

Sémantique distributionnelle, embeddings (et dong)

felipe@iro.umontreal.ca

RALI

Dept. Informatique et Recherche Opérationnelle
Université de **Montréal**



mars 2016, V0.00002 (pas terminé)



Plan

Taches typiques

(Before Deep) : modèle vectoriel

And then came the “Deep”

- Word2Vec

- Analogie

- Évaluation

- Idées intéressantes

- Le cas bilingue

Évaluation

Plan

Taches typiques

(Before Deep) : modèle vectoriel

And then came the “Deep”

- Word2Vec

- Analogie

- Évaluation

- Idées intéressantes

- Le cas bilingue

Évaluation

Crédit

- ▶ la plupart de ces ressources (et d'autres) sont listées ici :
<http://www.cs.cmu.edu/~mfaruqui/suite.html>



Rubenstein and Goodenough

[http://www.cs.cmu.edu/~mfaruqui/word-sim/
EN-RG-65.txt](http://www.cs.cmu.edu/~mfaruqui/word-sim/EN-RG-65.txt)

1	gem	jewel	3.94
2	midday	noon	3.94
3	automobile	car	3.92
4	cemetery	graveyard	3.88
5	cushion	pillow	3.84
6	...		
7	autograph	shore	0.06
8	fruit	furnace	0.05
9	noon	string	0.04
10	rooster	voyage	0.04
11	cord	smile	0.02

WordSimilarity-353

- ▶ <http://www.cs.technion.ac.il/~gabr/resources/data/wordsim353/>
 - ▶ **set1** 153 paires de mots : 13 jugements de **relatedness**
 - ▶ **set2** 200 paires de mots : 16 jugements

```

1 % more set1.csv | sort -t\, -k3,3n+
2 tiger,tiger,10.00,10,10,10,10,10,10,10,10,10,10,10,10,10,10
3 coast,shore,9.10,9,9.75,10,10,6,6,10,8,9.5,10,10,10,10,10
4 football,soccer,9.03,9,9.9,9,10,8,7,9.5,10,8,10,9,9,9,9
5 fuck,sex,9.44,10,9.75,10,10,8,9,9,10,9,10,8,10,10
6 journey,voyage,9.29,9,9.75,10,10,6,7,9.5,10,9.5,10,10,10,10
  
```

- ▶ split de <http://alfonseca.org/eng/research/wordsim353.html>

similarité 203 paires
relatedness 252 paires



À propos du split de *[Agirre et al., 2009]*

- ▶ 103 paires en commun dans les deux fichiers :

```
more wordsim_relatedness_goldstandard.txt \
  wordsim_similarity_goldstandard.txt | \
  sort | uniq -c | awk '$1 == 2 {print $0}' | \
  sort -k4,4gr | head -n 5+
```

1	2	cup	food	5,00
2	2	monk	oracle	5,00
3	2	journal	association	4,97
4	2	car	flight	4,94
5	2	street	children	4,94

MEN

[http:](http://clic.cimec.unitn.it/~elia.bruni/MEN.html)

[//clic.cimec.unitn.it/~elia.bruni/MEN.html](http://clic.cimec.unitn.it/~elia.bruni/MEN.html)

- ▶ 3000 paires de mots avec un score (entre 0 et 50) indiquant leur *relatedness*

cat-n feline-j	48	gun-n stair-n	2
copper-n metal-n	48	military-j tomato-n	2
flower-n garden-n	48	angel-n gasoline-n	1
grass-n lawn-n	48	feather-n truck-n	1
pregnancy-n pregnant-j	48	festival-n whisker-n	1

- ▶ paires vérifiant des contraintes de fréquences dans différentes ressources et échantillonnées de manière à générer une tâche équilibrée
- ▶ évaluation faite par comparaison binaire à 50 autres paires sur Mechanical Turk
- ▶ **tâche** : trouver les paires les plus similaires (ranking)



TOEFL Tests *[Freitag et al., 2005]*

- ▶ accessible ici :
<http://www.cs.cmu.edu/~dayne/wbst-nanews.tar.gz>
- ▶ chaque instance de test est de la forme :

stimulus

target : [synonym and 3 distractors]

ex :

comprehend : [encompass career admire pardon]

- ▶ 9887 instances de test pour les noms, 7398 pour les verbes

TOEFL Tests *[Freitag et al., 2005]*

force	violence	outburst encephalopathy sparta
force	power	plaintif vomiting deep
force	personnel	distress universe poe
force	strength	adrian character atlanta
force	effect	abuja flora share
buy	bribe	supply reinvent triumph
buy	corrupt	bar recount weight
buy	purchase	harp esteem forgive
record	commemorate	smack resort tremor
record	enter	ripple subpoena shirt
record	tape	drink pare pearl

Word2Vec datasets (analogies)

<http://word2vec.googlecode.com/svn/trunk>

questions-words.txt 19544 équations (avec solution) :

<p>: capital-common-countries Berlin Germany Islamabad Pakistan Baghdad Iraq Oslo Norway</p> <p>: capital-world Algiers Algeria Belmopan Belize Abuja Nigeria Doha Qatar</p> <p>: currency Armenia dram India rupee Angola kwanza Denmark krone</p> <p>: city-in-state Philadelphia Pennsylvania Atlanta Georgia Chicago Illinois Stockton California</p> <p>: family father mother husband wife brother sister prince princess</p> <p>: gram1-adjective-to-adverb cheerful cheerfully efficient efficiently apparent apparently free freely</p> <p>: gram2-opposite clear unclear impressive unimpressive aware unaware efficient inefficient</p>	<p>: gram3-comparative bright brighter simple simpler big bigger bright brighter</p> <p>: gram4-superlative bright brightest weird weirdest big biggest dark darkest</p> <p>: gram5-present-participle debug debugging code coding dance dancing enhance enhancing</p> <p>: gram6-nationality-adjective Australia Australian India Indian Albania Albanian Sweden Swedish</p> <p>: gram7-past-tense describing described playing played dancing danced vanishing vanished</p> <p>: gram8-plural bottle bottles man men bird birds bottle bottles</p> <p>: gram9-plural-verbs enhance enhances play plays describe describes increase increases</p>
--	--

Word2Vec datasets (analogies)

<http://word2vec.googlecode.com/svn/trunk>

questions-phrases.txt 3218 équations (avec solution) :

: newspapers

Boston Boston_Globe Salt_Lake Salt_Lake_Tribune
Baltimore Baltimore_Sun Cincinnati Cincinnati_Enquirer

: ice_hockey

Calgary Calgary_Flames Ottawa Ottawa_Senators
Boston Boston_Bruins Chicago Chicago_Blackhawks

: basketball

Dallas Dallas_Mavericks Boston Boston_Celtics
Boston Boston_Celtics New_York New_York_Knicks

: airlines

Finland Finnair Spain Spanair
Austria Austrian_Airlines Turkey Turkish_Airlines

: people-companies

Samuel_J._Palmisano IBM Jeff_Bezos Amazon
Eric_Schmidt Google Larry_Ellison Oracle



À propos des datasets Word2Vec

<http://word2vec.googlecode.com/svn/trunk>

- ▶ 5031 analogies impliquant la relation **capitale_of**, mais seulement 118 paires de pays / capitale ...
- ▶ 870 analogies impliquant la relation **plural-verbs**, mais seulement 30 paires de verbes ...
 - ▶ 27 avec la relation **+s** ; ex : **eat / eats**
 - ▶ et seulement 3 avec la relation **+es** : **go / goes**, **search / searches** et **vanish / vanishes**
 - ▶ **note** : [go: goes :: eat : eats] n'est pas une analogie formelle
- ▶ 1560 analogies impliquant la relation **past-tense**, mais seulement 40 paires de verbes ...
 - ▶ ex : **decreasing / decreased**, **knowing / knew**
 - ▶ dont la moitié (23) sont des verbes irréguliers ; ex : **sitted / sat** et **sleeped / slept**

MSR Sentence Completion dataset

<http://research.microsoft.com/en-us/projects/scc>

► 1040 questions comme celles-ci :

26 It may be that the solution of the one may _____ to be the solution of the other.

- a) prove
- b) learn
- c) choose
- d) forget
- e) succumb

27 We shall just be in time to have a little _____ with him.

- a) breakfast
- b) elegance
- c) garment
- d) basket
- e) dog

MSR Sentence Completion dataset

<http://research.microsoft.com/en-us/projects/scc>

► 1040 questions comme celles-ci :

26 It may be that the solution of the one may **prove** to be the solution of the other.

- a) prove
- b) learn
- c) choose
- d) forget
- e) succumb

27 We shall just be in time to have a little **breakfast** with him.

- a) breakfast
- b) elegance
- c) garment
- d) basket
- e) dog



BLESS

<https://sites.google.com/site/geometricalmodels/shared-evaluation>

- ▶ 26 554 lignes comme celle-ci :

```
frog-n amphibian_reptile attri green-j
```

qui indique qu'une grenouille est un amphibien ou un reptile et qu'un de ses attributs est d'être verte (gasp).

- ▶ 200 concepts (reptiles, mammifères, oiseaux, fruits, insects, armes, etc.) , 5 relations (attribut, co-hyponyme, hyperonyme, meronyme, événement)

On average, there are 14 attri, 18 coord, 19 event, 7 hyper and 15 mero relata per concept.

- ▶ 14 400 concept-relation-relatum + 12 154 control tuples (random)
- ▶ une tâche suggérée



BLESS/frog

<https://sites.google.com/site/geometricalmodels/shared-evaluation>

mero	blood-n bone-n eye-n foot-n gill-n head-n leg-n lung-n poison-n skin-n tongue-n wart-n
event	breathe-v catch-v chase-v croak-v die-v drink-v eat-v hop-v kill-v lay-v leap-v live-v sit-v swim-v
attri	amphibious-j aquatic-j brown-j colorful-j edible-j green-j little-j loud-j noisy-j old-j poisonous-j small-j ugly-j wild-j
coord	alligator-n bullfrog-n lizard-n snake-n toad-n turtle-n
hyper	amphibian-n animal-n beast-n chordate-n creature-n vertebrate-n

+ 50 relations aléatoires

frog-n amphibian_reptile random-n doyen-n



SemEval 2010

Cause-Effect (1003)

He had chest pains and `<e1>headaches</e1>` from `<e2>mold</e2>` in the bedrooms.

Component-Whole (941)

The `<e1>provinces</e1>` are divided into `<e2>counties</e2>` (shahrestan), and ...

Entity-Destination (845)

The prosecutors carried `<e1>guns</e1>` into the `<e2>court room</e2>`.

Product-Producer (717)

Next the `<e1>soldiers</e1>` put up a `<e2>wall</e2>` of stakes on the pile of dirt.

+ Entity-Origin (716), Member-Collection (690), Message-Topic (634), Content-Container (540), Instrument-Agency (504)

- ▶ **Tâche** : identifier la relation de 2717 paires de noms marquées dans des phrases

SemEval 2012 (pilot) / 2013 (main task)

- ▶ sur une échelle de 0 à 5, noter la similarité de paires de phrases :

MT	When we are faced with a potential risk, it is important to apply the precautionary principle.
5	When we are faced with a potential risk, it is important to put into practice the precautionary principle.
MT	This rule can be applied equally to all recreational but dangerous drugs when no one but the person consuming the drug is likely to be affected.
4.8	This rule can, moreover, apply to all drugs <i>récréationnelles</i> but dangerous provided that the consumer of drugs is the only run the risks.
MSR-VID	A man stares out a window.
4.8	A man looks out the window.
MSR-PAR	But at age 15, she had reached 260 pounds and a difficult decision : It was time to try surgery.
4.2	But at the age of 15, she weighed a whopping 117kg and came to a difficult decision : it was time to try surgery.

SemEval 2014 / Task3

► Phrase2Word (TRAIN : 500, TEST : 500) — échelle [0, 4]

4	the structural and functional unit of all known living organisms	DNA	Lexicographic
4	traditionally the land on which a college or university and related institutional buildings are situate	campus	Lexicographic
4	watering hole	pub	Idiomatic
4	wear out one's welcome	exhaust	Idiomatic
4	whole ball of wax	total	Idiomatic
	:		
0	to make from scratch	breath	Idiomatic
0	what a vivid imagination	insipid	Newswire
0	workouts to strengthen abdominal muscles	fractions	Search

SemEval 2014 / Task3

- ▶ Sentence2Phrase (TRAIN : 500, TEST : 500) — échelle [0, 4]

4	We completely agree and there's no daylight between us on the issue.	complete agreement	Idiomatic
4	do you now where i can watch free older movies online without download ?	streaming vintage movies for free	CQA

- ▶ Paragraph2Sentence (TRAIN : 500, TEST : 500) — échelle [0, 4]

4	so i bought an item a couple weeks ago on eBay and i still havent even been notified that my item has been shipped yet.. what do i do ?	What should I do when my eBay purchase still hasn't shipped after two weeks ?	CQA
---	---	---	-----

Plan

Taches typiques

(Before Deep) : modèle vectoriel

And then came the “Deep”

Word2Vec

Analogie

Évaluation

Idées intéressantes

Le cas bilingue

Évaluation



- ▶ If A and B have almost identical environments. . . we say that they are synonyms (Harris, 1954)
- ▶ you shall know a word by the company it keeps! (Firth, 1957)
- ▶ words which are similar in meaning occur in similar contexts (Rubenstein & Goodenough, 1965)
- ▶ . . . In other words, difference of meaning correlates with difference of distribution (Harris, 1970, p.786)
- ▶ words with similar meanings will occur with similar neighbors if enough text material is available (Schütze & Pedersen, 1995)
- ▶ a representation that captures much of how words are used in natural context will capture much of what we mean by meaning (Landauer & Dumais, 1997)
- ▶ in the proposed model, it will so generalize because “similar” words are expected to have a similar feature vector, and because the probability function is a smooth function of these feature values, a small change in the features will induce a small change in the probability. (Bengio et al, 2003)



Modèle vectoriel (Vector Space model)

- ▶ lire *[Turney and Pantel, 2010]* pour une introduction
- ▶ lire *[Baroni and Lenci, 2010]* pour une généralisation (tenseur)
 - 1 une matrice de “comptes” de **co-occurrences**
 - 2 un schéma de **pondération** (PMI, LLR, etc.)
 - 3 une politique de **réduction de dimensionnalité**
 - *singular value decomposition* *[Golub and Van Loan, 1996]*
 - *non-negative matrix factorization* *[Lee and Seung, 1999]*
 - aucune ! (très bon *baseline*)
 - etc.
- ▶ DISSECT offre les étapes 2 et 3

matrice de co-occurrence : terme \times document

- ▶ similarité de documents
- ▶ hypothèse **bag of word** : si une requête et un document ont des représentations (colonnes) similaires, alors ils véhiculent la même information [*Salton, 1975*]

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	1	8	15
soldier	2	2	12	36
fool	37	58	1	5
clown	5	117	0	0

Figure 19.1 The term-document matrix for four words in four Shakespeare plays. Each cell contains the number of times the (row) word occurs in the (column) document

- ▶ implémenté (par exemple) dans [Lucene](#)

. Pris de [*Jurafsky and Martin, 2015*]

matrice de co-occurrence : terme \times terme

- ▶ similarité de termes
- ▶ hypothèse **distributionnelle** : si deux termes ont des représentations (lignes) similaires, alors ils sont similaires

	aardvark	...	computer	data	pinch	result	sugar	...
apricot	0	...	0	0	1	0	1	
pineapple	0	...	0	0	1	0	1	
digital	0	...	2	1	0	1	0	
information	0	...	1	6	0	4	0	

Figure 19.4 Co-occurrence vectors for four words, computed from the Brown corpus, showing only six of the dimensions (hand-picked for pedagogical purposes). The vector for the word *digital* is outlined in red. Note that a real vector would have vastly more dimensions and thus be much sparser.

a

a. Pris de *[Jurafsky and Martin, 2015]*

matrice de co-occurrence : terme \times rel

- ▶ similarité de termes
- ▶ hypothèse **distributionnelle** : si deux termes ont des représentations (lignes) similaires, alors ils sont similaires

	<i>subj-of</i> , absorb	<i>subj-of</i> , adapt	<i>subj-of</i> , behave	...	<i>pobj-of</i> , inside	<i>pobj-of</i> , into	..	<i>nmod-of</i> , abnormality	<i>nmod-of</i> , anemia	<i>nmod-of</i> , architecture	..	<i>obj-of</i> , attack	<i>obj-of</i> , call	<i>obj-of</i> , come from	<i>obj-of</i> , decorate	..	<i>nmod</i> , bacteria	<i>nmod</i> , body	<i>nmod</i> , bone marrow
cell	1	1	1		16	30		3	8	1		6	11	3	2		3	2	2

Figure 19.13 Co-occurrence vector for the word *cell*, from Lin (1998), showing grammatical function (dependency) features. Values for each attribute are frequency counts from a 64-million word corpus, parsed by an early version of MINIPAR.

a

a. Pris de [\[Jurafsky and Martin, 2015\]](#)

matrice de co-occurrence : termes \times patron

- ▶ similarité de relations
- ▶ hypothèse : si deux paires de mots ont des représentations (lignes) similaires, alors elles sont similaires *X of Y*, *Y of X*, *X for Y*, *Y for X*, *X to Y*, et *Y to X*
 - ▶ une liste de 64 mots comme *of*, *for* ou *to*
 - ▶ formant 128 patrons (colonnes) contenant la paire (X,Y) :
[Turney, 2005]

X	Y	X of Y	Y used by X	X for Y	Y to X
mason	stone	0	3	0	3
carpenter	wood	0	7	0	2

Plan

Taches typiques

(Before Deep) : modèle vectoriel

And then came the “Deep”

Word2Vec

Analogie

Évaluation

Idées intéressantes

Le cas bilingue

Évaluation



Au menu

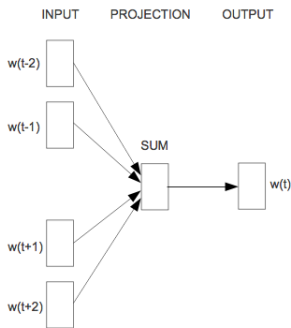
- ▶ un modèle vedette : Word2Vec *[Mikolov et al., 2013a]*
 - ▶ GloVe *[Pennington et al., 2014]*
 - ▶ FastText <https://fasttext.cc>
- ▶ propriétés des embeddings :
[Mikolov et al., 2013d, Mikolov et al., 2013c]
- ▶ des résultats : glory *[Baroni et al., 2014]*, modération
[Levy et al., 2015]
- ▶ cool works : *[Faruqui and Dyer, 2015, Faruqui et al., 2015b, Faruqui et al., 2015a]*
- ▶ modèles bilingues :
[Mikolov et al., 2013b, Chandar et al., 2014, Gouws et al., 2015, Coulmance et al., 2016]

Une révolution chez les “distributionnalistes” : Word2Vec *[Mikolov et al., 2013a]*

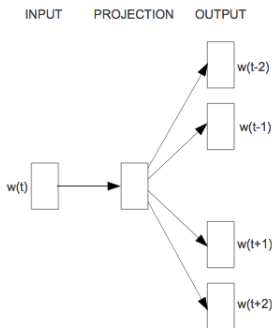
- ▶ un toolkit rapide implémentant deux modèles
 - ▶ <https://code.google.com/archive/p/word2vec/>
 - ▶ <https://radimrehurek.com/gensim/models/word2vec.html>
 - ▶ <https://github.com/dav/word2vec>
- ▶ des embeddings disponibles entraînés sur 6B de mots de Google News (180K mots) - dimension = 300
- ▶ directement utilisable dans de nombreuses applications

Les 2 modèles de Word2Vec

[Mikolov et al., 2013a]



CBOW



Skip-gram

- ▶ Skip-gram est le plus populaire (plus fiable pour les “petits” corpus)
- ▶ CBOW est plus rapide (bien pour les grands corpus)

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

- ▶ C un corpus d'entraînement, aka un ensemble D de paires (w, c) où w est un mot de C et c est un mot vu dans un contexte note : le modèle représente différemment les mots de contexte des mots du vocabulaire
- ▶ Soit (w, c) une paire de C . Appartient-elle à C ?
Soit $p(D = 1|w, c; \theta)$ la probabilité associée et $p(D = 0|w, c; \theta)$ le complément
- ▶ fonction objectif (optimisée par descente de gradient) :

$$\operatorname{argmax}_{\theta} \prod_{(w,c) \in D} p(D = 1|w, c; \theta) \prod_{(w,c) \in D'} 1 - p(D = 1|w, c; \theta)$$

avec $p(D = 1|w, c; \theta) = \frac{1}{1 + \exp^{-v_c \cdot v_w}}$

où v_c (resp. v_w) est le vecteur de c (resp. w)



Skip-gram *[Mikolov et al., 2013a]*

- ▶ Soit en prenant le log et en posant $\sigma(x) = \frac{1}{1+e^{-x}}$

$$\operatorname{argmax}_{\theta} \sum_{(w,c) \in D} \log \sigma(v_c \cdot v_w) + \sum_{(w,c) \in D'} \log \sigma(-v_c \cdot v_w)$$

- ▶ D' est construit en choisissant k paires aléatoirement selon les distributions unigrammes (des mots et des mots de contextes).
- ▶ les contextes sont définis par une fenêtre centrée autour du mot w considéré, et dont la taille est tirée aléatoirement (et uniformément sur un intervalle fixé)
- ▶ les mots les plus fréquents sont sous-échantillonnés (retirés aléatoirement de C) et les mots peu fréquents sont éliminés (*cut-off*).
- ▶ ça marche!!!
(lire *[Levy and Goldberg, 2014]* pour une explication)

Arithmétique analogique des représentations

[Mikolov et al., 2013d]

- ▶ $\text{vec}(\textit{Madrid}) - \text{vec}(\textit{Spain}) \simeq \text{vec}(\textit{Paris}) - \text{vec}(\textit{France})$
- ▶ permet de résoudre des équations analogiques : $[x: y :: z : ?]$
 - 1 calculer $t = \text{vec}(y) - \text{vec}(x) + \text{vec}(z)$ le vecteur cible
 - 2 rechercher dans V , le mot \hat{t} le plus proche de t :

$$\hat{t} = \underset{w}{\operatorname{argmax}} \frac{\text{vec}(w) \cdot \text{vec}(t)}{\|\text{vec}(w)\| \times \|\text{vec}(t)\|}$$

[Mikolov et al., 2013d]

- ▶ RNN entraîné sur 320M de mots ($V = 82k$)

Method	Adjectives	Nouns	Verbs	All
LSA-80	9.2	11.1	17.4	12.8
LSA-320	11.3	18.1	20.7	16.5
LSA-640	9.6	10.1	13.8	11.3
RNN-80	9.3	5.2	30.4	16.2
RNN-320	18.2	19.0	45.0	28.5
RNN-640	21.0	25.2	54.8	34.7
RNN-1600	23.9	29.2	62.2	39.6

Table 2: Results for identifying syntactic regularities for different word representations. Percent correct.

- ▶ test set de 8k analogies impliquant les mots les plus fréquents

[Mikolov et al., 2013c]

- ▶ 6B de mots de [Google News](#), 1M de mots les plus fréquents

Table 3: Comparison of architectures using models trained on the same data, with 640-dimensional word vectors. The accuracies are reported on our Semantic-Syntactic Word Relationship test set, and on the syntactic relationship test set of [20]

Model Architecture	Semantic-Syntactic Word Relationship test set		MSR Word Relatedness Test Set [20]
	Semantic Accuracy [%]	Syntactic Accuracy [%]	
RNNLM	9	36	35
NNLM	23	53	47
CBOV	24	64	61
Skip-gram	55	59	56

- ▶ le test syntaxique est le même que dans [Mikolov et al., 2013d]

[Mikolov et al., 2013c]

► Comparaison à d'autres modèles proposés

Table 4: Comparison of publicly available word vectors on the Semantic-Syntactic Word Relationship test set, and word vectors from our models. Full vocabularies are used.

Model	Vector Dimensionality	Training words	Accuracy [%]		
			Semantic	Syntactic	Total
Collobert-Weston NNLM	50	660M	9.3	12.3	11.0
Turian NNLM	50	37M	1.4	2.6	2.1
Turian NNLM	200	37M	1.4	2.2	1.8
Mnih NNLM	50	37M	1.8	9.1	5.8
Mnih NNLM	100	37M	3.3	13.2	8.8
Mikolov RNNLM	80	320M	4.9	18.4	12.7
Mikolov RNNLM	640	320M	8.6	36.5	24.6
Huang NNLM	50	990M	13.3	11.6	12.3
Our NNLM	20	6B	12.9	26.4	20.3
Our NNLM	50	6B	27.9	55.8	43.2
Our NNLM	100	6B	34.2	64.5	50.8
CBOW	300	783M	15.5	53.1	36.1
Skip-gram	300	783M	50.0	55.9	53.3

[Mikolov et al., 2013c]

- Big Data (plus de données, dimension plus élevée)

Table 6: Comparison of models trained using the DistBelief distributed framework. Note that training of NNLM with 1000-dimensional vectors would take too long to complete.

Model	Vector Dimensionality	Training words	Accuracy [%]			Training time [days x CPU cores]
			Semantic	Syntactic	Total	
NNLM	100	6B	34.2	64.5	50.8	14 x 180
CBOW	1000	6B	57.3	68.9	63.7	2 x 140
Skip-gram	1000	6B	66.1	65.1	65.6	2.5 x 125

Don't count, predict! [*Baroni et al., 2014*]

- plein de tâches, une étude des méta-paramètres de chaque méthode

name	task	measure	source	soa
rg	relatedness	Pearson	Rubenstein and Goodenough (1965)	Hassan and Mihalcea (2011)
ws	relatedness	Spearman	Finkelstein et al. (2002)	Halawi et al. (2012)
wss	relatedness	Spearman	Agirre et al. (2009)	Agirre et al. (2009)
wsr	relatedness	Spearman	Agirre et al. (2009)	Agirre et al. (2009)
men	relatedness	Spearman	Bruni et al. (2013)	Bruni et al. (2013)
toefl	synonyms	accuracy	Landauer and Dumais (1997)	Bullinaria and Levy (2012)
ap	categorization	purity	Almuhareb (2006)	Rothenhäusler and Schütze (2009)
esslli	categorization	purity	Baroni et al. (2008)	Katrenko and Adriaans (2008)
battig	categorization	purity	Baroni et al. (2010)	Baroni and Lenci (2010)
up	sel pref	Spearman	Padó (2007)	Herdağdelen and Baroni (2009)
mcrae	sel pref	Spearman	McRae et al. (1998)	Baroni and Lenci (2010)
an	analogy	accuracy	Mikolov et al. (2013a)	Mikolov et al. (2013c)
ansyn	analogy	accuracy	Mikolov et al. (2013a)	Mikolov et al. (2013a)
ansem	analogy	accuracy	Mikolov et al. (2013a)	Mikolov et al. (2013c)

Table 1: Benchmarks used in experiments, with type of task, figure of merit (measure), original reference (source) and reference to current state-of-the-art system (soa).

Don't count, predict! [*Baroni et al., 2014*]

- ▶ cnt = count vector, pre = word2Vec, dm = [*Baroni and Lenci, 2010*], cw = [*Collobert et al., 2011*]

	rg	ws	wss	wsr	men	toefl	ap	esslli	battig	up	mrae	an	ansyn	ansem
<i>best setup on each task</i>														
cnt	74	62	70	59	72	76	66	84	98	41	27	49	43	60
pre	84	75	80	70	80	91	75	86	99	41	28	68	71	66
<i>best setup across tasks</i>														
cnt	70	62	70	57	72	76	64	84	98	37	27	43	41	44
pre	83	73	78	68	80	86	71	77	98	41	26	67	69	64
<i>worst setup across tasks</i>														
cnt	11	16	23	4	21	49	24	43	38	-6	-10	1	0	1
pre	74	60	73	48	68	71	65	82	88	33	20	27	40	10
<i>best setup on rg</i>														
cnt	(74)	59	66	52	71	64	64	84	98	37	20	35	42	26
pre	(84)	71	76	64	79	85	72	84	98	39	25	66	70	61
<i>other models</i>														
soa	86	81	77	62	76	100	79	91	96	60	32	61	64	61
dm	82	35	60	13	42	77	76	84	94	51	29	NA	NA	NA
cw	48	48	61	38	57	56	58	61	70	28	15	11	12	9

Table 2: Performance of count (cnt), predict (pre), dm and cw models on all tasks. See Section 3 and Table 1 for figures of merit and state-of-the-art results (soa). Since dm has very low coverage of the an* data sets, we do not report its performance there.

Don't count, predict! [*Baroni et al., 2014*]

...

we set out to conduct this study because we were annoyed by the triumphalist overtones often surrounding predict models, despite the almost complete lack of proper comparison to count vectors. Our secret wish was to discover that it is all hype, and count vectors are far superior to their predictive counterparts.

...

we found that the predict models are so good that , while the triumphalist overtones still sound excessive, there are very good reasons to switch to the new architecture.



Représentation vectorielle binaire (non distributionnelle) *[Faruqui and Dyer, 2015]*

- ▶ en utilisant des ressources linguistiques (WordNet, PTB, FrameNet, etc.)
- ▶ vecteurs très creux (99.9%)
- ▶ comparables en performance aux modèles distributionnels état de l'art entraînés sur des billions de mots
- ▶ vecteurs disponibles (pour l'anglais) :
<https://github.com/mfaruqui/non-distributional>

Représentation vectorielle binaire (non distributionnelle) *[Faruqui and Dyer, 2015]*

Lexicon	Vocabulary	Features
WordNet	10,794	92,117
Supersense	71,836	54
FrameNet	9,462	4,221
Emotion	6,468	10
Connotation	76,134	12
Color	14,182	12
Part of Speech	35,606	20
Syn. & Ant.	35,693	75,972
Union	119,257	172,418

Table 1: Sizes of vocabulary and features induced from different linguistic resources.



Représentation vectorielle binaire (non distributionnelle) *[Faruqui and Dyer, 2015]*

Noun

- S: (n) [movie](#), [film](#), [picture](#), [moving picture](#), [moving-picture show](#), [motion picture](#), [motion-picture show](#), [picture show](#), [pic](#), [flick](#)
 - [direct hyponym](#) / [full hyponym](#)
 - S: (n) [telefilm](#)
 - S: (n) [feature](#), [feature film](#)
 - S: (n) [final cut](#)
 - S: (n) [home movie](#)
 - S: (n) [collage film](#)
 - S: (n) [coming attraction](#)
 - S: (n) [shoot-'em-up](#)
 - S: (n) [short subject](#)
 - S: (n) [documentary](#), [docudrama](#), [documentary film](#), [infotainment](#)
 - S: (n) [cinema verite](#)
 - S: (n) [film noir](#)
 - S: (n) [skin flick](#)
 - S: (n) [rough cut](#)
 - S: (n) [silent movie](#), [silent picture](#), [silents](#)
 - S: (n) [slow motion](#)
 - S: (n) [talking picture](#), [talkie](#)
 - S: (n) [three-D](#), [3-D](#), [3D](#)
 - S: (n) [musical](#), [musical comedy](#), [musical theater](#)
 - [part meronym](#)
 - [domain term category](#)
 - [direct hypernym](#) / [inherited hypernym](#) / [sister term](#)
 - [derivationally related form](#)
- S: (n) [film](#), [cinema](#), [celluloid](#)
- S: (n) [film](#), [photographic film](#)
- S: (n) [film](#)
- S: (n) [film](#), [plastic film](#)

Verb

- S: (v) [film](#), [shoot](#), [take](#)
- S: (v) [film](#)

features (binaires) induits
pour *film* :

SYNSET.FILM.V.01
SYNSET.FILM.N.01
HYPO.COLLAGEFILM.N.01
HYPER :SHEET.N.06

⋮

Représentation vectorielle binaire (non distributionnelle) [*Faruqui and Dyer, 2015*]

supersenses pour les noms, les verbes et les adjectifs

ex : *lioness* ⇒ SS.NOUN.ANIMAL

color lexique mot-couleur élaboré par crowdsourcing [?]

ex : *blood* ⇒ COLOR.RED

emotion lexique associant un mot à sa polarité (positif,négatif) et aux émotions (joie, peur, tristesse, etc.), élaboré par crowdsourcing [?]

ex : *cannibal* ⇒ POL.NEG, EMO.DISGUST et EMO.FEARCOLOR.RED

pos PTB part-of-speech tags

ex : *love* ⇒ PTB.NOUN, PTB.VERB



Représentation vectorielle binaire (non distributionnelle) *[Faruqui and Dyer, 2015]*

Word	POL.POS	COLOR.PINK	SS.NOUN.FEELING	PTB.VERB	ANTO.FAIR	...	CON.NO
love	1	1	1	1	0		1
hate	0	0	1	1	0		0
ugly	0	0	0	0	1		0
beauty	1	1	0	0	0		1
refundable	0	0	0	0	0		1

Table 2: Some linguistic word vectors. 1 indicates presence and 0 indicates absence of a linguistic feature.

- **note** : difficile à faire pour toutes les langues

Représentation vectorielle binaire (non distributionnelle) *[Faruqui and Dyer, 2015]*

Vector	Length (D)	Params.	Corpus Size	WS-353	RG-65	SimLex	Senti	NP
Skip-Gram	300	$D \times N$	300 billion	65.6	72.8	43.6	81.5	80.1
Glove	300	$D \times N$	6 billion	60.5	76.6	36.9	77.7	77.9
LSA	300	$D \times N$	1 billion	67.3	77.0	49.6	81.1	79.7
Ling Sparse	172,418	–	–	44.6	77.8	56.6	79.4	83.3
Ling Dense	300	$D \times N$	–	45.4	67.0	57.8	75.4	76.2
Skip-Gram \oplus Ling Sparse	172,718	–	–	67.1	80.5	55.5	82.4	82.8

Table 3: Performance of different type of word vectors on evaluation tasks reported by Spearman's correlation (first 3 columns) and Accuracy (last 2 columns). Bold shows the best performance for a task.

- ▶ Skip-Gram : pré-entraîné sur 300B de mots *[Mikolov et al., 2013a]*
- ▶ Glove : pré-entraîné sur 6B de mots *[Pennington et al., 2014]*
- ▶ LSA : obtenue à partir d'une matrice de co-occurrence calculée sur 1B de mots de Wikipedia *[Turney and Pantel, 2010]*
- ▶ Ling Dense : réduction de dimensionnalité avec SVD
- ▶ taches : similarité, sent. analysis (positif/négatif), NP-bracketing (*local (phone company) versus (local phone) company*)

Retrofitting de vecteurs à une ressource lexico-sémantique [*Faruqui et al., 2015a*]

- ▶ étape de post-traitement applicable à n'importe quelle représentation vectorielle de mots
- ▶ rapide (5 secondes pour 100k mots et dimension 300)
- ▶ **idée** : utiliser les informations lexico-sémantiques d'une ressource pour améliorer une représentation existante
- ▶ **comment** : encourager que les mots de distance similaire dans la représentation apprise soit proche de la représentation induite de la ressource (encodée sous forme de graphe).

Retrofitting de vecteurs à une ressource lexico-sémantique [*Faruqui et al., 2015a*]

- ▶ V le vocabulaire $\{w_1, \dots, w_{|V|}\}$
- ▶ \hat{Q} la collection de vecteurs $\hat{q}_i \in \mathbb{R}^d$ d'entrée, $i \in [1, \dots, |V|]$
- ▶ $E = \{(i, j) : i, j \in [1, |V|], \text{ et } w_i \mathcal{R} w_j \text{ selon la ressource}\}$
- ▶ chercher la nouvelle représentation Q qui minimise :

$$\Psi(Q) = \sum_i \left[\alpha_i \|q_i - \hat{q}_i\|^2 + \sum_{j:(i,j) \in E} \beta_{i,j} \|q_i - q_j\|^2 \right]$$

où :

- ▶ α_i et β_{ij} sont des méta-paramètres,
- ▶ ϕ est convexe en Q et se résout itérativement par : [?]

$$q_i = \frac{\sum_{j:(i,j) \in E} \beta_{ij} q_j + \alpha_i \hat{q}_i}{\sum_{j:(i,j) \in E} \beta_{ij} + \alpha_i}$$

- ▶ 10 itérations suffisent en pratique



Retrofitting de vecteurs à une ressource lexico-sémantique [*Faruqui et al., 2015a*]

Lexicon	MEN-3k	RG-65	WS-353	TOEFL	SYN-REL	SA
Glove	73.7	76.7	60.5	89.7	67.0	79.6
+PPDB	1.4	2.9	-1.2	5.1	-0.4	1.6
+WN _{syn}	0.0	2.7	0.5	5.1	-12.4	0.7
+WN _{all}	2.2	7.5	0.7	2.6	-8.4	0.5
+FN	-3.6	-1.0	-5.3	2.6	-7.0	0.0
SG	67.8	72.8	65.6	85.3	73.9	81.2
+PPDB	5.4	3.5	4.4	10.7	-2.3	0.9
+WN _{syn}	0.7	3.9	0.0	9.3	-13.6	0.7
+WN _{all}	2.5	5.0	1.9	9.3	-10.7	-0.3
+FN	-3.2	2.6	-4.9	1.3	-7.3	0.5
GC	31.3	62.8	62.3	60.8	10.9	67.8
+PPDB	7.0	6.1	2.0	13.1	5.3	1.1
+WN _{syn}	3.6	6.4	0.6	7.3	-1.7	0.0
+WN _{all}	6.7	10.2	2.3	4.4	-0.6	0.2
+FN	1.8	4.0	0.0	4.4	-0.6	0.2
Multi	75.8	75.5	68.1	84.0	45.5	81.0
+PPDB	3.8	4.0	6.0	12.0	4.3	0.6
+WN _{syn}	1.2	0.2	2.2	6.6	-12.3	1.4
+WN _{all}	2.9	8.5	4.3	6.6	-10.6	1.4
+FN	1.8	4.0	0.0	4.4	-0.6	0.2

- ▶ SG : Skip-Gram (100B words Google News, $d = 300$)
- ▶ GC : recursive NN (1B words of English Wikipedia $d = 50$)
- ▶ Multi : multilingue (360M EN-words WMT-2011, $d = 512$)

Table 2: Absolute performance changes with retrofitting. Spearman's correlation (3 left columns) and accuracy (3 right columns) on different tasks. Higher scores are always better. Bold indicates greatest improvement for a vector type.

- ▶ PPDB : 8M paraphrases au niveau des mots, FN : FrameNet

Une communauté qui s'organise

[Faruqui and Dyer, 2014]

- ▶ des embeddings déjà entraînés
- ▶ une suite de tests qui peuvent s'exécuter (similarité, analogie, complétion, etc.)
- ▶ une interface de visualisation

- ▶ **note** : pas certain que le site soit très populaire (ni mis à jour) pour le moment...
- ▶ <http://wordvectors.org/demo.php>

Mikolov strikes again [?]

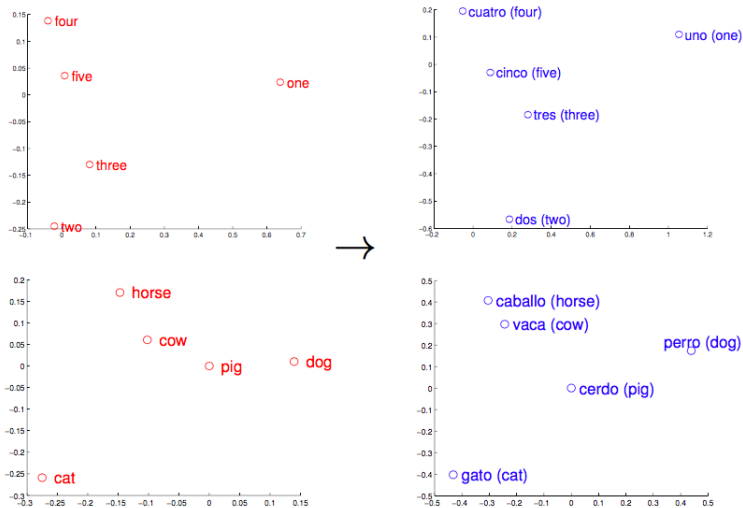


Figure 1: Distributed word vector representations of numbers and animals in English (left) and Spanish (right). The five vectors in each language were projected down to two dimensions using PCA, and then manually rotated to accentuate

Mikolov strikes again [?]

- ▶ on peut apprendre une transformation linéaire (rotation + scaling) d'un espace vers un autre avec un lexique bilingue $\{(x_i, z_i)\}$:

$$\hat{W} = \min_W \sum_i \|W\bar{x}_i - \bar{z}_i\|^2$$

où \bar{x}_i et \bar{z}_i désignent respectivement la représentation vectorielle source de x_i et cible de z_i

- ▶ W optimisée par descente de gradient sur un lexique d'environ 5k paires de mots
- ▶ au moment du test : *traduire* un mot x par \hat{z} :

$$\hat{z} = \operatorname{argmax}_z \cos(\bar{z}, \hat{W}\bar{x})$$



Mikolov strikes again [?]

Table 1: The sizes of the monolingual training datasets from WMT11. The vocabularies consist of the words that occurred at least five times in the corpus.

Language	Training tokens	Vocabulary size
English	575M	127K
Spanish	84M	107K
Czech	155M	505K

- ▶ 6K des most sources les plus fréquents traduits par GoogleTrans
 - ▶ premières 5K entrées pour calculer \hat{W}
 - ▶ 1K suivantes pour les tests
- ▶ baselines : edit-distance, ϵ -Rapp

Mikolov strikes again [?]

Table 2: Accuracy of the word translation methods using the WMT11 datasets. The Edit Distance uses morphological structure of words to find the translation. The Word Co-occurrence technique based on counts uses similarity of contexts in which words appear, which is related to our proposed technique that uses continuous representations of words and a Translation Matrix between two languages.

Translation	Edit Distance		Word Co-occurrence		Translation Matrix		ED + TM		Coverage
	P@1	P@5	P@1	P@5	P@1	P@5	P@1	P@5	
En → Sp	13%	24%	19%	30%	33%	51%	43%	60%	92.9%
Sp → En	18%	27%	20%	30%	35%	52%	44%	62%	92.9%
En → Cz	5%	9%	9%	17%	27%	47%	29%	50%	90.5%
Cz → En	7%	11%	11%	20%	23%	42%	25%	45%	90.5%

Plus de données ? (Google News)

- ▶ même split : 5K train / 1K test

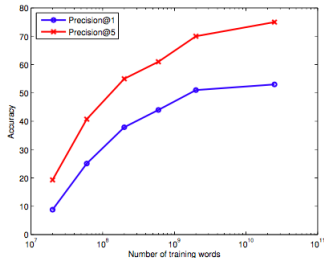


Figure 3: The Precision at 1 and 5 as the size of the monolingual training sets increase (EN→ES).

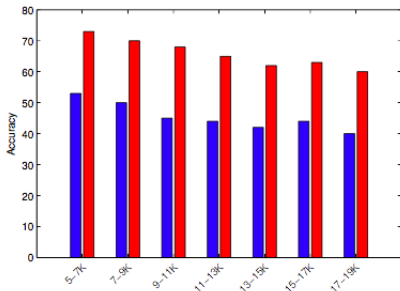


Figure 4: Accuracies of translation as the word frequency decreases. Here, we measure the accuracy of the translation on disjoint sets of 2000 words sorted by frequency, starting from rank 5K and continuing to 19K. In all cases, the linear transformation was trained on the 5K most frequent words and their translations. EN→ES.

Plan

Taches typiques

(Before Deep) : modèle vectoriel

And then came the “Deep”

- Word2Vec

- Analogie

- Évaluation

- Idées intéressantes

- Le cas bilingue

Évaluation

Sur la difficulté d'évaluer sans biais

[Levy et al., 2015]

- ▶ comparent 4 approches : matrice de co-occurrence (PMI), SVD, Skip-Gram, et GloVe
- ▶ étudient leurs paramètres en détail
- ▶ adaptent des choix faits dans Skip-Gram à d'autres méthodes lorsque possible
- ▶ **Bilan :**
 - ▶ match nul en performance (pas d'avantage clair d'une approche sur une autre)
 - ▶ Skip-Gram se comporte mieux (temps/mémoire) que les autres approches

Sur la difficulté d'évaluer sans biais

[Levy et al., 2015]

Method	WordSim Similarity	WordSim Relatedness	Bruni et al. MEN	Radinsky et al. M. Turk	Luong et al. Rare Words	Hill et al. SimLex	Google Add / Mul	MSR Add / Mul
PPMI	.755	.688	.745	.686	.423	.354	.553 / .629	.289 / .413
SVD	.784	.672	.777	.625	.514	.402	.547 / .587	.402 / .457
SGNS	.773	.623	.723	.676	.431	.423	.599 / .625	.514 / .546
GloVe	.667	.506	.685	.599	.372	.389	.539 / .563	.503 / .559
CBOw	.766	.613	.757	.663	.480	.412	.547 / .591	.557 / .598

Table 3: Performance of each method across different tasks using word2vec's recommended configuration: win = 2; dyn = with; sub = dirty; neg = 5; cds = 0.75; w+c = only w; eig = 0.0. CBOw is presented for comparison.

Method	WordSim Similarity	WordSim Relatedness	Bruni et al. MEN	Radinsky et al. M. Turk	Luong et al. Rare Words	Hill et al. SimLex	Google Add / Mul	MSR Add / Mul
PPMI	.755	.697	.745	.686	.462	.393	.553 / .679	.306 / .535
SVD	.793	.691	.778	.666	.514	.432	.554 / .591	.408 / .468
SGNS	.793	.685	.774	.693	.470	.438	.676 / .688	.618 / .645
GloVe	.725	.604	.729	.632	.403	.398	.569 / .596	.533 / .580

Table 4: Performance of each method across different tasks using the best configuration for that method and task combination, assuming win = 2.

Exemple d'observation [Levy et al., 2015]

- ▶ dans l'approche matrice de co-occurrences, un mot w et son contexte c est noté :

$$PMI(w, c) = \log \frac{p(w, c)}{p(w)p(c)}$$

- ▶ une approche courante est de mettre à 0 les valeurs de PMI lorsque $\#(w, c) = 0$ (plutôt que $-\infty$)
- ▶ une autre est de prendre : $PPMI(w, c) = \max(PMI(w, c), 0)$
- ▶ adaptation de choix faits dans Skip-Gram :

- ▶

$$SPPMI(w, c) = \max(PMI(w, c) - \log k, 0)$$

- ▶ sampling des k exemples négatifs (lissés avec $\alpha = 0.75$)

$$PMI_{\alpha}(w, c) = \log \frac{P(w, c)}{p(w) \cdot P_{\alpha}(c)} \text{ avec } P_{\alpha}(c) = \frac{\#(c)^{\alpha}}{\sum_c \#(c)^{\alpha}}$$



[Schnabel et al., 2015]

- recommandent de ne pas utiliser une tâche extrinsèque pour évaluer des embeddings pré-entraînés

	dev	test	p -value
Baseline	94.18	93.78	0.000
Rand. Proj.	94.33	93.90	0.006
GloVe	94.28	93.93	0.015
H-PCA	94.48	93.96	0.029
C&W	94.53	94.12	
CBOW	94.32	93.93	0.012
TSCCA	94.53	94.09	0.357

Table 4: F1 chunking results using different word embeddings as features. The p -values are with respect to the best performing method.

	test	p -value
BOW (baseline)	88.90	$7.45 \cdot 10^{-14}$
Rand. Proj.	62.95	$7.47 \cdot 10^{-12}$
GloVe	74.87	$5.00 \cdot 10^{-2}$
H-PCA	69.45	$6.06 \cdot 10^{-11}$
C&W	72.37	$1.29 \cdot 10^{-7}$
CBOW	75.78	
TSCCA	75.02	$7.28 \cdot 10^{-4}$

Table 5: F1 sentiment analysis results using different word embeddings as features. The p -values are with respect to the best performing embedding.

[Antoniak and Mimno, 2018]

- ▶ word2vec skipgram relancé plusieurs fois avec les mêmes paramètres

Run 1	Run 2	Run 3	Run 4	Run 5	Run 6	Run 7
viability pregnancies abortion abortions fetus gestation surgery expiration sudden fetal	fetus pregnancies gestation kindergarten viability headaches pregnant abortion pain bladder	trimester surgery visit tenure workday abortions hernia summer suicide abortion	surgery visit therapy pain hospitalization neck headaches trimester experiencing medications	trimester surgery incarceration visit arrival pain headaches birthday neck tenure	pregnancies occupation viability abortion tenure visit abortions pregnant birthday fetus	abdomen tenure stepfather wife groin throat grandmother daughter panic jaw

Table 5: The 10 closest words to the query term `pregnancy` are highly variable. None of the words shown appear in every run. Results are shown across runs of the `BOOTSTRAP` setting for the full corpus of the *9th Circuit*, the whole document size, and the `SGNS` model.

Et pour les mots peu fréquents ? [?]

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)

Fig. 1. Top-3 candidates produced by the two best approaches for a few test words.

Et pour les mots peu fréquents ? [?]

	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	—	—	—
<i>oracle</i>	4,6	—	19,0	31,8	—	57,6

- ▶ Wikipedia dump de juin 2013 (EN : 3.5M, FR : 1.3M articles)
- ▶ $V_{EN} = 7.3M$, $V_{FR} = 3.6M$
- ▶ 2 test sets : 1k-low (1k mots *rares*) 1k-high (1k mots non rares)
- ▶ rare = freq < 26 (92% des mots de V_{EN})



Agirre, E., Alfonseca, E., Hall, K., Kravalova, J., Paşca, M., and Soroa, A. (2009).

A study on similarity and relatedness using distributional and wordnet-based approaches.

In Proceedings of Human Language Technologies : The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, NAACL '09, pages 19–27.



Antoniak, M. and Mimno, D. (2018).

Evaluating the stability of embedding-based word similarities.
Transactions of the Association for Computational Linguistics, 6 :107–119.



Baroni, M., Dinu, G., and Kruszewski, G. (2014).

Don't count, predict ! a systematic comparison of context-counting vs. context-predicting semantic vectors.

In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers), pages 238–247, Baltimore, Maryland. Association for Computational Linguistics.

**Baroni, M. and Lenci, A. (2010).**

Distributional memory : A general framework for corpus-based semantics.

Comput. Linguist., 36(4) :673–721.

**Chandar, A. P. S., Lauly, S., Larochelle, H., Khapra, M. M., Ravindran, B., Raykar, V. C., and Saha, A. (2014).**

An autoencoder approach to learning bilingual word representations.

CoRR.

**Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011).**

Natural language processing (almost) from scratch.

Journal of Machine Learning Research, 12 :2493–2537.

**Coulmance, J., Marty, J., Wenzek, G., and Benhalloum, A. (2016).**

Trans-gram, fast cross-lingual word-embeddings.

CoRR, abs/1601.02502.



Faruqui, M., Dodge, J., Jauhar, S. K., Dyer, C., Hovy, E., and Smith, N. A. (2015a).

Retrofitting word vectors to semantic lexicons.

In *Proceedings of NAACL*.



Faruqui, M. and Dyer, C. (2014).

Community evaluation and exchange of word vectors at wordvectors.org.

In *Proceedings of ACL : System Demonstrations*.



Faruqui, M. and Dyer, C. (2015).

Non-distributional word vector representations.

In *Proceedings of ACL*.



Faruqui, M., Tsvetkov, Y., Yogatama, D., Dyer, C., and Smith, N. A. (2015b).

Sparse overcomplete word vector representations.

In *Proceedings of ACL*.



Freitag, D., Blume, M., Byrnes, J., Chow, E., Kapadia, S., Rohwer, R., and Wang, Z. (2005).

New experiments in distributional representations of synonymy.
In Proceedings of the Ninth Conference on Computational Natural Language Learning, pages 25–32.



Golub, G. H. and Van Loan, C. F. (1996).

Matrix Computations (3rd Ed.).

Johns Hopkins University Press.



Gouws, S., Bengio, Y., and Corrado, G. (2015).

Bilbowa : Fast bilingual distributed representations without word alignments.

In ICML.



Jurafsky, D. and Martin, J. H. (2015).

Speech and language processing.

(3rd ed. draft).



Lee, D. D. and Seung, H. S. (1999).

Learning the parts of objects by non-negative matrix factorization.

Nature, 401(6755) :788–791.

**Levy, O. and Goldberg, Y. (2014).**

Neural word embedding as implicit matrix factorization.

In Advances in Neural Information Processing Systems 27, pages 2177–2185.

**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.

CoRR, abs/1301.3781.

**Mikolov, T., Le, Q. V., and Sutskever, I. (2013b).**

Exploiting similarities among languages for machine translation.

CoRR, abs/1309.4168.



Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013c).

Distributed representations of words and phrases and their compositionality.

CoRR, abs/1310.4546.



Mikolov, T., tau Yih, W., and Zweig, G. (2013d).

Linguistic regularities in continuous space word representations.

In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies (NAACL-HLT-2013)*.



Pennington, J., Socher, R., and Manning, C. D. (2014).

Glove : Global vectors for word representation.

In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.



Salton, G. (1975).

Dynamic information and library processing / Gerard Salton.

Prentice-Hall Englewood Cliffs, N.J.



Schnabel, T., Labutov, I., Mimno, D. M., and Joachims, T. (2015).

Evaluation methods for unsupervised word embeddings.

In Màrquez, L., Callison-Burch, C., Su, J., Pighin, D., and Marton, Y., editors, *EMNLP*, pages 298–307. The Association for Computational Linguistics.



Turney, P. D. (2005).

Measuring semantic similarity by latent relational analysis.

CoRR.



Turney, P. D. and Pantel, P. (2010).

From frequency to meaning : Vector space models of semantics.

J. Artif. Int. Res., 37(1) :141–188.