Semi-supervised Learning for Multilingual Sentence Representation

Hedi Ben-younes - Alexandre Ramé

March 2016

Outline

We need to know what is being said on the web
$\Rightarrow$ understand web pages.

- Images,
- Links to other pages,
- **Texts**
- ...

Understand text = being able to represent documents in a semantic space.

Problem: text comes in different languages
We want to build a system that can compute similarity between texts:

- in different languages
- only based on semantics

$\Rightarrow$ We need a multilingual text embedding.

Two main problems are presented today:

1. How can we represent words and sentences using raw natural monolingual text ?

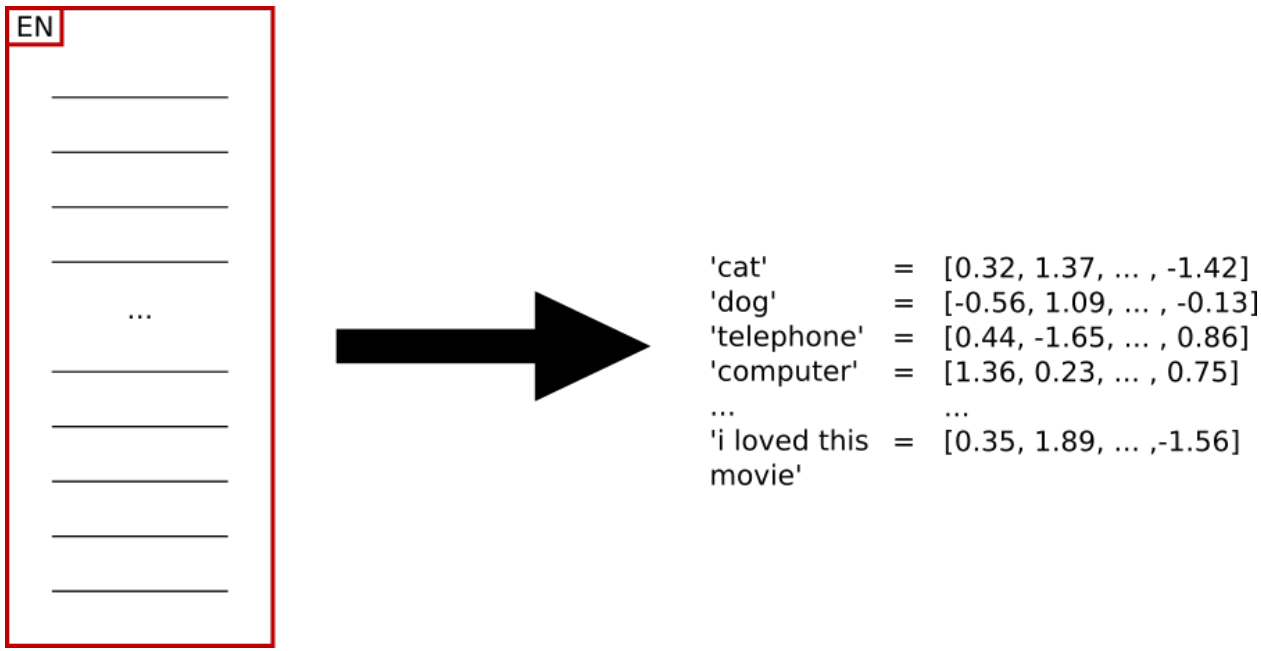


Figure: Monolingual embedding

2. How can we align multiple embedding spaces and project them into a unique multilingual semantic space ?

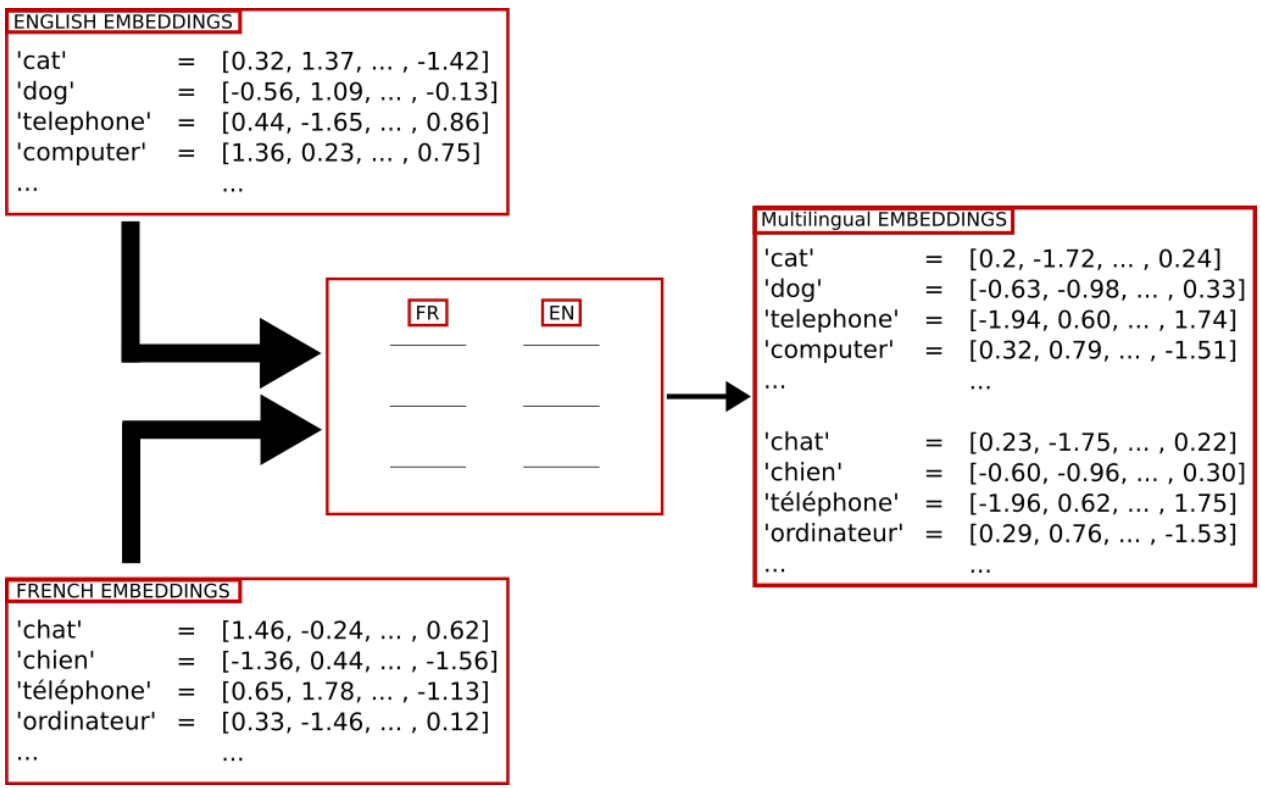


Figure: Multilingual embedding

How can we get a multilingual text embedding from

- a lot of raw unlabelled text,
- and a few aligned data ?

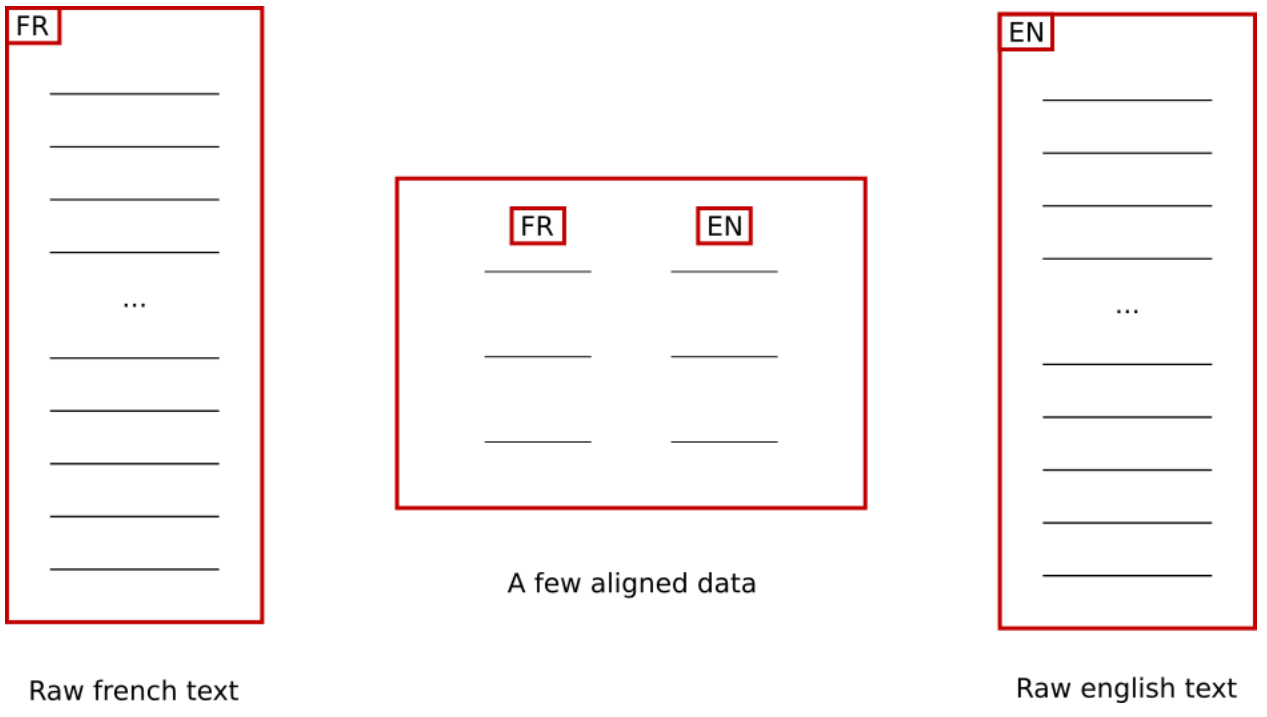


Figure: Textual data

Outline

Goal

Create as many monolingual embeddings as the number of languages.

How can we represent words and sentences in a semantic space ?

Word2Vec [Mikolov et al., 2013]

Word embeddings

- compact representation
- learnt from raw natural text
- close vectors = similar words
- linearize semantic relations

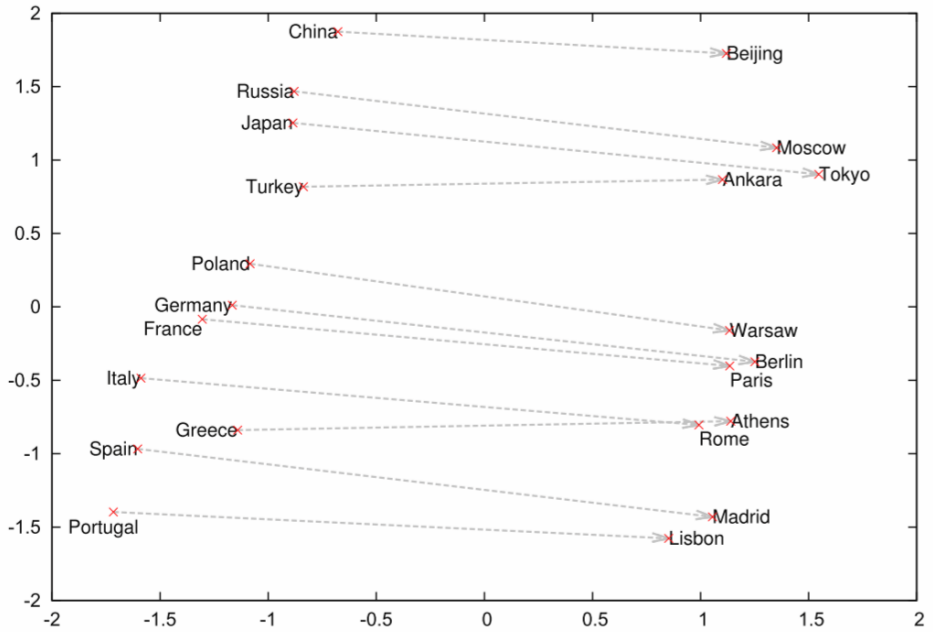


Figure: PCA of countries and their capital

Word2Vec [Mikolov et al., 2013] - Skip-gram

Given a word, we try to predict which words are nearby

$$p(w_O|w_I) = \frac{\exp\left(v'^{\top}_{w_O} v_{w_I}\right)}{\sum_{w=1}^{W} \exp\left(v'^{\top}_{w} v_{w_I}\right)}$$

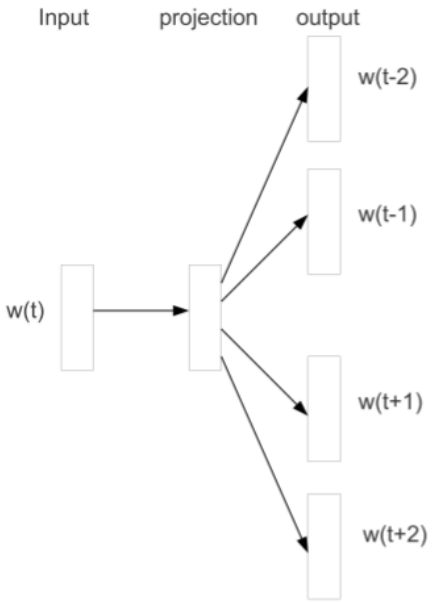


Figure: Skip-gram architecture

- The parameters of the model are the embeddings themselves
- Model trained with SGD
- Negative sampling for large vocabularies.

Sentence embeddings

How do we go from **word** embeddings to **sentence** embeddings ?
First solution: bag-of-word2vec

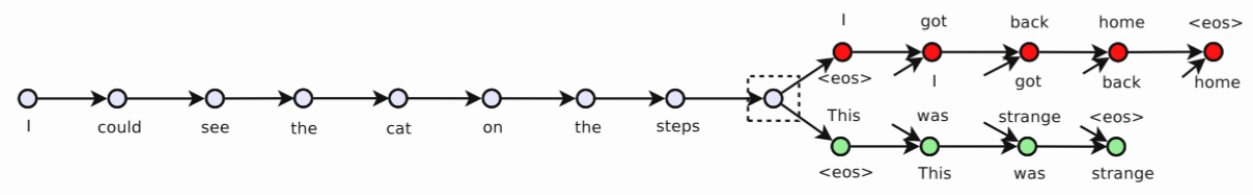
$$S = [w_1, ..., w_n]$$

$$vec(S) = \sum_{i=1}^{n} word2vec(w_i)$$

Skip-thought vectors [Kiros et al., 2015]

Directly learn a sentence embedding

- learn how to compose word embeddings to get a task-independant sentence embedding,
- Applies the Skip-Gram idea to the sentence level: given a sentence, we try to generate the previous and the next sentence



- Uses a list of consecutive sentences
- One GRU-encoder and two GRU-decoders

FastSent [Hill et al., 2016]

Simplified version of Skip-thought vectors

- Learns source u_w and target v_w embeddings
- Encoder: $\mathbf{s_i} = \sum_{w \in S_i} u_w$

Maximize

$$\sum_{w \in S_{i-1} \bigcup S_{i+1}} \Phi(\mathbf{s_i}, v_w)$$

where Φ is a softmax over the vocabulary

Conclusion

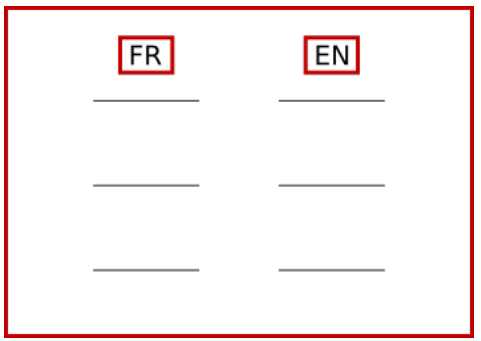
Now we can:

- represent words
- represent sentences
  - sum of Word2Vecs
  - Skip-thought vectors
  - FastSent
- our current sentence embedding is based upon a sum of **disambiguated** Word2Vecs
- we are currently switching to FastSent

Outline

Up to now, we have one embedding model per language.
We want to find projections from monolingual spaces to a multilingual space.
To do so, we need:

- aligned multi-lingual data

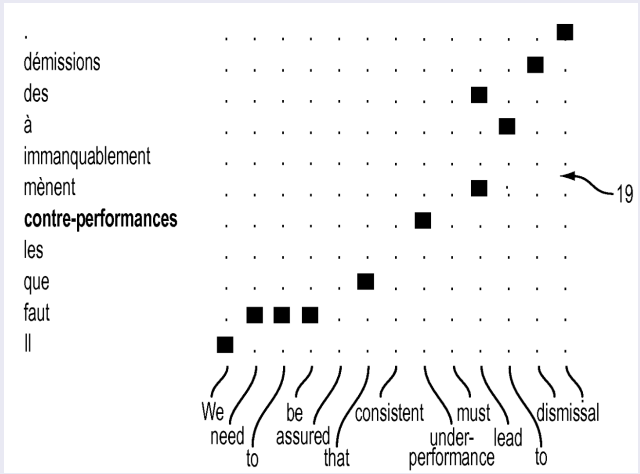


- an alignment model

Alignment Data

European Languages

From Europarl (dataset of aligned sentences) to aligned words with GIZA++ ([Koehn et al., 2007])




```
subprogrammes ||| programmes ||| 0.724452
eczema ||| eczéma ||| 1
curricula ||| programmes ||| 0.663559
propolis ||| propolis ||| 1
vaidere ||| vaidere ||| 0.791188
witnesses ||| violation ||| 1
violation ||| violation ||| 1
geografica ||| geografica ||| 1
programs ||| programmes ||| 0.830379
reemphasised ||| réaffirmée ||| 1
matador ||| matador ||| 1
edgardo ||| edgardo ||| 1
raucous ||| bruyante ||| 1
creativity ||| créativité ||| 0.927491
```

Other Languages

Most common english words translated into Chinese, Russian, ...

Canonical Correlation Analysis

CCA:

- proposed in [Hotelling, 1936]
- very well explained in [Hardoon et al., 2004]

Problem

Given two matrices $X \in \mathbb{R}^{n \times d_x}$ and $Y \in \mathbb{R}^{n \times d_y}$ representing the observed values x and y of paired centered multivariate random variables,
we seek for two projections $W \in \mathbb{R}^{d_x \times p}$ and $V \in \mathbb{R}^{d_y \times p}$ such as

- $\forall i \in [1, p], < W_{:,i}, x >$ and $< V_{:,i}, y >$ are highly correlated
- the dimensions in the target space are uncorrelated

Canonical Correlation Analysis

Let

$$\Sigma_{xy} = X'Y \in \mathbb{R}^{d_x \times d_y}$$
$$\Sigma_{xx} = X'X \in \mathbb{R}^{d_x \times d_x}$$
$$\Sigma_{yy} = Y'Y \in \mathbb{R}^{d_y \times d_y}$$

We find these projections by solving:

$$\max_{W,V} \mathrm{Tr}(W\Sigma_{xy}V) \qquad s.c \begin{cases} W'\Sigma_{xx}W = I \\ V'\Sigma_{yy}V = I \\ W'\Sigma_{xy}V * (\mathbf{1} - I) = \mathbf{0} \end{cases} \tag{1}$$

$\Rightarrow$ yields a close form

Canonical Correlation Analysis

To go further on CCA:

- Kernel CCA [Hardoon et al., 2004]
- Deep CCA [Andrew et al., 2013]

In practice, CCA doesn't scale for large alignment datasets:

- Need to go through the whole dataset to compute Σ_{kl}
- Kernel version requires storing / inverting Graam matrices

Correlational Network [Chandar et al., 2015]

CCA + Auto Encoder => Correlational Network

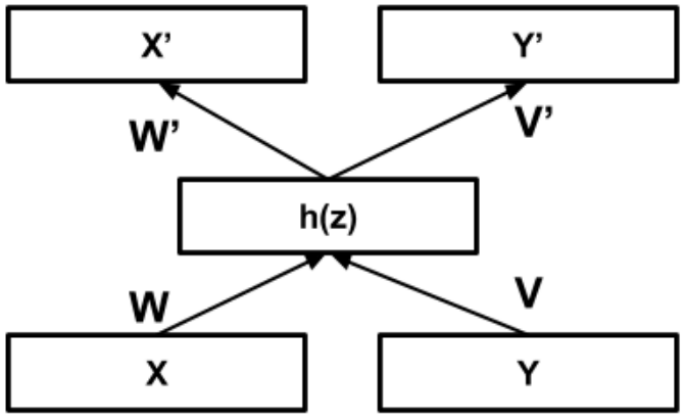


Figure: Correlational Network Architecture

The loss function of Correlational Network is the sum of:

- correlation of projections in the common subspace
- self-reconstruction terms
- cross-reconstruction terms

Correlational Network [Chandar et al., 2015]

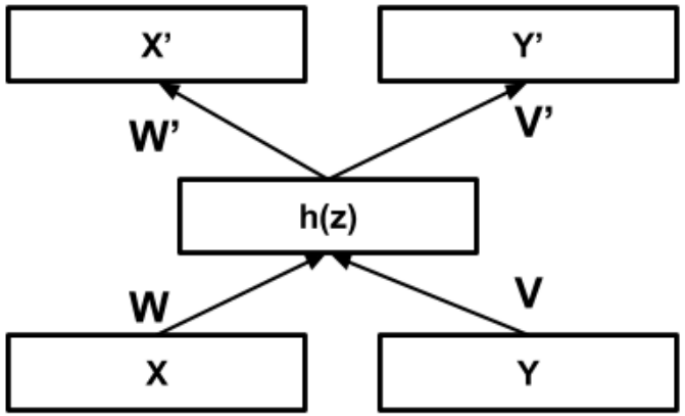


Figure: Correlational Network Architecture

$$\mathcal{J}_{\mathcal{Z}}(\theta) = \sum_{i=1}^{N} L((x_i, y_i), g(h((x_i, y_i)))) + L((x_i, y_i), g(h(x_i)))$$
$$+ L((x_i, y_i), g(h(y_i))) - \lambda corr(h(X), h(Y))$$

$$corr(h(X), h(Y)) = \frac{\sum_{i=1}^{N} <h(x_i) - \overline{h(X)} | h(y_i) - \overline{h(Y)}>}{\sqrt{\sum_{i=1}^{N} (h(x_i) - \overline{h(X)})^2 \sum_{i=1}^{N} (h(y_i) - \overline{h(Y)})^2}}$$

Hyperparameters

Hyperparameters

- λ of the correlation term
- mini-batch size
- regularization
- size hidden layer
- activation function: tanh
- add terms in the loss function
- tied or not
- contrastive term

Deep Correlational Network

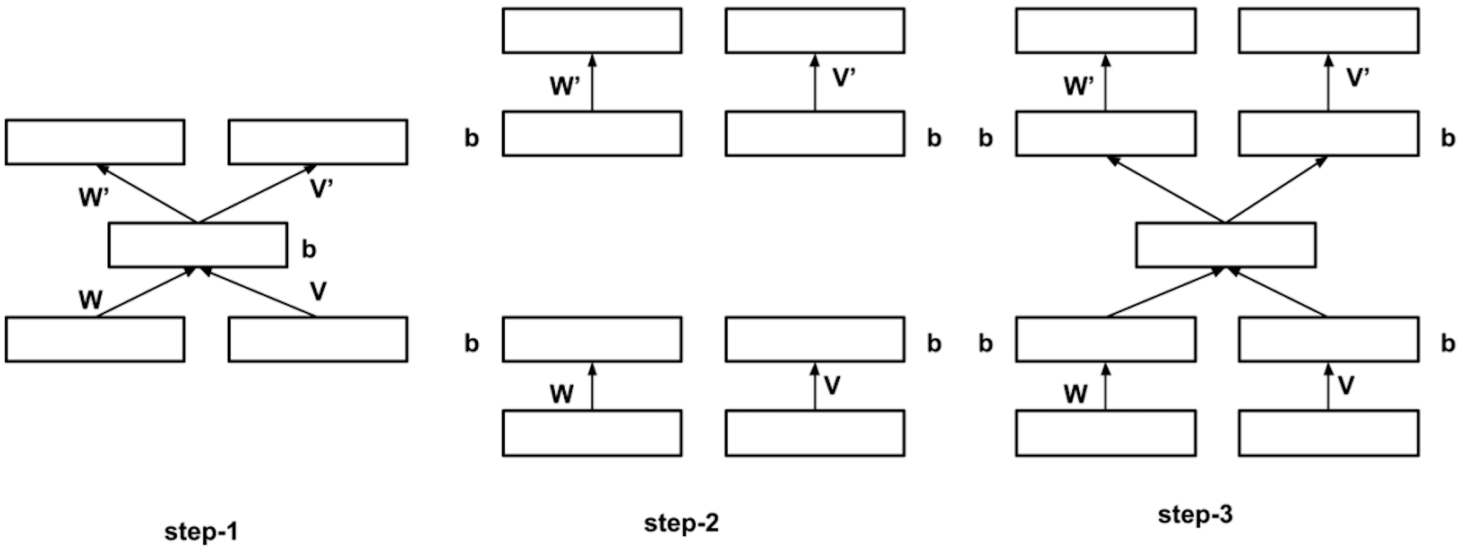


Figure: Deep Correlational Network Architecture

Training Procedure: similar to greedy layerwise pretraining of deep autoencoders

Bridge Correlational Network [Rajendran et al., 2015]

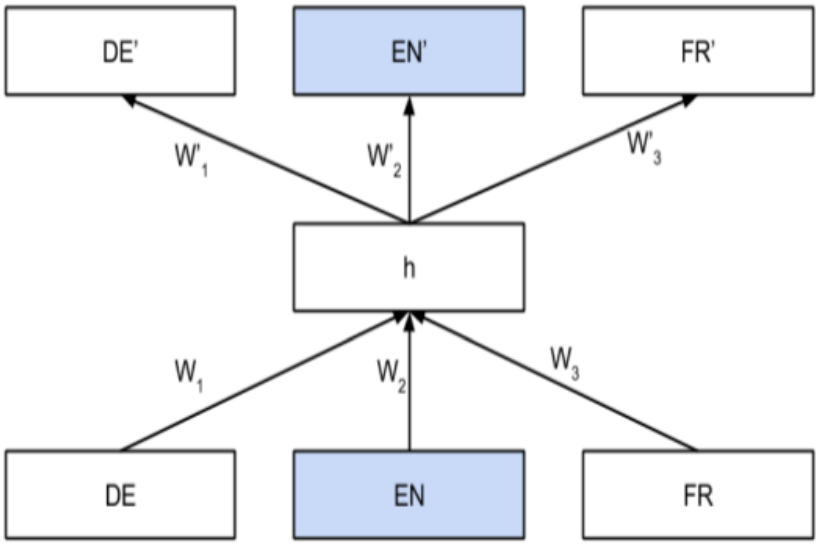


Figure: Bridge Correlational Network Architecture

- one pivot language
- multiple languages

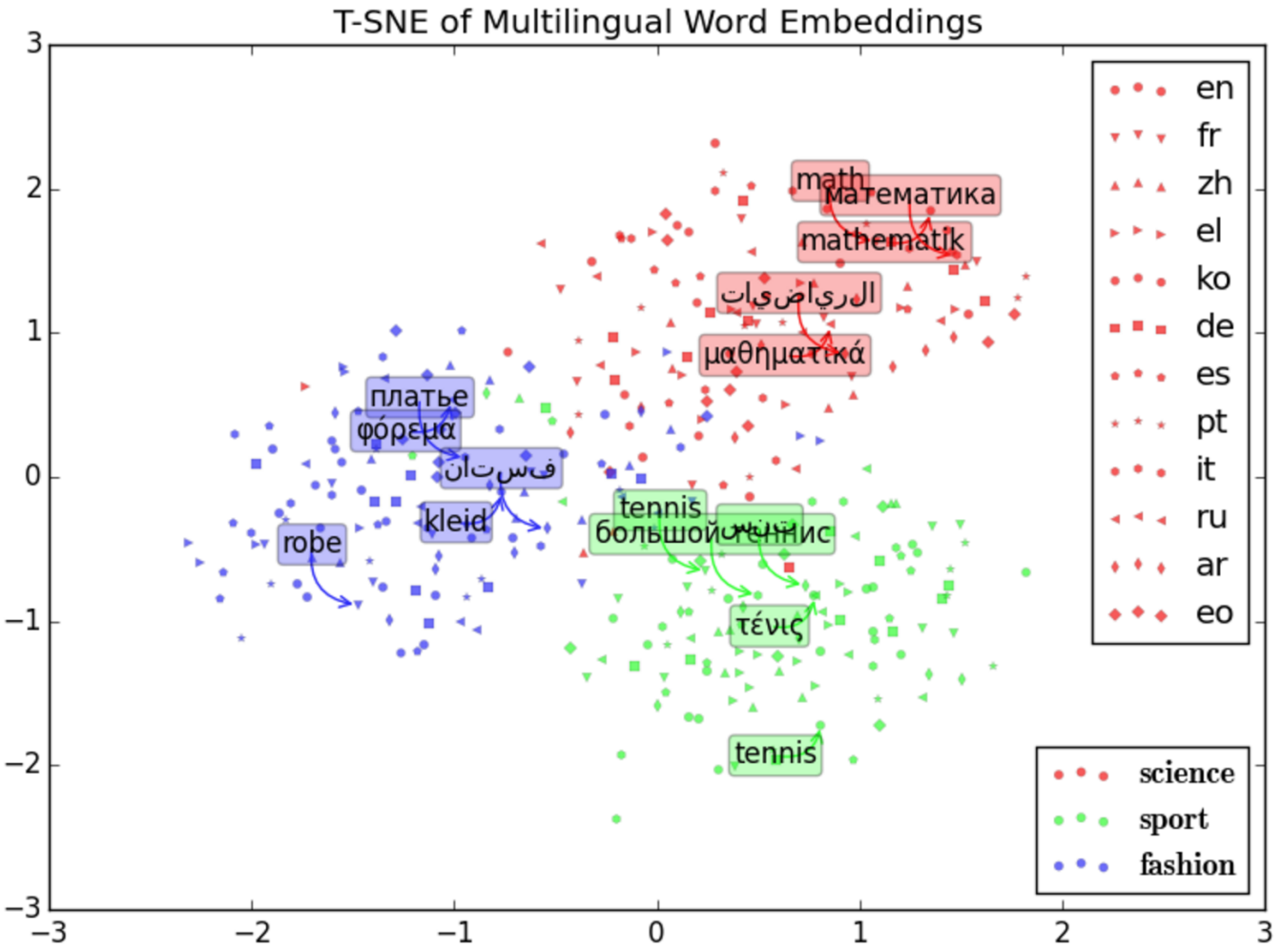
Outline

Transliteration Equivalence

English	French	Deutsch	Spanish	Greek	Chinese
machine learning:1.0	apprentissage automatique:0.83	algorithmen:0.81	computacional:0.82	επιστήμη υπολογιστών:0.78	模式 :0.76
pattern recognition :0.89	algorithmique:0.80	datenanalyse:0.79	aprendizaje automático:0.79	δηλωτική:0.75	算机 :0.75
algorithms:0.87	algorithmes:0.79	datenstrukturen:0.75	algoritmos:0.79	συνδυαστική:0.75	建模:0.73
data mining:0.87	modélisation:0.79	mustererkennung :0.75	redes neuronales:0.76	αλγορίθμων:0.74	化:0.73
natural language processing :0.86	heuristiques:0.78	numerische mathematik:0.75	aprendizaje automático:0.76	συνδυαστικής:0.74	形式化:0.71
Arab	**Korean**	**Russian**	**Italian**	**Portuguese**	**Esperanto**
كهوَرزميَت:0.78	응용:0.75	алгоритмов:0.81	computazionale:0.80	computacional:0.81	komputiko :0.74
تمدهج:0.78	데이터마이닝 :0.74	алгоритмы:0.79	computazionali:0.79	computacionais:0.80	komputado:0.73
اَلهَسوبي :0.78	알고리즘과:0.73	моделирования:0.78	algoritmi:0.77	algoritmos:0.78	komputiko:0.72
اَلهَسوبي :0.74	신경망:0.70	вычислений:0.77	statistica:0.77	otimização:0.76	algoritmoj:0.71
هَسوبي:0.74	최적화:0.70	многомерных:0.77	computazione:0.76	linguagens de programação :0.76	analitiko:0.71

The top 5 closest words to 'machine learning' in cosine similarity

T-SNE


T-SNE of Multilingual Word Embeddings
math
математика
mathematik
الرياضيات
μαθηματικά
платье
φόρεμα
فستان
kleid
robe
tennis
большой теннис
τένις
tennis
en
fr
zh
el
ko
de
es
pt
it
ru
ar
eo
science
sport
fashion

English Word Similarity Evaluation

smart	stupid	5.81
wood	forest	7.73
money	cash	9.15
king	queen	8.58
king	rook	5.92

OPEC	country	5.63
day	summer	3.94
day	dawn	7.53
country	citizen	7.31
planet	people	5.75

Figure: Wordsim353 similarity and relatedness

English Model	WS-Similarity	WS-Relatedness	WS
Our SGNS[1] Model in the English space	0.742	0.621	0.67
Our SGNS[1] Model in the multilingual space	**0.765**	0.627	0.685
SGNS[1]	0.737[2]-0.798[3]	0.592[2]-0.700[3]	
Glove	0.651[2]-0.746[3]	0.541[2]-0.643[3]	
Swivel	0.748[2]	0.616[2]	

Spearman's ρ correlation on English Word Similarity tasks

[1] Skip Gram Negative Sampling
[2] According to [Shazeer et al., 2016]
[3] According to [Levy et al., 2015]

Other Languages' Word Similarity Evaluation

```
intelligent stupide 5.81
bois forêt 7.73
argent monnaie 9.15
roi reine 8.58
roi tour 5.92
```

```
聪明的 愚蠢的 5.81
木材 林 7.73
钱 货币 9.15
景 女王 8.58
景 圆 5.92
```

Figure: French and Chinese Wordsim

Lang	WS	Lang	WS
en	**0.685**	fr	0.594
de	0.603	eo	0.584
ru	0.617	es	0.597
pt	0.587	ko	0.573
it	0.596	zh	0.563
el	0.562	ar	0.566

Spearman's ρ correlation on Word Similarity task for different languages

Bilingual Word Similarity Evaluation

```
smart stupide 5.81
intelligent stupid 5.81
king reine 8.58
roi queen 8.58
```

```
聪明的 stupide 5.81
intelligent 笨 5.81
投资回报率 reine 8.58
roi 女王 8.58
```

Figure: English-French and Chinese-French Wordsim

Langs	WS	Langs	WS
en/fr	**0.613**	fr/fr	0.594
de/fr	0.554	eo/fr	0.526
ru/fr	0.561	es/fr	0.578
pt/fr	0.542	ko/fr	0.524
it/fr	0.574	zh/fr	0.513
el/fr	0.524	ar/fr	0.518

Spearman's ρ Correlation on Bilingual Word Similarity task for different languages with French

Sentence Similarity Evaluation: SICK

SICK(Sentences Involving Compositional Knowledge) dataset

- "A group of children is playing in the house and there is no man standing in the background"
- "A group of kids is playing in a yard and an old man is standing in the background"
- similarity=3.2

Model	Sick
Our English Model	0.59
Skip-thought	0.57[1]
FastSent	0.61[1]

Spearman's ρ correlation on Sentence Similarity task

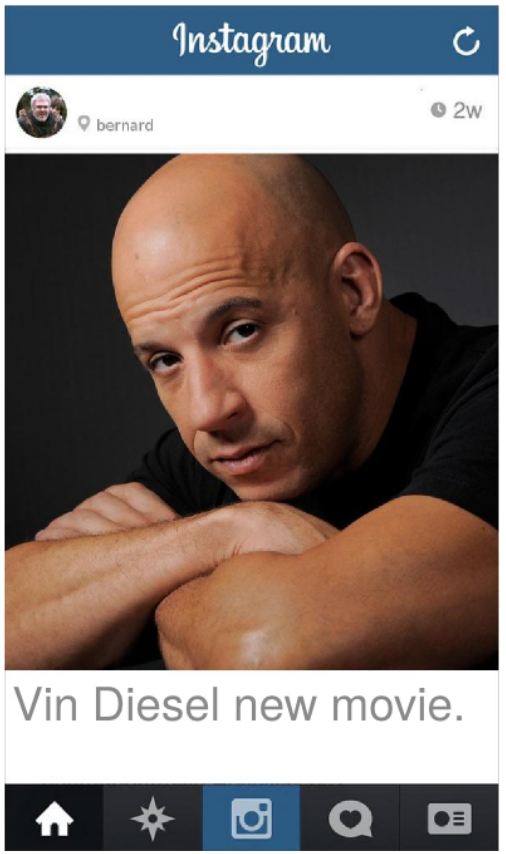
[1] According to [Hill et al., 2016]

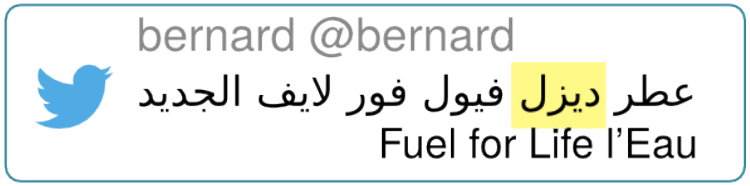
Sentence Multilingual Application

- Transfer Learning
  - Cross Language Topic Classification
  - Cross Language Moderation
- Multilingual Information Retrieval
- Named Entity Recognition: detection of persons, brands, topics, ...
- Multilingual Disambiguation

Multilingual Search Engine

Diesel (fragrance)





Conclusion

Summary

We are able to represent in a **unique** semantic space:

- words and sentences
- in multiple languages (12 up to now)

=> understanding of texts from the web

Ongoing Works

- Character-aware embeddings
- Add images in our multimodal space

Bibliographie I

Andrew, G., Arora, R., Livescu, K., and Bilmes, J. (2013).
Deep canonical correlation analysis.
In *International Conference on Machine Learning (ICML)*, Atlanta, Georgia.

Chandar, S., Khapra, M. M., Larochelle, H., and Ravindran, B. (2015).
Correlational neural networks.
Neural computation.

Hardoon, D. R., Szedmak, S. R., and Shawe-taylor, J. R. (2004).
Canonical correlation analysis: An overview with application to learning methods.
Neural Comput., 16(12):2639–2664.

Hill, F., Cho, K., and Korhonen, A. (2016).
Learning distributed representations of sentences from unlabelled data.
CoRR, abs/1602.03483.

Hotelling, H. (1936).
Relations between two sets of variates.
Biometrika, 28:321–377.

Kiros, R., Zhu, Y., Salakhutdinov, R., Zemel, R., Torralba, A., Urtasun, R., and Fidler, S. (2015).
Skip-thought vectors.
In *NIPS*.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., et al. (2007).
Moses: Open source toolkit for statistical machine translation.
In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, pages 177–180. Association for Computational Linguistics.

Bibliographie II

Levy, O., Goldberg, Y., and Dagan, I. (2015).
Improving distributional similarity with lessons learned from word embeddings.
Transactions of the Association for Computational Linguistics, 3:211–225.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013).
Distributed representations of words and phrases and their compositionality.
In *NIPS 2013*, pages 3111–3119.

Rajendran, J., Khapra, M. M., Chandar, S., and Ravindran, B. (2015).
Bridge correlational neural networks for multilingual multimodal representation learning.
arXiv preprint arXiv:1510.03519.

Shazeer, N., Doherty, R., Evans, C., and Waterson, C. (2016).
Swivel: Improving embeddings by noticing what's missing.
arXiv preprint arXiv:1602.02215.