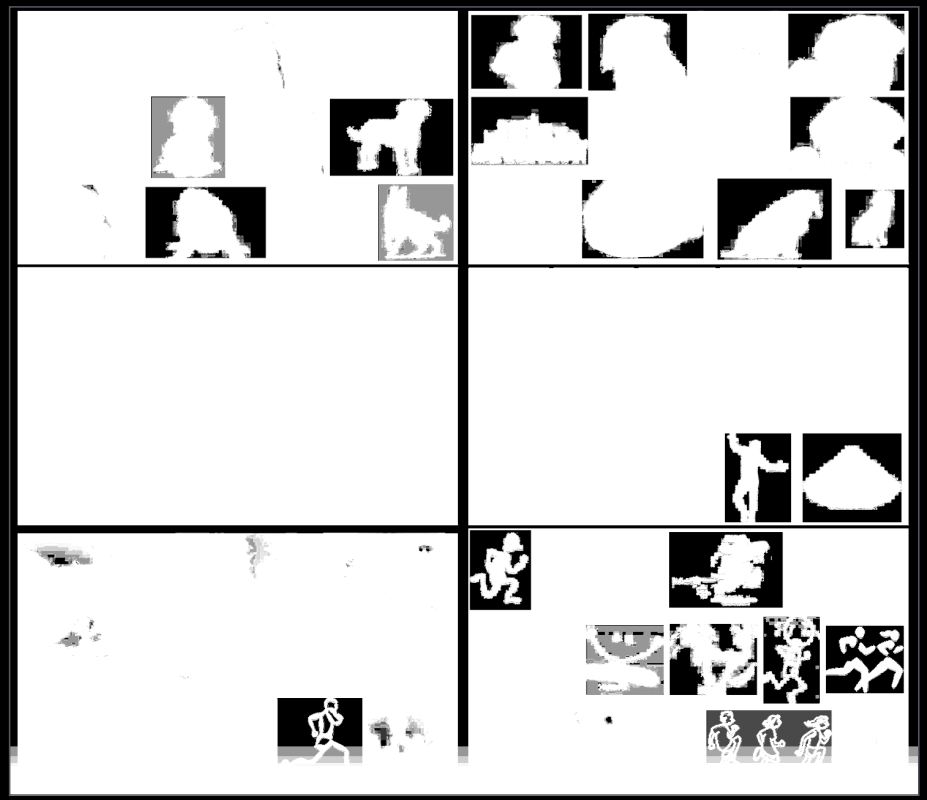
løbe

Search & SIFT

dog

red

run



hund

rød

løbe

Search & CNNs

dog

red

run

hund

rød

løbe

Word-object PMI

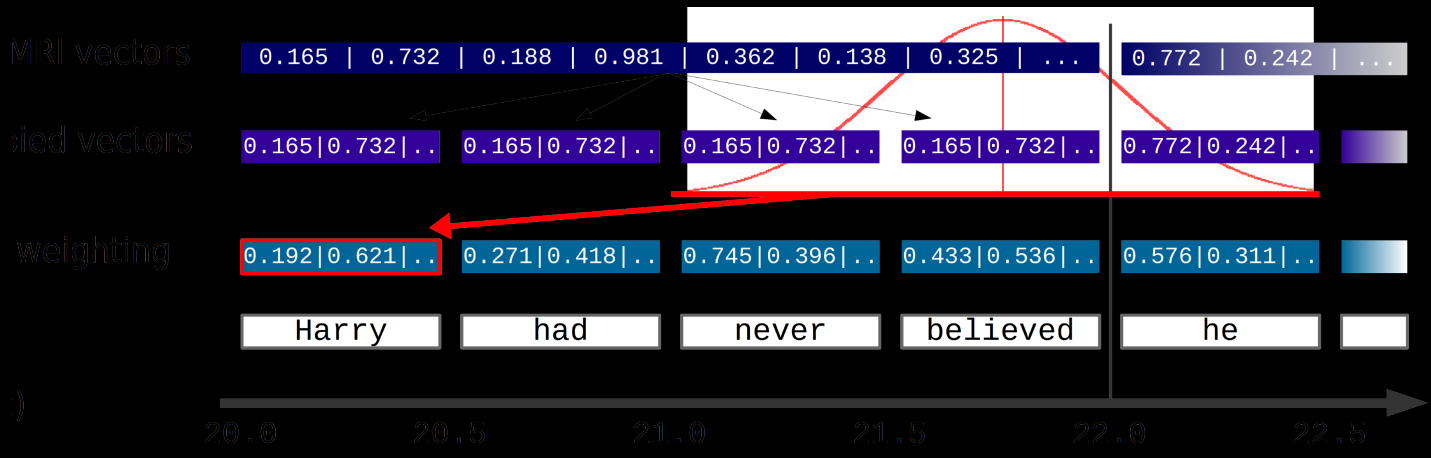
The brown dog is running after the black dog.

Ein brauner Hund und ein schwarzer Hund.

Petrén Hansen, Hartmann and Søgaard (2017)

Gaze data & fMRI

Gaze data & fMRI



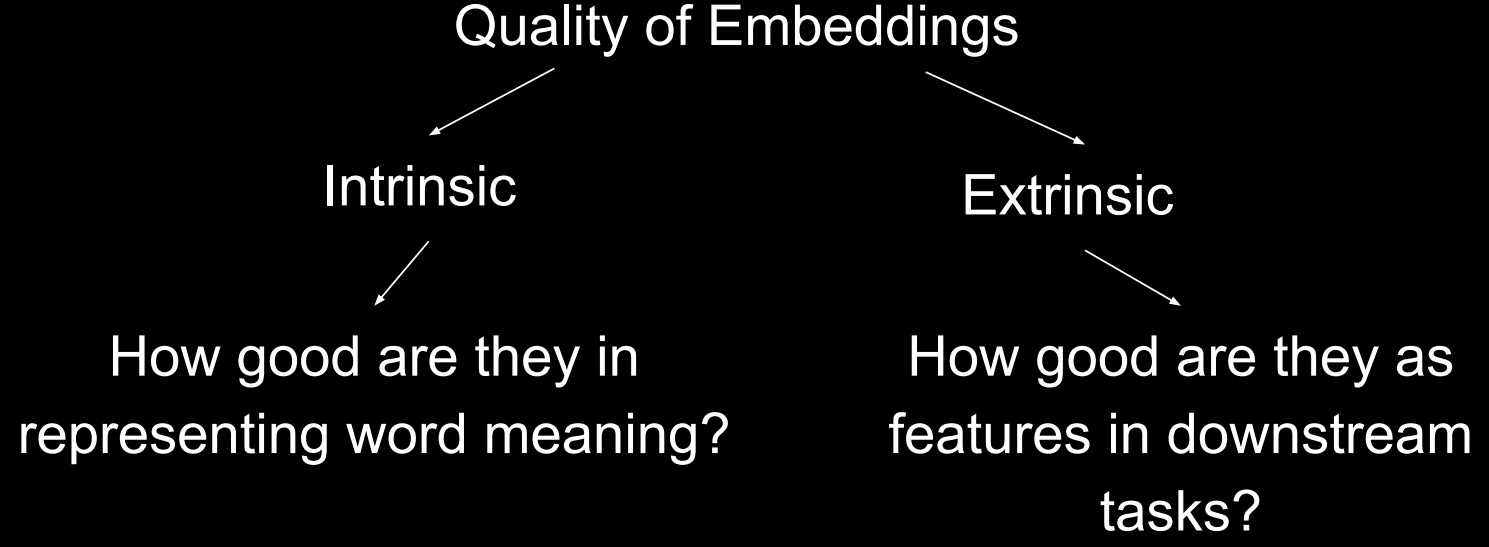
Target	EW30	fMRI	Okay?
students	teachers	mistake	No
creep	drift	long	No
peace	death	eat	Maybe
tight	nasty	hold	Maybe
squeak	twisted	broke	Yes
admiring	cursing	stunned	Yes
amazed	delighted	impressed	Yes

In sum

SOA 	SOA 	
 Premature	Close, but no cigar 	Close, but no cigar 

Part V: Evaluation and Application

Evaluation


Quality of Embeddings
Intrinsic
Extrinsic
How good are they in representing word meaning?
How good are they as features in downstream tasks?

Evaluation

Intrinsic >> Extrinsic?

Intrinsic << Extrinsic?

Intrinsic ⟶ Extrinsic?

Most Ideal but hardly true!

[Schnabel et al, EMNLP 2015; Tsvetkov et al, EMNLP 2015]

Intrinsic Evaluation

Properties of Intrinsic Evaluation Metric:

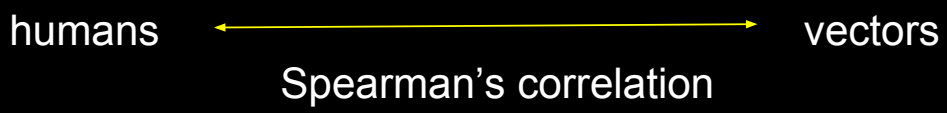
1. Correlated with downstream applications
2. Fast and easy to use
3. Provides insights on the model, facilitates error analysis
4. Approximates a range of related tasks

Intrinsic Evaluation

Word Similarity Tasks

tiger	tiger	10
king	cabbage	0.2
...	...	
sugar	approach	0.8

tiger	tiger	1
king	cabbage	0.3
...	...	
sugar	approach	0.4



[Rubenstein & Goodenough, 1965; Finkelstein et al, 2002; Bruni et al, 2012; Hill et al 2014]

Intrinsic Evaluation

Cross-lingual Word Similarity

CROSS-LINGUAL			
DE-ES	Sessel	taburete	3.08
DE-FA	Lawine	برف	2.25
DE-IT	Taifun	ciclone	3.46
EN-DE	pancreatic cancer	Chemotherapie	1.75
EN-ES	Jupiter	Mercurio	3.25
EN-FA	film	پوچ‌گرایی	0.25
EN-IT	island	penisola	3.08
ES-FA	duna	بیابان	2.25
ES-IT	estrella	pianeta	2.83
IT-FA	avvocato	نمایشگر	0.08

Words of different languages annotated for similarity!

[SemEval 2017 Task 2: Camacho-Collados et al 2017]

Problems with Word Similarity

Low correlation with extrinsic tasks

Subjectivity: Relatedness & Similarity

Frequency & hubness effect in
vector-space affecting cosine similarity

Inability to account for polysemy

[Faruqui et al, RepEval 2016]

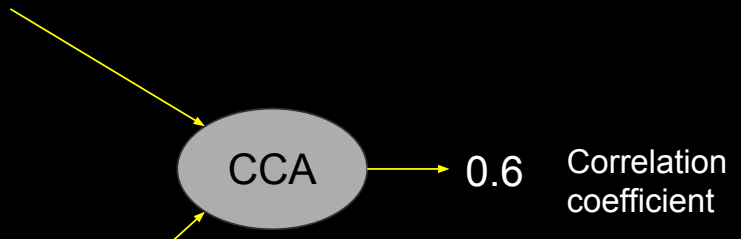
Intrinsic Evaluation: QVec

WORD	d1	d2	d3	d4	d5	...	d50
summer	0.12	0.21	0.07	0.25	0.33	...	0.51
spring	0.19	0.57	0.99	0.30	0.02	...	0.73
fall	0.53	0.77	0.43	0.20	0.29	...	0.85
light	0.00	0.68	0.84	0.45	0.11	...	0.03
clear	0.27	0.50	0.21	0.56	0.25	...	0.32

Word vector matrix

WORD	PTB.NN	PTB.VB	...	PTB.JJ
summer	0.96	0.00	...	0.00
spring	0.94	0.02	...	0.00
fall	0.49	0.43	...	0.00
light	0.52	0.02	...	0.41
clear	0.00	0.10	...	0.87

Linguistic vector matrix



- Grounds vectors in linguistic properties!
- High correlation shown with extrinsic tasks like classification, parsing etc.

[Tsvetkov et al, EMNLP 2015]

Intrinsic Evaluation: QVec

WORD	d1	d2	d3	d4	d5	...	d50
summer	0.12	0.21	0.07	0.25	0.33	...	0.51
spring	0.19	0.57	0.99	0.30	0.02	...	0.73
fall	0.53	0.77	0.43	0.20	0.29	...	0.85
light	0.00	0.68	0.84	0.45	0.11	...	0.03
clear	0.27	0.50	0.21	0.56	0.25	...	0.32

Word vector matrix

WORD	PTB.NN	PTB.VB	...	PTB.JJ
summer	0.96	0.00	...	0.00
spring	0.94	0.02	...	0.00
fall	0.49	0.43	...	0.00
light	0.52	0.02	...	0.41
clear	0.00	0.10	...	0.87

Linguistic vector matrix

Inextensible easily to other languages!

Similarity-Based vs Feature-Based?

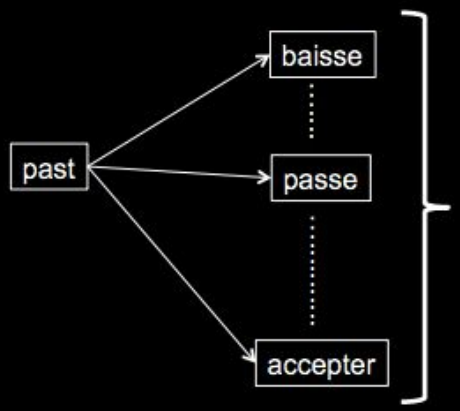
Parallel to the division to intrinsic and extrinsic tasks, there is another perspective:

Similarity-based use of embeddings: applications that rely on (constrained) nearest neighbour search in the cross-lingual word embedding graph

Feature-based use of embeddings: use cross-lingual embeddings directly as features: information retrieval, cross-lingual transfer?

Cross-lingual Dictionary Induction

- For a word in English, find top-10 neighbors in foreign language
- Evaluate these neighbours against a gold dictionary



Gold Dictionary

(white, blanc)
…
(past, passe)
…
(watch, garde)
…
(school, école)
…

Candidate Generation

Matching with Gold data

[Vulić and Moens, NAACL 2013]

Cross-Lingual Lexicon Induction

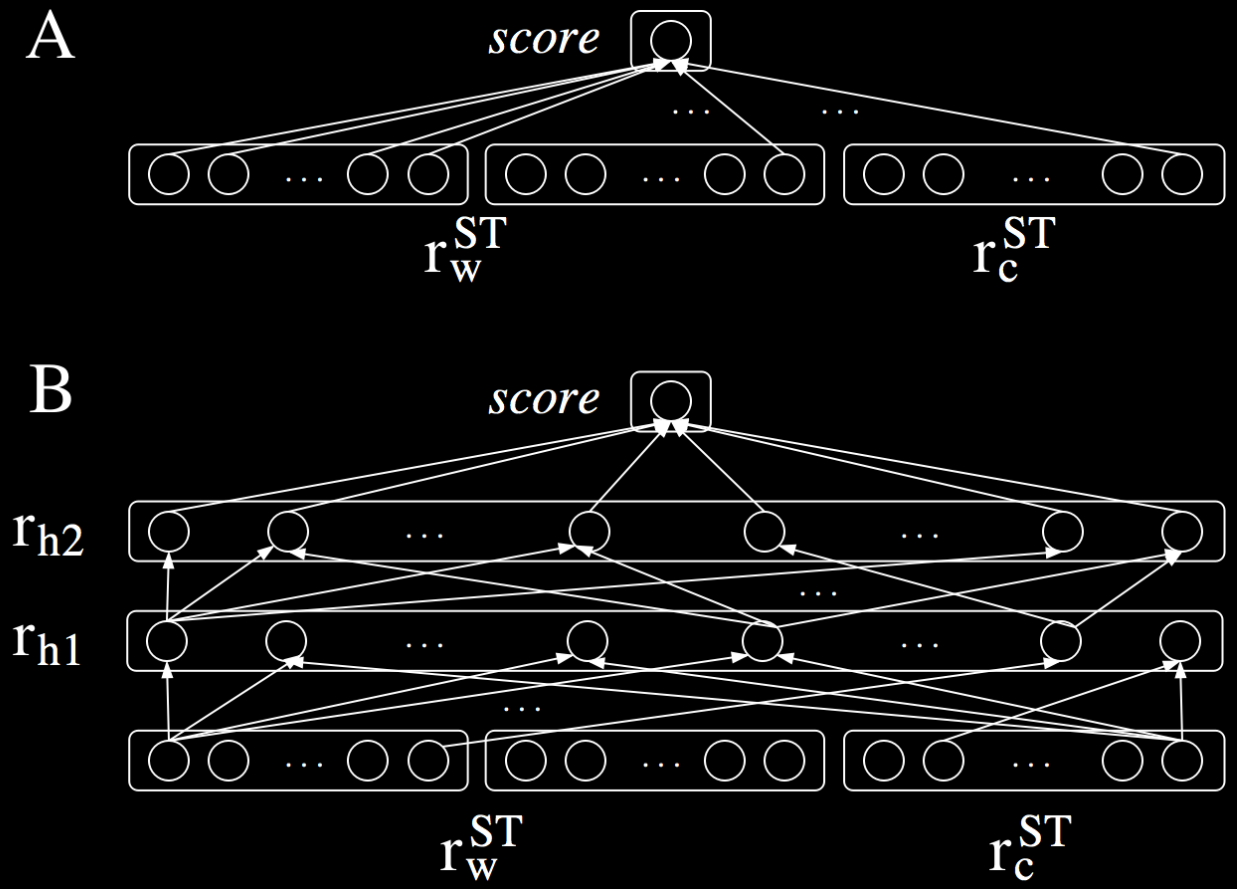
Framed as cross-lingual synonymy in a shared embedding space?

gravedad$_{es}$		*tassazione*$_{it}$	
es	en	it	en
gravitacional	**gravity***	tasse	**taxation***
gravitatoria	gravitation*	fiscale	taxes
aceleracin	acceleration	tassa	tax*
gravitacin	non-gravitational	imposte	levied
inercia	inertia	imposta	fiscal
gravity	centrifugal	fiscali	low-tax
msugra	free-falling	l'imposta	revenue
centrífuga	gravitational	tonnage	levy
curvatura	free-fall	tax	annates
masa	newton	accise	evasion

Metrics: Acc_N, MRR?

Cross-Lingual Lexicon Induction

Similarity-based vs. classification-based lexicon induction

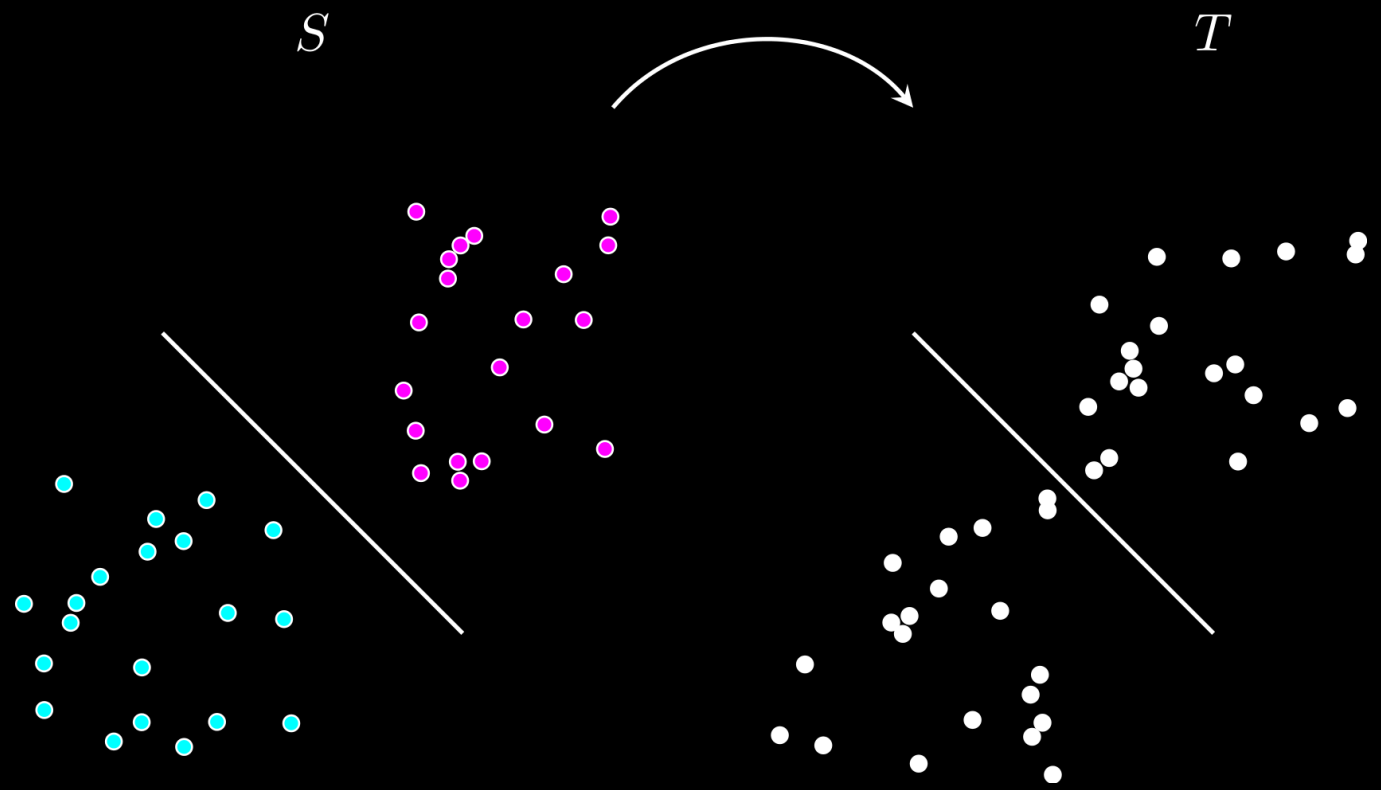


Word embeddings used as features for classification

Extrinsic Evaluation

Cross-lingual Document Classification

Quite popular in the CLWE literature

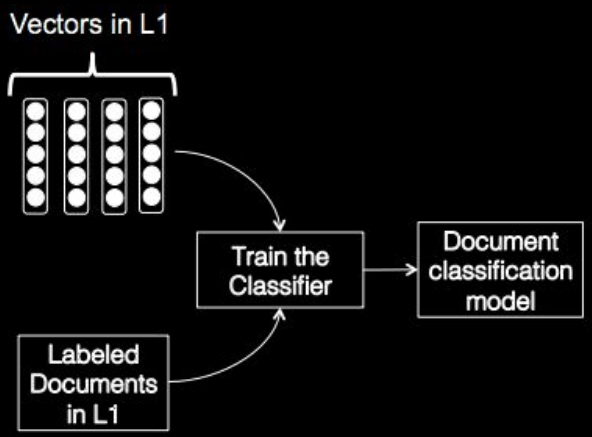


Cross-lingual knowledge transfer?

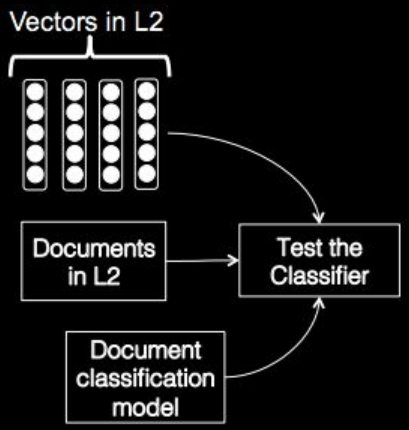
[Klementiev et al., COLING 2012; Heyman et al., DAMI 2015]

Extrinsic Evaluation

Cross-lingual Document Classification



TRAIN classifier in L1



TEST classifier in L2

Tests the transferability of vectors across languages!

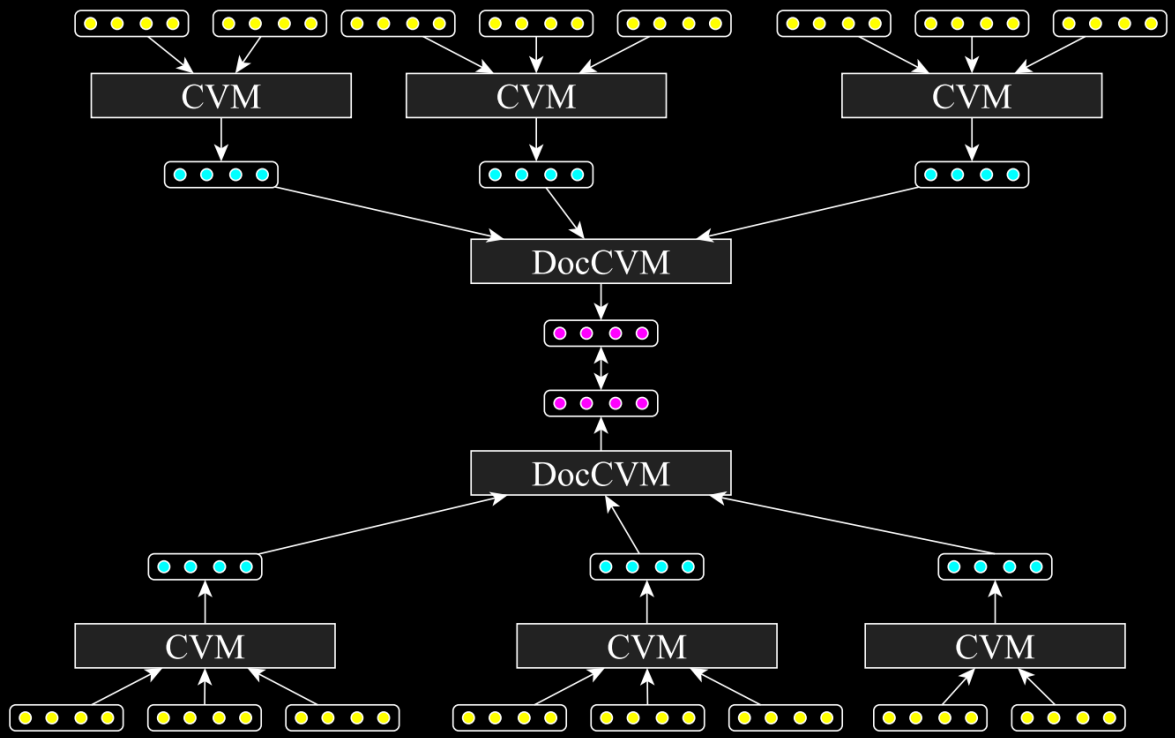
https://github.com/shyamupa/biling-survey

[Klementiev et al., COLING 2012]

Extrinsic Evaluation

Cross-lingual Document Classification

Using word-level information composed into a document-level representation



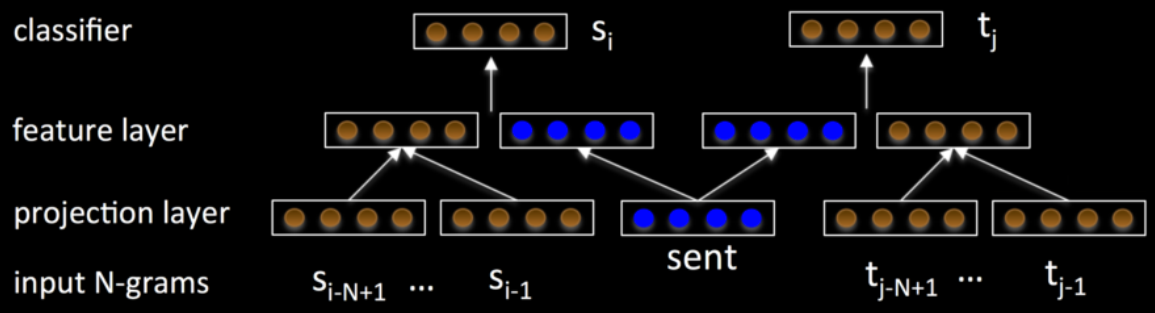
[Hermann and Blunsom, ACL 2014]

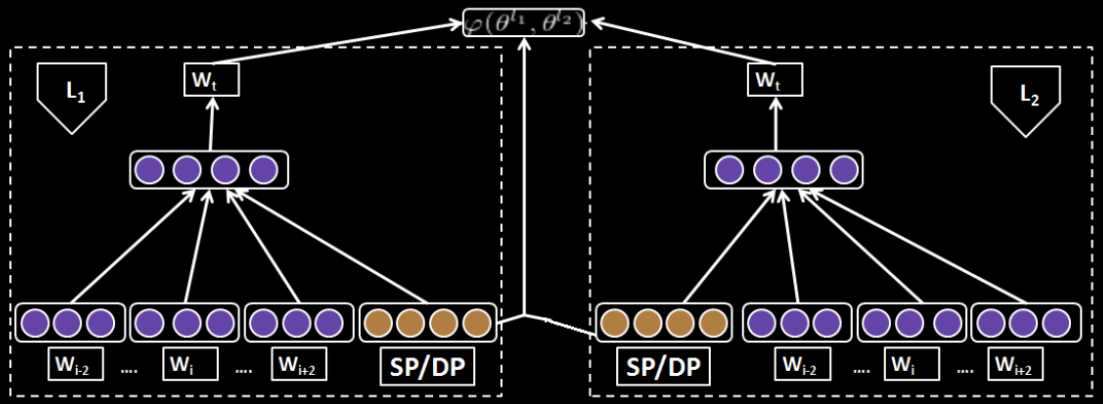
1. Is this really an optimal way to approach this task?
2. Is this really an optimal way to evaluate word embeddings?

Digression

Cross-lingal Representation Beyond the Word Level

Towards (direct) bilingual phrase and sentence representations





[Pham et al., NAACL 2015; Mogadala and Rettinger, NAACL 2016]

But this is a completely different story...

Extrinsic Evaluation

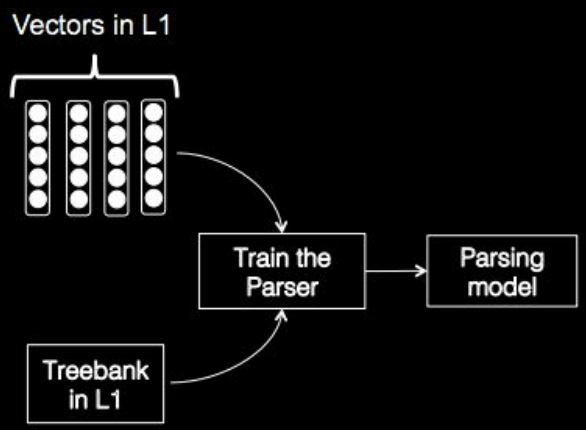
Cross-lingual Document Classification

Some results... (finally?)

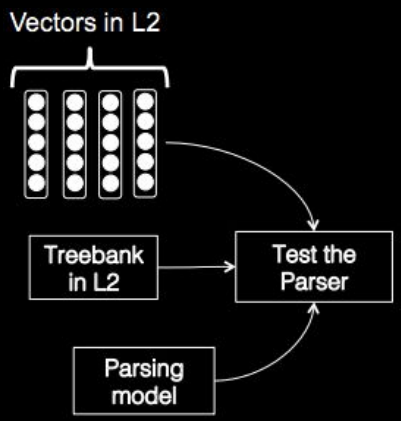
Method	en-de	de-en	en-fr	fr-en	en-es	es-en
Majority class	46.8	46.8	22.5	25.0	15.3	22.2
MT	68.1	67.4	76.3	71.1	52.0	58.4
Klementiev et al. (2012)	77.6	71.1	74.5	61.9	31.3	63.0
Chandar et al. (2014)	91.8	74.2	84.6	74.2	49.0	64.4
Hermann and Blunsom (2014)	86.4	74.7	-	-	-	-
Kocisky et al. (2014)	83.1	75.4	-	-	-	-
Gouws et al. (2015)	86.5	75.0	-	-	-	-
Luong et al. (2015)	87.6	77.8	-	-	-	-
Coulmance et al. (2015)	87.8	78.7	-	-	-	-
Mogadala and Rettinger (2016)	88.1	78.9	79.2	77.8	56.9	67.6

Extrinsic Evaluation

Cross-lingual Dependency Parsing



TRAIN parser in L1



TEST parser in L2

Tests the transferability of vectors across languages!

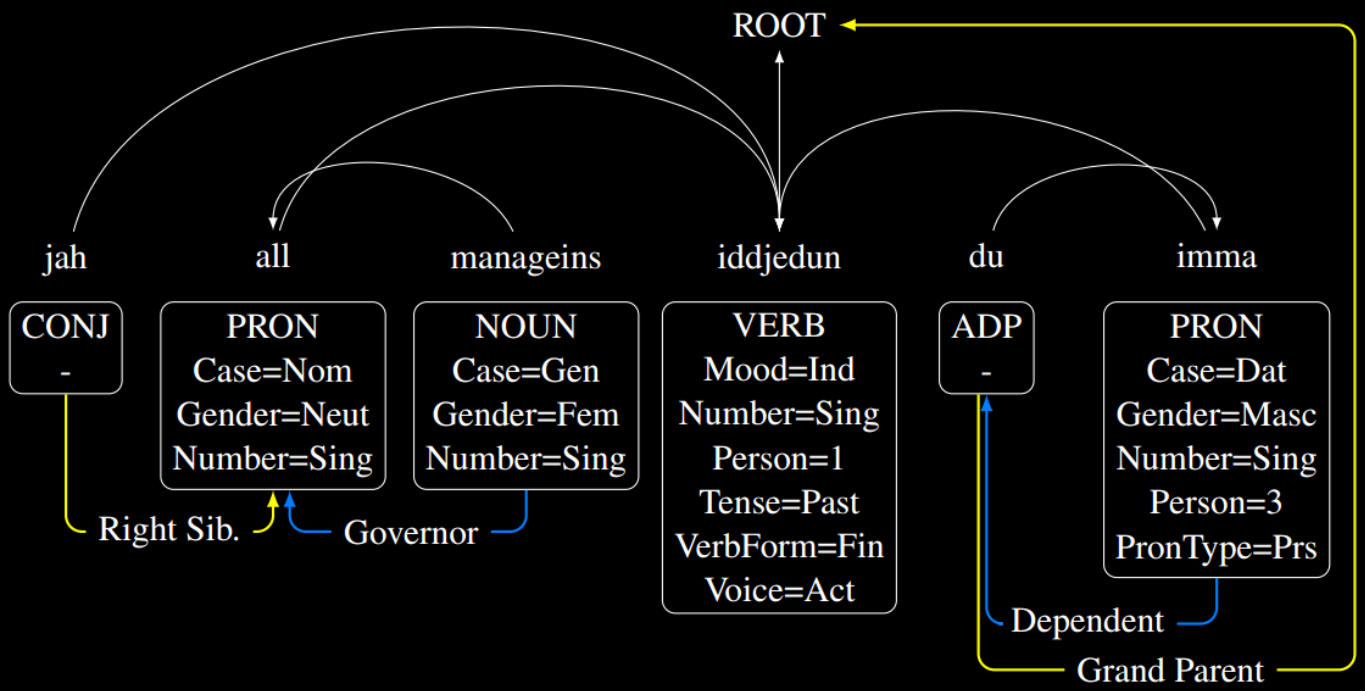
https://github.com/shyamupa/biling-survey

[Guo et al., ACL 2015]

Extrinsic Evaluation

Cross-lingual Dependency Parsing

Delexicalized transfer plus lexical features: clusters, embeddings, dictionaries



[Søgaard et al., ACL 2015; Guo et al., JAIR 2016; Upadhyay et al., ACL 2016; Dehouck and Denis, EACL 2017]

Extrinsic Evaluation and Application

Cross-lingual Task X, Cross-lingual Task Y, ...

Cross-lingual lexical and textual entailment

[Vyas and Carpuat, NAACL 2016; Agić and Schluter, arXiv 2017]

English-English	English-French
affection $\rightarrow$ feeling	affection $\rightarrow$ sentiment
aspirin $\rightarrow$ drug	aspirin $\not\rightarrow$ drogue
water $\rightarrow$ wet	water $\rightarrow$ humide
feeling $\not\rightarrow$ nostalgia	feeling $\not\rightarrow$ nostalgie

Extrinsic Evaluation and Application

Cross-lingual Task X, Cross-lingual Task Y, ...

Cross-lingual supersense tagging

[Gouws and Søgaard, NAACL 2015]

Word alignment

[Levy et al., EACL 2017]

POS tagging

[Søgaard et al., ACL 2015; Zhang et al., NAACL 2016]

Wikification

[Tsai and Roth, NAACL 2016]

Extrinsic Evaluation and Application

Cross-lingual Task X, Cross-lingual Task Y, ...

Cross-lingual discourse parsing

[Braud et al., EACL 2017]

Frame-semantic parsing

[Johannsen et al., EMNLP 2013]

Semantic role labeling

[Kozhevnikov and Titov, ACL 2013; Akbik et al., ACL 2016]

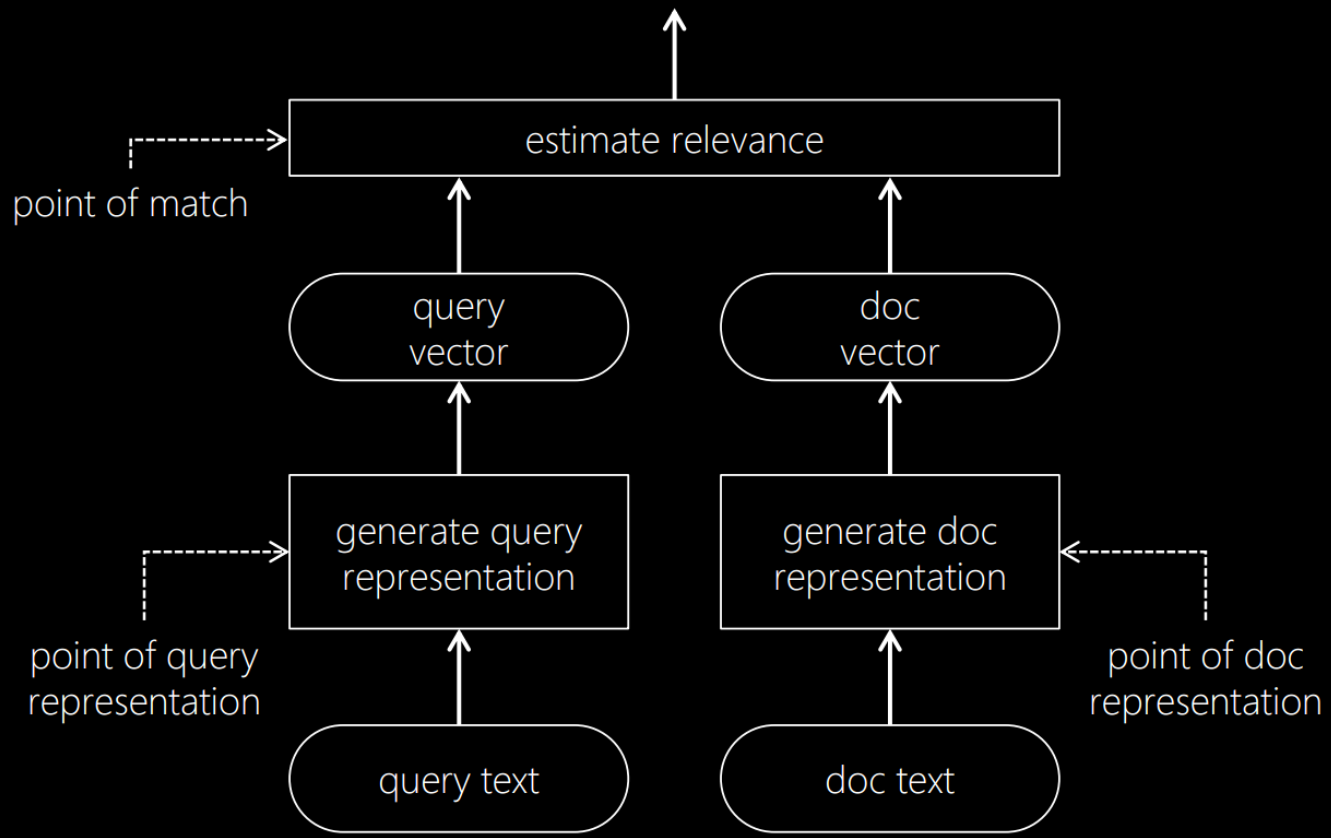
Sentiment analysis

[Zhou et al., ACL 2016; Abdalla and Hirst, arXiv 2017]

Extrinsic Evaluation and Application

Cross-lingual Information Retrieval

Semantic representations for cross-lingual information IR



[Vulić and Moens, SIGIR 2015; Mitra and Craswell, arXiv 2017]

Extrinsic Evaluation and Application

Cross-lingual Information Retrieval

Document and Query Embeddings

Composition of word embeddings:
[Mitchell and Lapata, ACL 2008]

$$\overrightarrow{d} = \overrightarrow{w_1} + \overrightarrow{w_2} + \ldots + \overrightarrow{w_{|N_d|}}$$

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

$$\overrightarrow{d} = [f_{d,1}, \ldots, f_{d,k}, \ldots, f_{d,dim}]$$

Extrinsic Evaluation and Application

Cross-lingual Information Retrieval

Document and Query Embeddings

$\rightarrow$ The same principles with **queries**

$$\overrightarrow{Q} = \overrightarrow{q_1} + \overrightarrow{q_2} + \ldots + \overrightarrow{q_m}$$

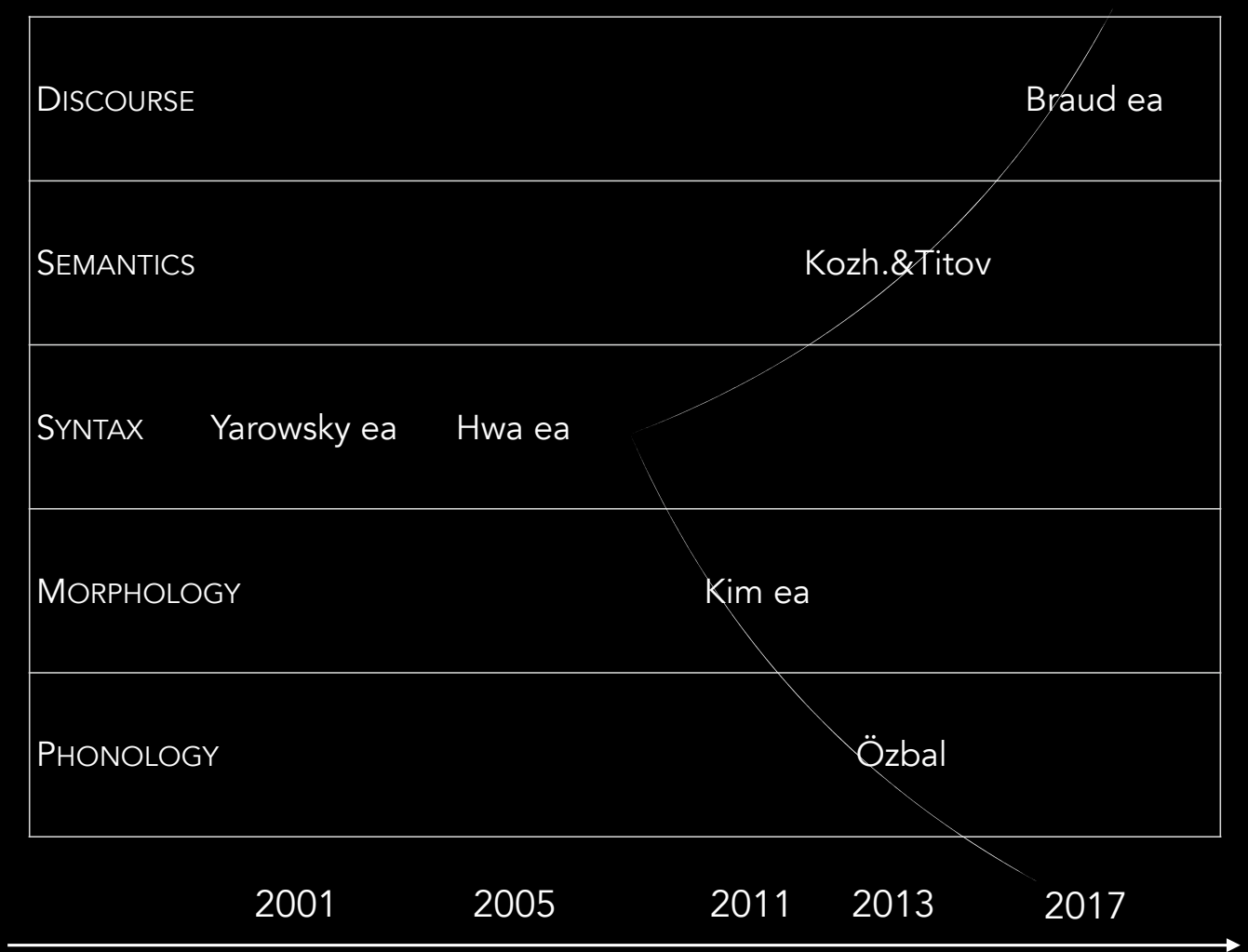
The dim-dimensional **query embedding** in the same bilingual word embedding space:

$$\overrightarrow{Q} = [f_{Q,1}, \ldots, f_{Q,k}, \ldots, f_{Q,dim}]$$

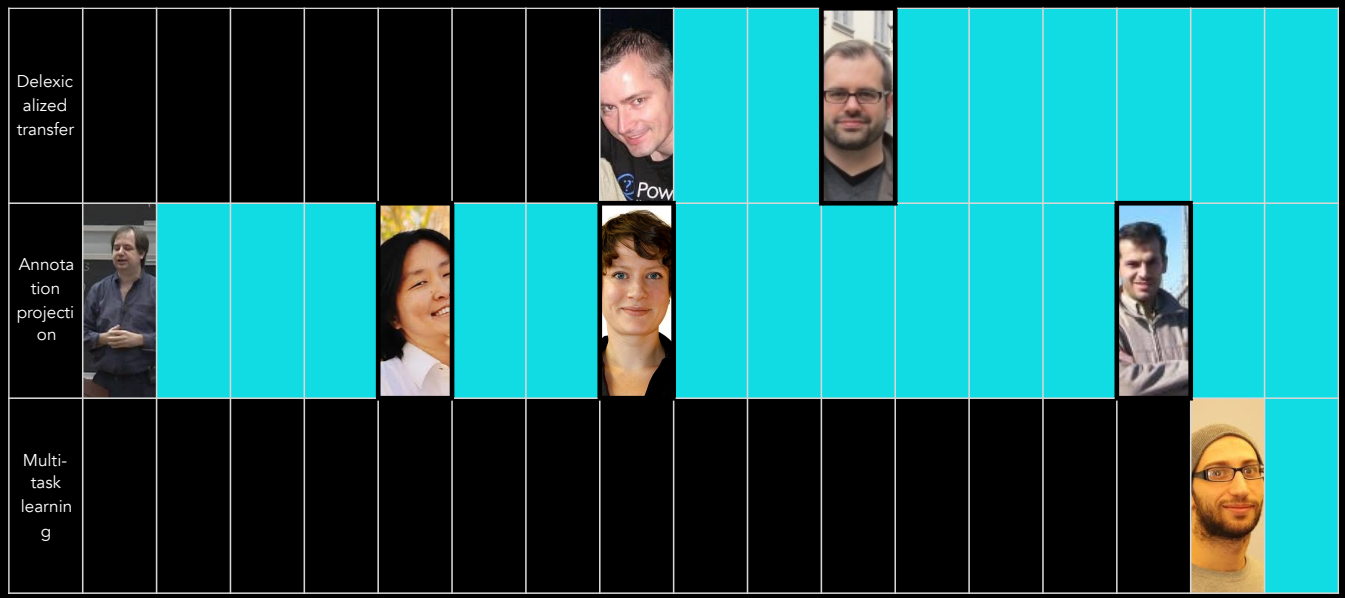
Cross-lingual transfer

Anders Søgaard
http://cst.dk/anders/

UNIVERSITY OF
COPENHAGEN


DISCOURSE
Braud ea
SEMANTICS
Kozh.&Titov
SYNTAX
Yarowsky ea
Hwa ea
MORPHOLOGY
Kim ea
PHONOLOGY
Özbal
2001
2005
2011
2013
2017

Approaches

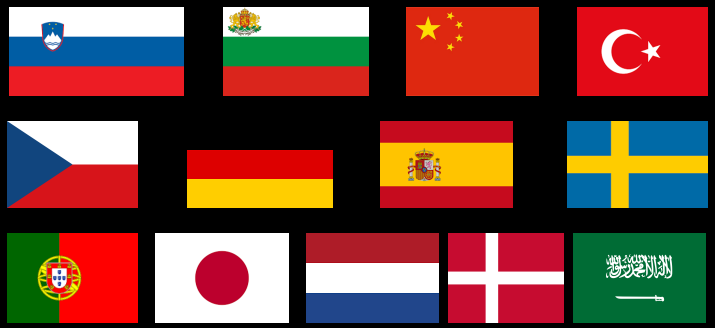

Delexicalized transfer
Annotation projection
Multi-task learning
2001
2005
2011
2013
2017

Downstream [✓]

Diversity [✗]

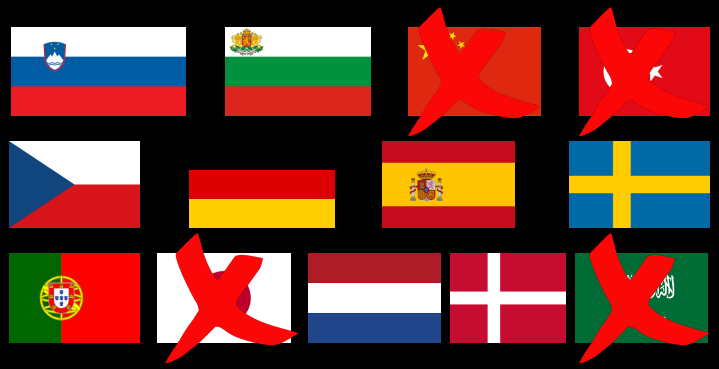
Translations

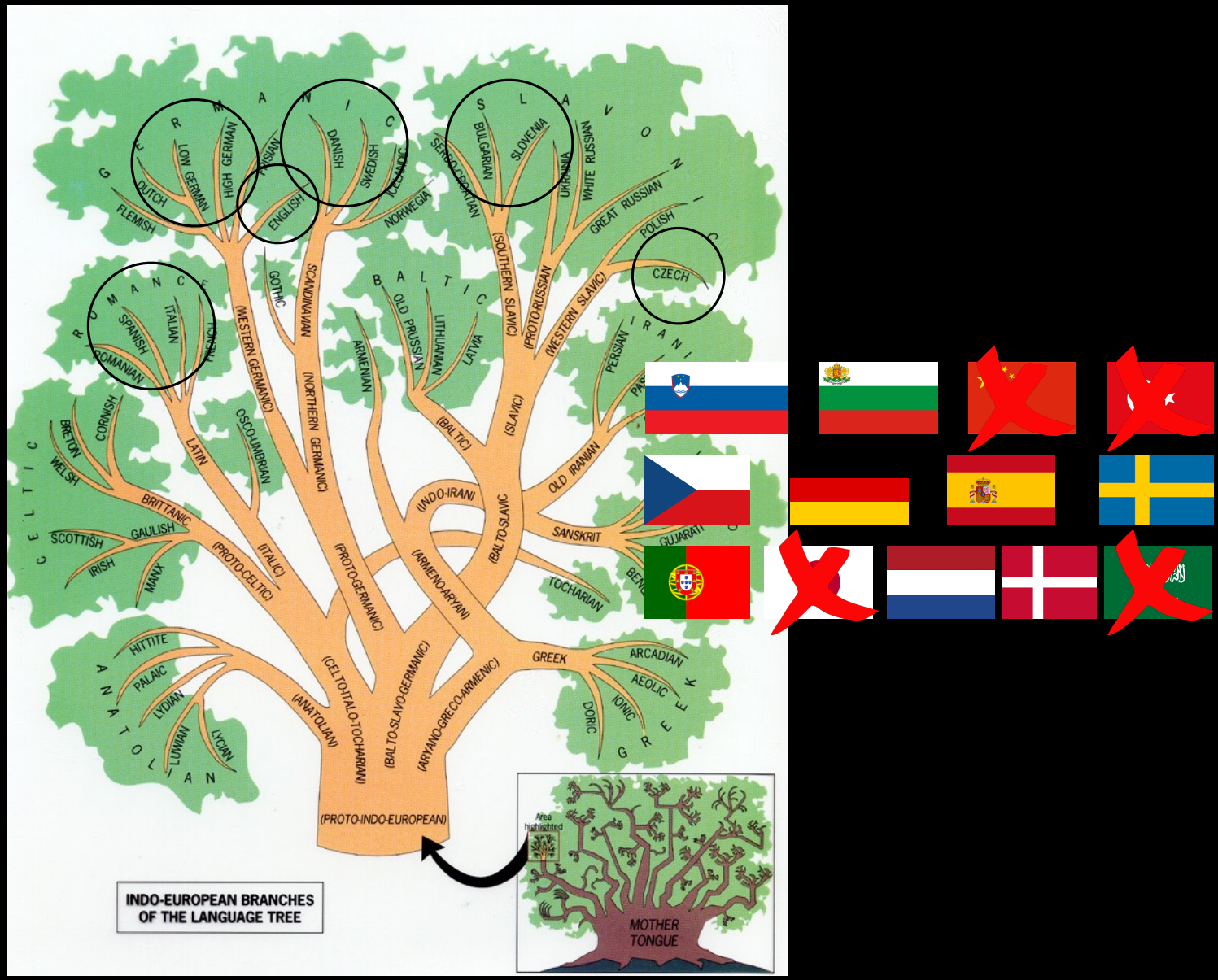
Treebanks



Translations

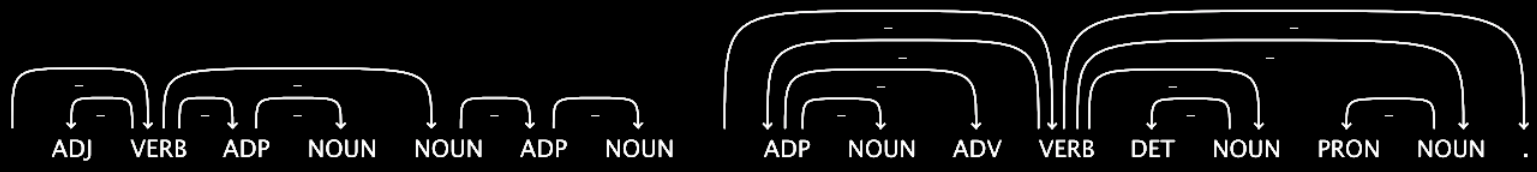
Treebanks

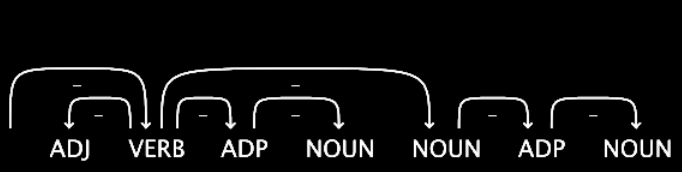



GERMANIC
ROMANCE
CELTIC
ANATOLIAN
GREEK
BALTIC
SLAVONIC
IRANIC
DUTCH
FLEMISH
LOW GERMAN
HIGH GERMAN
FRISIAN
ENGLISH
DANISH
SWEDISH
ICELANDIC
NORWEGIA
SERBO-CROATIAN
BULGARIAN
SLOVENIA
UKRAINA
WHITE RUSSIAN
GREAT RUSSIAN
POLISH
CZECH
ROMANIAN
SPANISH
ITALIAN
FRENCH
GOTHIC
SCANDINAVIAN
(WESTERN GERMANIC)
(NORTHERN GERMANIC)
OSCO-UMBRIAN
LATIN
BRETON
CORNISH
WELSH
SCOTTISH
GAULISH
BRITTANIC
IRISH
MANX
(ITALIC)
(PROTO-CELTIC)
(PROTO-GERMANIC)
OLD PRUSSIAN
LITHUANIAN
LATVIA
ARMENIAN
(BALTIC)
(SLAVIC)
(SOUTHERN SLAVIC)
(PROTO-RUSSIAN)
(WESTERN SLAVIC)
PERSIAN
(INDO-IRANI
(ARMENO-ARYAN)
(BALTO-SLAVIC)
OLD IRANIAN
SANSKRIT
GUJARATI
TOCHARIAN
HITTITE
PALAIC
LYDIAN
LUWIAN
LYCIAN
(ANATOLIAN)
(CELTO-ITALO-TOCHARIAN)
(BALTO-SLAVO-GERMANIC)
(ARYANO-GRECO-ARMENIC)
GREEK
ARCADIAN
AEOLIC
IONIC
DORIC
(PROTO-INDO-EUROPEAN)
Area highlighted
MOTHER TONGUE
INDO-EUROPEAN BRANCHES
OF THE LANGUAGE TREE

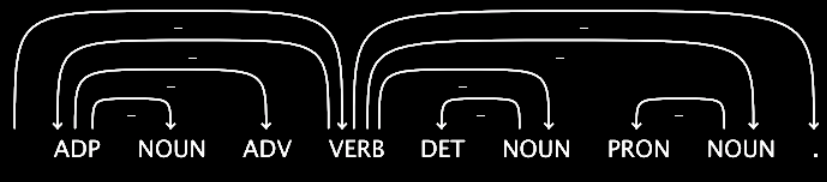
Approaches

- Annotation projection
- Cross-lingual re-lexicalized transfer
- Combinations of both

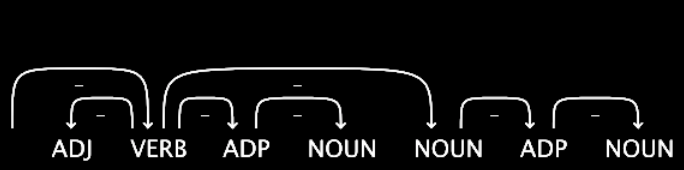

ADJ VERB ADP NOUN NOUN ADP NOUN
ADP NOUN ADV VERB DET NOUN PRON NOUN .

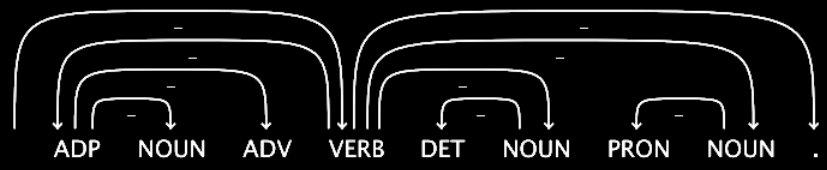

ADJ
VERB
ADP
NOUN
NOUN
ADP
NOUN

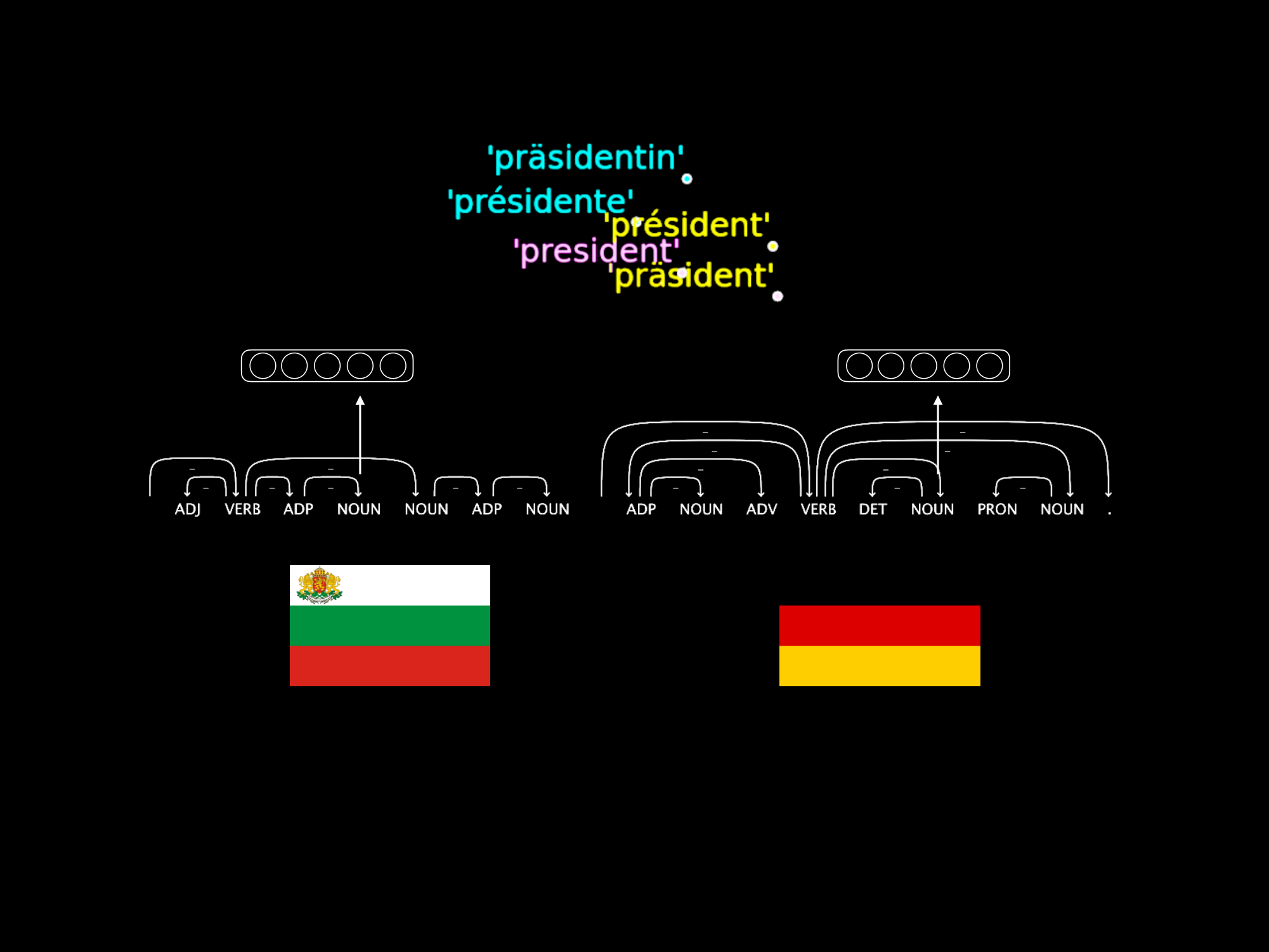
?


ADP
NOUN
ADV
VERB
DET
NOUN
PRON
NOUN
.

?


ADJ
VERB
ADP
NOUN
NOUN
ADP
NOUN


ADP
NOUN
ADV
VERB
DET
NOUN
PRON
NOUN
.


'präsidentin'
'présidente'
'président'
'president'
'präsident'
ADJ VERB ADP NOUN NOUN ADP NOUN
ADP NOUN ADV VERB DET NOUN PRON NOUN .

DISCOURSE	
SEMANTICS	
SYNTAX	Søgaard et al. (2015), Upadhyay et al. (2016), Ammar et al. (2016)
MORPHOLOGY	Gouws and Søgaard (2015), Søgaard et al. (2015)
PHONOLOGY	

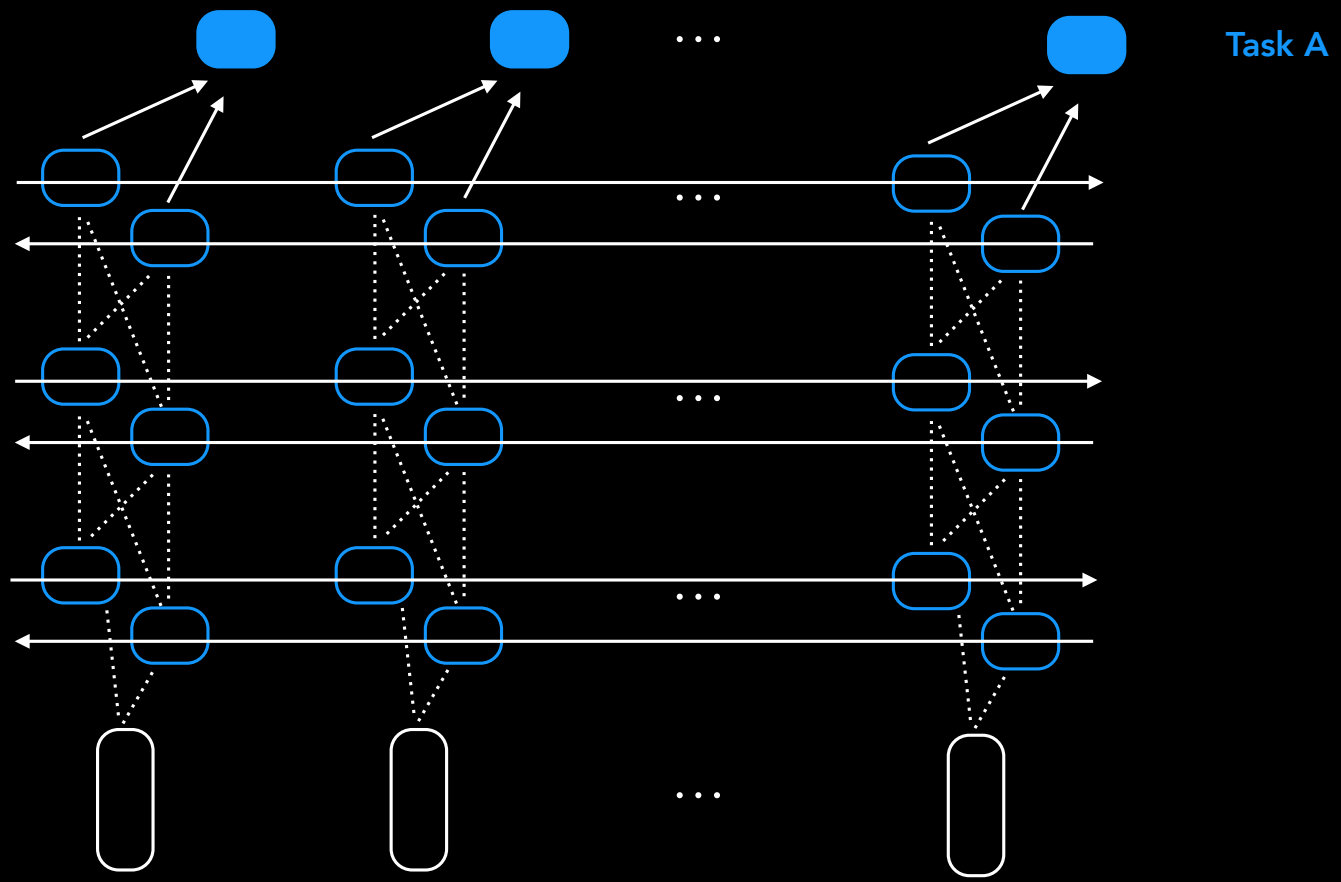
Cross-lingual embeddings
with downstream loss?

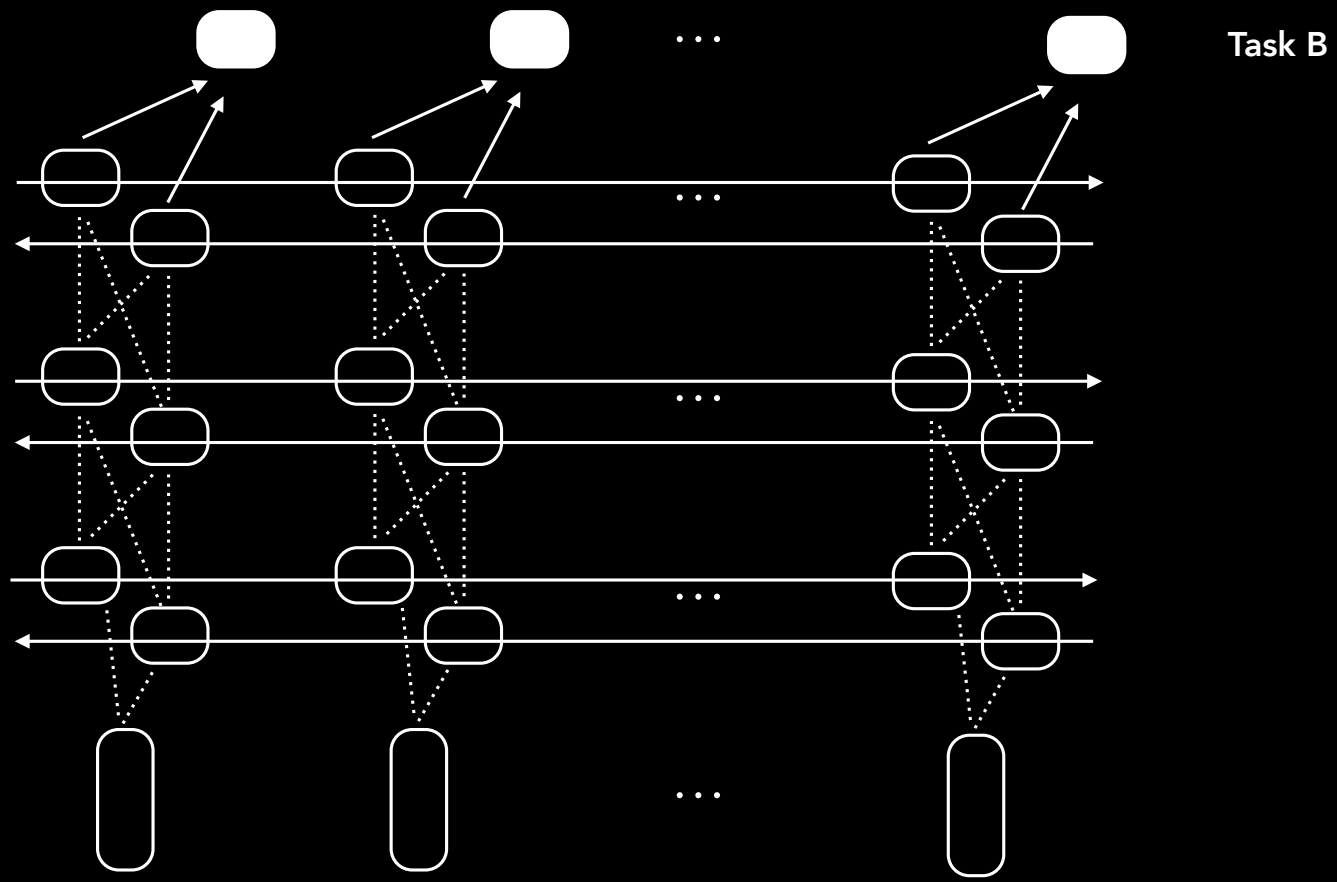
Approaches

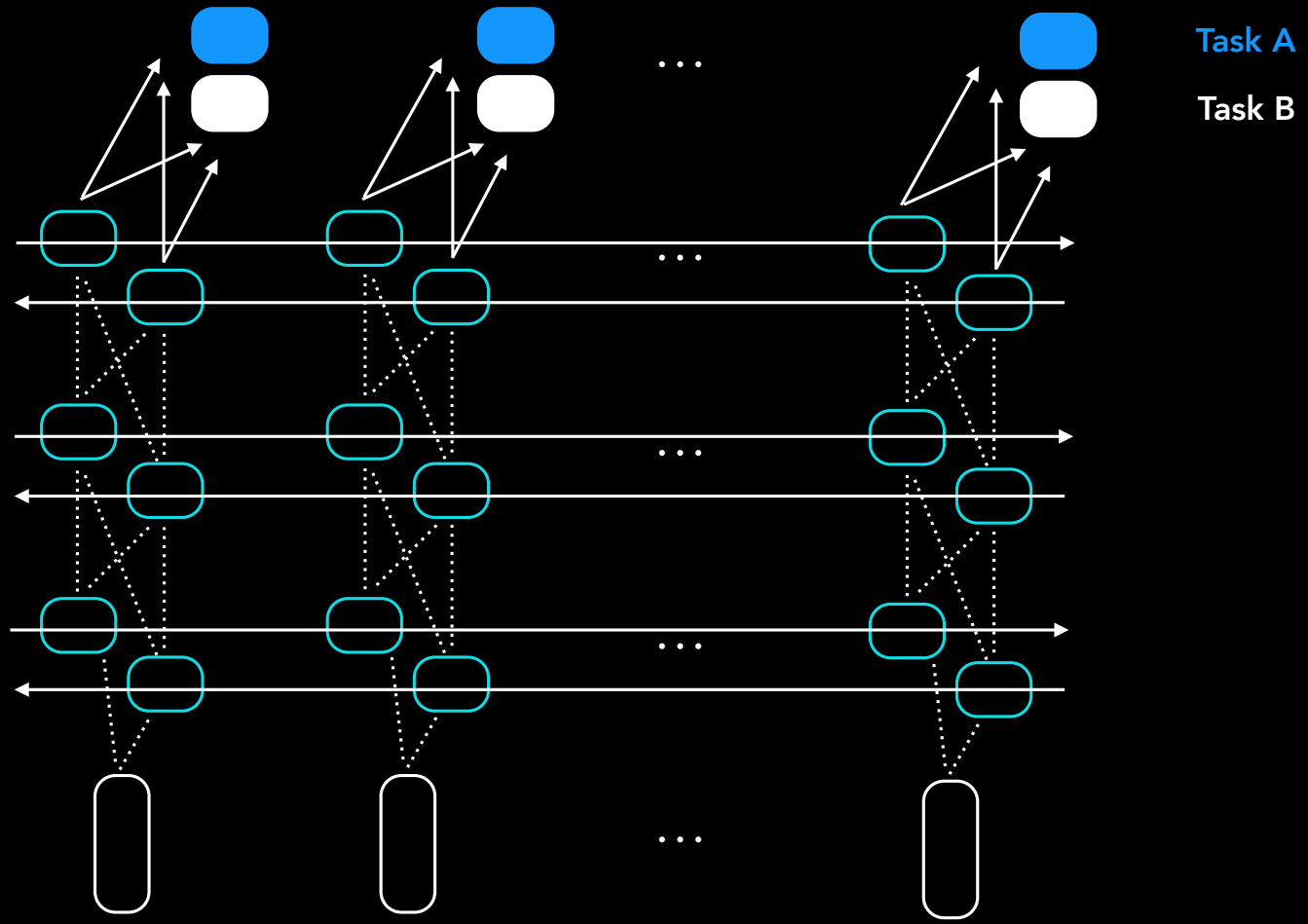
- Fine-tuning pretrained models (e.g, NMT)
- Downstream training with posterior regularization (e.g., Ferreira et al., 2016)
- Multi-task learning

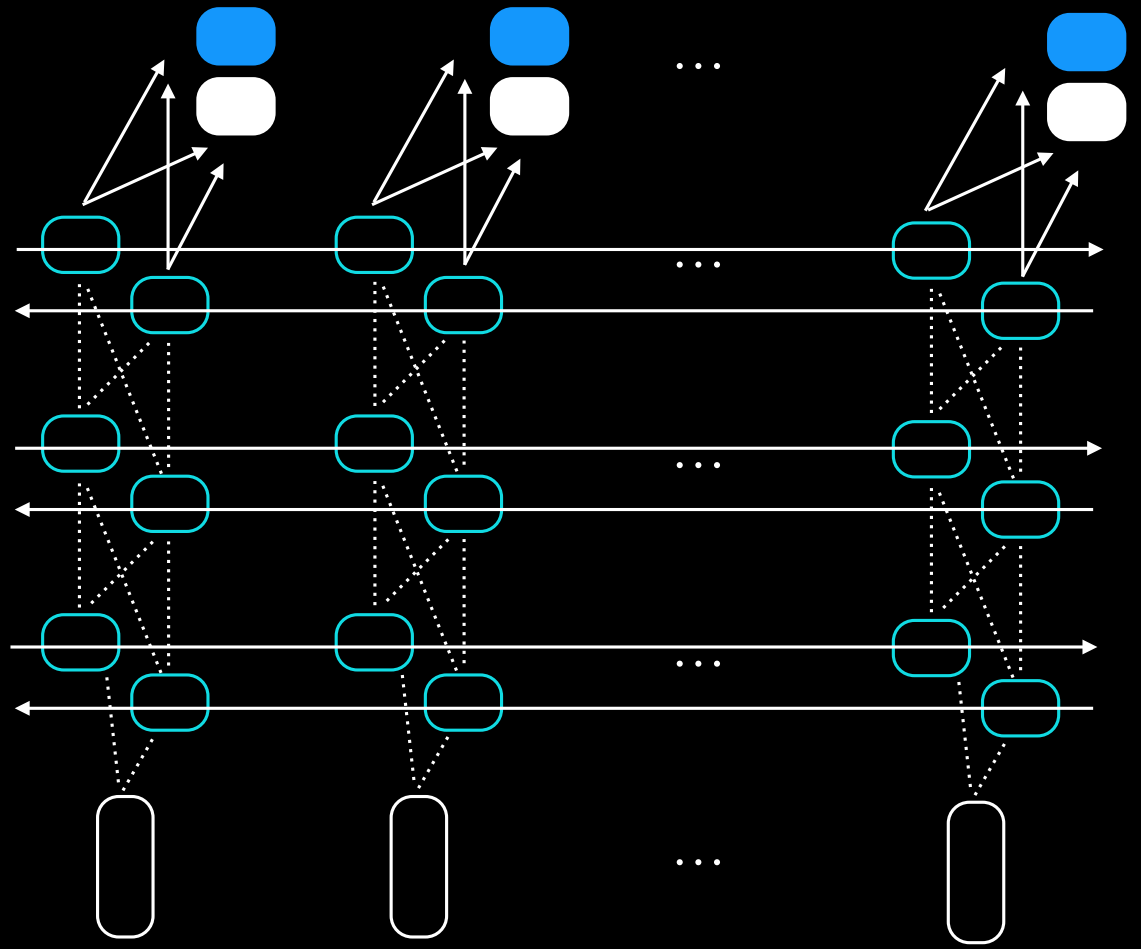
“When humans tackle new problems, they bring to bear what they have learned before for related problems."

Rich Caruana


Task A


Task B

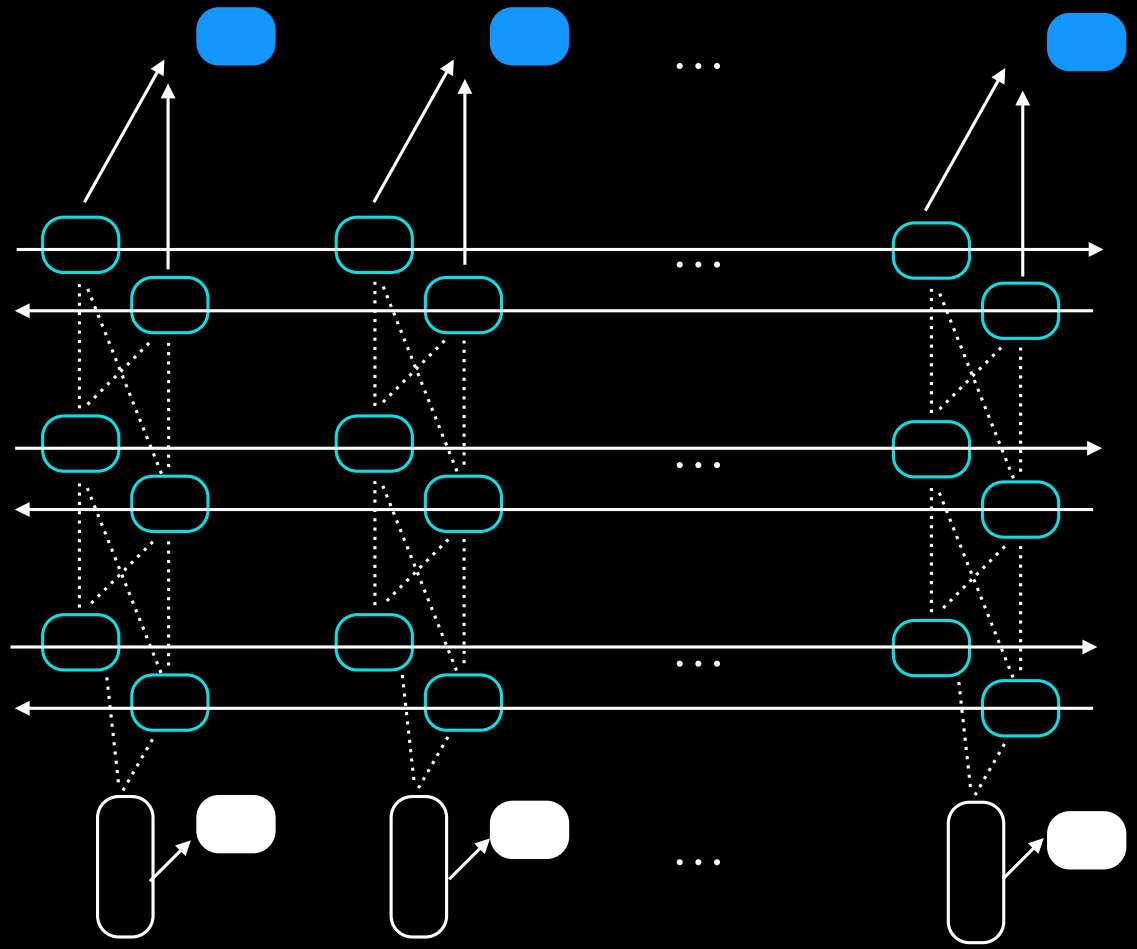

Task A
Task B



Task A

Task B

During training, we randomly choose a dataset, then a data point. Note that therefore the datasets can be disjoint.



Task A

Task B

You can do the same with embeddings, in theory, but open question whether it works, and with what curriculum.

Work on multi-task learning of monolingual word embeddings

- Collobert et al. (2011): *NLP (almost) from scratch*
- Luo et al. (2014): *Pre-trained Multi-View Word Embedding Using Two-side Neural Network*
- Liu et al. (2015): *Representation Learning Using Multi-Task Deep Neural Networks for Semantic Classification and Information Retrieval*

Extrinsic Evaluation and Application

Language Understanding: Multilingual Dialog State Tracking

The **Neural Belief Tracker** is a novel DST model/framework which aims to satisfy the following design goals:

1. End-to-end learnable (no SLU modules or semantic dictionaries).
2. Generalisation to unseen slot values.
3. Capability of leveraging the semantic content of pre-trained word vector spaces without human supervision.

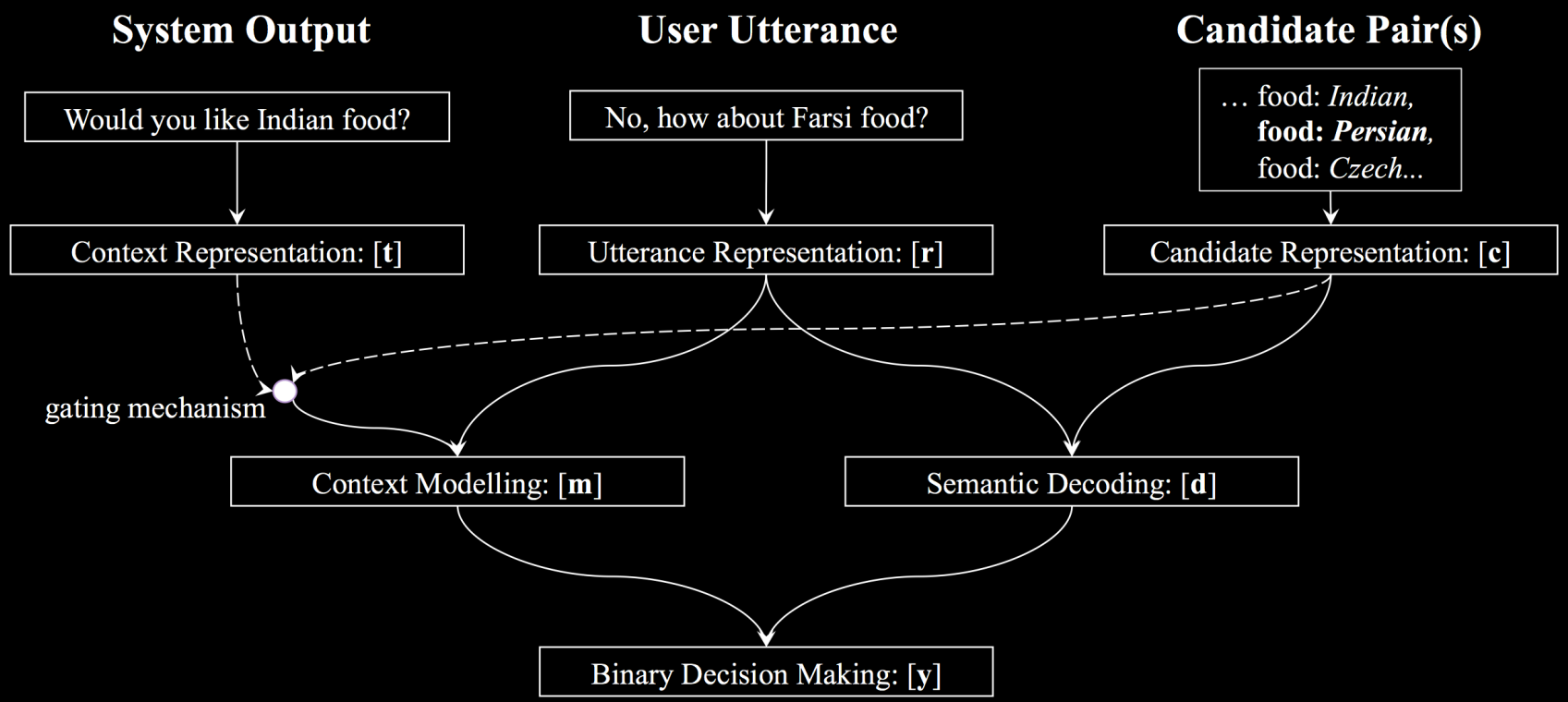
[Mrkšić et al., ACL 2017]

Extrinsic Evaluation and Application

Language Understanding: Multilingual Dialog State Tracking

Representation Learning + Label Embedding + Separate Binary Decisions

To overcome data sparsity, NBT models use *label embedding* to decompose multi-class classification into many binary ones.



Extrinsic Evaluation and Application

Language Understanding: Multilingual Dialog State Tracking

DST in Italian and German

Multilingual WOZ 2.0 (Mrkšić et al., 2017); Italian and German

The 1,200 dialogues in WOZ 2.0 were translated by native Italian and German speakers instructed to consider preceding dialogue context.

Word Vector Space	EN	IT	DE
Best Baseline Vector Space	81.6	71.8	50.5
Monolingual Distributional Vectors	77.6	71.2	46.6
+ Monolingual Specialisation	80.9	72.7	52.4
++ Cross-Lingual Specialisation	80.3	75.3	55.7
+++ Multilingual DST Model	**82.8**	**77.1**	**57.7**

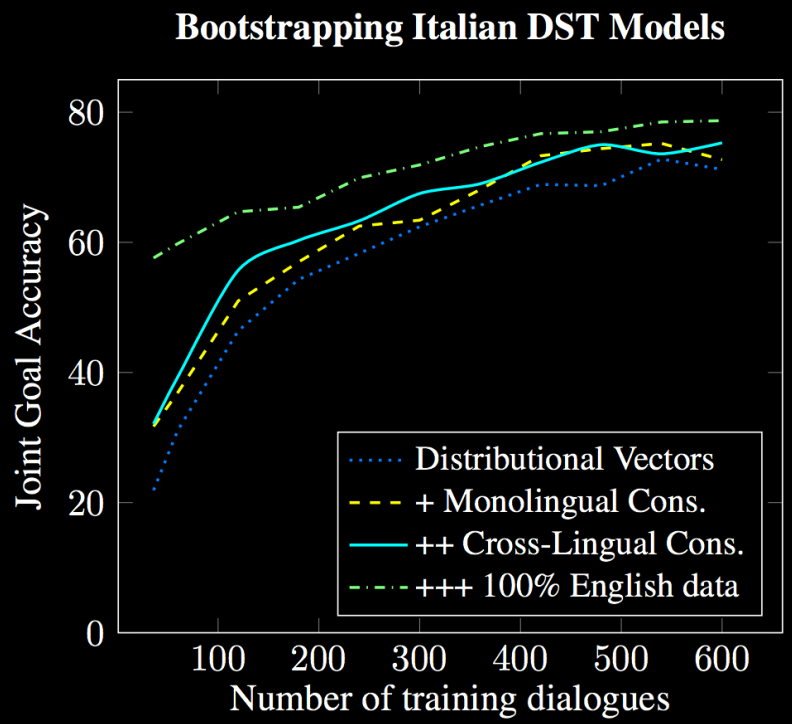
Extrinsic Evaluation and Application

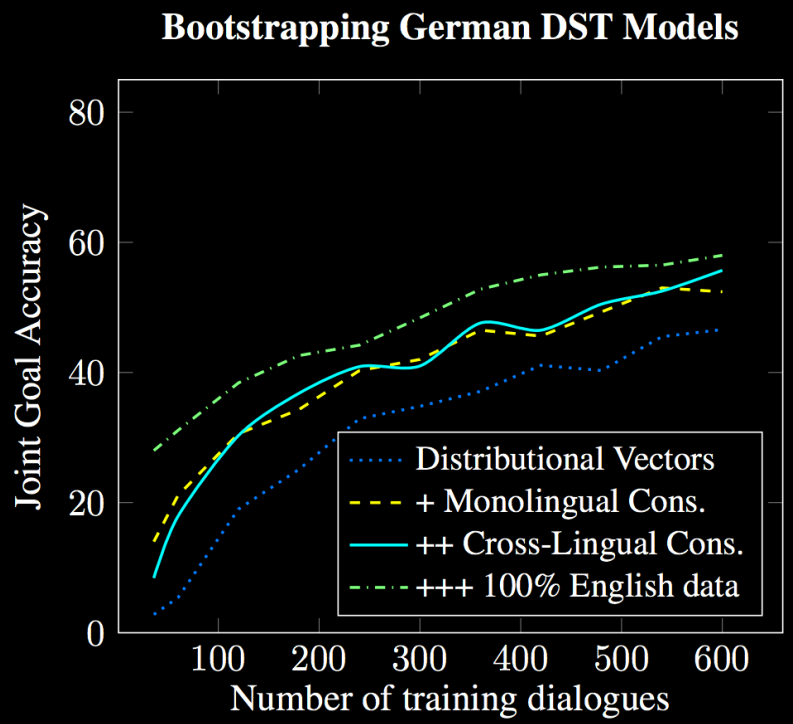
Language Understanding: Multilingual Dialog State Tracking

DST Models for Resource-Poor Languages

Ontology Grounding: Multilingual DST Models

The domain ontology (i.e. the concepts it expresses) is language agnostic, which means that 'labels' persist across languages. Using training data for two (or more) languages, and cross-lingual vectors of high quality, we train the first-ever *multilingual* DST model.





Take-Home Messages

1. Cross-Lingual Word Embedding Models are More Similar to Old(er) NLP Ideas than It Seems

But they are simple, efficient, and state-of-the-art...

2. CLWE Models are More Similar than It Seems

We have addressed similarities between a plethora of cross-lingual word embedding models

Take-Home Messages

1. Cross-Lingual Word Embedding Models are More Similar to Old(er) NLP Ideas than It Seems

But they are simple, efficient, and state-of-the-art...

2. CLWE Models are More Similar than It Seems

We have addressed similarities between a plethora of cross-lingual word embedding models

3. Data is Crucial (more than the Chosen Algorithm)

Optimization and various modeling tricks matter, but the chosen multilingual signal is the most important factor

Take-Home Messages

1. Cross-Lingual Word Embedding Models are More Similar to Old(er) NLP Ideas than It Seems

But they are simple, efficient, and state-of-the-art...

2. CLWE Models are More Similar than It Seems

We have addressed similarities between a plethora of cross-lingual word embedding models

3. Data is Crucial (more than the Chosen Algorithm)

Optimization and various modeling tricks matter, but the chosen multilingual signal is the most important factor

4. CL Word Embeddings Support Cross-Lingual NLP

They are useful across a wide variety of cross-lingual NLP tasks, for cross-lingual (re-lexicalized) transfer, language understanding, IR, ...

Further Work / Research Directions

1. Intrinsic Evaluation → Downstream Performance?

Performance on intrinsic evaluation datasets correlates with downstream tasks such as DST. However, substantial intrinsic gains do not always lead to large downstream gains. Why?

Further Work / Research Directions

1. Intrinsic Evaluation → Downstream Performance?

Performance on intrinsic evaluation datasets correlates with downstream tasks such as DST. However, substantial intrinsic gains do not always lead to large downstream gains. Why?

2. Other (and Better) Evaluation Protocols

Is CLDC really the best extrinsic task we can come up with? Language understanding task seem much more interesting...

Further Work / Research Directions

1. Intrinsic Evaluation $\rightarrow$ Downstream Performance?

Performance on intrinsic evaluation datasets correlates with downstream tasks such as DST. However, substantial intrinsic gains do not always lead to large downstream gains. Why?

2. Other (and Better) Evaluation Protocols

Is CLDC really the best extrinsic task we can come up with? Language understanding task seem much more interesting...

3. Conflating and De-Conflating Word Vectors

Generalizing multi-prototype and multi-sense word embeddings to cross-lingual settings. How?

Further Work / Research Directions

1. Intrinsic Evaluation $\rightarrow$ Downstream Performance?

Performance on intrinsic evaluation datasets correlates with downstream tasks such as DST. However, substantial intrinsic gains do not always lead to large downstream gains. Why?

2. Other (and Better) Evaluation Protocols

Is CLDC really the best extrinsic task we can come up with? Language understanding task seem much more interesting...

3. Conflating and De-Conflating Word Vectors

Generalizing multi-prototype and multi-sense word embeddings to cross-lingual settings. How?

4. Combining Distributional and Resource-Based Information

Recent initiative with strong results. Lexical resources are abundant, we should not shy away from using them (e.g., BabelNet, PanLex)

Useful Software I

Simple Python implementation of the basic mapping approach

[Mikolov et al., arXiv 2013; Dinu et al., ICLR 2015]
http://clic.cimec.unitn.it/~georgiana.dinu/down/

Python implementation of an extended (iterative) mapping approach

[Artetxe et al., EMNLP 2016; Artetxe et al., ACL 2017]
https://github.com/artetxem/vecmap

CLWEs (mapping based on fastText vectors) for 78 languages

[Smith et al., ICLR 2017]
https://github.com/Babylonpartners/fastText_multilingual

SOTA (cross-lingual) semantic specialisation model (in TensorFlow) plus bilingual vectors for 52 languages

[Mrkšić et al., TACL 2017; Vulić et al., ACL 2017]
https://github.com/nmrksic/attract-repel

Useful Software II

BilBOWA model implementation

[Gouws et al., ICML 2015]
https://github.com/gouwsmeister/bilbowa

BiSkip model implementation and some vectors

[Luong et al., NAACL 2015]
https://nlp.stanford.edu/~lmthang/bivec/

multiCluster, multiCCA, multiSkip models plus some vectors

[Ammaer et al., arXiv 2016]
http://128.2.220.95/multilingual/data/

Test data and best practices for several tasks

[Upadhyay et al., ACL 2016]
https://github.com/shyamupa/biling-survey

Useful Software III

(Re)implementations of several CLWE models

[Bérard et al., LREC 2016]
https://github.com/eske/multivec

Bilingual training of multi-sense embeddings

[Šuster et al., NAACL 2016]
https://github.com/rug-compling/bimu

CCA-based model implementation

[Faruqui and Dyer, EACL 2014]
https://github.com/mfaruqui/crosslingual-cca

A CLWE which exploits bilingual dictionaries

[Duong et al., EMNLP 2016]
https://github.com/longdt219/XlingualEmb

Cross-Lingual Space of Thankyous

Book in preparation: Morgan & Claypool Synthesis Lectures

The three of us + Sebastian Ruder

References I


Oliver Adams, Adam Makarucha, Graham Neubig, Steven Bird, and Trevor Cohn. 2017.
Cross-lingual word embeddings for low-resource language modeling.
In *EACL*. pages 937–947.


Željko Agić, Dirk Hovy, and Anders Søgaard. 2015.
If all you have is a bit of the Bible: Learning POS taggers for truly low-resource languages.
In *ACL*. pages 268–272.


Željko Agić, Anders Johannsen, Barbara Plank, Héctor Martínez Alonso, Natalie Schluter, and Anders Søgaard. 2016.
Multilingual projection for parsing truly low-resource languages.
Transactions of the Association for Computational Linguistics 4:301–312.


Waleed Ammar, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A. Smith. 2016.
Massively multilingual word embeddings.
CoRR abs/1602.01925.

References II


Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2016.
Learning principled bilingual mappings of word embeddings while preserving monolingual invariance.
In *EMNLP*. pages 2289–2294.


Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017.
Learning bilingual word embeddings with (almost) no bilingual data.
In *ACL*. pages 451–462.


Alexandre Berard, Christophe Servan, Olivier Pietquin, and Laurent Besacier. 2016.
MultiVec: A multilingual and multilevel representation learning toolkit for NLP.
In *LREC*.


Shane Bergsma and Benjamin Van Durme. 2011.
Learning bilingual lexicons using the visual similarity of labeled web images.
In *IJCAI*. pages 1764–1769.


Chloé Braud, Maximin Coavoux, and Anders Søgaard. 2017.
Cross-lingual RST discourse parsing.
In *EACL*. pages 292–304.

References III


Sarath A.P. Chandar, Stanislas Lauly, Hugo Larochelle, Mitesh M. Khapra, Balaraman Ravindran, Vikas C. Raykar, and Amrita Saha. 2014.
An autoencoder approach to learning bilingual word representations.
In *NIPS*. pages 1853–1861.


Jocelyn Coulmance, Jean-Marc Marty, Guillaume Wenzek, and Amine Benhalloum. 2015.
Trans-gram, fast cross-lingual word embeddings.
In *EMNLP*. pages 1109–1113.


Mathieu Dehouck and Pascal Denis. 2017.
Delexicalized word embeddings for cross-lingual dependency parsing.
In *EACL*. pages 241–250.


Georgiana Dinu, Angeliki Lazaridou, and Marco Baroni. 2015.
Improving zero-shot learning by mitigating the hubness problem.
In *ICLR Workshop Papers*.


Long Duong, Hiroshi Kanayama, Tengfei Ma, Steven Bird, and Trevor Cohn. 2016.
Learning crosslingual word embeddings without bilingual corpora.
In *EMNLP*. pages 1285–1295.

References IV


Long Duong, Hiroshi Kanayama, Tengfei Ma, Steven Bird, and Trevor Cohn. 2017.
Multilingual training of crosslingual word embeddings.
In *EACL*. pages 894–904.


Greg Durrett, Adam Pauls, and Dan Klein. 2012.
Syntactic transfer using a bilingual lexicon.
In *EMNLP*. pages 1–11.


Manaal Faruqui and Chris Dyer. 2013.
An information theoretic approach to bilingual word clustering.
In *ACL*. pages 777–783.


Manaal Faruqui and Chris Dyer. 2014.
Improving vector space word representations using multilingual correlation.
In *EACL*. pages 462–471.


Éric Gaussier, Jean-Michel Renders, Irina Matveeva, Cyril Goutte, and Hervé Déjean. 2004.
A geometric view on bilingual lexicon extraction from comparable corpora.
In *ACL*. pages 526–533.

References V


Stephan Gouws, Yoshua Bengio, and Greg Corrado. 2015.
BilBOWA: Fast bilingual distributed representations without word alignments.
In *ICML*. pages 748–756.


Stephan Gouws and Anders Søgaard. 2015.
Simple task-specific bilingual word embeddings.
In *NAACL-HLT*. pages 1386–1390.


Jiang Guo, Wanxiang Che, David Yarowsky, Haifeng Wang, and Ting Liu. 2015.
Cross-lingual dependency parsing based on distributed representations.
In *ACL*. pages 1234–1244.


Aria Haghighi, Percy Liang, Taylor Berg-Kirkpatrick, and Dan Klein. 2008.
Learning bilingual lexicons from monolingual corpora.
In *ACL*. pages 771–779.


Bradley Hauer, Garrett Nicolai, and Grzegorz Kondrak. 2017.
Bootstrapping unsupervised bilingual lexicon induction.
In *EACL*. pages 619–624.


Karl Moritz Hermann and Phil Blunsom. 2014a.
Multilingual distributed representations without word alignment.
In *ICLR*.

References VI


Karl Moritz Hermann and Phil Blunsom. 2014b.
Multilingual models for compositional distributed semantics.
In *ACL*. pages 58–68.


Geert Heyman, Ivan Vulić, and Marie-Francine Moens. 2017.
Bilingual lexicon induction by learning to combine word-level and character-level representations.
In *EACL*. pages 1085–1095.


Kejun Huang, Matt Gardner, Evangelos Papalexakis, Christos Faloutsos, Nikos Sidiropoulos, Tom Mitchell, Partha P. Talukdar, and Xiao Fu. 2015.
Translation invariant word embeddings.
In *EMNLP*. pages 1084–1088.


Anders Johannsen, Héctor Martínez Alonso, and Anders Søgaard. 2015.
Any-language frame-semantic parsing.
In *EMNLP*. pages 2062–2066.


Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. 2012.
Inducing crosslingual distributed representations of words.
In *COLING*. pages 1459–1474.

References VII


Tomáš Kočiský, Karl Moritz Hermann, and Phil Blunsom. 2014.
Learning bilingual word representations by marginalizing alignments.
In *ACL*. pages 224–229.


Angeliki Lazaridou, Georgiana Dinu, and Marco Baroni. 2015a.
Hubness and pollution: Delving into cross-space mapping for zero-shot learning.
In *ACL*. pages 270–280.


Angeliki Lazaridou, Nghia The Pham, and Marco Baroni. 2015b.
Combining language and vision with a multimodal skip-gram model.
In *NAACL-HLT*. pages 153–163.


Omer Levy, Anders Søgaard, and Yoav Goldberg. 2017.
A strong baseline for learning cross-lingual word embeddings from sentence alignments.
In *EACL*. pages 765–774.


Thang Luong, Hieu Pham, and Christopher D. Manning. 2015.
Bilingual word representations with monolingual quality in mind.
In *Workshop on Vector Space Modeling for Natural Language Processing*. pages 151–159.

References VIII


Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013a.
Exploiting similarities among languages for machine translation.
CoRR abs/1309.4168.


Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013b.
Distributed representations of words and phrases and their compositionality.
In *NIPS*. pages 3111–3119.


David M. Mimno, Hanna M. Wallach, Jason Naradowsky, David A. Smith, and Andrew McCallum. 2009.
Polylingual topic models.
In *EMNLP*. pages 880–889.


Aditya Mogadala and Achim Rettinger. 2016.
Bilingual word embeddings from parallel and non-parallel corpora for cross-language text classification.
In *NAACL-HLT*. pages 692–702.


Nikola Mrkšić, Diarmuid Ó Séaghdha, Tsung-Hsien Wen, Blaise Thomson, and Steve Young. ????
Neural Belief Tracker: Data-driven dialogue state tracking.
In *ACL*.

References IX


Nikola Mrkšić, Ivan Vulić, Diarmuid Ó Séaghdha, Ira Leviant, Roi Reichart, Milica Gašić, Anna Korhonen, and Steve Young. 2017.
Semantic specialisation of distributional word vector spaces using monolingual and cross-lingual constraints.
Transactions of the ACL .


Yves Peirsman and Sebastian Padó. 2010.
Cross-lingual induction of selectional preferences with bilingual vector spaces.
In *NAACL*. pages 921–929.


Janarthanan Rajendran, Mitesh M. Khapra, Sarath Chandar, and Balaraman Ravindran. 2016.
Bridge correlational neural networks for multilingual multimodal representation learning.
In *NAACL-HLT*. pages 171–181.


Sebastian Ruder. 2017.
A survey of cross-lingual embedding models.
CoRR abs/1706.04902.


Tianze Shi, Zhiyuan Liu, Yang Liu, and Maosong Sun. 2015.
Learning cross-lingual word embeddings via matrix co-factorization.
In *ACL*. pages 567–572.

References X


Samuel L. Smith, David H. P. Turban, Steven Hamblin, and Nils Y. Hammerla. 2017.
Offline bilingual word vectors, orthogonal transformations and the inverted softmax .


Anders Søgaard, Željko Agić, Héctor Martínez Alonso, Barbara Plank, Bernd Bohnet, and Anders Johannsen. 2015.
Inverted indexing for cross-lingual NLP.
In *ACL*. pages 1713–1722.


Hubert Soyer, Pontus Stenetorp, and Akiko Aizawa. 2015.
Leveraging monolingual data for crosslingual compositional word representations.
In *ICLR*.


Oscar Täckström, Ryan T. McDonald, and Jakob Uszkoreit. 2012.
Cross-lingual word clusters for direct transfer of linguistic structure.
In *NAACL-HLT*. pages 477–487.


Chen-Tse Tsai and Dan Roth. 2016.
Cross-lingual wikification using multilingual embeddings.
In *NAACL-HLT*. pages 589–598.

References XI


Shyam Upadhyay, Manaal Faruqui, Chris Dyer, and Dan Roth. 2016.
Cross-lingual models of word embeddings: An empirical comparison.
In *ACL*. pages 1661–1670.


Simon Šuster, Ivan Titov, and Gertjan van Noord. 2016.
Bilingual learning of multi-sense embeddings with discrete autoencoders.
In *NAACL-HLT*. pages 1346–1356.


Ivan Vulić. 2017.
Cross-lingual syntactically informed distributed word representations.
In *EACL*. pages 408–414.


Ivan Vulić, Douwe Kiela, Stephen Clark, and Marie-Francine Moens. 2016.
Multi-modal representations for improved bilingual lexicon learning.
In *ACL*. pages 188–194.


Ivan Vulić and Anna Korhonen. 2016.
On the role of seed lexicons in learning bilingual word embeddings.
In *ACL*. pages 247–257.

References XII


Ivan Vulić and Marie-Francine Moens. 2013.
A study on bootstrapping bilingual vector spaces from non-parallel data (and nothing else).
In *EMNLP*. pages 1613–1624.


Ivan Vulić and Marie-Francine Moens. 2015a.
Bilingual word embeddings from non-parallel document-aligned data applied to bilingual lexicon induction.
In *ACL*. pages 719–725.


Ivan Vulić and Marie-Francine Moens. 2015b.
Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings.
In *SIGIR*. pages 363–372.


Ivan Vulić and Marie-Francine Moens. 2016.
Bilingual distributed word representations from document-aligned comparable data.
Journal of Artificial Intelligence Research 55:953–994.

References XIII


Ivan Vulić, Nikola Mrkšić, and Anna Korhonen. 2017.
Cross-lingual induction and transfer of verb classes based on word vector space specialisation.
In *EMNLP*.


Ivan Vulić, Wim De Smet, and Marie-Francine Moens. 2011.
Identifying word translations from comparable corpora using latent topic models.
In *ACL Short Papers*. pages 479–484.


Yogarshi Vyas and Marine Carpuat. 2016.
Sparse bilingual word representations for cross-lingual lexical entailment.
In *NAACL-HLT*. pages 1187–1197.


Min Xiao and Yuhong Guo. 2014.
Distributed word representation learning for cross-lingual dependency parsing.
In *CoNLL*. pages 119–129.


Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. 2015.
Normalized word embedding and orthogonal transform for bilingual word translation.
In *NAACL-HLT*. pages 1006–1011.

References XIV


Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017.
Adversarial training for unsupervised bilingual lexicon induction.
In *ACL*. pages 1959–1970.


Yuan Zhang, David Gaddy, Regina Barzilay, and Tommi Jaakkola. 2016.
Ten pairs to tag – Multilingual POS tagging via coarse mapping between embeddings.
In *NAACL-HLT*. pages 1307–1317.


Will Y. Zou, Richard Socher, Daniel Cer, and Christopher D. Manning. 2013.
Bilingual word embeddings for phrase-based machine translation.
In *EMNLP*. pages 1393–1398.