Représentations de Mots

Plongements de mots

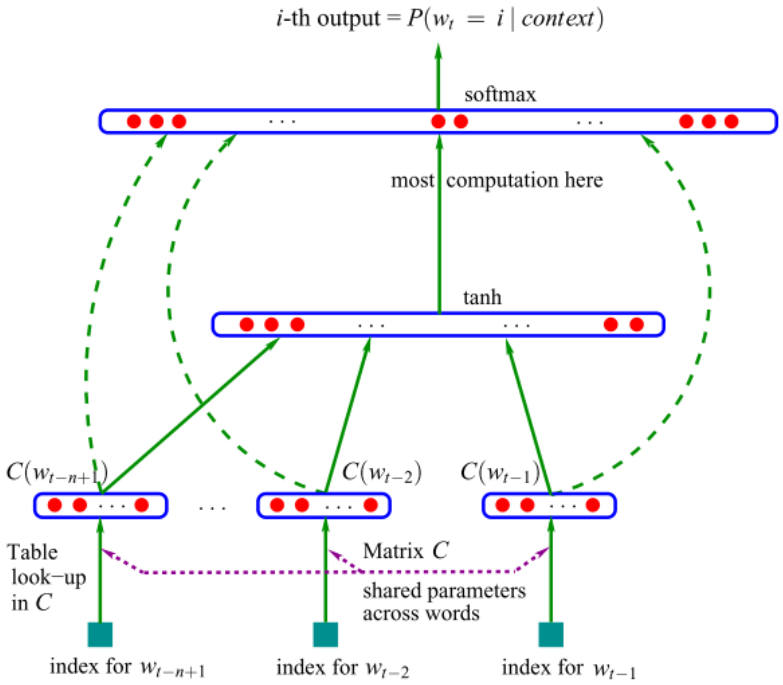
BNP PARIBAS

SCIENCES SORBONNE UNIVERSITÉ

1.1 Fighting the Curse of Dimensionality with Distributed Representations

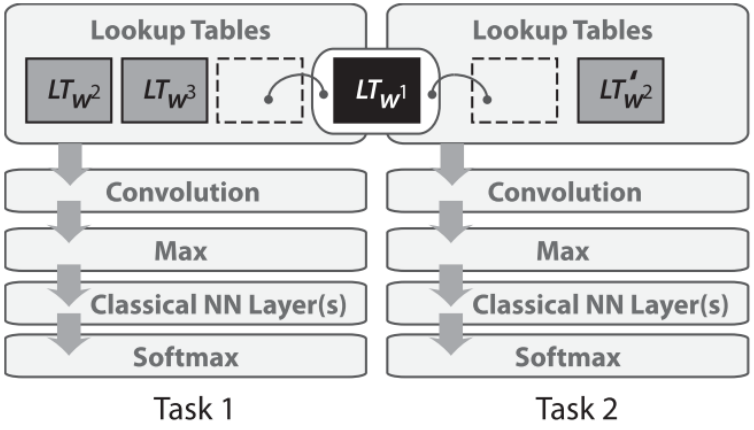
In a nutshell, the idea of the proposed approach can be summarized as follows:

1. associate with each word in the vocabulary a distributed *word feature vector* (a real-valued vector in $\mathbb{R}^m$),
2. express the joint *probability function* of word sequences in terms of the feature vectors of these words in the sequence, and
3. learn simultaneously the *word feature vectors* and the parameters of that *probability function*.



(Bengio 2003) A Neural Probabilistic Language Model, JMLR 2003

Pré-entraînement : SENNA



SENNA

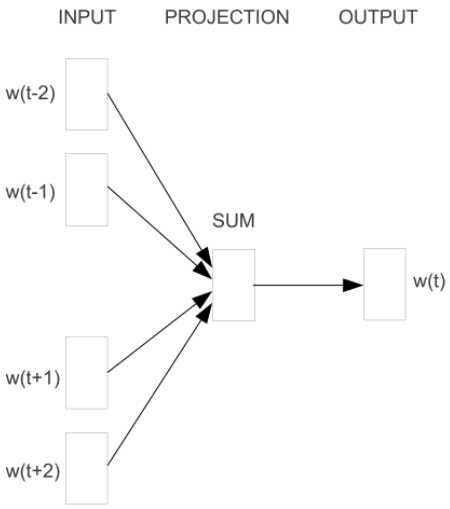
- Pré-entraînement d'un modèle de langue non supervisé
- Contextes gauche et droit
- Fonction de coût simplifiée

Table 2. A Deep Architecture for SRL improves by learning auxiliary tasks that share the first layer that represents words as wsz-dimensional vectors. We give word error rates for wsz=15, 50 and 100 and various shared tasks.

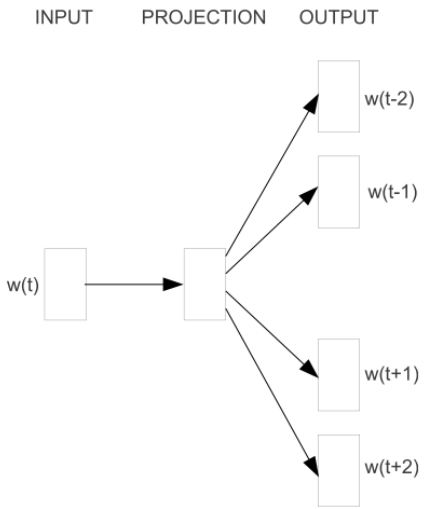
	wsz=15	wsz=50	wsz=100
SRL	16.54	17.33	18.40
SRL + POS	15.99	16.57	16.53
SRL + Chunking	16.42	16.39	16.48
SRL + NER	16.67	17.29	17.21
SRL + Synonyms	15.46	15.17	15.17
SRL + Language model	14.42	14.30	14.46
SRL + POS + Chunking	16.46	15.95	16.41
SRL + POS + NER	16.45	16.89	16.29
SRL + POS + Chunking + NER	16.33	16.36	16.27
SRL + POS + Chunking + NER + Synonyms	15.71	14.76	15.48
SRL + POS + Chunking + NER + Language model	14.63	14.44	14.50

(Collobert 2008) A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning, ICML 2008

Pré-entraînement : Word2Vec



CBOW



Skip-gram

Word2Vec

- Réduire la complexité de l'apprentissage de représentations
- Le séparer du modèle de langue
- Enlever la couche cachée + réduire le contexte
- Prédiction mot vers contexte
- Negative Sampling

(Mikolov 2013) Efficient Estimation of Word Representations in Vector Space

Pré-entraînement : Word2Vec

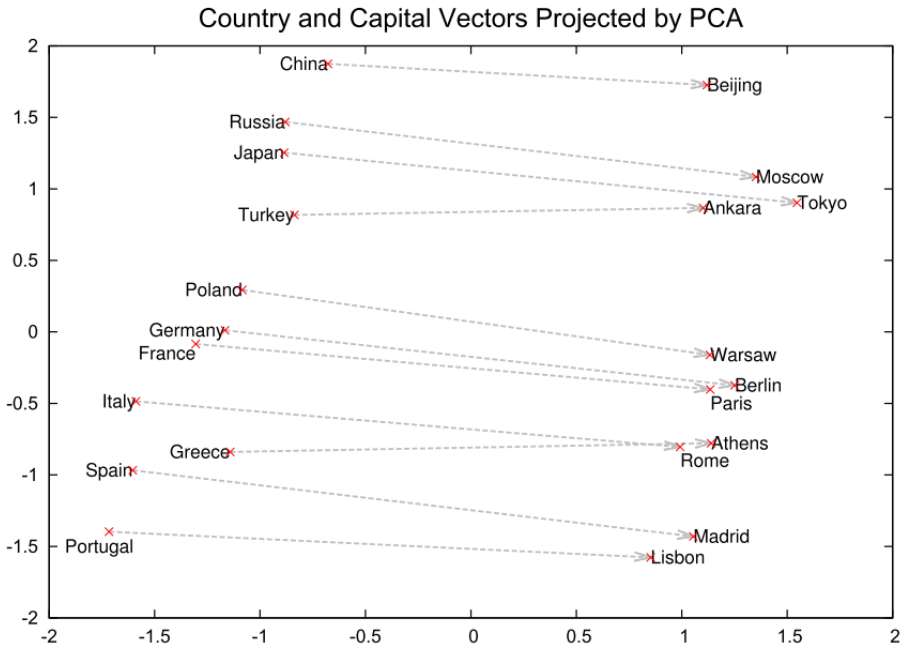


Figure 2: Two-dimensional PCA projection of the 1000-dimensional Skip-gram vectors of countries and their capital cities. The figure illustrates ability of the model to automatically organize concepts and learn implicitly the relationships between them, as during the training we did not provide any supervised information about what a capital city means.

Word Analogy Dataset

- « king – man + woman = queen »
- Relations sémantiques et syntactiques

(Mikolov 2013) Efficient Estimation of Word Representations in Vector Space

Figure de (Mikolov 2013) Distributed Representations of Words and Phrases and their Compositionality, NIPS 2013

Pré-entraînement : Word2Vec

Table 1: *Examples of five types of semantic and nine types of syntactic questions in the Semantic-Syntactic Word Relationship test set.*

Type of relationship	Word Pair 1		Word Pair 2	
Common capital city	Athens	Greece	Oslo	Norway
All capital cities	Astana	Kazakhstan	Harare	Zimbabwe
Currency	Angola	kwanza	Iran	rial
City-in-state	Chicago	Illinois	Stockton	California
Man-Woman	brother	sister	grandson	granddaughter
Adjective to adverb	apparent	apparently	rapid	rapidly
Opposite	possibly	impossibly	ethical	unethical
Comparative	great	greater	tough	tougher
Superlative	easy	easiest	lucky	luckiest
Present Participle	think	thinking	read	reading
Nationality adjective	Switzerland	Swiss	Cambodia	Cambodian
Past tense	walking	walked	swimming	swam
Plural nouns	mouse	mice	dollar	dollars
Plural verbs	work	works	speak	speaks

Word Analogy Dataset

- « king – man + woman = queen »
- Relations sémantiques et syntactiques

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

(Mikolov 2013) Efficient Estimation of Word Representations in Vector Space
(Nissim 2019) Fair is Better than Sensational: Man is to Doctor as Woman is to Doctor

Pré-entraînement : GloVe

Probability and Ratio	$k = solid$	$k = gas$	$k = water$	$k = fashion$
$P(k\|ice)$	1.9×10^{-4}	6.6×10^{-5}	3.0×10^{-3}	1.7×10^{-5}
$P(k\|steam)$	2.2×10^{-5}	7.8×10^{-4}	2.2×10^{-3}	1.8×10^{-5}
$P(k\|ice)/P(k\|steam)$	8.9	8.5×10^{-2}	1.36	0.96

$$F(w_i, w_j, \tilde{w}_k) = \frac{P_{ik}}{P_{jk}}$$

$$F\left((w_i - w_j)^T \tilde{w}_k\right) = \frac{P_{ik}}{P_{jk}}$$

$$J = \sum_{i,j=1}^{V} f\left(X_{ij}\right)\left(w_i^T \tilde{w}_j + b_i + \tilde{b}_j - \log X_{ij}\right)^2$$

GloVe (Global Vectors)

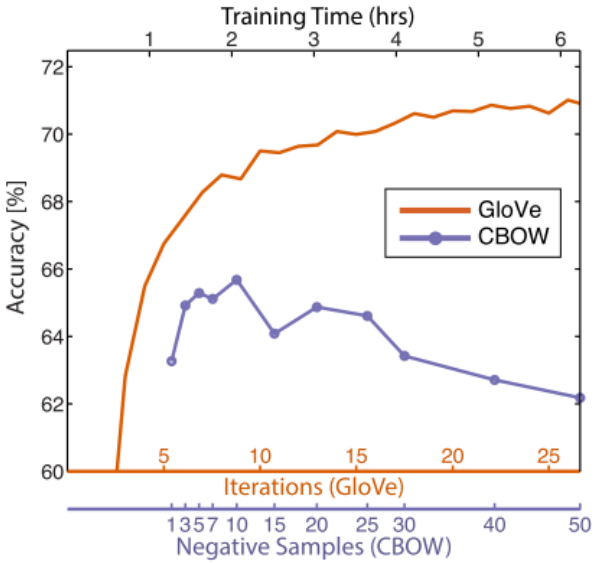
- Cooccurrences « globales » (vs fenêtre locale)
- Statistique (vs prédictif)

X_{ij} = nb d'occurrences du mot i dans le contexte du mot j

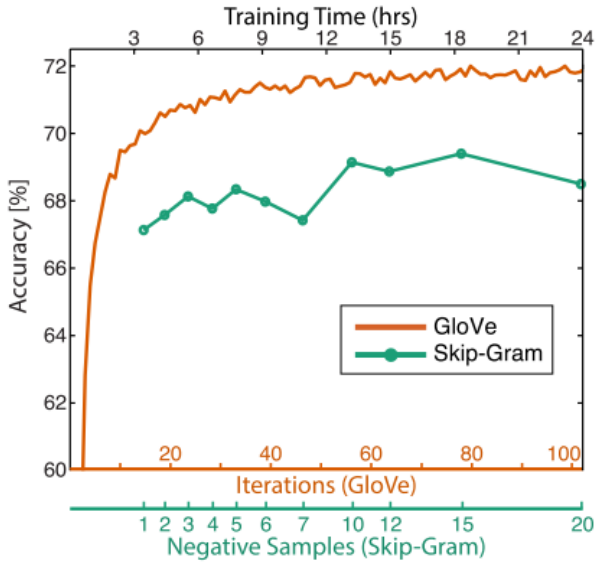
f = pénalisation des cooccurrences trop rares et fréquentes

(Pennington 2014) GloVe: Global Vectors for Word Representation, EMNLP 2014

Pré-entraînement : GloVe



(a) GloVe vs CBOW



(b) GloVe vs Skip-Gram

Figure 4: Overall accuracy on the word analogy task as a function of training time, which is governed by the number of iterations for GloVe and by the number of negative samples for CBOW (a) and skip-gram (b). In all cases, we train 300-dimensional vectors on the same 6B token corpus (Wikipedia 2014 + Gigaword 5) with the same 400,000 word vocabulary, and use a symmetric context window of size 10.

(Pennington 2014) GloVe: Global Vectors for Word Representation, EMNLP 2014