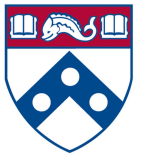
Cross-lingual Word Representations

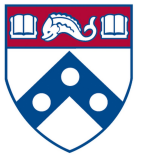
Shyam Upadhyay

CIS 530 – Computational Linguistics (Guest Lecture)

Who is this guy?



- 5th year PhD student with Prof. Dan Roth.
- Recently defended my thesis.
  - Thesis Title: Exploiting Cross-lingual Representations for NLP.
- Working on cross-lingual NLP problems since 2015-ish.

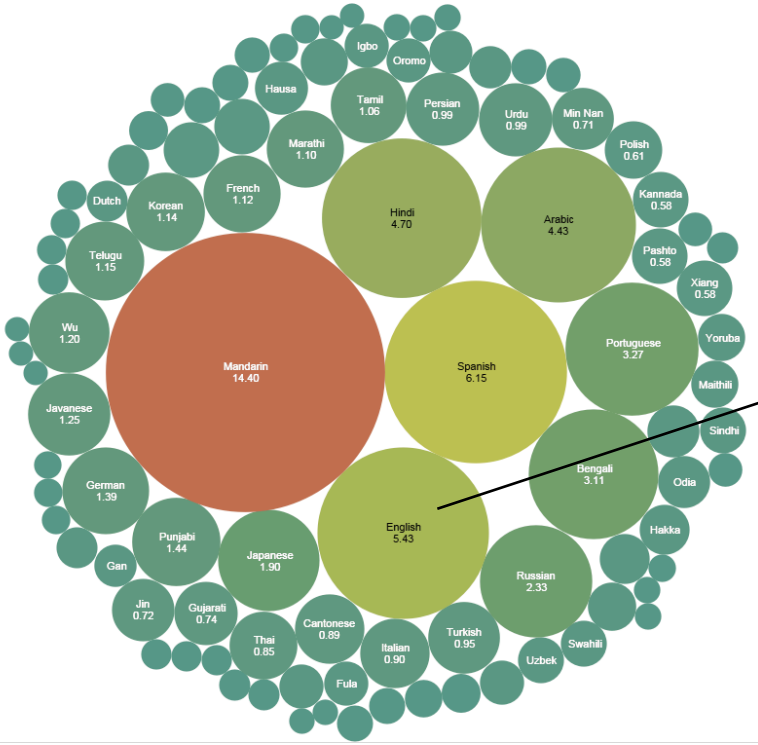


Today's theme: NLP for other languages

Why should I care about other languages?

Natural Language Processing != "English Language Processing"

- 25 languages with >50M native speakers.
- >40% of the Web is **NOT** English.
- Yet most NLP resources are available predominantly in English.



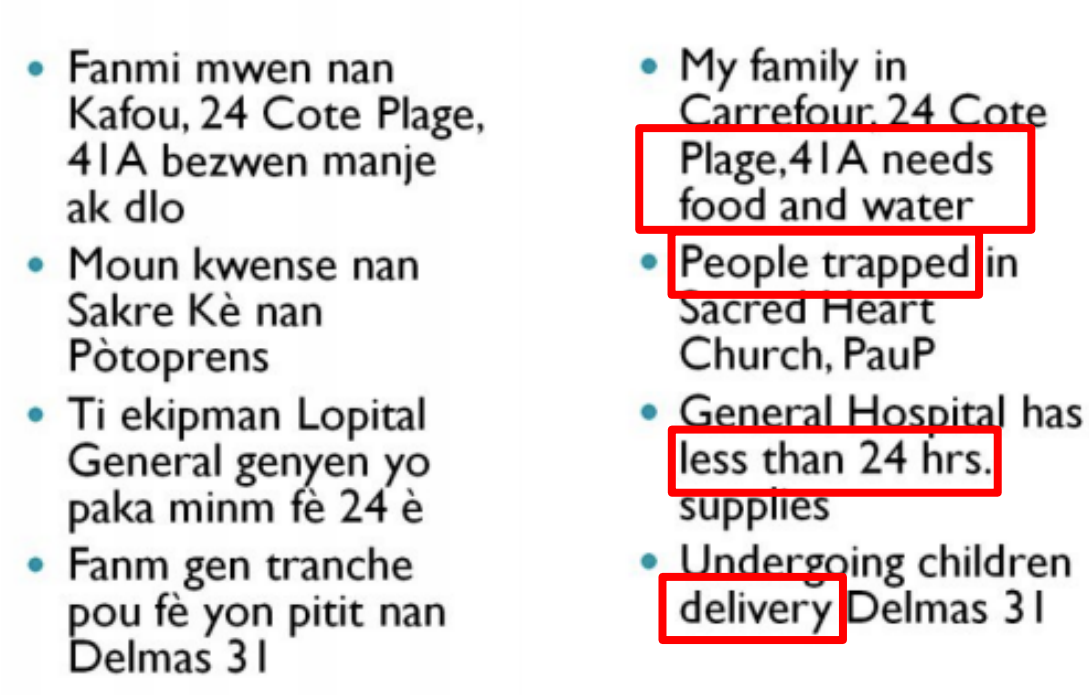
English 5% native speakers

https://en.wikipedia.org/wiki/Languages_used_on_the_Internet
https://en.wikipedia.org/wiki/List_of_languages_by_number_of_native_speakers

Why should I care about other languages?

Social Impact

- Early Warning Systems, Disaster relief, Response to unexpected events.



Example from
Yulia Tsvetkov's JSALT slides (CMU)

Improving NLP in General

- Test-beds for evaluating existing NLP techniques used in English.

Example: Document Classification in French

French

?

- Science
- Politics
- Sports
- Entertainment

Let's see how one would solve this in English first…

Document Classification in English

English

"NASA"

"shuts"

"down"

"Mars"

"rover"

We need to represent a document using features of smaller units (the words).

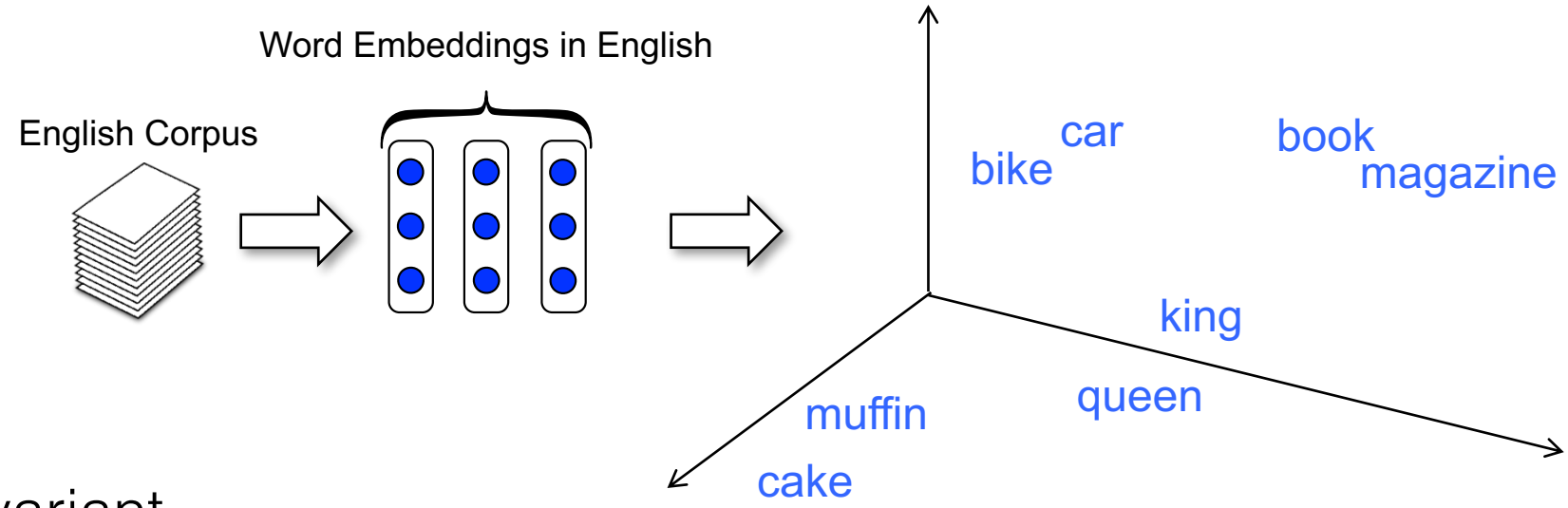
Science

Politics

Sports

Entertainment

(Monolingual) Word Representations



- Task Invariant.
  - NER, Dependency Parsing, POS tagging, SRL …
- Easy to train.
  - Fast implementations like word2vec
  - Raw text extremely easy to find.
- Compact representations.
  - Large vocabularies can be represented using 100 dimensional vectors.

A Simple Document Classifier

English

"NASA"
"shuts"
"down"
"Mars"
"rover"

Document's
Feature vector

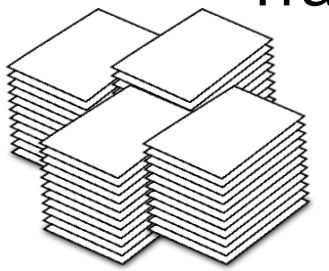
Learnable
Weight Vector
(one for each class)

Science
Politics
Sports
Entertainment

Training Data
$(\mathbf{x}, \mathbf{y})$

Back to building the French classifier …

Attempt 1: Traditional Approach



Training Data in English

$(\mathbf{x}, \mathbf{y})$

Feature Extraction $\mathbf{x} \rightarrow \Phi(\mathbf{x})$

Learning Algorithm $\mathcal{A} : (\Phi(\mathbf{x}), \mathbf{y}) \rightarrow \mathbf{M}$

Learnt Model $\mathbf{M} : \Phi(\mathbf{x}) \rightarrow \mathbf{y}$

Training Data in French

$(\mathbf{x}, \mathbf{y})$

$\mathbf{x} \rightarrow \Phi(\mathbf{x})$

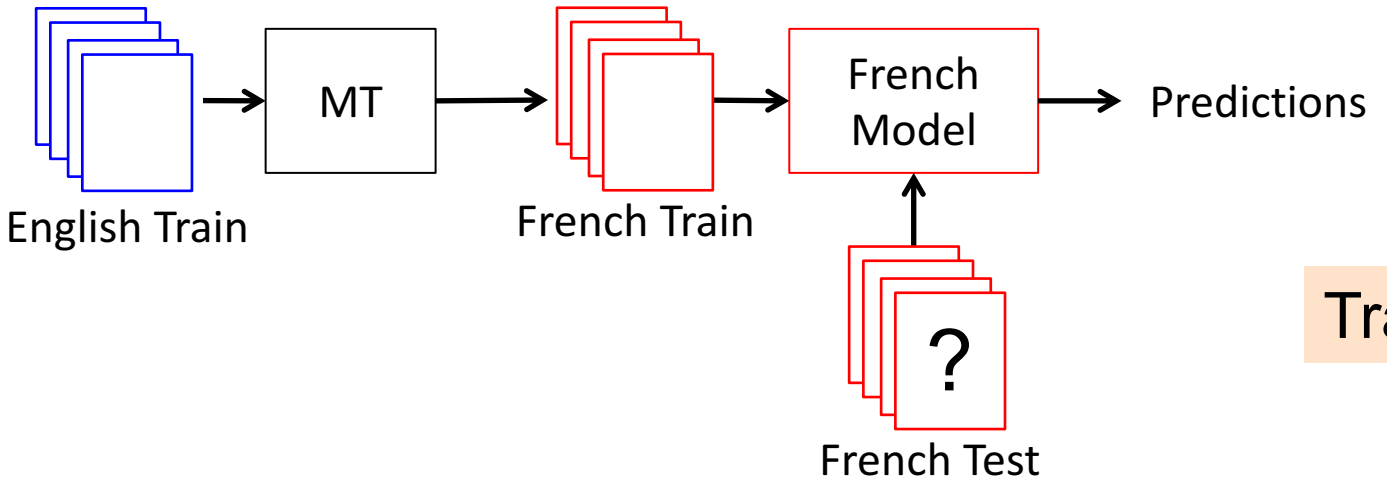
$\mathcal{A} : (\Phi(\mathbf{x}), \mathbf{y}) \rightarrow \mathbf{M}$

$\mathbf{M} : \Phi(\mathbf{x}) \rightarrow \mathbf{y}$

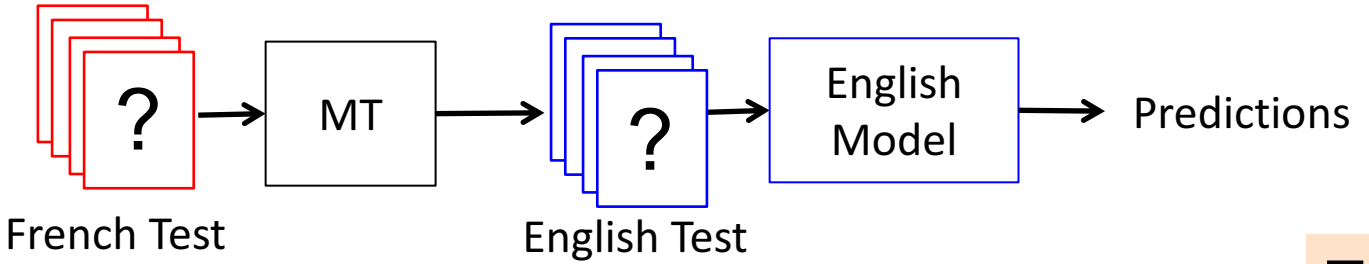
Limitations of Attempt 1

- No sharing of information between languages.
  - E.g., if “NASA” appears in a document ...
- Cannot work with little supervision in French
  - Need to annotate French documents from scratch ...
  - Worst case – No supervision in French.

Attempt 2: Machine Translation (MT)



Translate Train



Translate Test

Limitations of Attempt 2

- Good MT models are hard to develop
  - Expensive resources, training etc.
  - Prevents fast scaling to new languages.
  - MT is unlikely to be perfect … mismatch in train and test conditions.

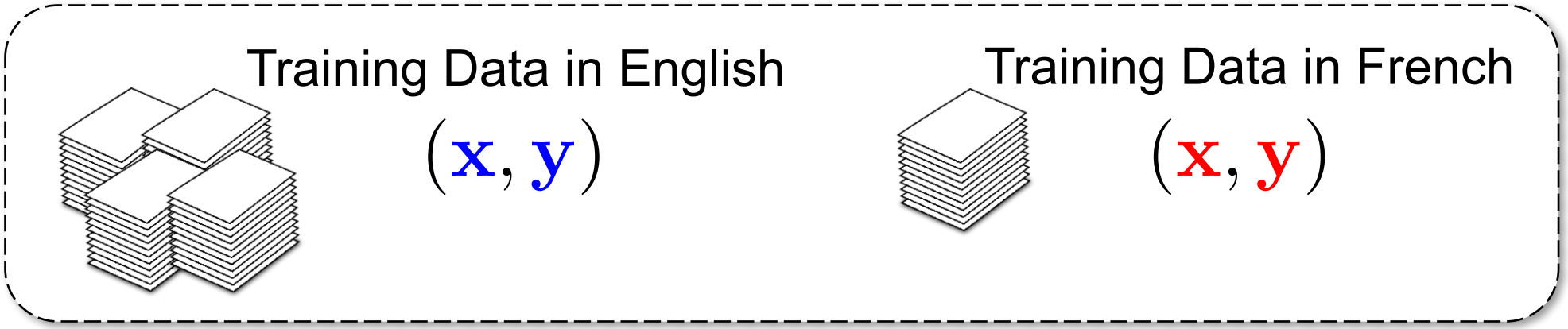
- Do you really need MT for something like document classification?

Elon Musk: le vaisseau spatial de nouvelle génération
de SpaceX sera construit et lancé en Floride

Downing street bittet Brüssel förmlich um Verspätung,
da May auf einen neuen Deal drängen will

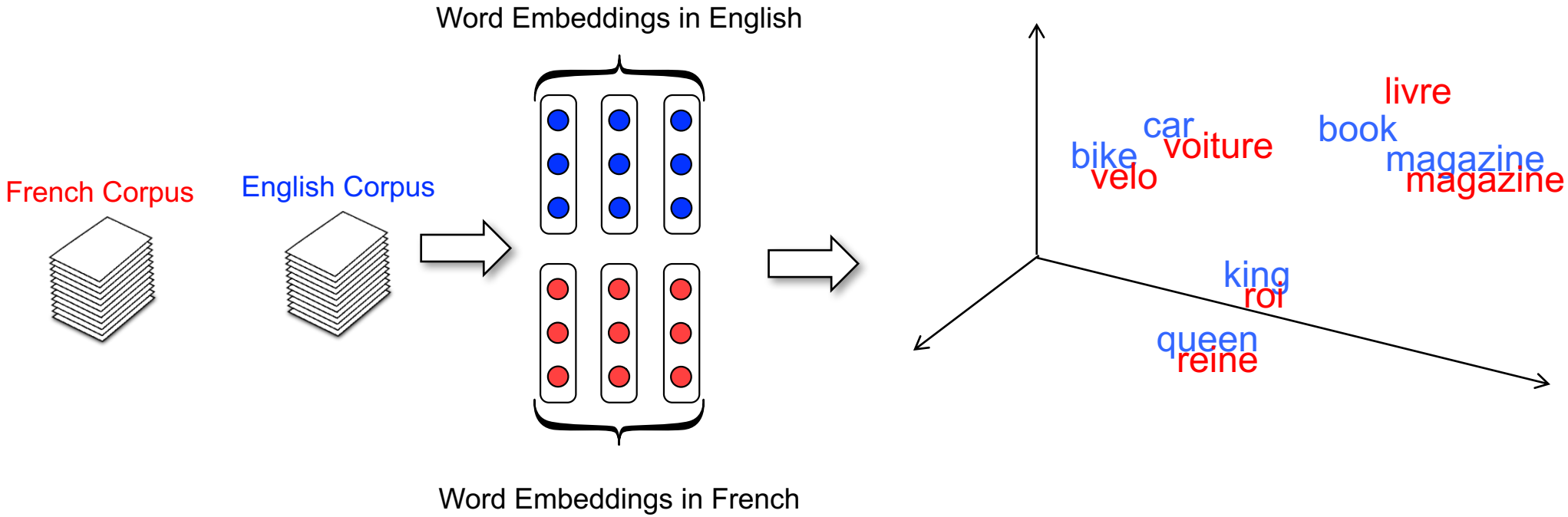
Ideally …

- We should combine all available supervision, and share information across languages (what we learn for English could be applied to French)



To share supervision, we need features (word embeddings) that are language-invariant.

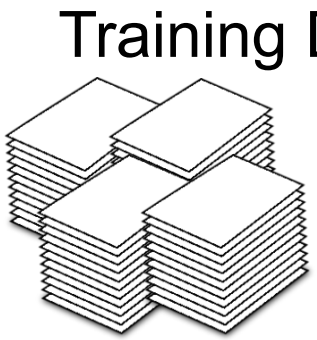
Language Invariant Word Representations



With language invariant representations,
the vector for book and livre look (almost) identical to a classifier.

Building NLP Models using Language-invariant (Cross-lingual) Spaces

Direct Model Transfer (through Shared Cross-lingual Space)



Training Data in English

$(\mathbf{x}, \mathbf{y})$

Φ

$(\Phi(\mathbf{x}), \mathbf{y})$

Test Data in French

$(\mathbf{x}, \mathbf{y})$

Φ

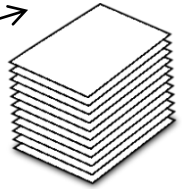
$(\Phi(\mathbf{x}), \mathbf{y})$

Shared Cross-lingual Space

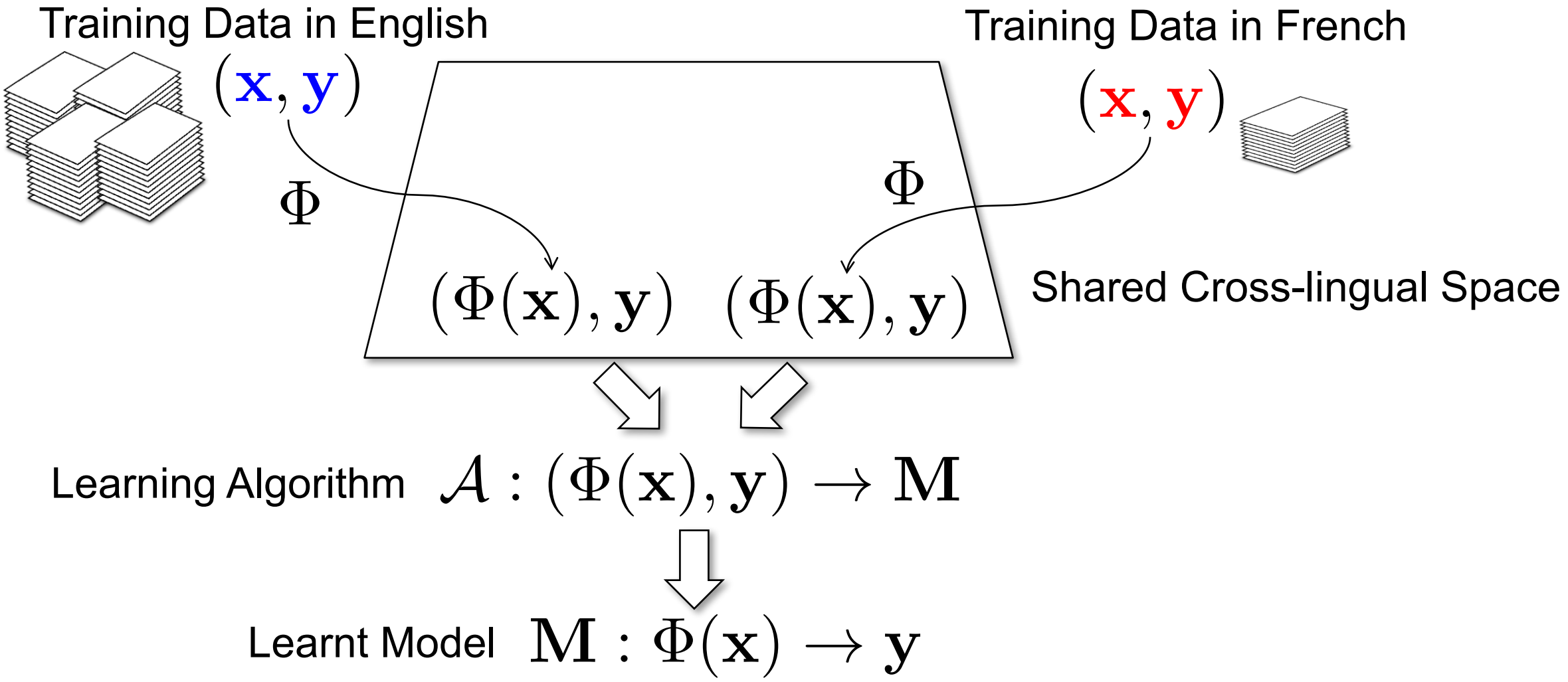
Learning Algorithm $\mathcal{A} : (\Phi(\mathbf{x}), \mathbf{y}) \rightarrow \mathbf{M}$

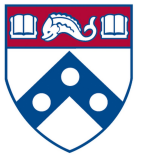
Learnt Model $\mathbf{M} : \Phi(\mathbf{x}) \rightarrow \mathbf{y}$

$\hat{\mathbf{y}}$



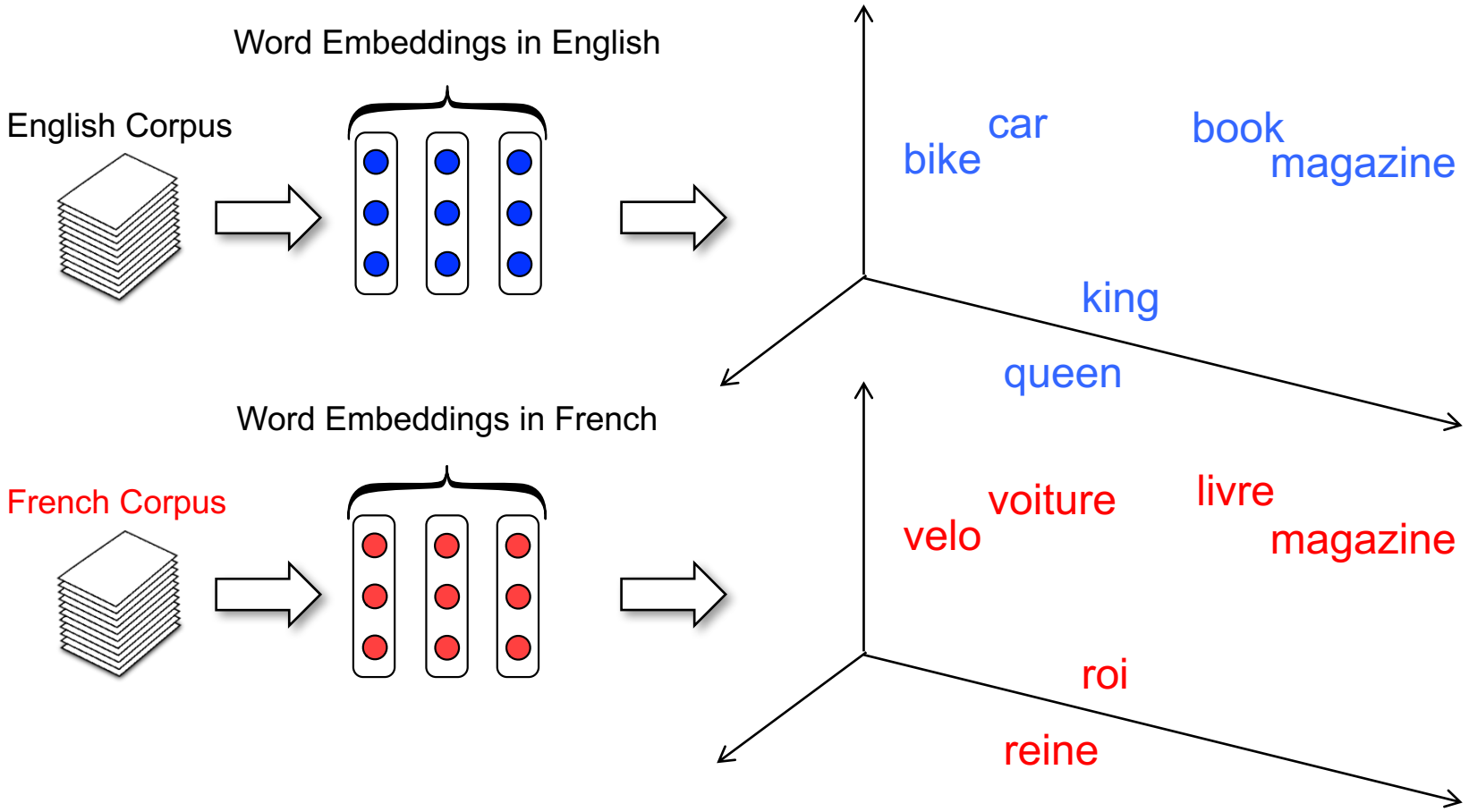
Joint Training (through Shared Cross-lingual Space)



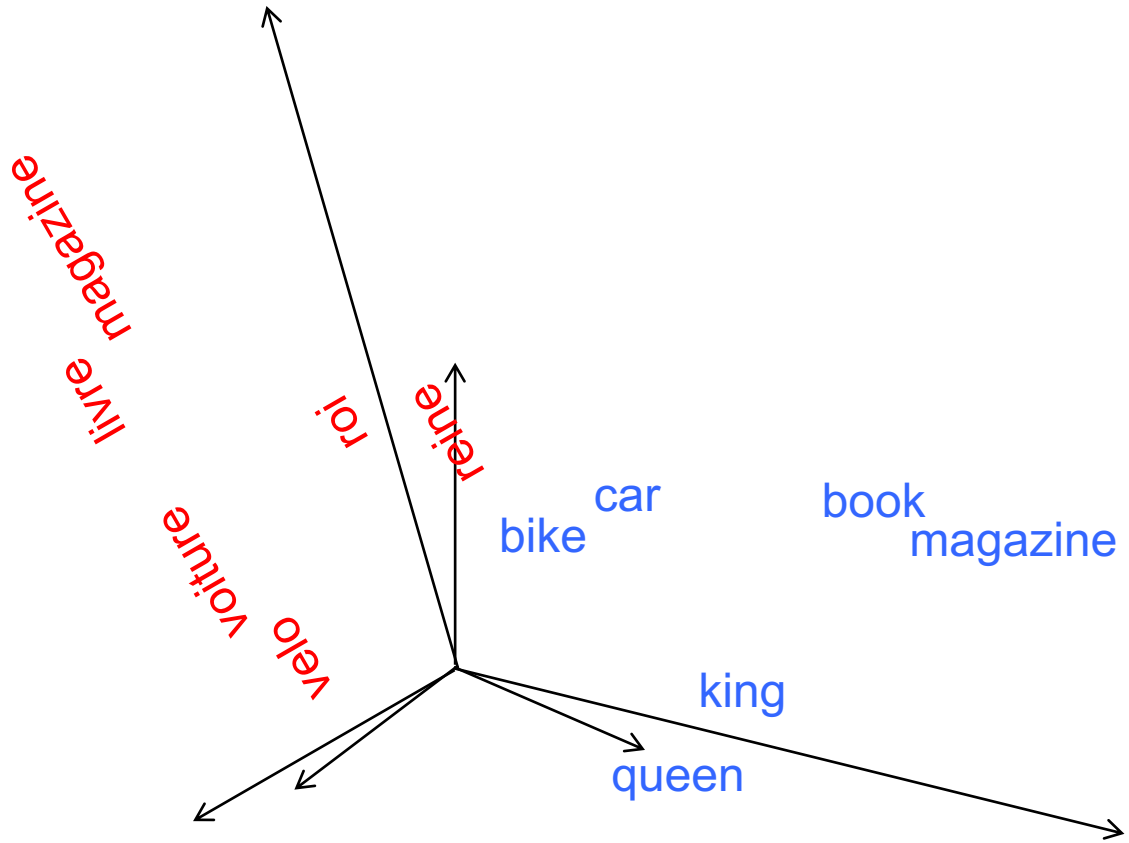


How to learn cross-lingual spaces?

What's wrong with this?



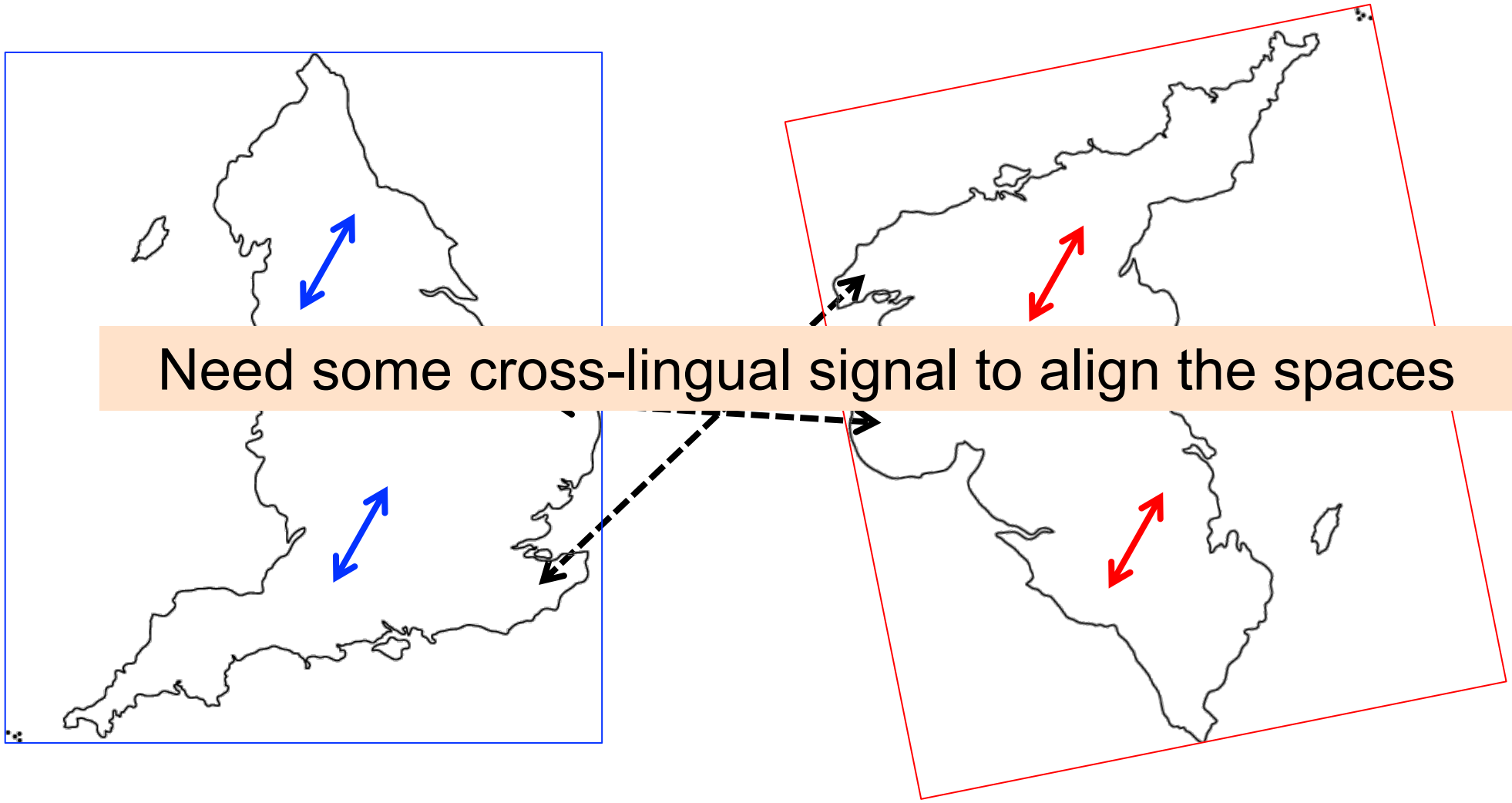
Things are more complicated …



Limitations

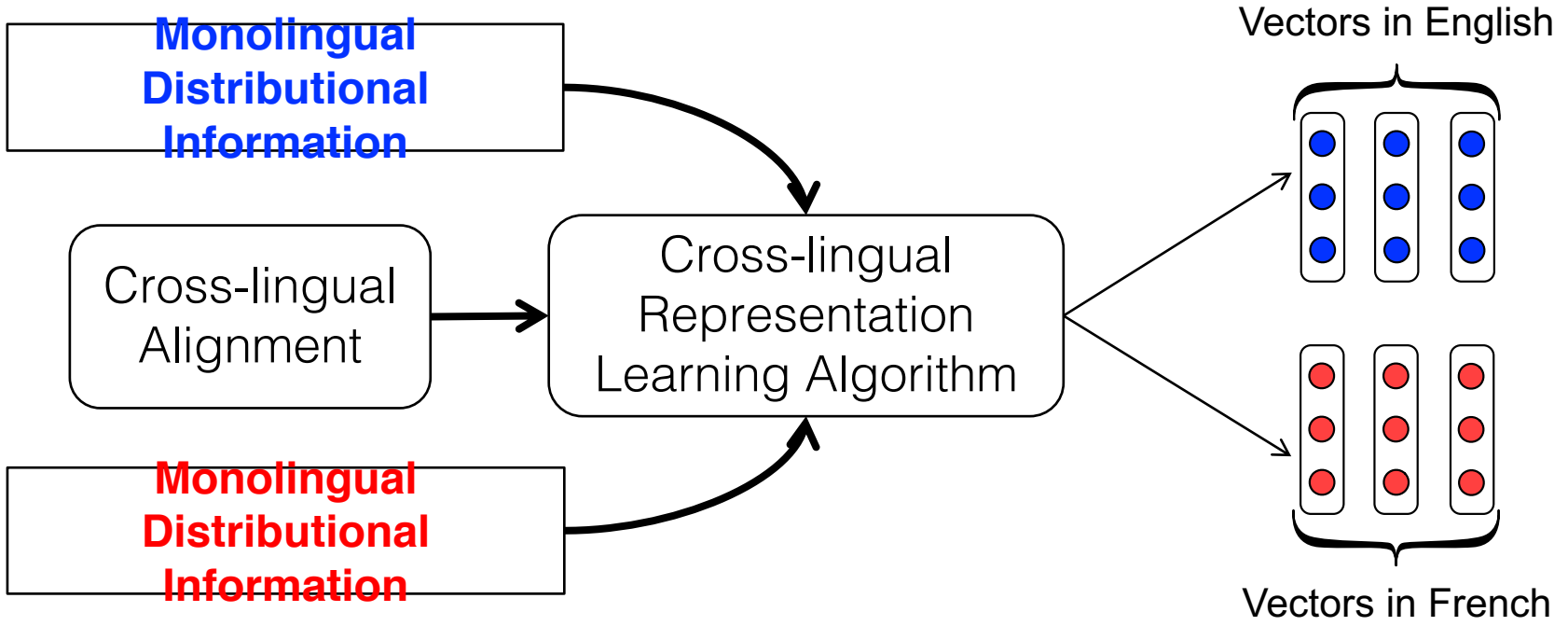
- Unable to capture related-ness between words *across languages*.
- Book and Livre will look very different to the classifier, defeating the purpose.

Another way to think about it …



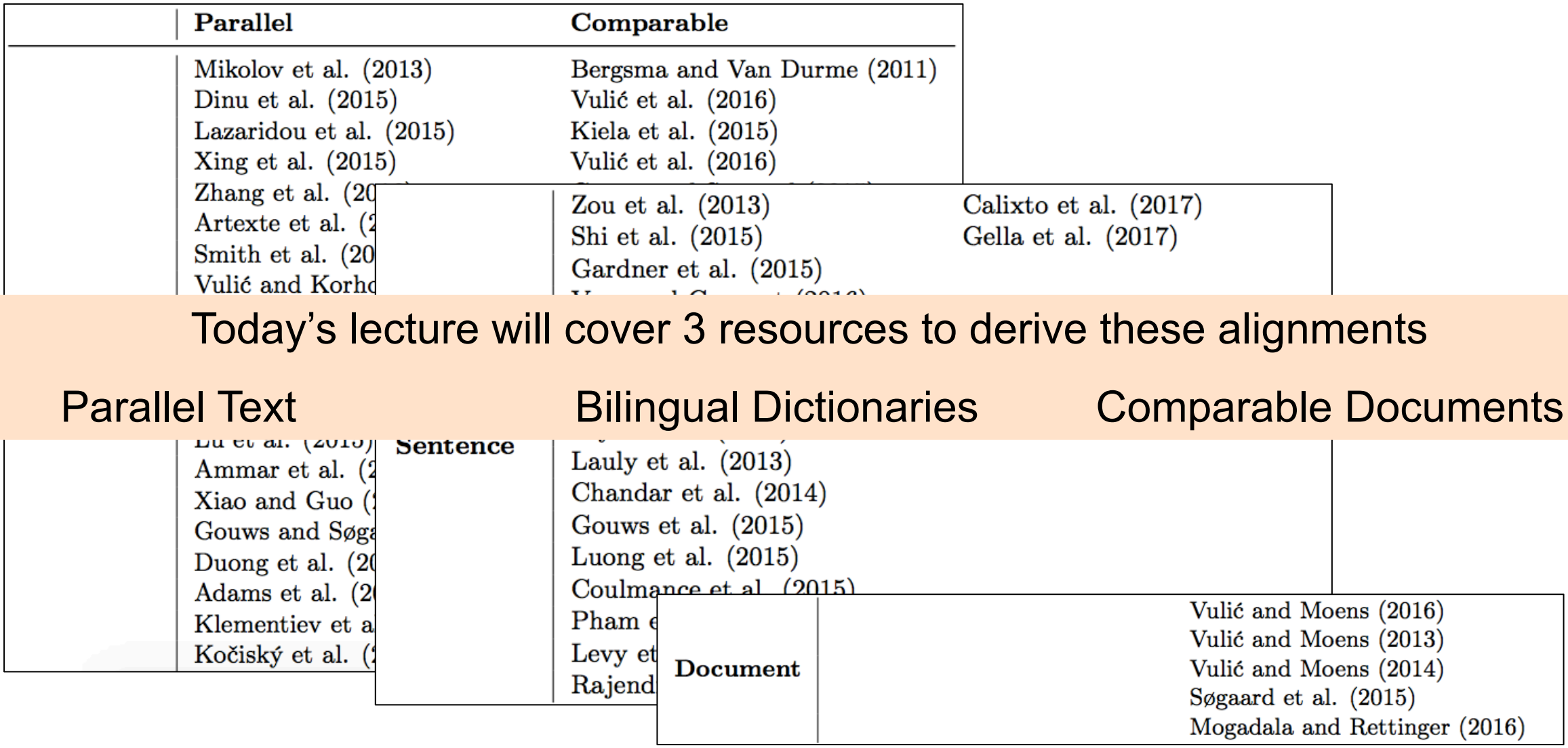
from https://www.samtalksml.net/aligning-vector-representations/

General Idea



$$\underbrace{\mathbf{W}^*}_{\text{Vectors for English}}, \overbrace{\mathbf{V}^*}^{\text{Vectors for French}} \leftarrow \operatorname{argmin} \underbrace{\alpha A(\mathbf{W})}_{\text{Mono. Obj. for English}} + \overbrace{\beta B(\mathbf{V})}^{\text{Mono. Obj. for French}} + \underbrace{C(\mathbf{W}, \mathbf{V})}_{\text{Cross-lingual Obj.}}$$

Different Forms of Cross-lingual Alignments



	Parallel	Comparable
	Mikolov et al. (2013)	Bergsma and Van Durme (2011)
	Dinu et al. (2015)	Vulić et al. (2016)
	Lazaridou et al. (2015)	Kiela et al. (2015)
	Xing et al. (2015)	Vulić et al. (2016)
	Zhang et al. (20	
	Artexte et al. (2	
	Smith et al. (20	
	Vulić and Korhc	

	Zou et al. (2013)	Calixto et al. (2017)
	Shi et al. (2015)	Gella et al. (2017)
	Gardner et al. (2015)	

	Lu et al. (2015)	
	Ammar et al. (2	
	Xiao and Guo (	
	Gouws and Søga	
	Duong et al. (20	
	Adams et al. (2	
	Klementiev et a	
	Kočiský et al. (	
Sentence	Lauly et al. (2013)	
	Chandar et al. (2014)	
	Gouws et al. (2015)	
	Luong et al. (2015)	
	Coulmance et al. (2015)	
	Pham e	
	Levy et	
	Rajend	

Document		Vulić and Moens (2016)
		Vulić and Moens (2013)
		Vulić and Moens (2014)
		Søgaard et al. (2015)
		Mogadala and Rettinger (2016)

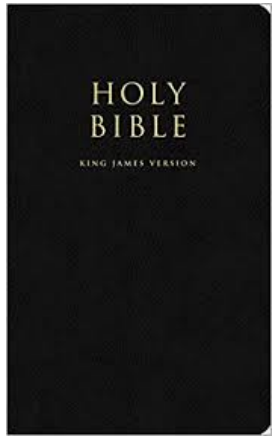
Today's lecture will cover 3 resources to derive these alignments

Parallel Text Bilingual Dictionaries Comparable Documents

Ruder et al. (2018)

Parallel Text

Consists of sentence pairs that are translations of each other.

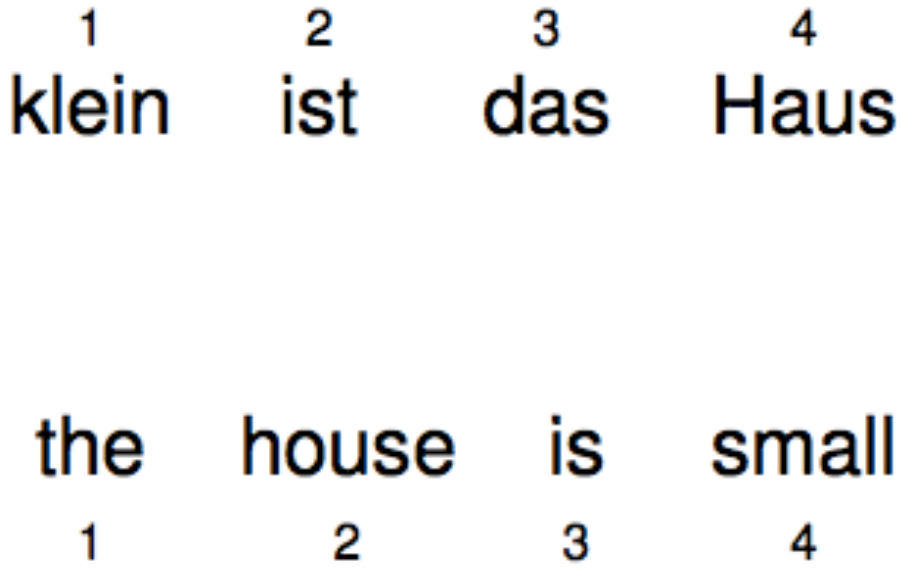


English: In the beginning God created the heavens and the earth.
Spanish: En el principio Dios creó los cielos y la tierra.

English: The chair recognizes the representative from Belgium.
Spanish: El presidente reconoce al representante de Bélgica.

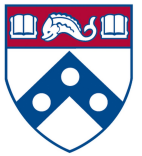
English: No, I am your father.
Spanish: No, yo soy tu padre.

Deriving Word Alignments from Parallel Text



Borrowed from Philipp Koehn's slides

Word Alignments from Parallel Text

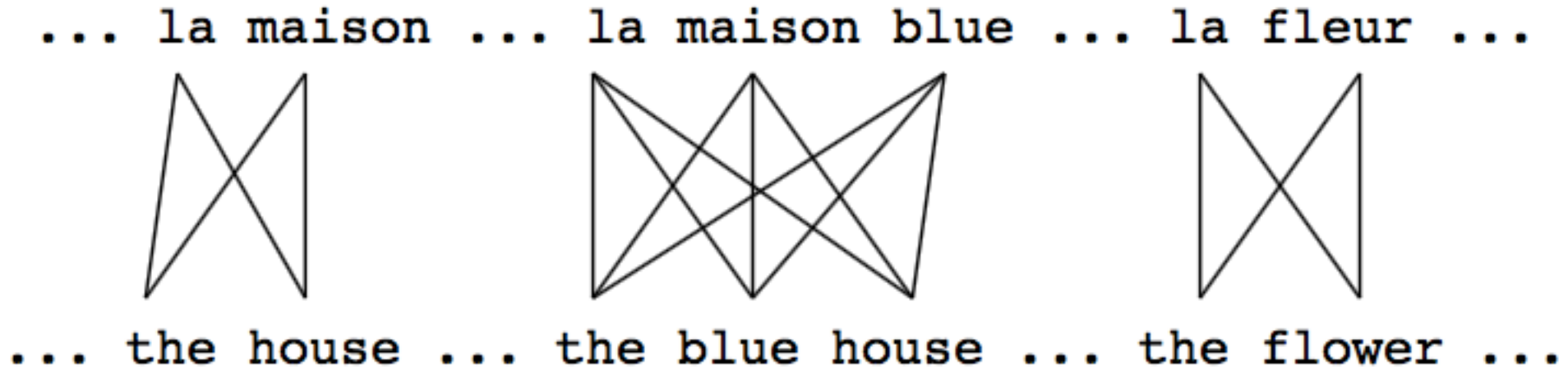


... la maison ... la maison blue ... la fleur ...

... the house ... the blue house ... the flower ...

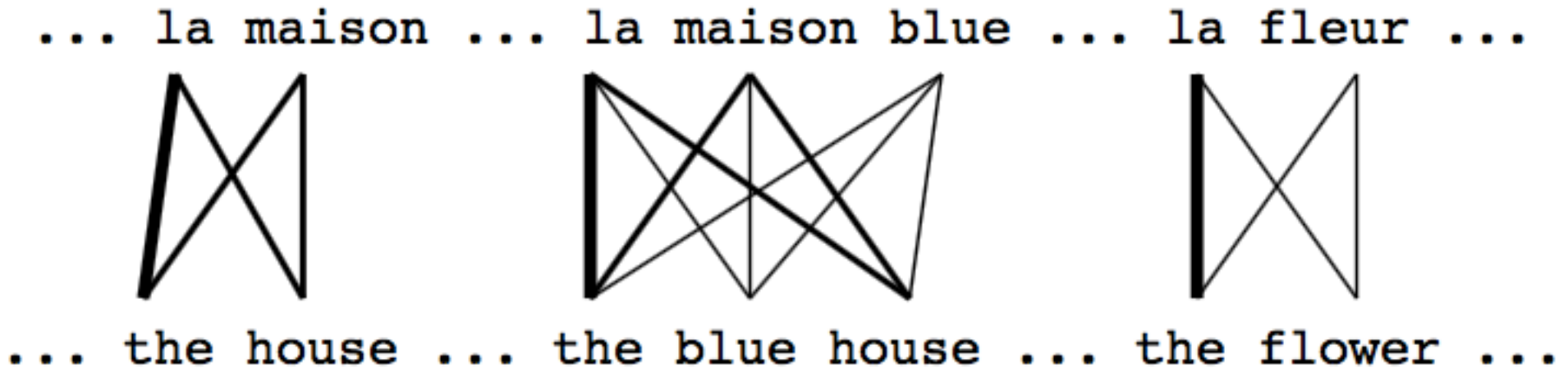
Borrowed from Philipp Koehn's slides

Word Alignments from Parallel Text



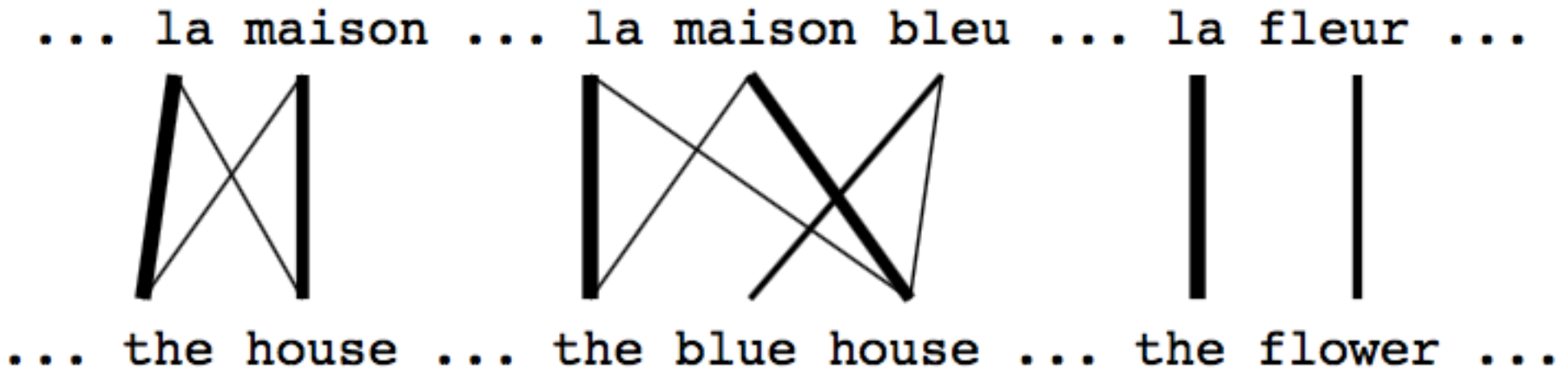
Borrowed from Philipp Koehn's slides

Word Alignments from Parallel Text



Borrowed from Philipp Koehn's slides

Word Alignments from Parallel Text



Borrowed from Philipp Koehn's slides

Word Alignments from Parallel Text

Borrowed from Philipp Koehn's slides

First Attempt – Cross-lingual Word Clusters

Decide the number of clusters.
Randomly initialize clusters in English and Spanish.

Cluster words in English monolingually.

Project words in Spanish to English clusters using word alignments.

Cluster words in Spanish monolingually.

Project words in English to Spanish clusters using w alignments.

Repeat.

A word in Spanish is assigned to the English cluster with which its most often aligned.

Täckström et al. (2012)

Cross-lingual Word Clusters

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, ...

Cluster

Täckström et al. (2012)

Issues with Cluster Based Approaches

- Slow to train
  - Scales **quadratically** in the number of clusters!
  - Expressivity tradeoff.
- "
  - Similarity between words is discretely defined.

Vector-based word embeddings do not suffer from these issues.

- The cluster assignment function is usually discontinuous
  - Cannot do end-to-end learning (e.g., cannot backpropagate to fine-tune word vectors)

The Skip-gram Formulation (word2vec)

$$-\sum_{w} \sum_{w_c \in \text{Neighbors}(w)} \log \Pr(w_c \mid w)$$

Mikolov et al. (2013)

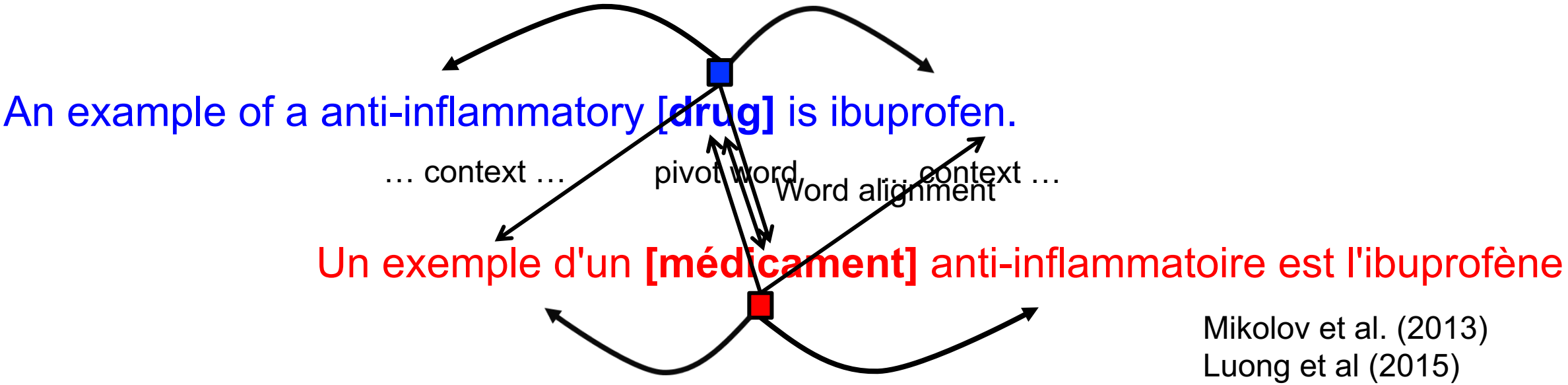
Using Parallel Text – Word and Sentence Alignments

$$D_{11}(\mathbf{V}) = -\sum_{v} \sum_{v_c \in \text{Neighbors}(v)} \log \Pr(v_c \mid v)$$

$$D_{22}(\mathbf{W}) = -\sum_{w} \sum_{w_c \in \text{Neighbors}(w)} \log \Pr(w_c \mid w)$$

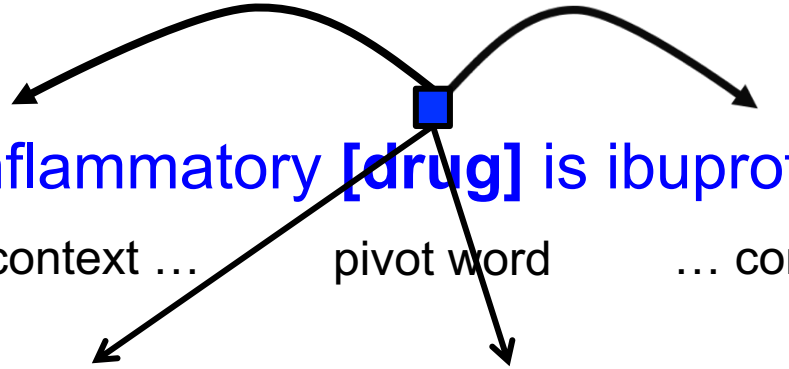
$$D_{12}(\mathbf{W}, \mathbf{V}) = -\sum_{(w,v) \in Q} \sum_{w_c \in \text{Neighbors}(w)} \log \Pr(w_c \mid v)$$

$$D_{21}(\mathbf{W}, \mathbf{V}) = -\sum_{(w,v) \in Q} \sum_{v_c \in \text{Neighbors}(v)} \log \Pr(v_c \mid w)$$



Mikolov et al. (2013)
Luong et al (2015)

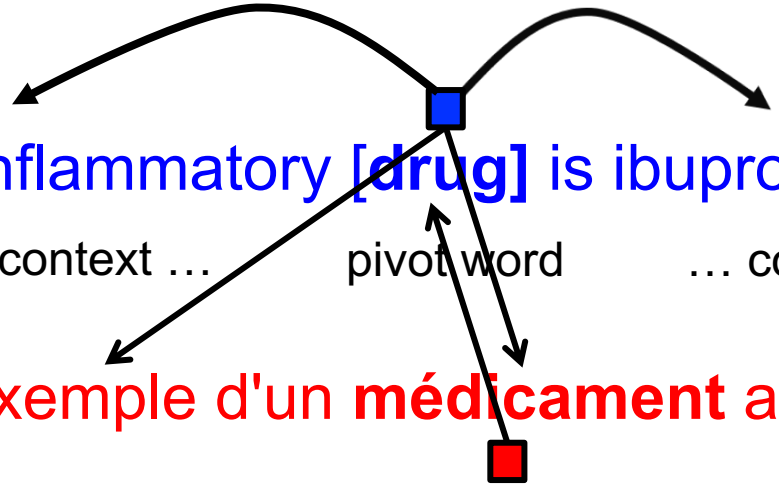
Why do we need all these terms?



An example of a anti-inflammatory **[drug]** is ibuprofen.

… context … pivot word … context …

Un exemple d'un **médicament** anti-inflammatoire est l'ibuprofène



An example of a anti-inflammatory **[drug]** is ibuprofen.

… context … pivot word … context …

Un exemple d'un **médicament** anti-inflammatoire est l'ibuprofène

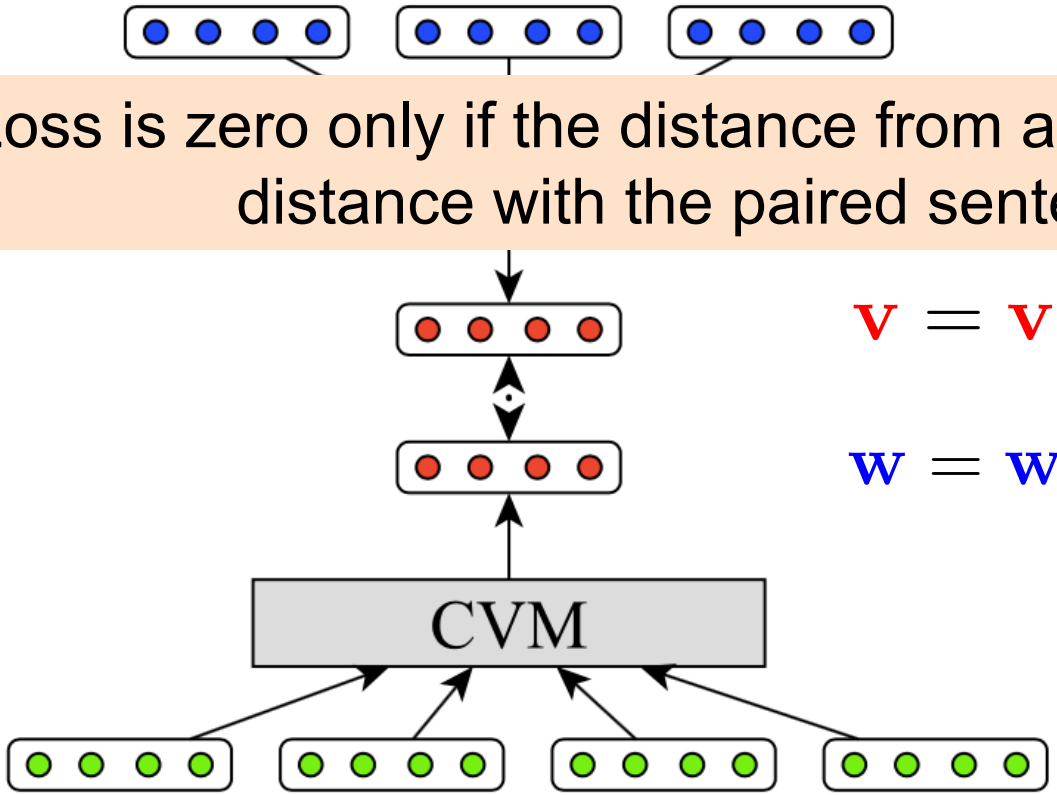
Using Parallel Text – Sentence Alignment Only

$$E(\mathbf{v}, \mathbf{w}) = \|\mathbf{v} - \mathbf{w}\|^2$$

$$Loss(\mathbf{v}, \mathbf{w}, \underbrace{\mathbf{w}^n}) = \max\left(\delta + E(\mathbf{v}, \mathbf{w}) - E(\mathbf{v}, \mathbf{w}^n), 0\right)$$

Random English Sentence

Loss is zero only if the distance from a random sentence is more than distance with the paired sentence by at least delta.

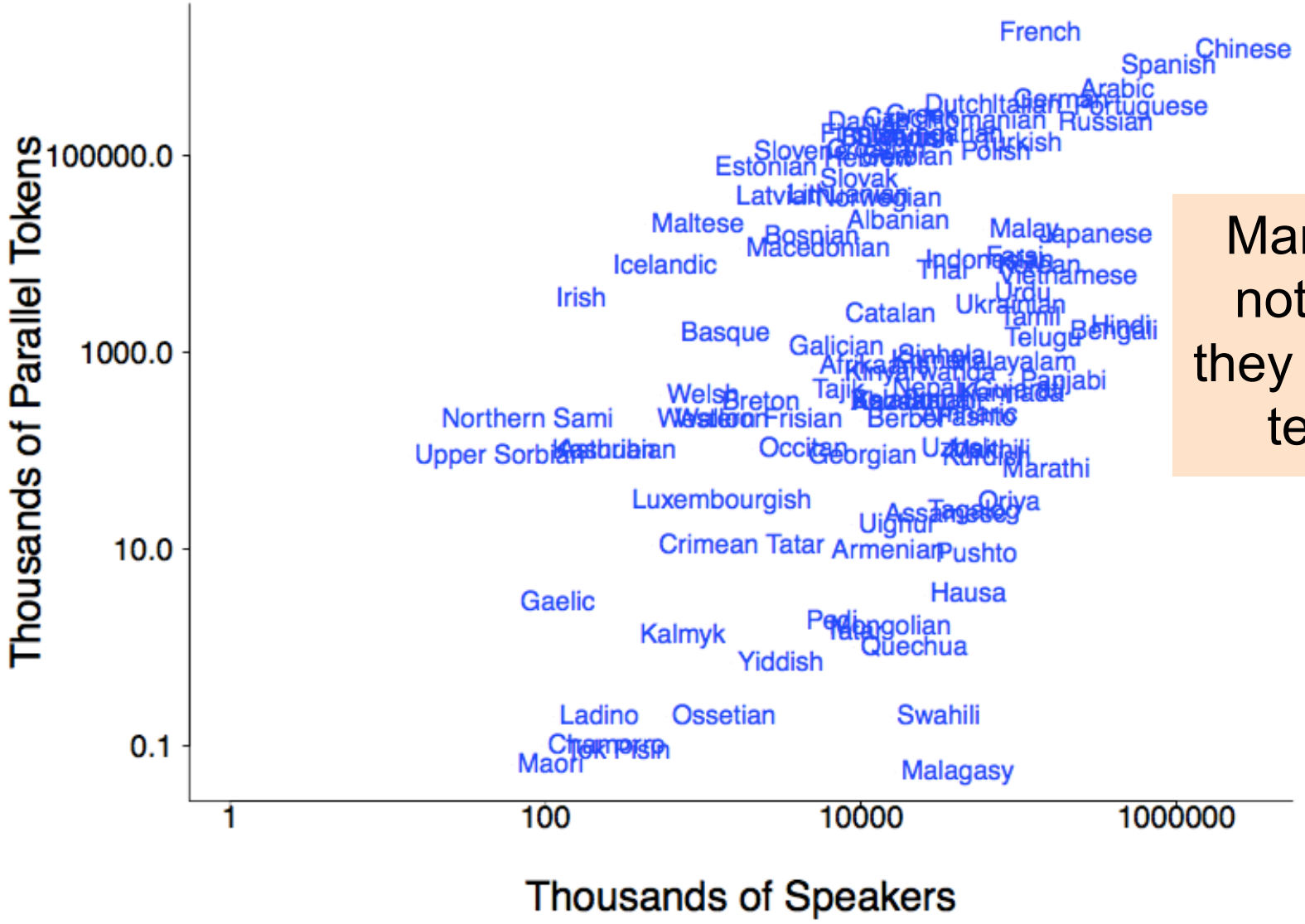


$$\mathbf{v} = \mathbf{v_1} + \mathbf{v_2} + \cdots + \mathbf{v_n}$$

$$\mathbf{w} = \mathbf{w_1} + \mathbf{w_2} + \cdots + \mathbf{w_m}$$

Hermann and Blunsom (2013)

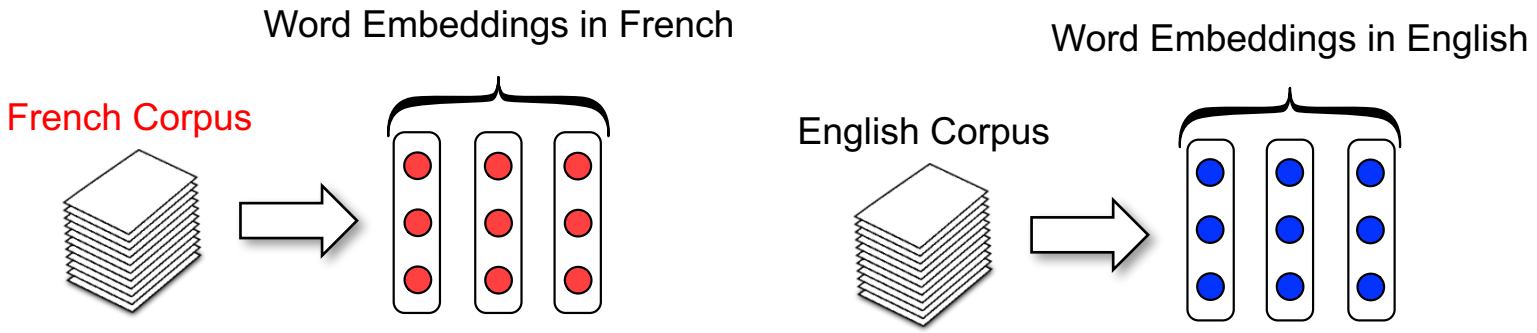
Parallel Text is Scarce



Many languages are not plotted because they don't have parallel text with English!

Ann Irvine's Thesis

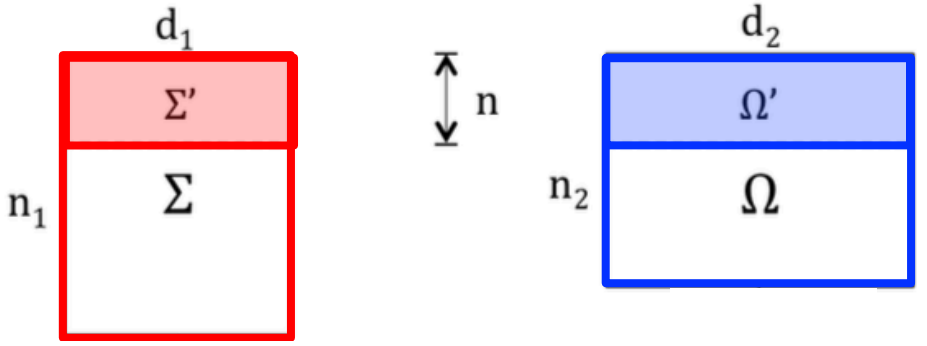
Using Monolingual Text and Bilingual Dictionary



English	French
children	enfants
law	loi
money	argent
world	monde
peace	paix
market	marche

Bilingual Dictionary

Learning Projections for each language using CCA



Projection directions

$$\overbrace{\mathbf{V}, \mathbf{W}} = \mathrm{CCA}(\boldsymbol{\Sigma}', \boldsymbol{\Omega}')$$

$$\boldsymbol{\Sigma}^*, \boldsymbol{\Omega}^* = \boldsymbol{\Sigma}\mathbf{V}, \boldsymbol{\Omega}\mathbf{W}$$

CCA finds directions for English and Spanish spaces such that after projection they are “aligned”.

Faruqui and Dyer (2014)

Canonical Correlation Analysis (CCA)

aligned

$\mathbf{x} \in \mathbf{X}$ $\mathbf{y} \in \mathbf{Y}$

Project the aligned vectors in direc

$\mathbf{x}' = \mathbf{x}\mathbf{v}$ $\mathbf{y}' = \mathbf{y}\mathbf{w}$

Efficient ways to compute CCA are available.
(e.g., canoncorr in Matlab)

"Correlation" of the projected vectors

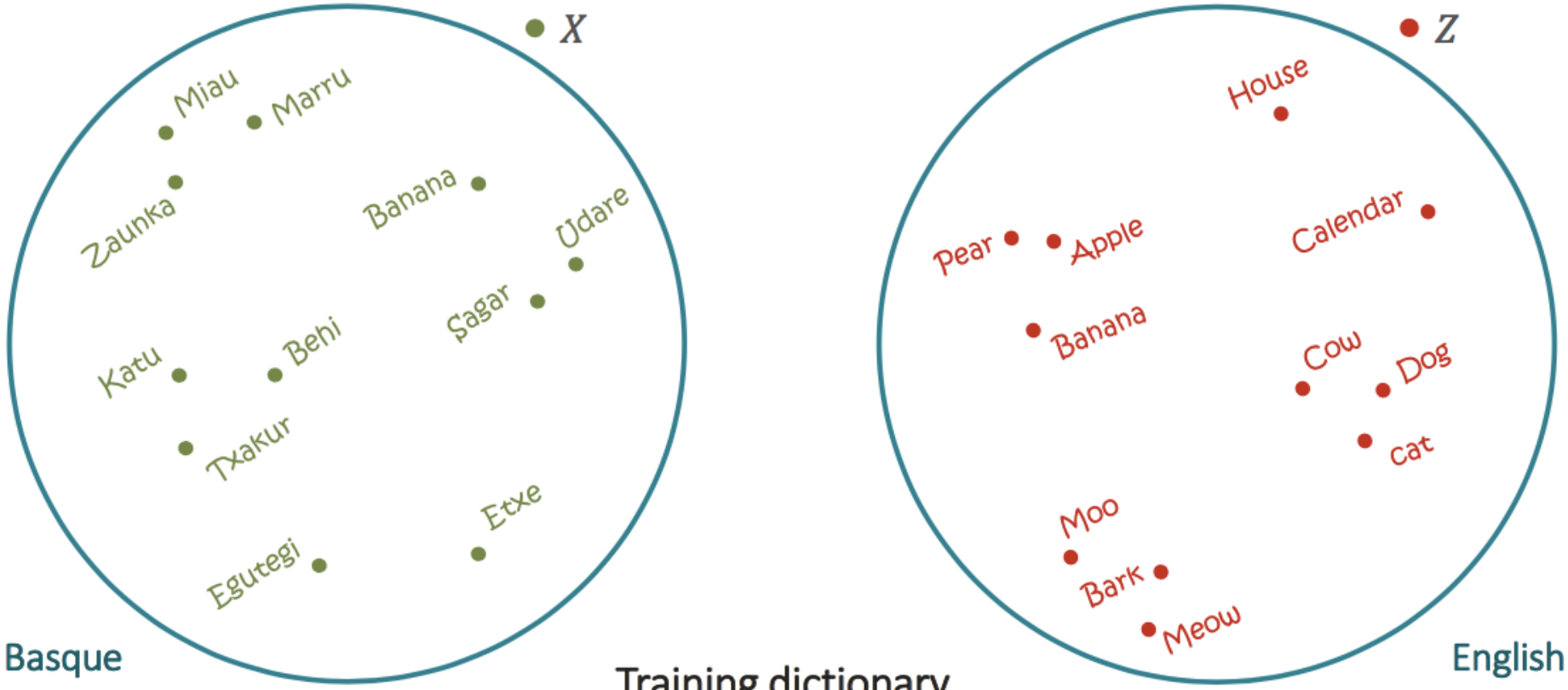
$$\rho(\mathbf{x}\mathbf{v}, \mathbf{y}\mathbf{w}) = \frac{E[\langle \mathbf{x}\mathbf{v}, \mathbf{y}\mathbf{w} \rangle]}{\sqrt{E[\langle \mathbf{x}\mathbf{v}, \mathbf{x}\mathbf{v} \rangle] E[\langle \mathbf{y}\mathbf{w}, \mathbf{y}\mathbf{w} \rangle]}}$$

CCA finds directions that maximize the correlation

$$\mathbf{v}, \mathbf{w} = \mathrm{CCA}(\mathbf{x}, \mathbf{y})$$
$$= \arg\max_{\mathbf{v},\mathbf{w}} \rho(\mathbf{x}\mathbf{v}, \mathbf{y}\mathbf{w})$$

Hotelling (1936)

Learning a Single Projection

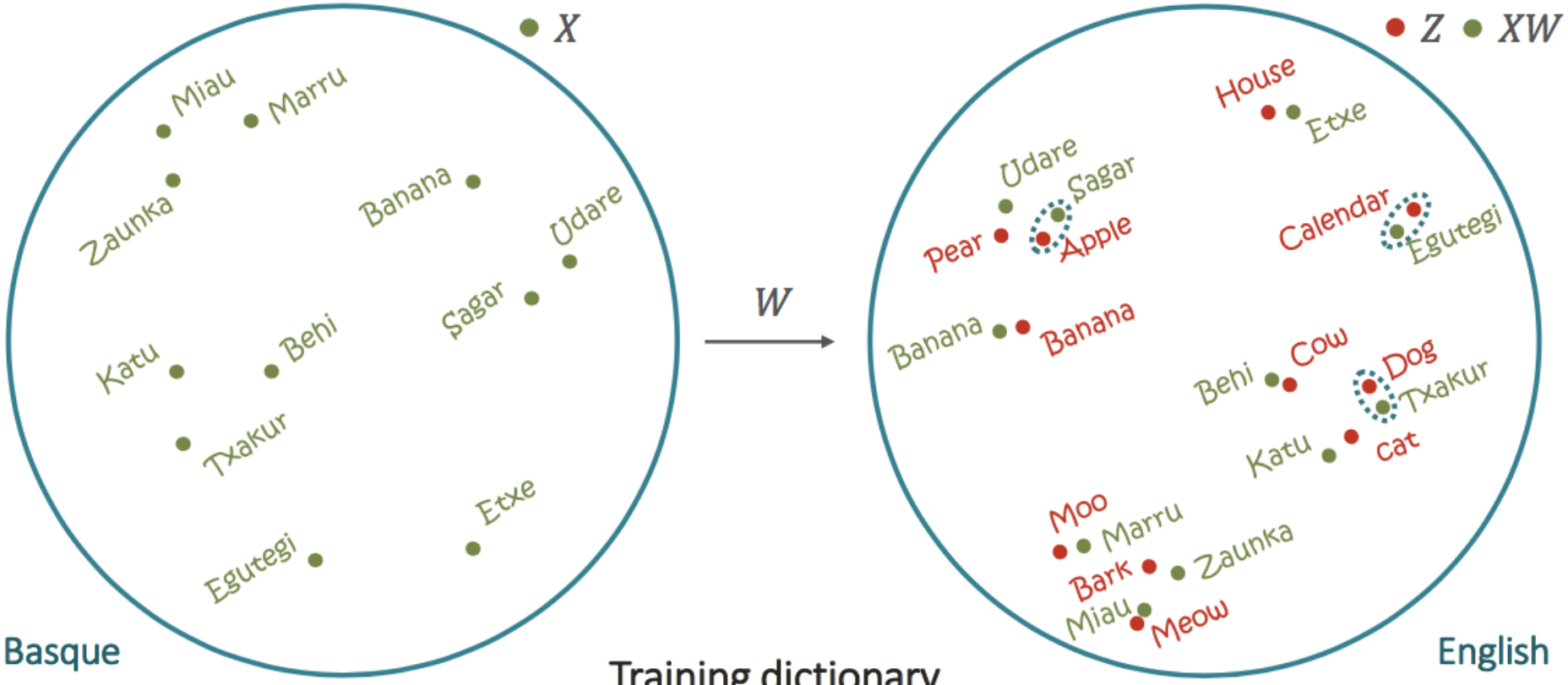


Training dictionary

Txakur	Dog
Sagar	Apple
⋮	⋮
Egutegi	Calendar

from Mikel Artexte's talk

Learning a Single Projection

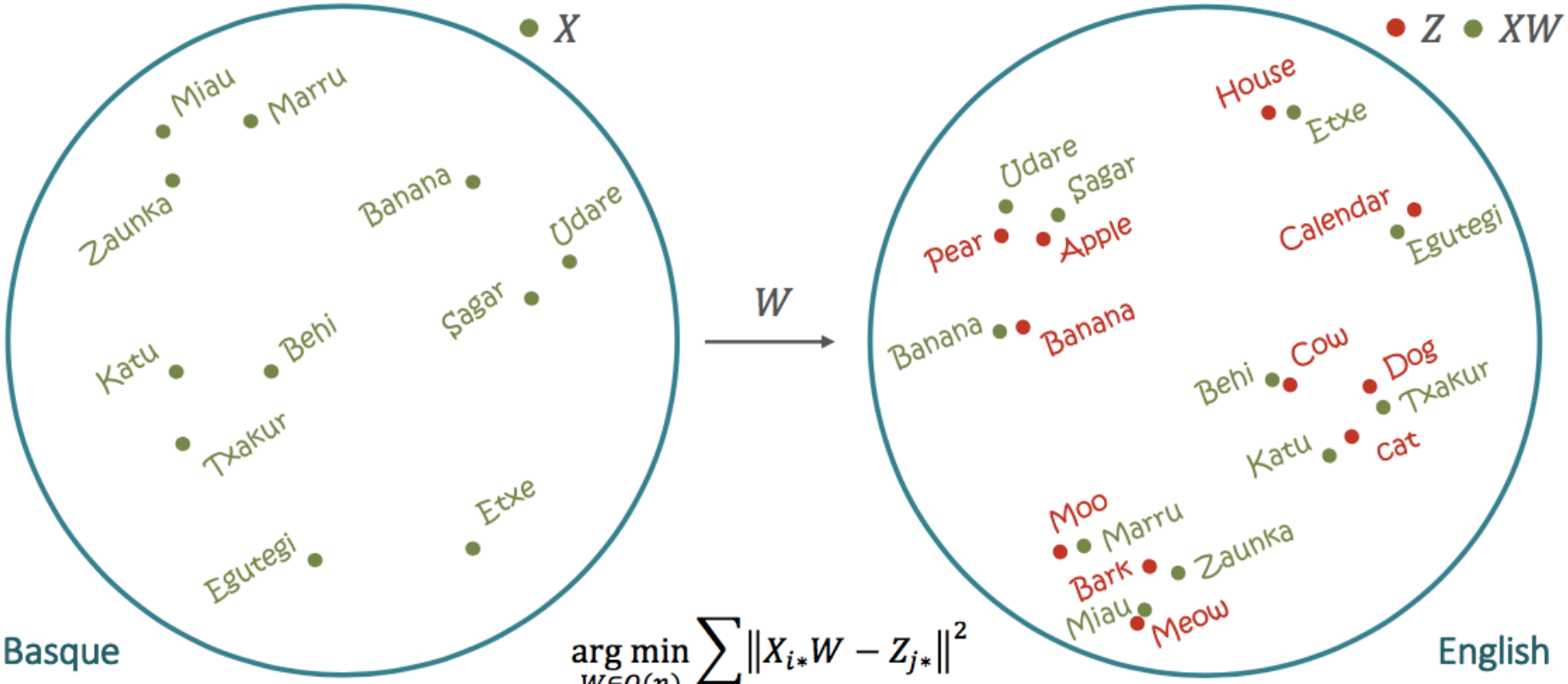


Training dictionary

Txakur	Dog
Sagar	Apple
⋮	⋮
Egutegi	Calendar

from Mikel Artexte's talk

Learning a Single Projection

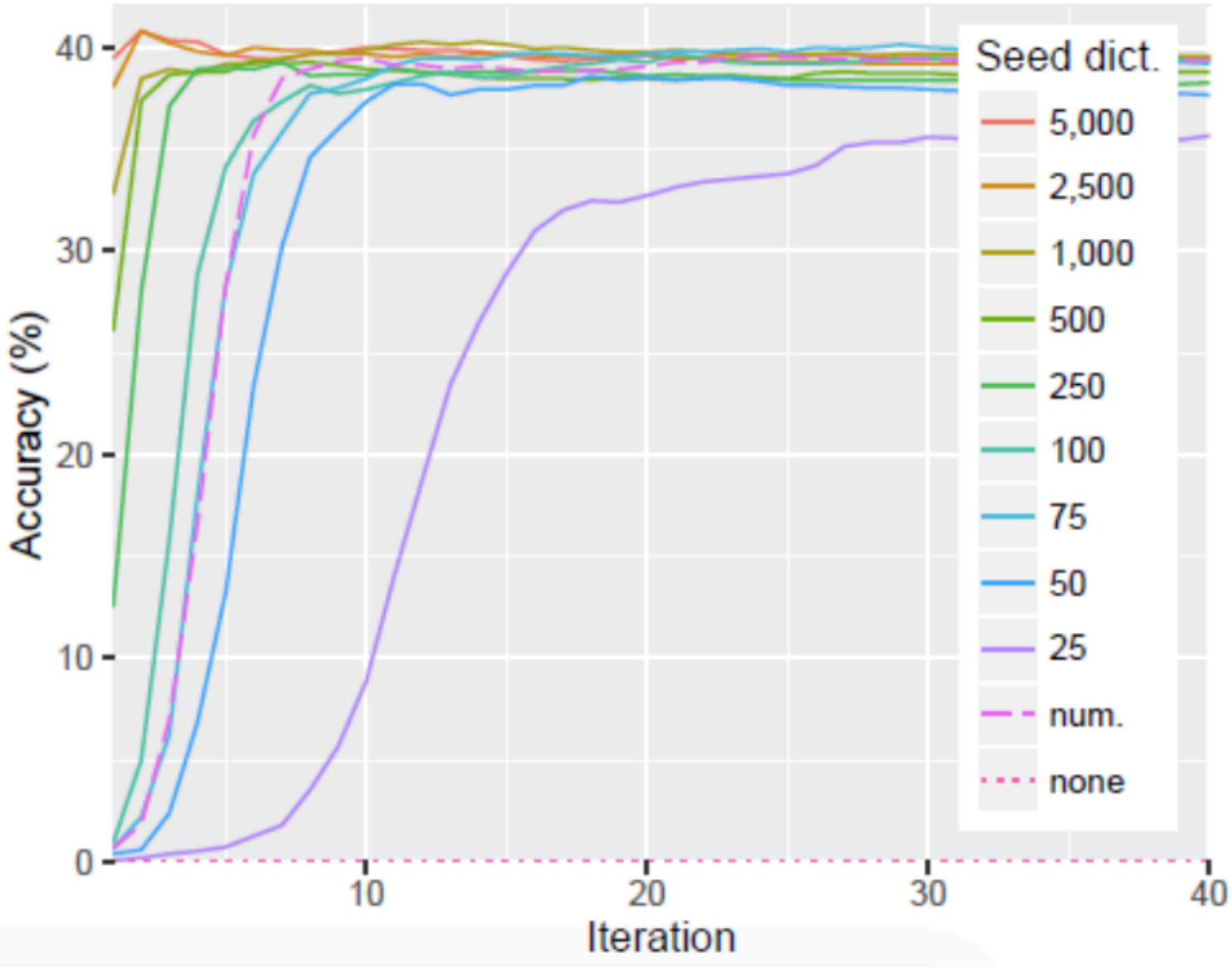


$$\arg\min_{W \in O(n)} \sum_i \|X_{i*}W - Z_{j*}\|^2$$

$$\begin{matrix} \text{Txakur} \\ \text{Sagar} \\ \vdots \\ \text{Egutegi} \end{matrix} \begin{bmatrix} X_{1,*} \\ X_{2,*} \\ \vdots \\ X_{n,*} \end{bmatrix} [W] \approx \begin{bmatrix} Z_{1,*} \\ Z_{2,*} \\ \vdots \\ Z_{n,*} \end{bmatrix} \begin{matrix} \text{Dog} \\ \text{Apple} \\ \vdots \\ \text{Calendar} \end{matrix}$$

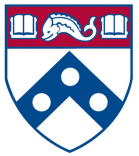
from Mikel Artexte's talk

How big a dictionary is needed?



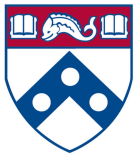
from Mikel Artexte's talk

How big a dictionary is needed?



- Usually ~5k word pairs suffice.
  - Of course, depends on the quality.
- Will bigger dictionaries help?
  - Yes, but not by much.
- Also see DeepCCA (Andrew et al. 2013) for non-linear projections.

Which dictionary will lead to better embeddings?



(Assume translations are perfect.)

superconductor, superconductores

pomegranate, granada

porpoise, marsopa

commandant, comandante

cardiologist, cardióloga

computer, computadora

apple, manzana

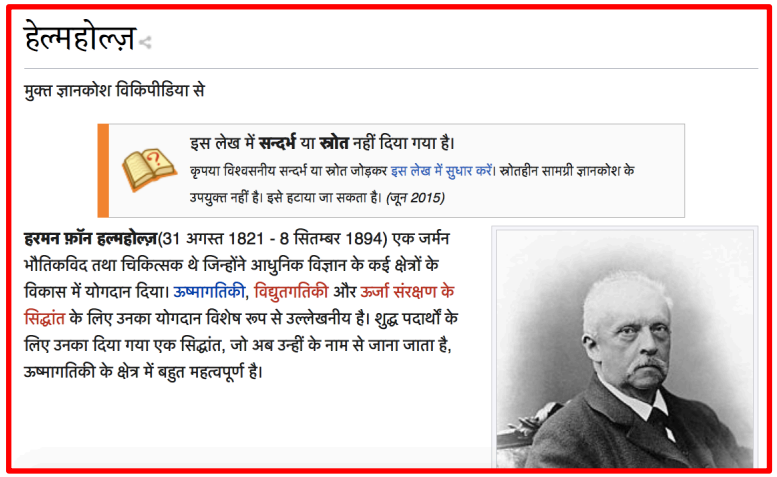
dog, perro

love, amor

heaven, cielo

Using Comparable Documents

Comparable documents are only ***thematically*** aligned.
(topic, genre, time etc.)



हेल्महोल्ज़

मुक्त ज्ञानकोश विकिपीडिया से

इस लेख में **सन्दर्भ** या **स्रोत** नहीं दिया गया है।
कृपया विश्वसनीय सन्दर्भ या स्रोत जोड़कर इस लेख में सुधार करें। स्रोतहीन सामग्री ज्ञानकोश के उपयुक्त नहीं है। इसे हटाया जा सकता है। *(जून 2015)*

हरमन फ़ॉन हल्महोल्ज़(31 अगस्त 1821 - 8 सितम्बर 1894) एक जर्मन भौतिकविद तथा चिकित्सक थे जिन्होंने आधुनिक विज्ञान के कई क्षेत्रों के विकास में योगदान दिया। ऊष्मागतिकी, विद्युतगतिकी और ऊर्जा संरक्षण के सिद्धांत के लिए उनका योगदान विशेष रूप से उल्लेखनीय है। शुद्ध पदार्थों के लिए उनका दिया गया एक सिद्धांत, जो अब उन्हीं के नाम से जाना जाता है, ऊष्मागतिकी के क्षेत्र में बहुत महत्वपूर्ण है।

Hermann von Helmholtz

From Wikipedia, the free encyclopedia

"Helmholtz" redirects here. For other uses, see Helmholtz (disambiguation).

Hermann Ludwig Ferdinand von Helmholtz (August 31, 1821 – September 8, 1894) was a German physician and physicist who made significant contributions in several scientific fields. The largest German association of research institutions, the Helmholtz Association, is named after him.[5]

In physiology and psychology, he is known for his mathematics of the eye, theories of vision, ideas on the visual perception of space, color vision research, and on the sensation of tone, perception of sound, and empiricism in the physiology of perception.

In physics, he is known for his theories on the conservation of energy, work in electrodynamics, chemical thermodynamics, and on a mechanical foundation of thermodynamics.

Hermann von Helmholtz
ForMemRS

Trump urged Spain to 'build a wall' across Sahara, says minister

La solución de Trump para la crisis migratoria de Europa: construir un muro

(Trump's Solution for Europe's Migrant Crisis - Build a Wall)

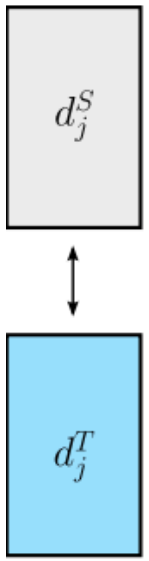
English: The brown dog is running after the black dog.
German: Ein brauner Hund und ein schwarzer Hund.

Advantages / Disadvantages

- Relatively easier to obtain than Parallel Corpus.
  - Lots of News Sources - BBC, VoA …
  - Wikipedia

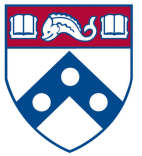
- Cross-lingual alignment is a bit fuzzy.
  - One document may contain more/less information than other.
  - “Unprocessed form of parallel text”

Using Comparable Corpora with word2vec



Vulic and Moens (2016)

Using Comparable Documents from Wikipedia

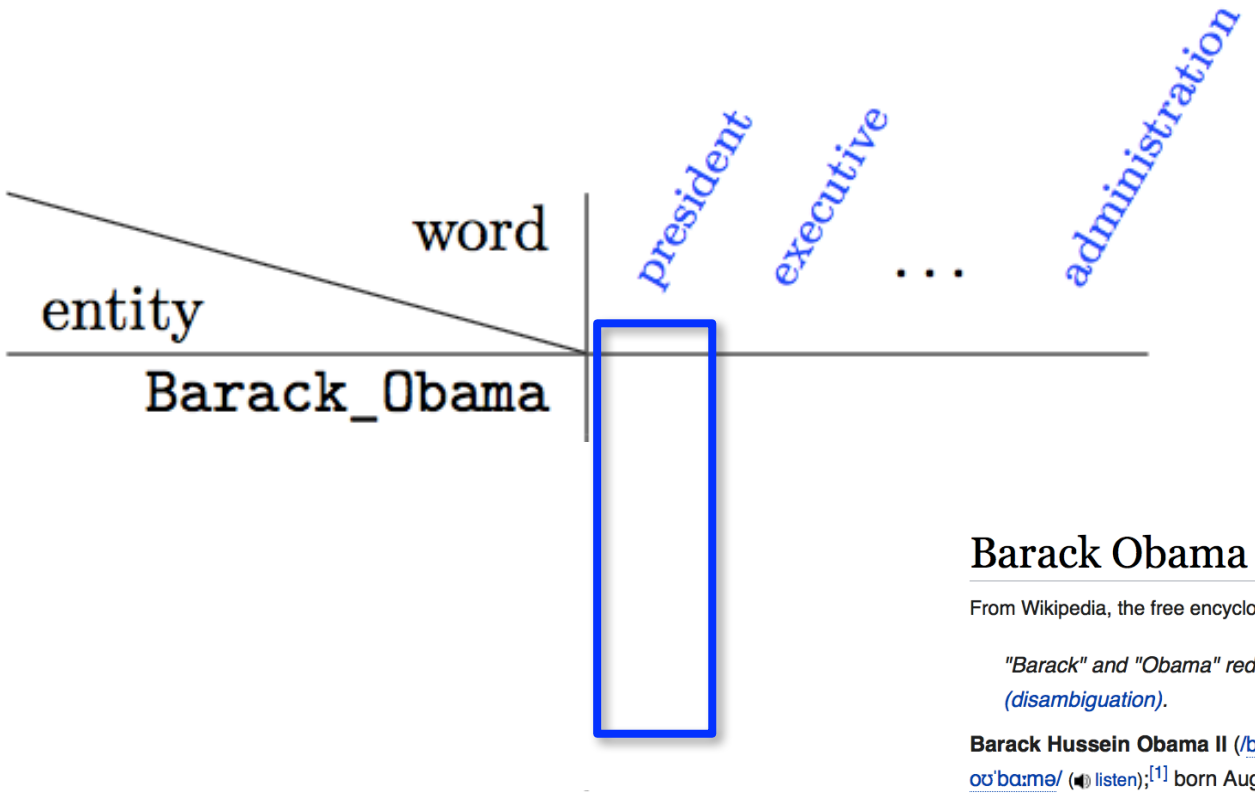


Term-document matrix

Jurafsky and Martin slides

Monolingual Representation from Wikipedia

Also see Explicit Semantic Analysis from Gabrilovich & Markovitch (2009)



Barack Obama

From Wikipedia, the free encyclopedia

"Barack" and "Obama" redirect here. For other uses, see Barack (disambiguation) and Ob (disambiguation).

Barack Hussein Obama II (/bəˈrɑːk huːˈseɪn oʊˈbɑːmə/ (listen);[1] born August 4, 1961) is an American attorney and politician who served as the 44th president of the United States from 2009 to 2017. A member of the Democratic Party, he was the first African American to be elected to the presidency. He previously served as a U.S. senator from Illinois from 2005 to 2008.

Barack Obama

Using Comparable Documents from Wikipedia

$\mathcal{A}$ inter-language links $\mathcal{B}$

en.wikipedia.org/wiki/Albert_Einstein

hi.wikipedia.org/wiki/अल्बर्ट_आइंस्टीन

Albert Einstein

From Wikipedia, the free encyclopedia

"Einstein" redirects here. For the musicologist, see Alfred Einstein. For other people, see Einstein (surname). F and Einstein (disambiguation).

Albert Einstein (/ˈaɪnstaɪn/;[4] German: [ˈalbɛɐ̯t ˈʔaɪnʃtaɪn] (listen); 14 March 1879 – 18 April 1955) was a German-theoretical physicist[5] who developed the theory of relativity, one of the two pillars of modern physics (alongside quantum mechanics).[3][6]:274 His work is also known for its influence on the philosophy of science.[7][8] He is best k to the general public for his mass–energy equivalence formula $E = mc^2$, which has been dubbed "the world's most famous equation".[9] He received the 1921 Nobel Prize in Physics "for his services to theoretical physics, and espe for his discovery of the law of the photoelectric effect",[10] a pivotal step in the development of quantum theory.

Near the beginning of his career, Einstein thought that Newtonian mechanics was no longer enough to reconcile th of classical mechanics with the laws of the electromagnetic field. This led him to develop his special theory of relati during his time at the Swiss Patent Office in Bern (1902–1909), Switzerland. However, he realized that the principle relativity could also be extended to gravitational fields, and he published a paper on general relativity in 1916 with h theory of gravitation. He continued to deal with problems of statistical mechanics and quantum theory, which led to explanations of particle theory and the motion of molecules. He also investigated the thermal properties of light whi laid the foundation of the photon theory of light. In 1917, he applied the general theory of relativity to model the stru of the universe.[11][12]

Languages

Deutsch

★ Español

Français

한국어

Italiano

Русский

Tagalog

★ Tiếng Việt

中文

During h

many lar

the Patie

referred

"Affordal

Reform a

Ask, Dor

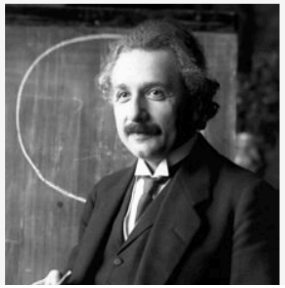
Recover

र्च १८७९ - १८ अप्रैल १९५५) एक विश्वप्रसिद्ध सैद्धांतिक भौतिकविद्
रण E = mc² के लिए जाने जाते हैं। उन्हें सैद्धांतिक भौतिकी,
९२१ में नोबेल पुरस्कार प्रदान किया गया।

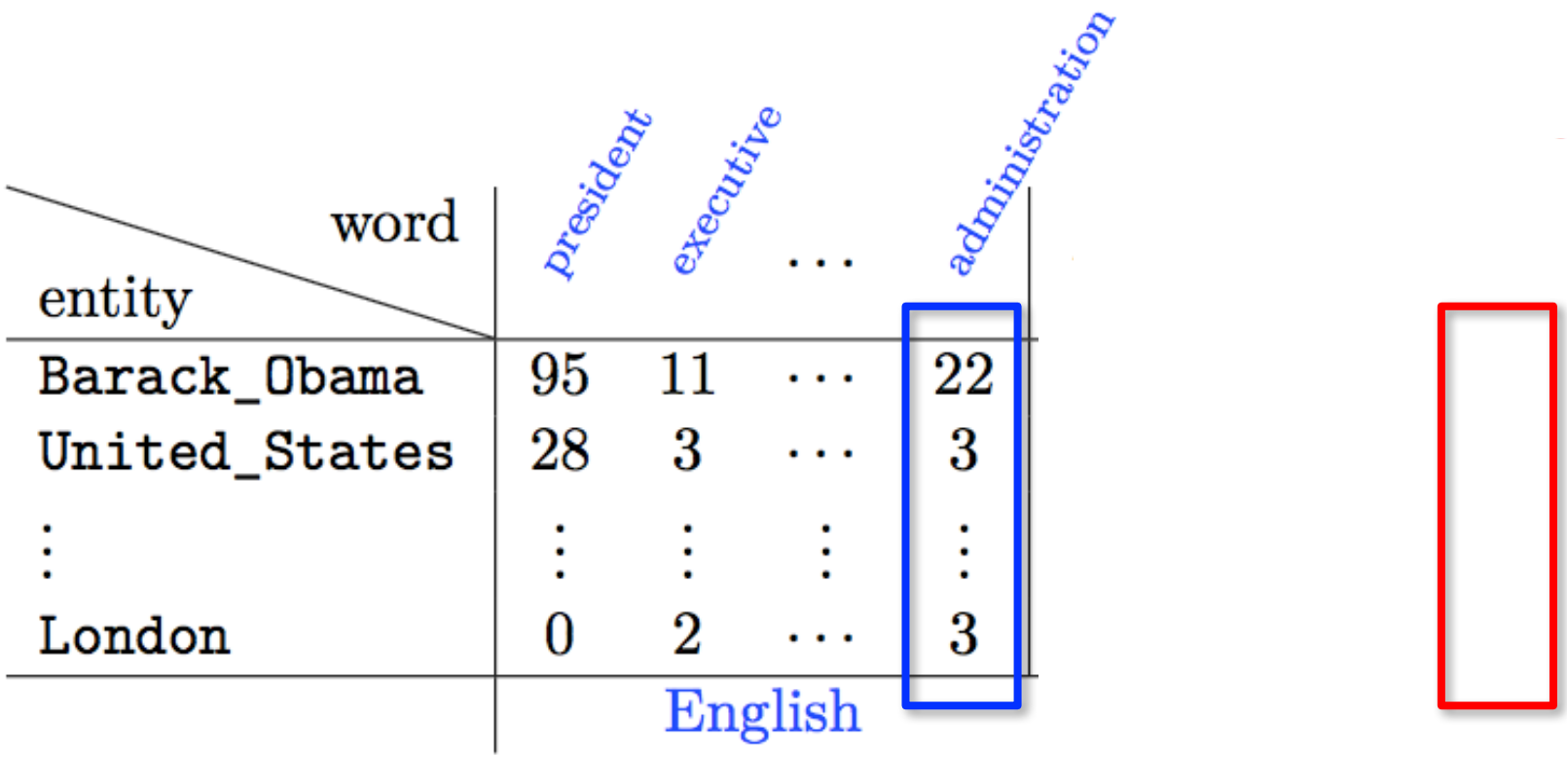
मान्य आपेक्षिकता के सिद्धांत (१९१६) सहित कई योगदान दिए। उनके
तिक उपच्छाया, सांख्यिक मैकेनिक्स की समस्याएँ, अणुओं का
अणु वाले गैस का क्वांटम सिद्धांत, कम विकिरण घनत्व वाले
क्षेत्र सिद्धांत और भौतिकी का ज्यामितीकरण शामिल है। आइंस्टीन
किताबें लिखीं। १९९९ में टाइम पत्रिका ने शताब्दी-पुरुष घोषित किया।
ानिक माने गए।

प्रकाशन किया। 5 दिसंबर 2014 को विश्वविद्यालयों और
द्वितीय दस्तावेज एवं पत्र की प्रदर्शन की घोषणा की हैं। आइंस्टीन के
को "बुद्धिमान" का पर्याय बना दिया है।[2]

अल्बर्ट आइंस्टीन
Albert Einstein



Using Comparable Documents from Wikipedia



What can go wrong here?

Language	# Articles
German	2.06M
French	1.87M
Italian	1.36M

The size of the intersection with English Wikipedia matters!
Smaller intersection means fewer entries in the entity-word matrix.

Arabic	521k
Turkish	292k
Tamil	104k
Tagalog	69k

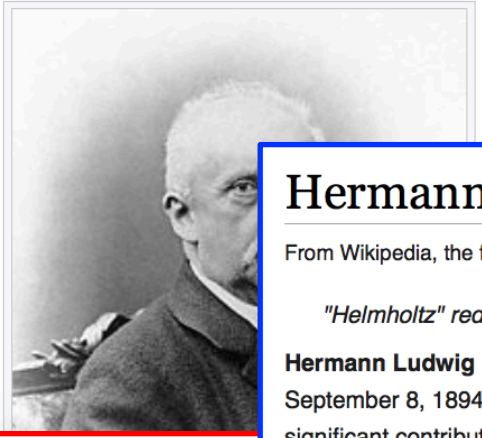
Quality of the Wikipedia Documents …

हेल्महोल्ज़

मुक्त ज्ञानकोश विकिपीडिया से

इस लेख में **सन्दर्भ** या **स्रोत** नहीं दिया गया है।
कृपया विश्वसनीय सन्दर्भ या स्रोत जोड़कर इस लेख में सुधार करें। स्रोतहीन सामग्री ज्ञानकोश के उपयुक्त नहीं है। इसे हटाया जा सकता है। *(जून 2015)*

हरमन फ़ॉन हल्महोल्ज़(31 अगस्त 1821 - 8 सितम्बर 1894) एक जर्मन भौतिकविद तथा चिकित्सक थे जिन्होंने आधुनिक विज्ञान के कई क्षेत्रों के विकास में योगदान दिया। ऊष्मागतिकी, विद्युतगतिकी और ऊर्जा संरक्षण के सिद्धांत के लिए उनका योगदान विशेष रूप से उल्लेखनीय है। शुद्ध पदार्थों के लिए उनका दिया गया एक सिद्धांत, जो अब उन्हीं के नाम से जाना जाता है, ऊष्मागतिकी के क्षेत्र में बहुत महत्वपूर्ण है।



Quality and length of Wikipedia articles varies with language (Lewoniewski et al., 2017).

Hermann von Helmholtz

From Wikipedia, the free encyclopedia

"Helmholtz" redirects here. For other uses, see Helmholtz (disambiguation).

Hermann Ludwig Ferdinand von Helmholtz (August 31, 1821 – September 8, 1894) was a German physician and physicist who made significant contributions in several scientific fields. The largest German association of research institutions, the Helmholtz Association, is named after him.[5]

In physiology and psychology, he is known for his mathematics of the eye, theories of vision, ideas on the visual perception of space, color vision research, and on the sensation of tone, perception of sound, and empiricism in the physiology of perception.

In physics, he is known for his theories on the conservation of energy, work in electrodynamics, chemical thermodynamics, and on a mechanical foundation of thermodynamics.

Hermann von Helmholtz
ForMemRS

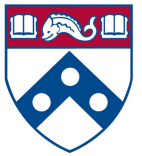
Large Wikipedia != Good Quality

1 000 000+ articles

№	Language	Language (local)	Wiki	Articles	Total	Edits	Admins	Users	Active Users
1	English	English	en	**5,826,451**	47,356,234	883,765,692	1,183	35,933,830	140,965
2	Cebuano	Sinugboanong Binisaya	ceb	**5,375,709**	9,113,133	25,826,696	6	56,347	177
3	Swedish	Svenska	sv	**3,749,463**	7,699,541	45,180,005	66	662,444	2,901
4	German	Deutsch	de	**2,283,569**	6,399,279	185,519,085	188	3,147,655	20,506
5	French	Français	fr	**2,090,641**	10,068,855	157,211,879	158	3,392,574	19,450
6	Dutch	Nederlands	nl	**1,961,492**	4,063,187	53,268,479	44	983,718	4,399
7	Russian	Русский	ru	**1,535,875**	5,844,084	98,429,801	84	2,491,920	11,742
8	Italian	Italiano	it	**1,514,293**	6,134,422	103,133,326	111	1,797,653	8,755
9	Spanish	Español	es	**1,511,708**	6,607,841	114,225,492	73	5,359,668	17,657
10	Polish	Polski	pl	**1,325,258**	2,991,496	56,014,806	102	946,939	4,503
11	Waray-Waray	Winaray	war	**1,263,504**	2,877,153	6,198,266	3	39,613	83
12	Vietnamese	Tiếng Việt	vi	**1,204,360**	14,511,525	50,726,293	21	656,937	1,805
13	Japanese	日本語	ja	**1,143,908**	3,388,264	71,846,299	42	1,456,676	13,502
14	Chinese	中文	zh	**1,049,708**	5,715,225	53,335,459	81	2,700,601	8,394
15	Portuguese	Português	pt	**1,020,039**	4,803,214	54,358,053	81	2,224,987	5,747

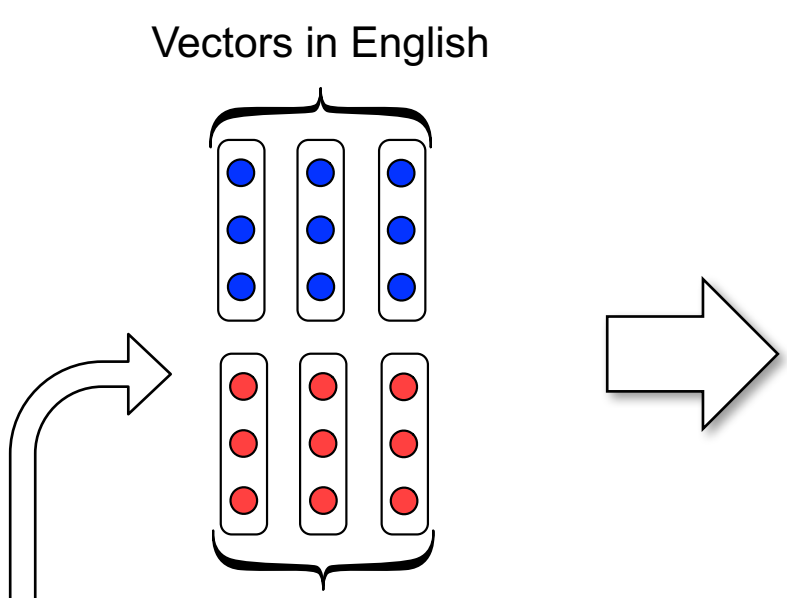
Turns out (most of) Cebuano and Waray-Waray Wikipedia were automatically created.

https://meta.wikimedia.org/wiki/List_of_Wikipedias



Quick Recap

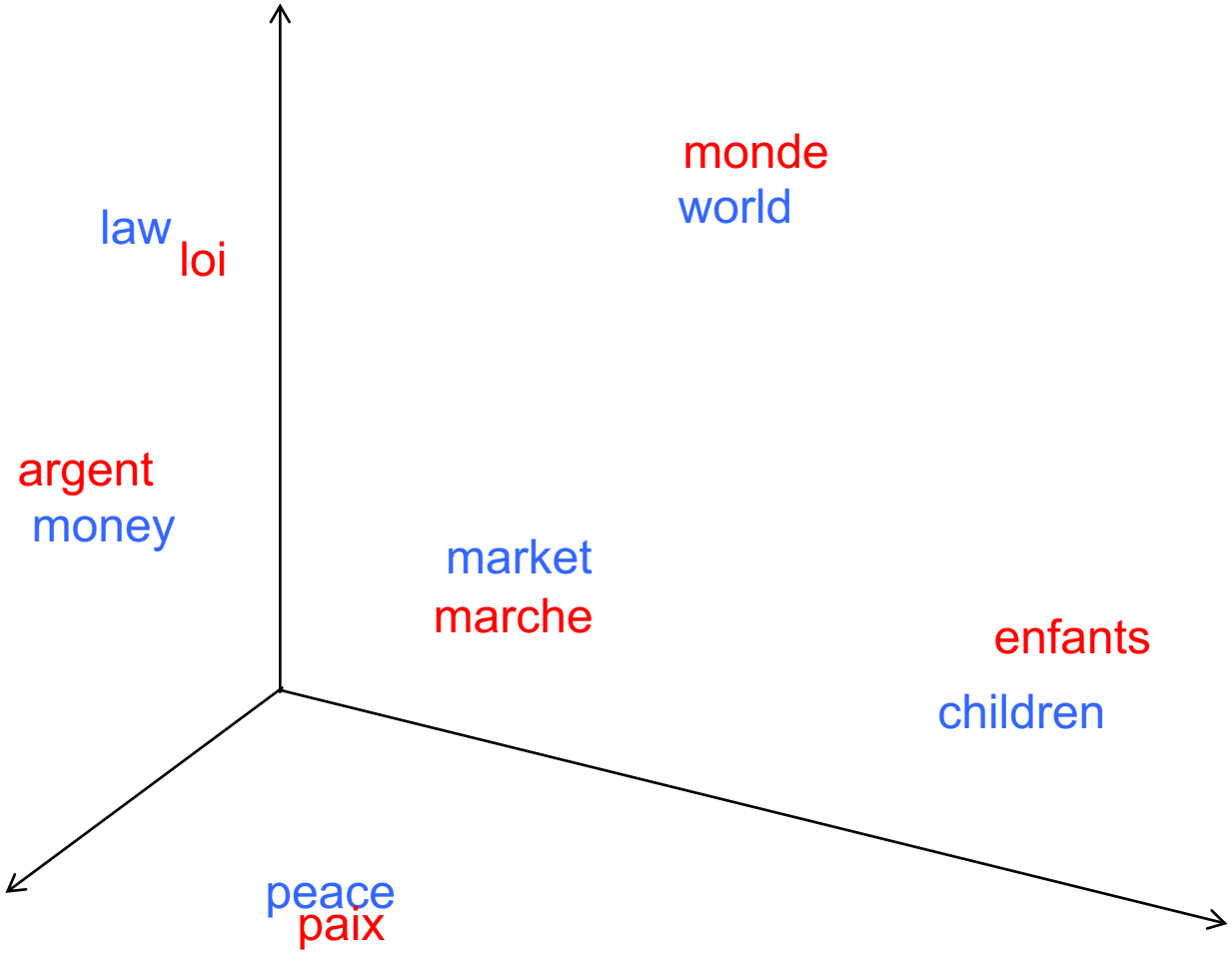
Cross-lingual Representations



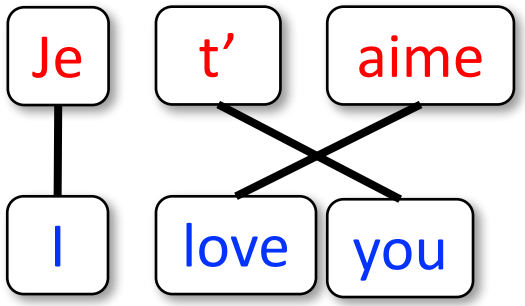
English	French
children	enfants
law	loi
money	argent
world	monde
peace	paix
market	marche

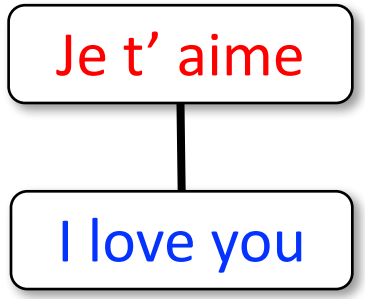
English Corpus

French Corpus



Summary of the Cross-lingual Alignments used

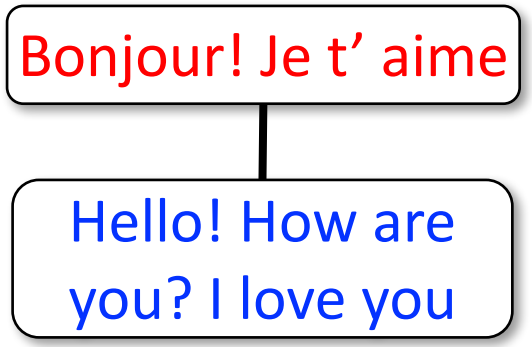




(You, t’)

(Love, aime)

(I, je)



Nature of Cross-lingual Alignment

word + sentence

sentence

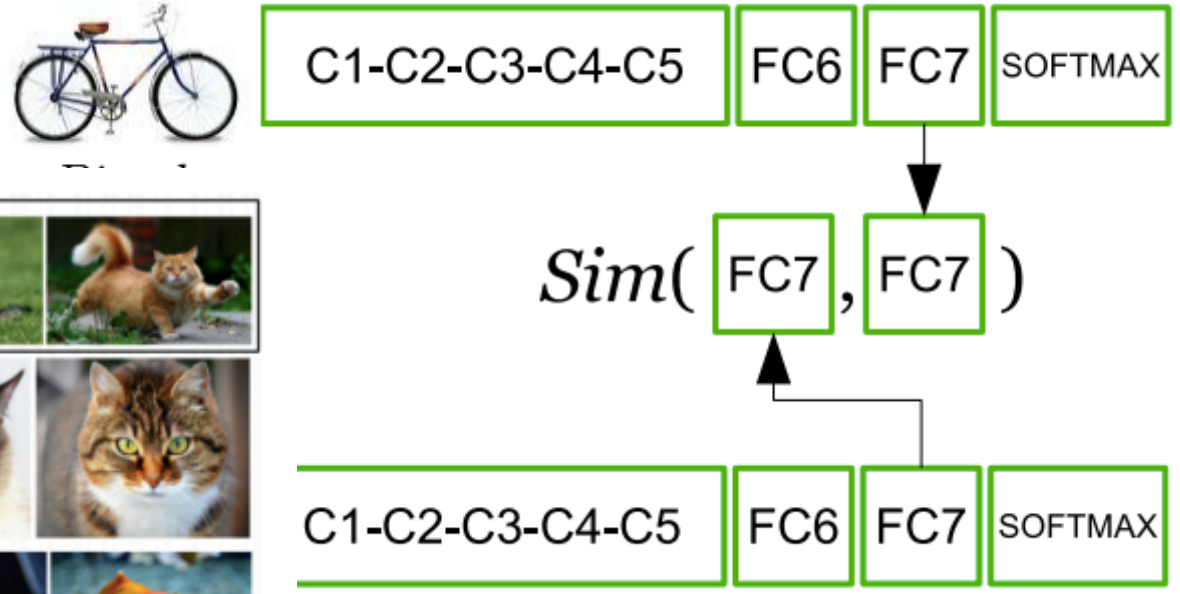
word

document

Upadhyay et al. (2016)

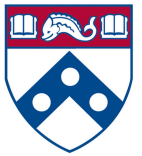
Not covered today …

- Visual Alignment



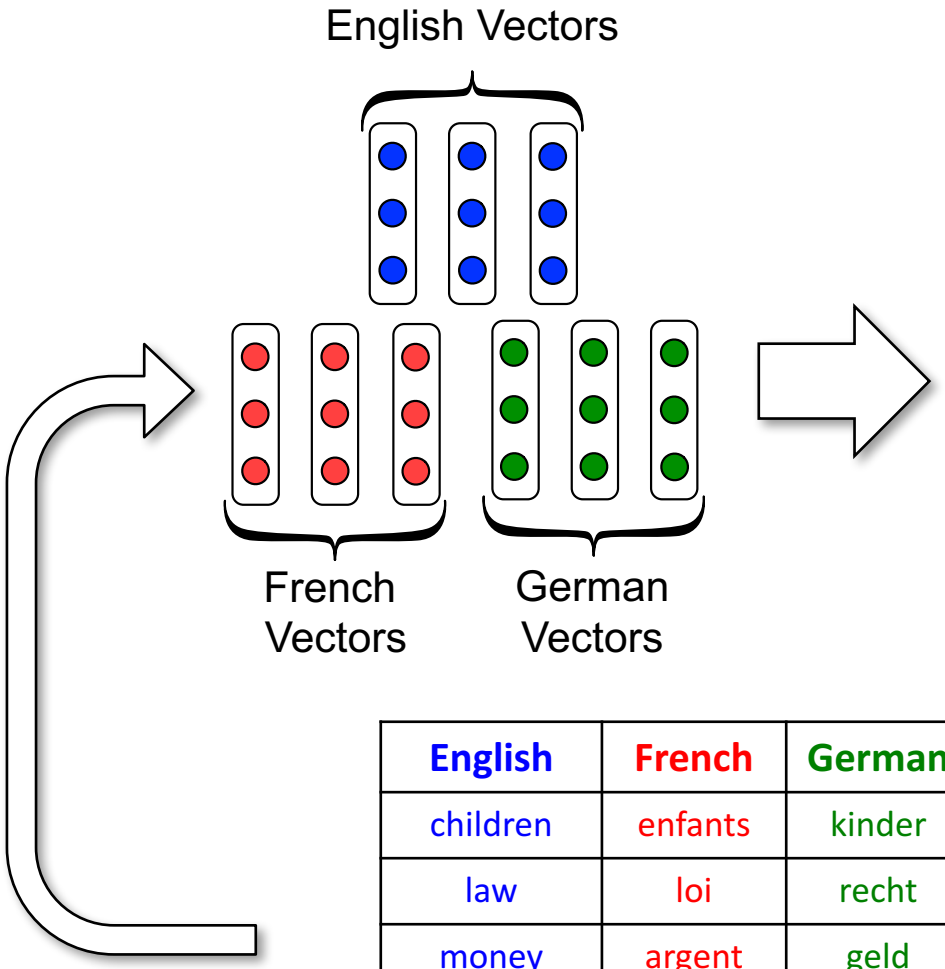
Kiela et al. (2015)

Hewitt*, Ippolito* et al. (2018). Work done at UPenn with Chris!

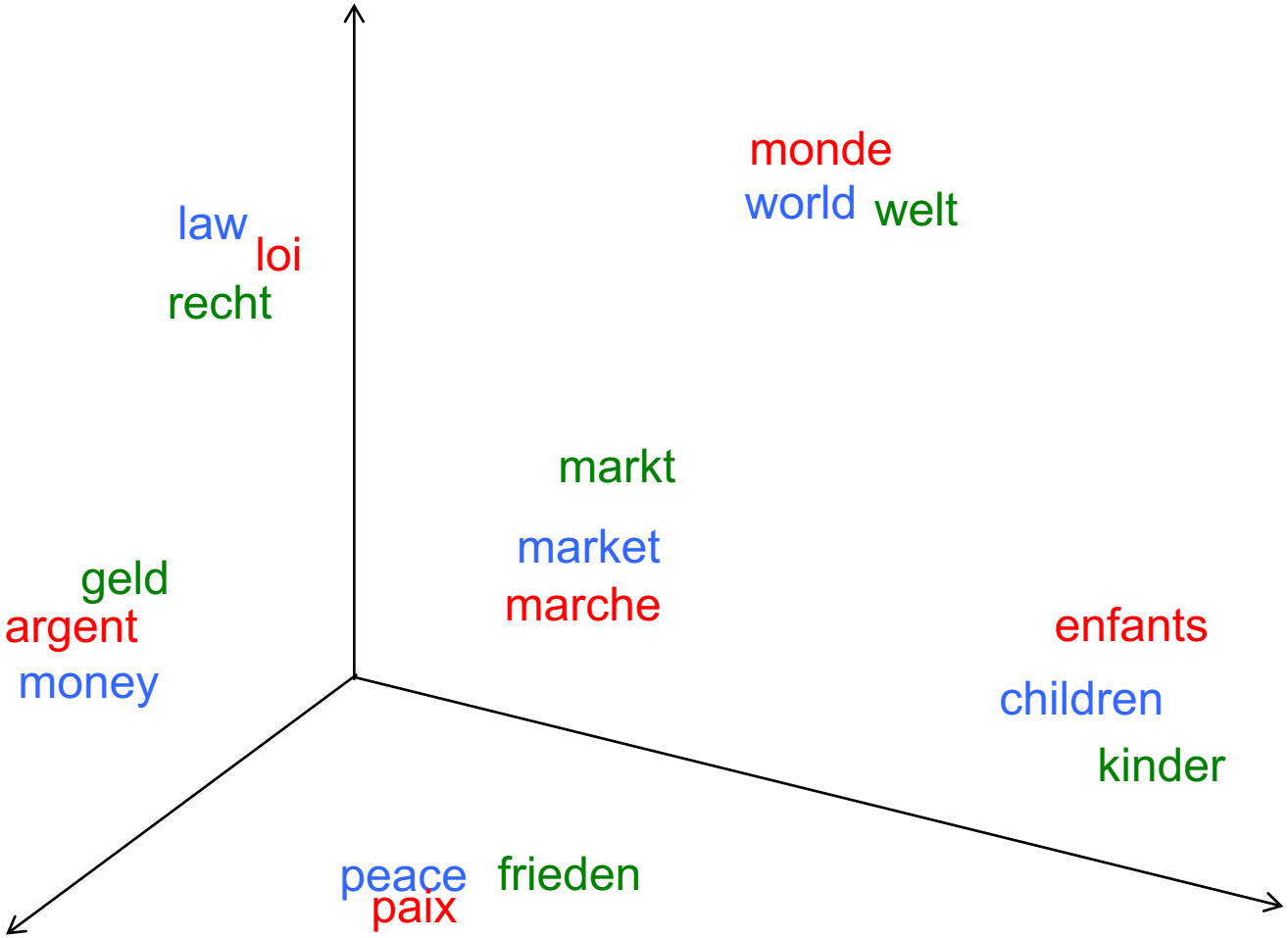


Going Multilingual

Multilingual Word Embeddings

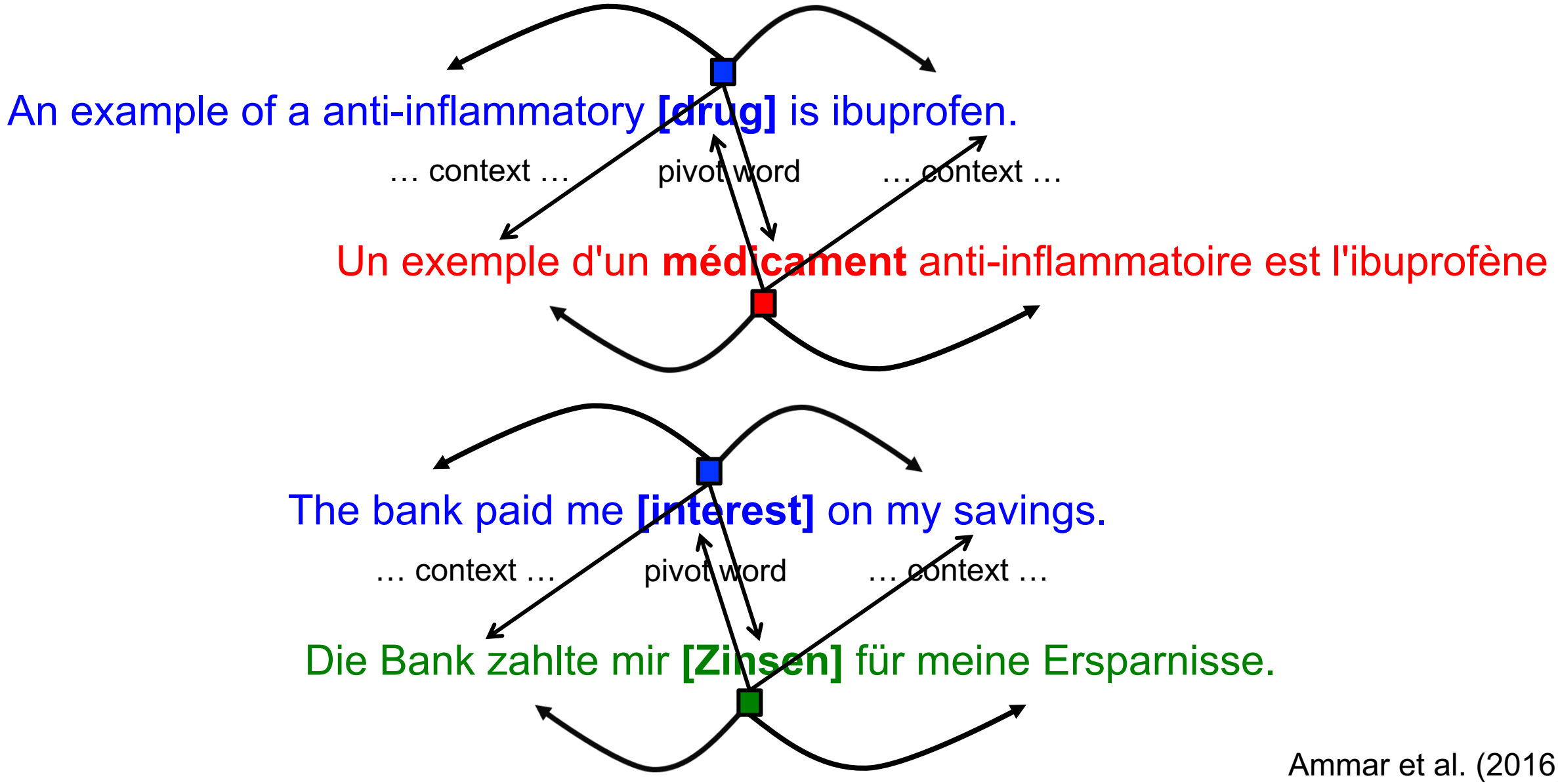


English	French	German
children	enfants	kinder
law	loi	recht
money	argent	geld
world	monde	welt
peace	paix	frieden
market	marche	markt



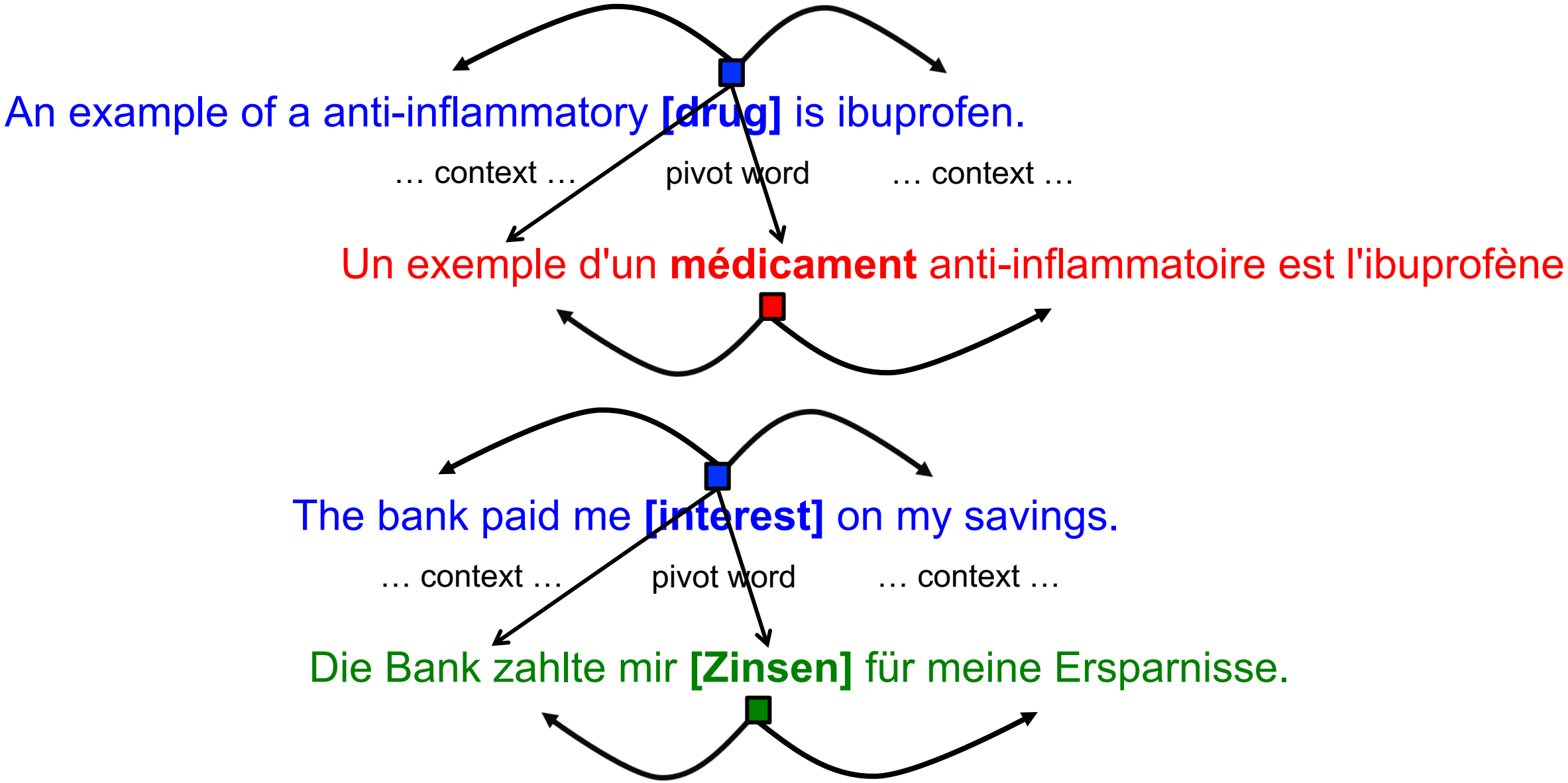
Ammar et al. (2016), Smith et al. (2017)

MultiSkip



Ammar et al. (2016)

Why do I need cross-lingual terms in both directions?



Which method should I choose?

- I want to train Multilingual vectors for
  English (en), German (de), Czech (cz), Hindi (hi), Arabic (ar)

- Option 1: en-de, en-cz, en-hi, en-ar (English as the bridge)

- Option 2: en-de, de-cz, cz-hi, hi-ar (serial bridging)

- Option 3: just do all pairs!

The choice of method usually depends on what is available.

Recap - Canonical Correlation Analysis (CCA)

$$\mathbf{V}, \mathbf{W} = \mathrm{CCA}(\mathbf{X}, \mathbf{Y})$$

$$\mathbf{X}^*, \mathbf{Y}^* = \mathbf{V}\mathbf{X}, \mathbf{W}\mathbf{Y}$$

$$\mathbf{x} \rightarrow \mathbf{V}\mathbf{x}$$

$$\mathbf{y} \rightarrow \mathbf{W}\mathbf{y}$$

MultiCCA

$$\langle \mathbf{Vx}, \mathbf{Wy} \rangle$$

$$\langle \mathbf{Vz}, \mathbf{Wy} \rangle = \langle \mathbf{W}^\intercal \mathbf{Vz}, \mathbf{y} \rangle$$

Perform CCA for English and Spanish.
Project Spanish vectors into the shared space, do not modify English vectors.
Perform CCA for English and German.
Project German vectors into the shared space, do not modify English vectors.

...

Ammar et al. (2016)

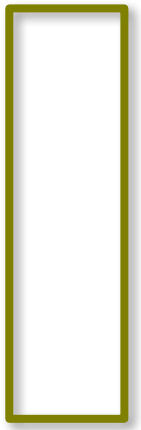
Questions

- Say I have aligned 100 languages using MultiSkip or MultiCCA.
  - What happens when I want to add a new language?
  - Which method is preferable?

- Both MultiSkip and MultiCCA (as described today) use English as the bridge.
  - Say I want to align Russian, Ukranian, Slovenian and English together.
  - What can go wrong?

Using Multilingual Encyclopedic Resources

entity \ word	president	executive	…	administration	torre	grande	…	politico
Barack_Obama	95	11	…	22	1	1	…	17
United_States	28	3	…	3	0	10	…	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
London	0	2	…	3	10	12	…	2
		English				Spanish		



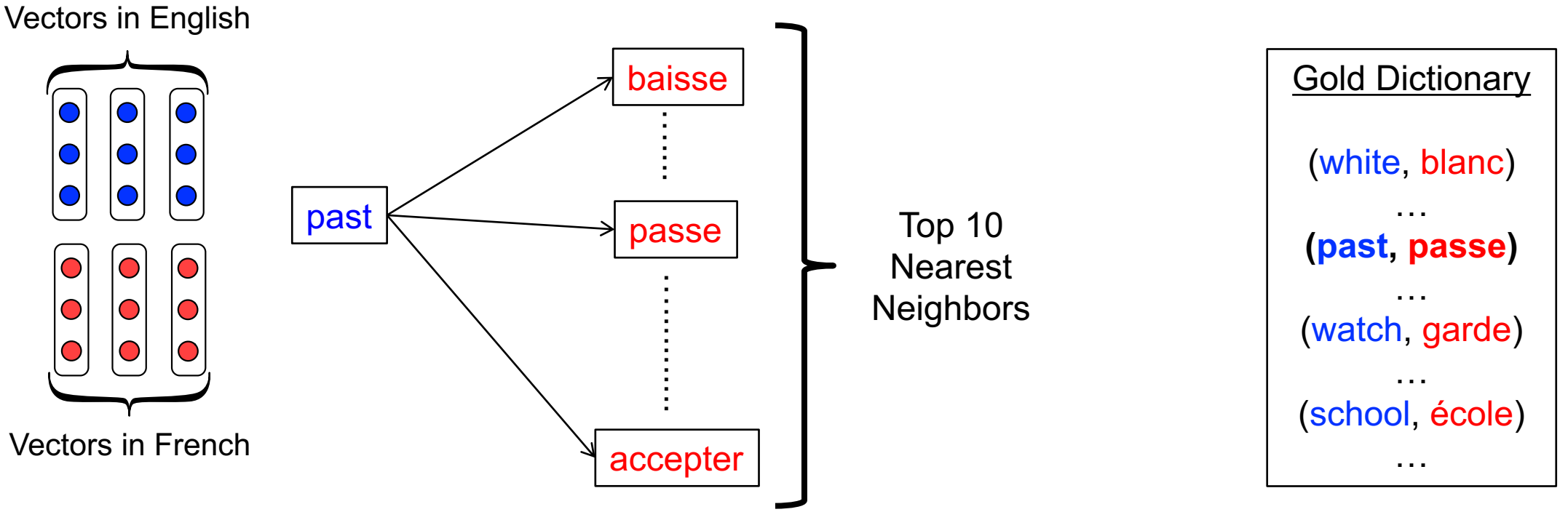
Evaluation of Cross-lingual Representations

Two forms of Evaluation

- Intrinsic
  - How good are vectors at representing word meaning and relationships?
  - Involves probing vectors to see if they encode known relationships.
  - E.g., can we identify the correct translation using the vectors?

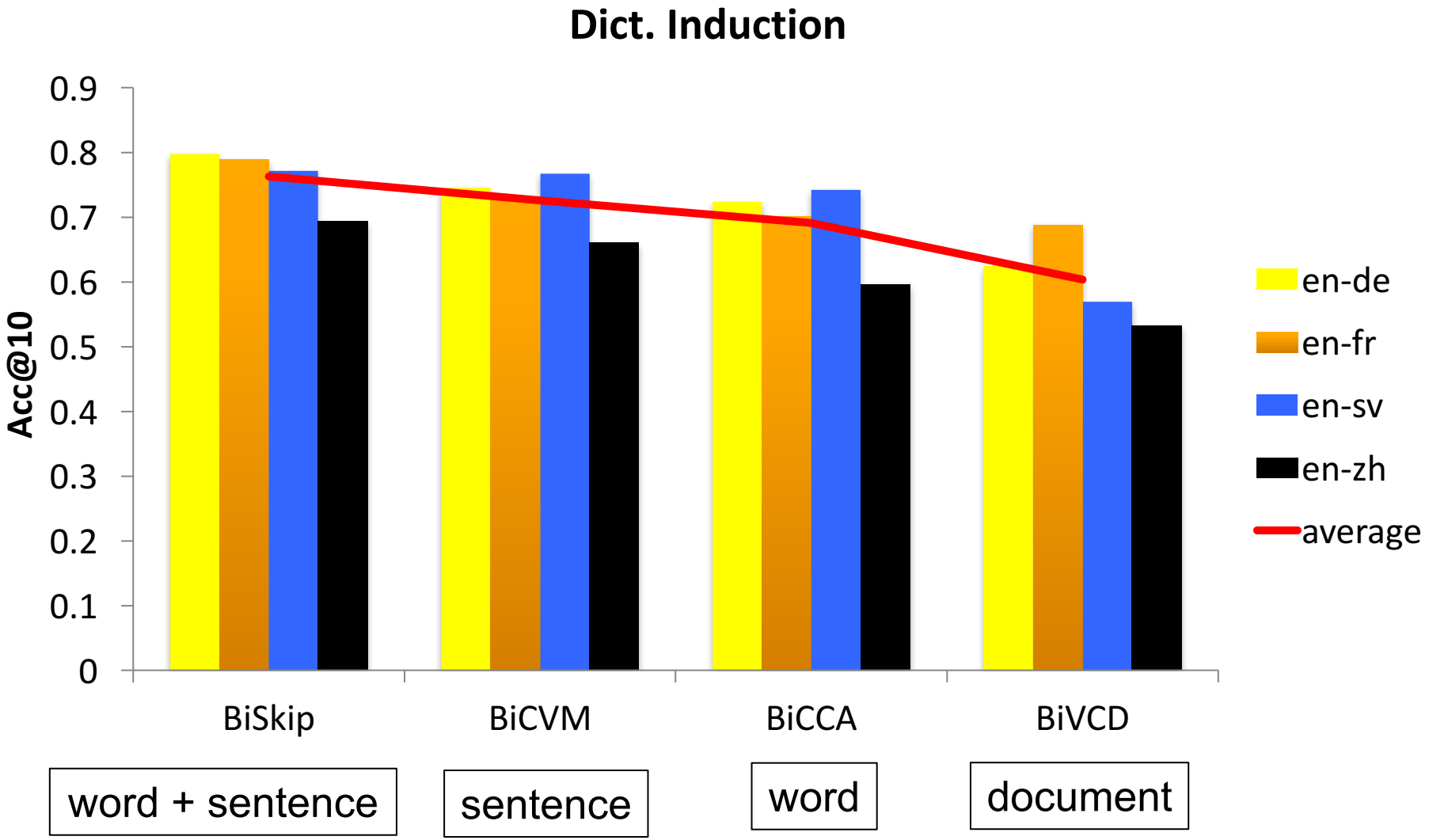
- Extrinsic
  - How good are vectors when used as features in downstream tasks?
  - Involves using the vectors as features in a model performing a different task
  - E.g., use word vectors for doing NER.

Bilingual Dictionary Induction

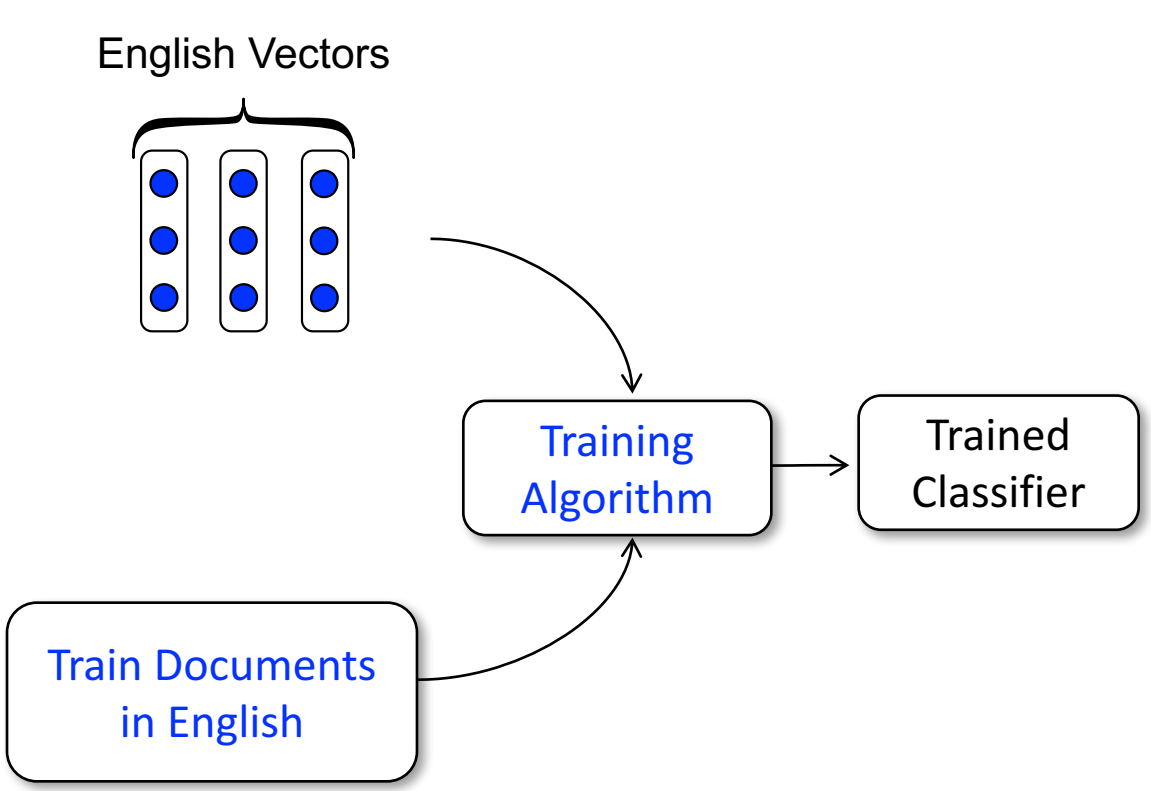


Accuracy@10 – How often is the correct translation in the top-10 neighbors?

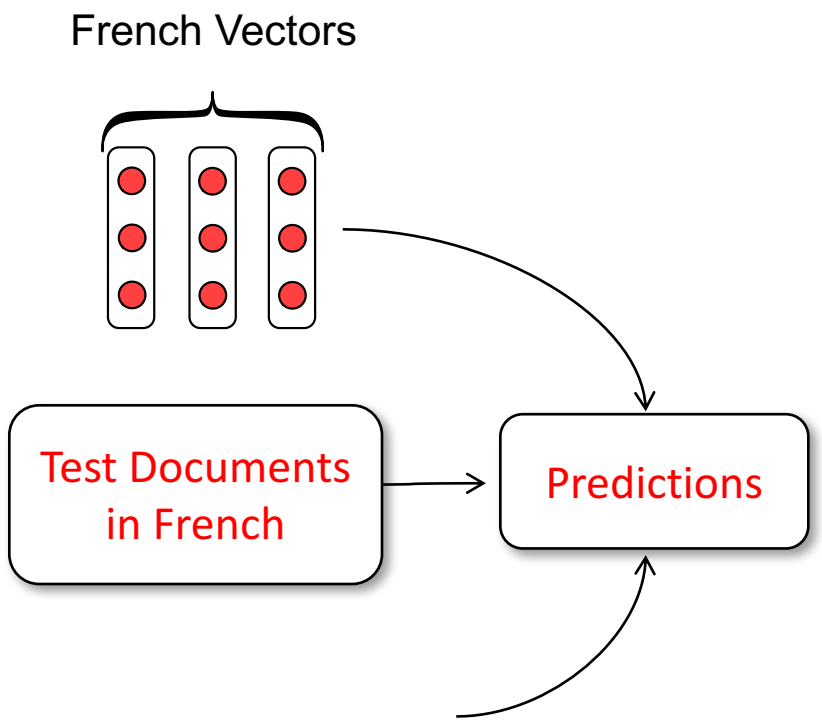
Some Results



Cross-lingual Document Classification

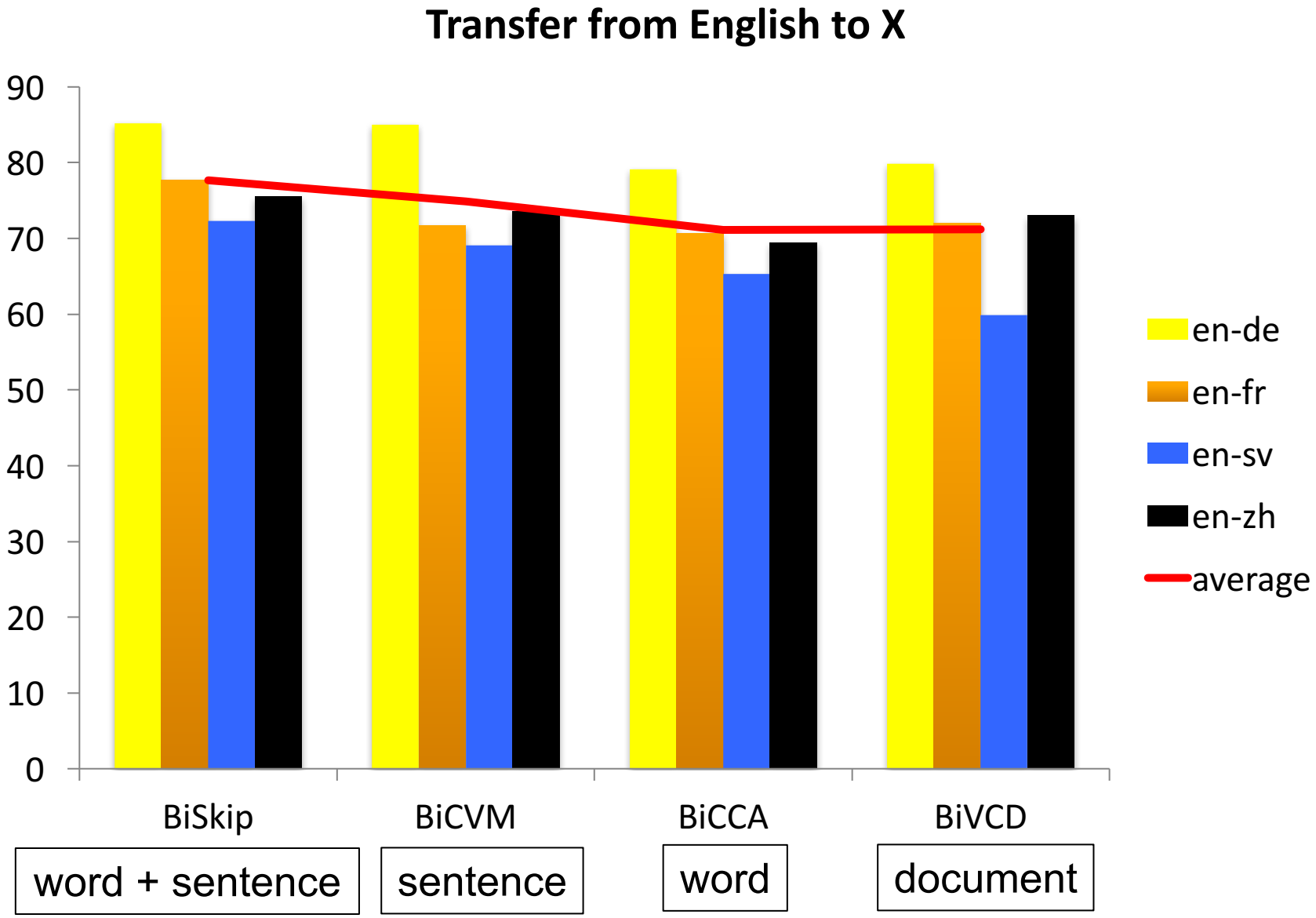


Training on English

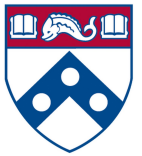


Testing on French

Some Results



Which evaluation is better? Intrinsic or Extrinsic?



- Hope is "*good performance intrinsic evaluation leads to good performance on extrinsic evaluation*".
- Rarely true in practice.
- Intrinsic evaluation is still useful (but not sufficient).

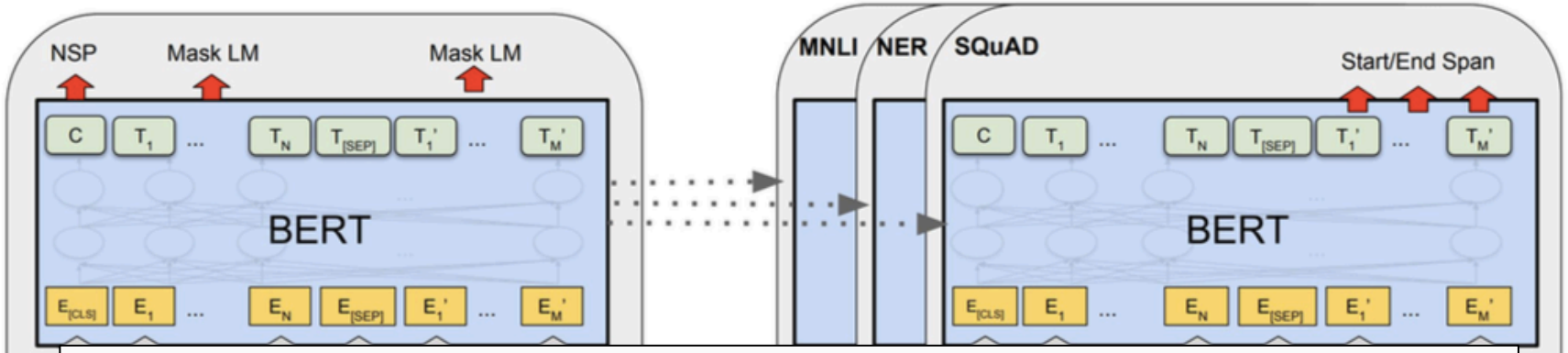
The Limitations of Cross-language Word Embeddings Evaluation

Amir Bakarov[†*] **Roman Suvorov**[*] **Ilya Sochenkov**[‡*]
[†]National Research University Higher School of Economics,
[*]Federal Research Center 'Computer Science and Control' of the Russian Academy of Sciences,
[‡]Skolkovo Institute of Science and Technology (Skoltech),
Moscow, Russia
amirbakarov@gmail.com, rsuvorov@isa.ru, ivsochenkov@gmail.com

Problems With Evaluation of Word Embeddings Using Word Similarity Tasks

Manaal Faruqui[1] **Yulia Tsvetkov**[1] **Pushpendre Rastogi**[2] **Chris Dyer**[1]
[1]Language Technologies Institute, Carnegie Mellon University
[2]Department of Computer Science, Johns Hopkins University
{mfaruqui,ytsvetko,cdyer}@cs.cmu.edu, pushpendre@jhu.edu

Recent Work – Multilingual BERT



- Trained single model on 104 languages from Wikipedia. Shared 110k WordPiece vocabulary.

System	English	Chinese	Spanish
XNLI Baseline - Translate Train	73.7	67.0	68.8
XNLI Baseline - Translate Test	73.7	68.4	70.7
BERT - Translate Train	81.9	76.6	77.8
BERT - Translate Test	81.9	70.1	74.9
BERT - Zero Shot	81.9	63.8	74.3

- XNLI is MultiNLI translated into multiple languages. Always evaluate on human-translated Test.
- Translate Train: MT English Train into Foreign, then fine-tune.
- Translate Test: MT Foreign Test into English, use English model.
- Zero Shot: Use Foreign test on English model.

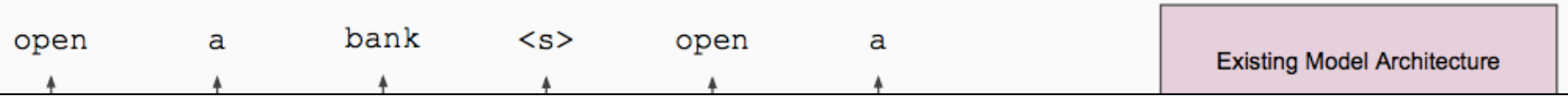
We discussed some of these techniques earlier today

Recent Work – Multilingual ELMo

- *ELMo: Deep Contextual Word Embeddings*, AI2 & University of Washington, 2017

Train Separate Left-to-Right and Right-to-Left LMs

Apply as "Pre-trained Embeddings"



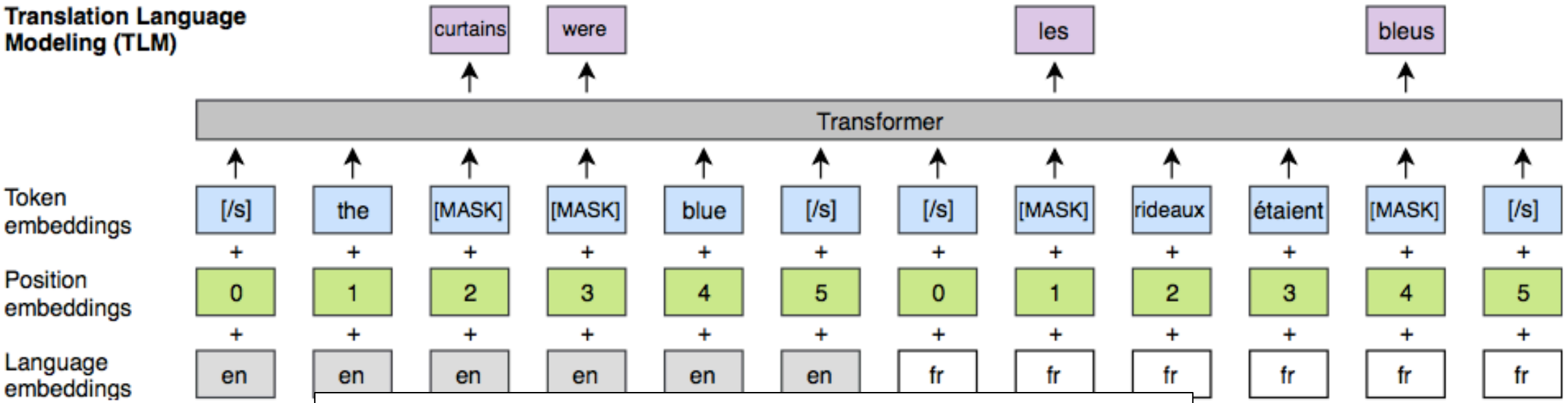
Polyglot Contextual Representations Improve Crosslingual Transfer

Phoebe Mulcaire[♡] **Jungo Kasai**[♡] **Noah A. Smith**[♡♢]
[♡]University of Washington, Seattle, WA, USA
[♢]Allen Institute for Artificial Intelligence, Seattle, WA, USA
{pmulc,jkasai,nasmith}@cs.washington.edu

Recent Work – Pre-training with Parallel Data

The curtains were blue

Les rideaux etaient bleus

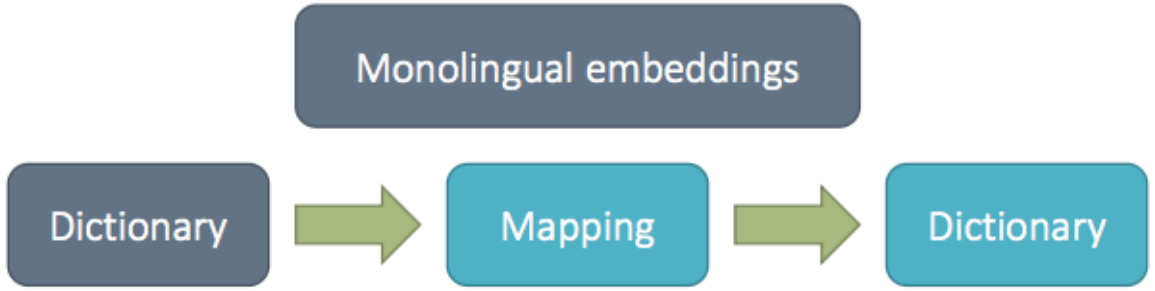


Cross-lingual Language Model Pretraining

Guillaume Lample*
Facebook AI Research
Sorbonne Universités
glample@fb.com

Alexis Conneau*
Facebook AI Research
Université Le Mans
aconneau@fb.com

Recent Work – Self-learning Embeddings



Now used as initialization step for doing **Unsupervised(!!)** Machine Translation in recent works

from Mikel Artexte's talk

Research Directions

- Handling Out of Vocabulary (OOV) Words.

Donaudampfschifffahrtsgesellschaftskapitän

“Danube steamship company captain”

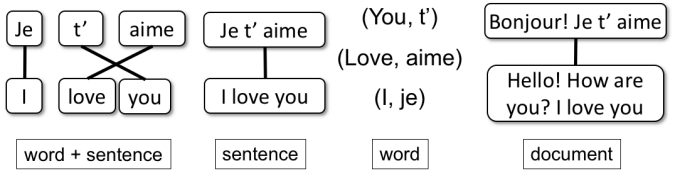
- Good Cross-lingual **Sentence** Representations.

English	You don’t have to stay there. You can leave.	Face-To-Face	Entailment
French	La figure 4 montre la courbe d’offre des services de partage de travaux. Les services de partage de travaux ont une offre variable.	Government	Entailment

Some of My Work in Cross-lingual NLP

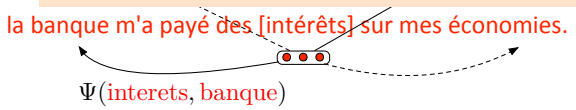
Lexical Semantics

Value of different forms of cross-lingual alignments



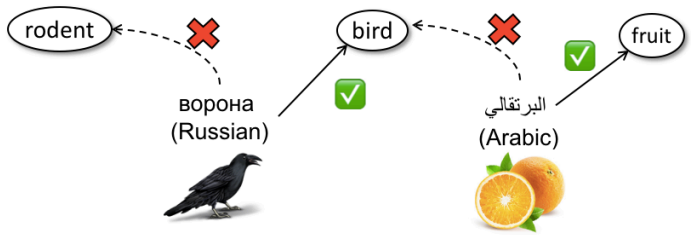
Upadhyay et al. (ACL 2016)

Multi-sense Embeddings with Cross-lingual Signals



Upadhyay et al. (Repl4NLP 2017, **Best Paper**)

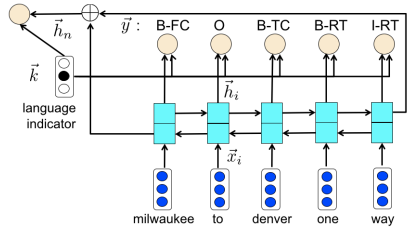
Cross-lingual Lexical Entailment



Upadhyay et al. (NAACL 2018)

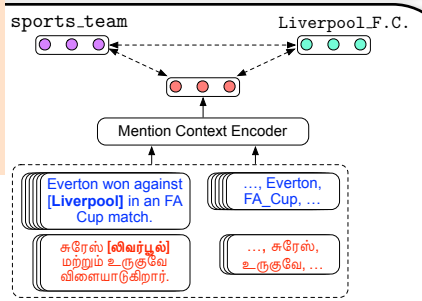
Information Extraction

Cross-lingual Spoken Language Understanding



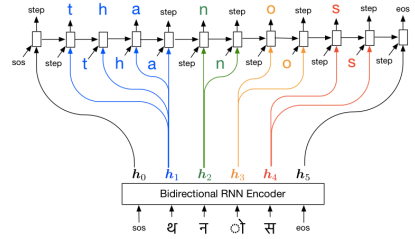
Upadhyay et al. (ICASSP 2018)

Entity Linking



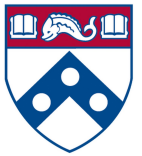
Upadhyay et al. (EMNLP 2018)

Low Resource Transliteration



Upadhyay et al. (EMNLP 2018)

Catch me after the class if you are interested in discussing these



Questions?