A Comparison of Methods for Identifying the Translation of Words in a Comparable Corpus: Recipes and Limits

Laurent Jakubina
jakubinl@iro.umontreal.ca

dirigé par
Philippe Langlais

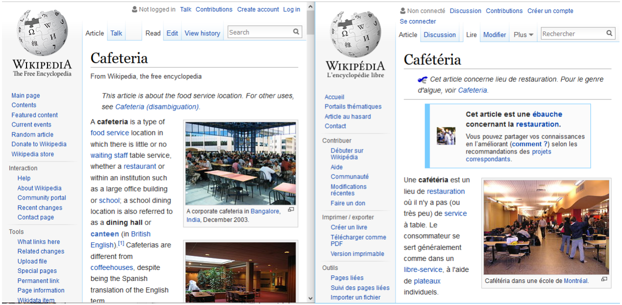
RALI - DIRO
Université de Montréal

RALI-OLST 24 February 2016

Introduction

Definition

A comparable corpus is a pair of text(s) in two (or more) different languages, that talk about the same subject (domain, event(s), person(s), etc.) but are **not** the literal translation of each others.



Well-known Example : Wikipedia (aligned-document)

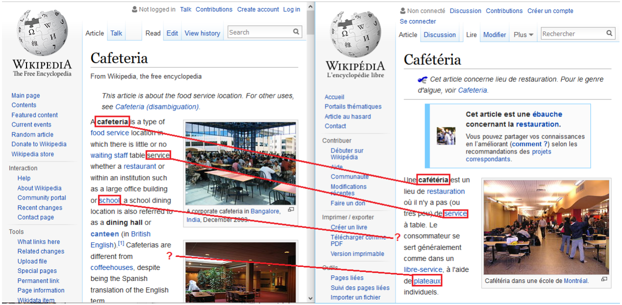
BUCC : Building and Using Comparable Corpora

An International Conference and a Community working on and with Comparable Corpora.
Interested ? Their State-Of-The-Art Book [Sharoff et al., 2013]

Introduction

Definition

A comparable corpus is a pair of text(s) in two (or more) different languages, that talk about the same subject (domain, event(s), person(s), etc.) but are **not** the literal translation of each others.



Well-known Example : Wikipedia (aligned-document)

BUCC : Building and Using Comparable Corpora

An International Conference and a Community working on and with Comparable Corpora.
Interested ? Their State-Of-The-Art Book [Sharoff et al., 2013]

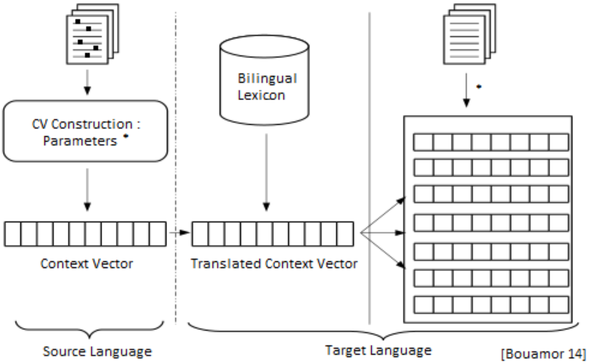
Plan

Approaches

Context-Based Projection (`context`)

Assumption :

If two words co-occur more often than expected from chance in a source language, then theirs translations must co-occur more often than expected from chance in a targuet language. [Rapp, 1995]



Steps of the `context` approach in a nutshell

Context-Based Projection - Construction

Lorem [Ipsum is simply dummy text of the] printing and typesetting industry. Lorem Ipsum has been [the industry's standard dummy text ever since] the 1500s.

		dummy	lorem	printing	standard	**text**	ipsum	the	⋮
		2	2	1	1	2	2	3	…
dummy	2	X	0	0	1	**2**	1	2	6

Cooccurrence Vector with a Contextual Window of Size 3.

Context-Based Projection - Construction

Lorem Ipsum is simply dummy text of the printing and typesetting industry.
Lorem Ipsum has been the industry's standard dummy text ever since the 1500s.

		dummy	lorem	printing	standard	text	ipsum	the	⋮
		2	2	1	1	2	2	3	...
dummy	2	X	**1**	**1**	1	2	**2**	2	**11**

Cooccurrence Matrix with a Contextual Window of Size 10.

Parameters :

- Contextual Window Sizes : `1 (3), 3 (7), 7 (15), ..., 15 (31), ...`
- (Number of visited occurrences per word : unlimited)

Context-Based Projection - Construction

		dummy	lorem	printing	standard	text	ipsum	the	⋮
		2	2	1	1	2	2	3	...
dummy	2	X	0	0	1	2	1	2	6

Cooccurrence Vector of word *dummy* with a Contextual Window of Size 3.

Assocation Measure

Function using (cooccurrence) frequencies of words w_1 and w_2 to return a single real number for the pair (w_1, w_2).

Observed Frequencies

	w_1	$\neg w_1$
w_2	O_{11} **(2)**	O_{12} **(2)**
$\neg w_2$	O_{21} **(2)**	O_{22}

$$e.g.\ ORD(w_1, w_2) = \log \frac{(O_{11} + \frac{1}{2})(O_{22} + \frac{1}{2})}{(O_{12} + \frac{1}{2})(O_{21} + \frac{1}{2})} \quad (1)$$

	lorem	printing	standard	text	ipsum	the	⋮
dummy	0.672	1.183	2.282	2.282	1.771	1.945	...

Context Vector of word *dummy*.

Context-Based Projection - Construction

Parameters :

- **Association Measures** : `Point-Wise Mutual Information (PMI), Odd Ratio Discontinu (ORD), Log Likelihood Ratio (LLR), Chi-Square (CHI)` [Evert, 2005]

cafeteria

PMI [1]	ORD	LLR	CHI
lemell (17)	portio (13)	gymnasium (2412441)	roenbergensi (1129770)
kaffitar (17)	lemell (13)	room (2411686)	gymnasium (845933)
374,429 (17)	kaffitar (13)	library (2411679)	auditorium (585119)
roseteria (17)	374,429 (13)	auditorium (2411541)	britling (574579)
hyangjeokdang (17)	roseteria (13)	school (2411263)	portio (360902)
obbolaawwanii (17)	hyangjeokdang (13)	restaurant (2410799)	uhlhornsweg (324810)
library.in (17)	obbolaawwanii (13)	classroom (2410680)	gym (282499)
amraai (17)	library.in (13)	gym (2410296)	eszpresszó (240600)
albergus (17)	amraai (13)	student (2410014)	classroom (212006)
portio (17)	albergus (13)	building (2409730)	lemell (180451)
seulkimaru (17)	seulkimaru (13)	shop (2409718)	kaffitar (180451)
coffito (17)	coffito (13)	office (2409647)	374,429 (180451)
and1954 (17)	and1954 (13)	new (2409616)	roseteria (180451)
chauhaus (17)	chauhaus (13)	hall (2409553)	hyangjeokdang (180451)
...	...	...	...

1. [Levy et al., 2015] - p3 : A well-known shortcoming of PMI, is its bias towards infrequent events (Turney and Pantel, 2010).

Context-Based Projection - Projection

cafeteria (K)	lemell (17)	kaffitar (17)	...	sundeck (12)	deck (12)	...
	lounge (9)	pool (6)	...			

Seed Bilingual Lexicon

English Terms		French Reference Translation(s)					
...	...	...	...	...	...	...	...
diplodocus	332	diplodocus	189				
invested	15610	constitué	32062	constituée	20902	placé	23171
mat	15907	carpette	71	mat	3066	mate	790
loyal	24843	loyal	1649				
...	...	...	...	...	...	...	...

cafeteria (K)	lemell (17)	kaffitar (17)	...	véranda (12)	pont (12)	...
	sportsplex (9)	flâner (9)	...			

cafeteria (-K)	véranda (12)	pont (12)	...	flâner (9)	prélasser (9)	...
	trust (6)	piscine (6)	...			

Parameters (among) :

- Size : `See later`
- Keep (K) or not (-K) the Unknown Words (!)

Context-Based Projection - Alignment

Before

Do *Construction Step* for all French Vocabulary Words ($\approx 3.5M$). No *Projection Step*.

Context (Words) Alignment

Projected Source Context Vector :

cafeteria	scolaire (24.11)	$\boxed{\text{restaurant}}$	(24.10)	$\boxed{\text{étudiante}}$	(24.10)	$\boxed{\text{construire}}$	(24.09)	...

Target Context Vectors :

			....				
cafétaria	intermarché (24.08)	$\boxed{\text{étudiante}}$ (24.08)		terrasse (24.08)	$\boxed{\text{restaurant}}$	(24.08)	...
			....				
supérette	écomarché (20.06)	déchetterie (18.43)		cybermarché (17.96)	$\boxed{\text{construire}}$	(16.83)	...
			....				

Similarity Measure

Function returning a single real number according to common members between *Projected Source* and *Target* Context Vectors.

$$e.g.\ cos(v_{src}, v_{trg}) = \frac{v_{src} \cdot v_{trg}}{||v_{src}|| \cdot ||v_{trg}||} \tag{2}$$

Context-Based Projection - Alignment

Outputs/Results

			....		
cafeteria	cafétéria (0.053)	cafeteria (0.051)	supérette (0.050)	buanderie (0.045)	...
sinusoid	sinusoïdale (0.081)	susceptible (0.078)	sinusoïde (0.076)	longitudinale (0.073)	...
explanatory	explicatif (0.064)	supplémentaire (0.056)	épistémologique (0.055)	explicative (0.054)	...
stereo	stereo (0.047)	magnétoscope (0.042)	flanger (0.039)	égaliseur (0.038)	...
			....		

Parameters :

- Target Vocabulary Size : `All (French Words in Wikipedia)`
- (Context Vector Size : `unlimited`)
- Similarity Measure : `Cosine Similarity` [Laroche and Langlais, 2010]

Document-Based Alignment (document)

- [Prochasson and Fung, 2011]
- Initially proposed for handling the translation of **rare words**.
- Context (Words) Vectors → Context "Documents" Vectors

Aligned Comparable Corpora (E.g. Wikipedia)

English ⟷ **French**

Coffeehouse ⟷ **Café (établissement)**

A coffeehouse may share some of the same characteristics of a bar or restaurant, but it is different from a cafeteria . Many coffee houses . . .

Les synonymes varient selon l'ancrage culturel de leur public ou de leur implantation géographique : bar, bistrot, cafétaria , troquet, estaminet, . . .

Wayside (TV series)

. . .the sixteenth floor contains the cafeteria and kitchen ;[1] the (technically nonexistent) nineteenth floor contains a chute . . .

Ikea ⟷ **Ikea**

. . .at the exit café (cafetaria) as well as beef hot dogs, while in United Kingdom . . .

. . . cafétaria . . . supérette . . .

Kellogg's

. . . cafétaria . . .

Document-Based Alignment (`document`)

Document Alignment

Source Context Vector :

cafeteria	$\boxed{\textbf{Coffeehouse}}$ (1.0)	**Wayside (TV series)** (1.0)	$\boxed{\textbf{Ikea}}$ (1.0)	...	...		

Target Context Vectors :

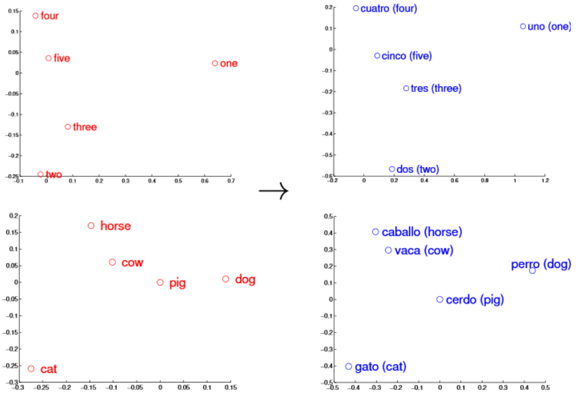
		....			
cafétaria	$\boxed{\textbf{Café}}$ (1.0)	$\boxed{\textbf{Ikea}}$ (1.0)	**Kellogg's** (1.0)	...	...
		....			
supérette	**Ikea** (1.0)	...	...	...	...
		....			

Parameters :

- Document Pairs : `All` (750 000) vs 20 000 [Prochasson and Fung, 2011]
- Target Vocabulary Size : `All` (3M) vs 120K [Prochasson and Fung, 2011]

Word Embedding Alignment (`embedding`)

- [Mikolov et al., 2013b] + [Dinu and Baroni, 2014]
- Continuous representation(s) of words (Embedding) show projection similarities between Languages.



[Mikolov et al., 2013b]

Word Embedding Alignment (`embedding`)

Embedding **Construction**, for both English and French language

We used the `Word2Vec`[2] toolkit [Mikolov et al., 2013a].

English Embeddings

...	...	...	...	...	...
text	0.22	0.09	−0.77	−0.12	...
cafeteria	−0.32	−0.28	−0.08	0.25	...
stereo	1.97	−0.30	−0.35	−0.22	...
dummy	0.28	0.24	−0.36	−0.07	...
one	−0.36	0.23	−0.52	−0.05	...
...	...	...	...	...	...

French Embeddings

...	...	...	...	...	...
texte	0.70	1.56	0.57	0.27	...
un	0.06	−1.44	0.14	−0.24	...
cafeteria	0.00	0.36	−1.07	0.45	...
ville	1.75	−1.43	1.15	−0.32	...
saint	1.31	−0.03	0.69	0.24	...
...	...	...	...	...	...

Parameters :

- Neural Network Architecture : `Context Bag-of-Word` or `Skip-Gram`.
- Optimized Training Algorithm : `Hierarchical Softmax` or `Sampling (5,10)`.
- Dimensionnality (Vector Size).
- Contextual Window Size : like `context` approach.

2. `https://code.google.com/p/word2vec/`

Word Embedding Alignment (`embedding`)

Embedding Alignment

= Learning a projection (linear mapping) from English Embeddings to French Embeddings.

English Embeddings					Projection				French Embeddings			
...	...	...	...		...	...	...		...	...	...	...
text	0.22	0.09	...		1.73	4.83	...		*texte*	0.70	1.56	...
cafeteria	−0.32	−0.28	...		−7.21	−3.93	...		*un*	0.06	−1.44	...
stereo	1.97	−0.30	...	↔	5.07	−3.41	...	↔	*cafeteria*	0.00	0.36	...
dummy	0.28	0.24	...		5.42	8.51	...		*ville*	1.75	−1.43	...
one	−0.36	0.23	...		1.57	−1.25	...		*saint*	1.31	−0.03	...
...	...	...	...		...	...	...		...	...	...	...

We used the implementation from [Dinu and Baroni, 2014].

Parameters :

- Size
- Nature (e.g. Highest Frequencies)

Experimental Protocol

(Aligned) Comparable Corpora

- `Wikipedia` dump of June 2013 in both English and French.
- 757 287 paired documents (by inter-language links).
- Used without any particular cleaning ($\neq$ similar studies).

	English Wikipedia	French Wikipedia
# Docs	3 539 093	1 334 116
# Voc	7 321 576	3 652 871
# Tokens	1 204 699 806	330 886 854

Summary Statistics for the English and French Wikipedia (2013)

Seed Bilingual Lexicon

- `context` and `embedding` both require a seed bilingual lexicon : we used an in-house one.
- We recover the frequency of each (English and French) word in Wikipedia.

English Terms		French Reference Translation(s)					
...	...	...	...	...	...	...	...
diplodocus	332	diplodocus	189				
invested	15610	constitué	32062	constituée	20902	placé	23171
mat	15907	carpette	71	mat	3066	mate	790
loyal	24843	loyal	1649				
...	...	...	...	...	...	...	...

- For `embedding` ;
  - **5k-high** : Top 5 000 entries with highest frequencies [Mikolov et al., 2013a, Dinu and Baroni, 2014].
  - **5k-rand** : 5 000 entries randomly picked.
  - **2k-low** : 2 000 entries involving *rare*[3] English words.
- For `context` ;
  - **All** : 107 799 entries.
  - "More is the Best"

3. Words occurring at most 25 times

Test Sets

- 2 list of English source terms and their reference (French) translation.
  - **1k-low** : 1 000 *rare* English words (and their translations).
  - **1k-high** : 1 000 "frequent" English words (and their translations).
- Why rare words ? 6.8 million words (92%) in English Wikipedia occur less than 26 times.
- Half of the test words have only one reference translation, the remainder having an average of 3 translations.

Examples (5 entries in each test set)

	English Term	French Reference Translation(s)		
Frequent	coloration	coloration		
	tempestuous	orageux	tempétueux	
	hinny	bardeau	bardot	
	malpractice	malfaçon	malversation	négligence
	compile	compiler		

	English Term	French Reference Translation(s)
Rare	veratrine	vératrine
	centiliter	centilitre
	rescindable	résiliable
	mundanely	prosaïquement
	filmsetter	photocomposeuse

Metrics

- Each approach produces a ranked list of (at most) 20 (French) translation candidates for each (English) test word.
- Performance = accuracy at rank 1, 5 and 20 (`TOP@i`).
- `TOP@i` = the percentage of test words for which a reference translation is identified in the first *i* proposed candidates.

Example

- test-word$_1$ = [**cand**$_{11}$, cand$_{12}$, cand$_{13}$, cand$_{14}$, cand$_{15}$] (reference is the 1 candidate)
- test-word$_2$ = [cand$_{21}$, cand$_{22}$, **cand**$_{23}$, cand$_{24}$, cand$_{25}$] (reference is the 3 candidate)
- test-word$_3$ = [cand$_{31}$, cand$_{32}$, cand$_{33}$, cand$_{34}$, **cand**$_{35}$] (reference is the 5 candidate)
- ————————————————————————————-
- `TOP@1` = 33%
- `TOP@5` = 100%

Results and Recipes

All Results

- Best variant for each approach according to `TOP@1`.
- An *Oracle* shows that approaches are complementary.
- Disappointment for the poor performance of the `document` approach which was specifically designed to handle rare words.

	1k-low			**1k-high**		
	TOP@1	TOP@5	TOP@20	TOP@1	TOP@5	TOP@20
`embedding`	**2,2**	**6,1**	**11,9**	**21,7**	**34,2**	**44,9**
`context`	2,0	4,3	7,6	19,0	32,7	44,3
`document`	0,7	2,3	5,0	*10,0*	*19.0*	*24.0*
oracle	4,6	10,5	19,0	31,8	46,8	57,6

Why so bad ?

- "Thesaurus Effect".
- Morphological variations.
- Correct candidates. . .but not in our reference.

donut		beigne		
context	-	aromatisé (0.05)	donut (0.05)	beignet (0.04)
embedding	-	liper (0.54)	babalous (0.53)	savonnettes (0.52)
brilliantly		brillamment		
context	-	imaginatif (0.05)	captivant (0.05)	rusé (0.05)
embedding	-	éclatant (0.69)	pathétique (0.67)	émouvant (0.66)
gentle		doucet,	doux,	délicat
context	-	enjoué (0.05)	serviable (0.05)	affable (0.04)
embedding	-	colérique (0.76)	enjoué (0.75)	espiègle (0.75)
pathologically		pathologiquement		
context	-	cordonale (0.05)	pathologique (0.05)	diagnostiqué (0.05)
embedding	-	psychosexuel (0.60)	psychoaffectif (0.60)	piloérection (0.59)

Why so bad ?

- "Thesaurus Effect".
- Morphological variations.
- Correct candidates. . .but not in our reference.

donut		beigne		
context	-	aromatisé (0.05)	donut (0.05)	beignet (0.04)
embedding	-	liper (0.54)	babalous (0.53)	savonnettes (0.52)
brilliantly		brillamment		
context	-	imaginatif (0.05)	captivant (0.05)	rusé (0.05)
embedding	-	éclatant (0.69)	pathétique (0.67)	émouvant (0.66)
gentle		doucet,	doux,	délicat
context	-	enjoué (0.05)	serviable (0.05)	affable (0.04)
embedding	-	colérique (0.76)	enjoué (0.75)	espiègle (0.75)
pathologically		pathologiquement		
context	-	cordonale (0.05)	pathologique (0.05)	diagnostiqué (0.05)
embedding	-	psychosexuel (0.60)	psychoaffectif (0.60)	piloérection (0.59)

Why so bad ?

- "Thesaurus Effect".
- Morphological variations.
- Correct candidates. . .but not in our reference.

donut		beigne		
context	-	aromatisé (0.05)	donut (0.05)	beignet (0.04)
embedding	-	liper (0.54)	babalous (0.53)	savonnettes (0.52)
brilliantly		brillamment		
context	-	imaginatif (0.05)	captivant (0.05)	rusé (0.05)
embedding	-	éclatant (0.69)	pathétique (0.67)	émouvant (0.66)
gentle		doucet,	doux,	délicat
context	-	enjoué (0.05)	serviable (0.05)	affable (0.04)
embedding	-	colérique (0.76)	enjoué (0.75)	espiègle (0.75)
pathologically		pathologiquement		
context	-	cordonale (0.05)	pathologique (0.05)	diagnostiqué (0.05)
embedding	-	psychosexuel (0.60)	psychoaffectif (0.60)	piloérection (0.59)

Why so bad ?

- "Thesaurus Effect".
- Morphological variations.
- Correct candidates. . .but not in our reference.

donut		beigne		
context	-	aromatisé (0.05)	donut (0.05)	beignet (0.04)
embedding	-	liper (0.54)	babalous (0.53)	savonnettes (0.52)
brilliantly		brillamment		
context	-	imaginatif (0.05)	captivant (0.05)	rusé (0.05)
embedding	-	éclatant (0.69)	pathétique (0.67)	émouvant (0.66)
gentle		doucet,	doux,	délicat
context	-	enjoué (0.05)	serviable (0.05)	affable (0.04)
embedding	-	colérique (0.76)	enjoué (0.75)	espiègle (0.75)
pathologically		pathologiquement		
context	-	cordonale (0.05)	pathologique (0.05)	diagnostiqué (0.05)
embedding	-	psychosexuel (0.60)	psychoaffectif (0.60)	piloérection (0.59)

Best hyper-parameters on **1k-high** (for `context`)

⇒ Best Model (in 25 xps) : `Window Size of 3, PMI, keeping English words in context.`

AM	T@1	T@20
PMI	**19.0**	**44.3**
ORD	18.6 [17.9]	38.0 [42.5]
LLR	2.4	7.8
CHI	1.5 [0.7]	6.3 [8.3]

WS	T@1	T@20
3	**19.0**	**44.3**
5	18.8	42.1
7	18.6	38.0
1	7.1	18.7

Keep (**K**) or Not	T@1	T@20
K (7,PMI)	**19.0**	**44.3**
¬ **K** (7,PMI)	9.6	31.7
K (15,ORD)	**18.6**	**39.0**
¬ **K** (15,ORD)	5.7	19.5

- [Jakubina and Langlais, 2015]
- **K** vs ¬ **K** : interesting discovery.
- some configs are very close.

Best hyper-parameters on **1k-low** (for `context`)

⇒ Best Model (in 50 xps) : `Window Size of 15, ORD, keeping English words in context.`

AM	T@1	T@20
ORD	**2.0**	**7.6**
PMI	1.8 [1.6]	7.6 [8.0]
LLR	1.1	2.8
CHI	0.8	2.5

WS	T@1	T@20
15 (ORD)	**2.0**	7.6
10 (PMI)	1.6	**8.0**
7 (ORD)	1.0	7.1
5 (PMI)	0.8	4.2
3 (CHI)	0.6	4.6
1 (CHI)	0.8	2.5

Keep (**K**) or Not	T@1	T@20
K (31,ORD)	**2.0**	**7.6**
¬ **K** (31,ORD)	1.0	4.3
K (15,LLR)	**0.3**	**2.1**
¬ **K** (15,LLR)	0.1	1.6

- Same tendency except window size.

Disappointing (for `document`)

- Investigation of only a few configurations.
- Sanity check : same Target Voc Size as [Prochasson and Fung, 2011].

TGS	T@1	T@20
all ($\simeq$3M)	0.7	5.0
low (120k)	4.9	20.2

- $\Rightarrow$ The approach does not scale well to large datasets.

Best hyper-parameters on **1k-high** (for embedding)

⇒ Best Model (in 50 xps) : CBOW model, negative sampling (10 samples), dimensionnality of 200[4], window size of 5 and **5k-high**.

WS	T@1	T@20
5	17.9	35.2
3	14.6	33.7
15	14.0	31.3

Training Set	T@1	T@20
5k-high	21.7	44.9
5k-rand	18.2	40.5
2k-low	1.00	10.3

- Confirms both [Mikolov et al., 2013a, Dinu and Baroni, 2014].
  - Our TOP@1 (22%) lower than TOP@1 of [Mikolov et al., 2013a] (30%) . . .
  - Our Target Voc Size is 3 millions against theirs of hundred thousands.

4. The largest dimensionality for which we managed to train for frequent words.

Best hyper-parameters on **1k-low** (for `embedding`)

⇒ Best Model (in 80 xps) : `Skip-Gram model, hierarchical softmax, dimensionnality of 250`[5]`, window size of 10 and` **`5k-rand`**.

WS	T@1	T@20
10 (200)	**1.2**	**7.1**
3 (200)	1.0	5.9
15 (200)	0.9	6.9

Training Set	T@1	T@20
5k-rand (skg,hs,250,10)	**2.2**	**11.9**
2k-low (skg,hs,250,10)	1.2	8.7
5k-rand (skg,hs,200,10)	**1.3**	**7.1**
2k-low (skg,hs,200,10)	0.7	5.5
5k-high (skg,hs,200,10)	0.4	3.2

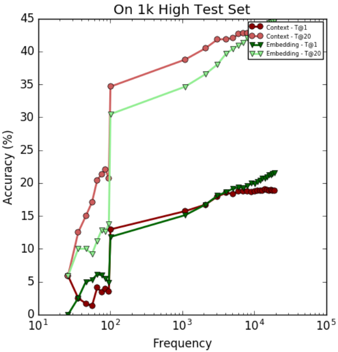
- Confirms both [Mikolov et al., 2013a, Dinu and Baroni, 2014].

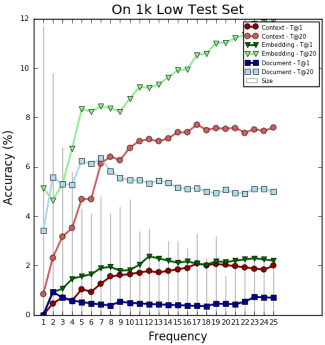
5. The largest dimensionality for which we managed to train for rare words.

Analysis

Frequency

- Performance when translating subsets of test words with a frequency [6] below a given threshold.
- The frequency bias is clearly observable.
- For some ranges of frequencies, `context` might be the good approach to go with.





6. Frequency of the source English word, in Wikipedia.

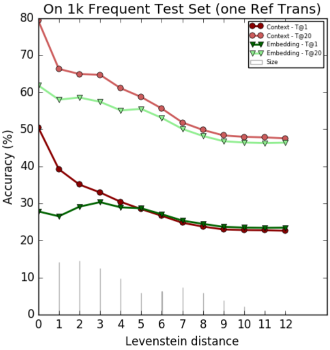
String Similarity

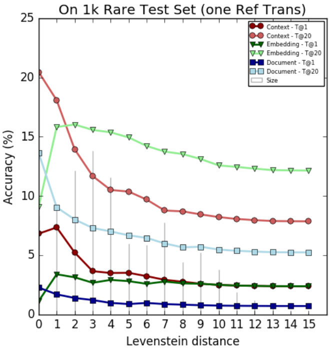
Examples :

Short Levenshtein Dist	
veratrine	vératrine

Long Levenshtein Dist	
filmsetter	photocomposeuse

- A decrease of performance for words which reference translation is dissimilar.





Medical Terms

- Why ? Often studied (e.g. [Morin and Prochasson, 2011] [Hazem and Morin, 2012] [Kontonatsios et al., 2014])
- Cross our test words with an in-house list of medical terms.
- Only 22 in **1k-low** and 88 in **1k-high** : maybe not so representative but...

	1k-low		**1k-high**	
	TOP@1	TOP@20	TOP@1	TOP@20
`embedding`	4.5 (+2.7)	13.6 (+1.7)	27.5 (+5.8)	53.7 (+8.8)
`context`	0.0 (-2.0)	4.5 (-3.1)	48.7 (+29.7)	72.5 (+28.3)
`document`	4.5 (+3.8)	22.7 (+17.7)	—	—

Conclusion

Conclusion

- Comparison of 3 approaches for identifying translations in comparable corpora.
- Extensive study of how their hyper-parameters impact performances.
- **Without reducing (somehow arbitrarily) the size of the target vocabulary.**
- Analyses of some properties, coming from (source word – target translation) pairs that we feel are worth reporting when conducting such a task.

Discussions

- On Frequent Words : `context` $(44,3) \simeq$ `embedding` $(44,9)$
  - Echoes [Levy et al., 2015]
- on *Rare* Words : `embedding` $(11,9) >$ `context` $(7,6) >>$ `document` $(5,0)$
  - Definitely, translating rare words is a challenge that deserves further investigations.
- Combinaison $= (+2,4)$ to $(+7,1)$ on *rare* words and $(+10,1)$ to $(+13,0)$ on frequent words.
  - Some evidences that the approaches we tested are complementary and that combining their outputs should be fruitful.
- According to some properties of test words (nature, frequency) and some results from hyper-parameters study, combining different variants of the same approach should lead to better performance.

Thank You !

Questions ?

Bibliography I

Dinu, G. and Baroni, M. (2014).
Improving zero-shot learning by mitigating the hubness problem.
ResearchGate.

Evert, S. (2005).
The statistics of word cooccurrences.
PhD thesis, Dissertation, Stuttgart University.

Hazem, A. and Morin, E. (2012).
Adaptive Dictionary for Bilingual Lexicon Extraction from Comparable Corpora.
In *LREC*, pages 288–292.

Jakubina, L. and Langlais, P. (2015).
Projective methods for mining missing translations in DBpedia.
ACL-IJCNLP 2015, page 23.

Kontonatsios, G., Korkontzelos, I., Tsujii, J., and Ananiadou, S. (2014).
Combining String and Context Similarity for Bilingual Term Alignment from Comparable Corpora.
In *EMNLP*, pages 1701–1712.

Bibliography II

Laroche, A. and Langlais, P. (2010).
Revisiting Context-based Projection Methods for Term-translation Spotting in Comparable Corpora.
In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 617–625. Association for Computational Linguistics.

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., Chen, K., Corrado, G., and Dean, J. (2013a).
Efficient estimation of word representations in vector space.

Mikolov, T., Le, Q. V., and Sutskever, I. (2013b).
Exploiting similarities among languages for machine translation.

Morin, E. and Prochasson, E. (2011).
Bilingual lexicon extraction from comparable corpora enhanced with parallel corpora.
In *Proceedings of the 4th workshop on building and using comparable corpora : comparable corpora and the web*, pages 27–34. Association for Computational Linguistics.

Bibliography III

Prochasson, E. and Fung, P. (2011).
Rare Word Translation Extraction from Aligned Comparable Documents.
In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies - Volume 1*, HLT '11, pages 1327–1335. Association for Computational Linguistics.

Rapp, R. (1995).
Identifying Word Translations in Non-parallel Texts.
In *Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics*, ACL '95, pages 320–322. Association for Computational Linguistics.

Sharoff, S., Rapp, R., and Zweigenbaum, P. (2013).
Overviewing Important Aspects of the Last Twenty Years of Research in Comparable Corpora.
In Sharoff, S., Rapp, R., Zweigenbaum, P., and Fung, P., editors, *Building and Using Comparable Corpora*, pages 1–17. Springer Berlin Heidelberg.