

NATURAL LANGUAGE PROCESSING KNOWLEDGE MANAGEMENT

Agro 3–IODAA–Semestre 1

Vincent Guigue



HISTORIQUE & ÉVOLUTION

(pré)Historique très rapide

1962 ACL (Journal since 1965)

1966 ALPAC report / 1st AI Winter

1978 SIGIR

1987 MUC

1992 CIKM

1973 SEQUEL/SQL

1999 RDF

2008 SparQL

Les problématiques de gestion de l'information apparaissent en même temps que l'informatique moderne (¿50')

- Manipulation du texte
- Stockage & indexation des informations
- Structuration des connaissances
- Extraction des connaissances

Les différentes tâches en NLP

■ Tâches historiques

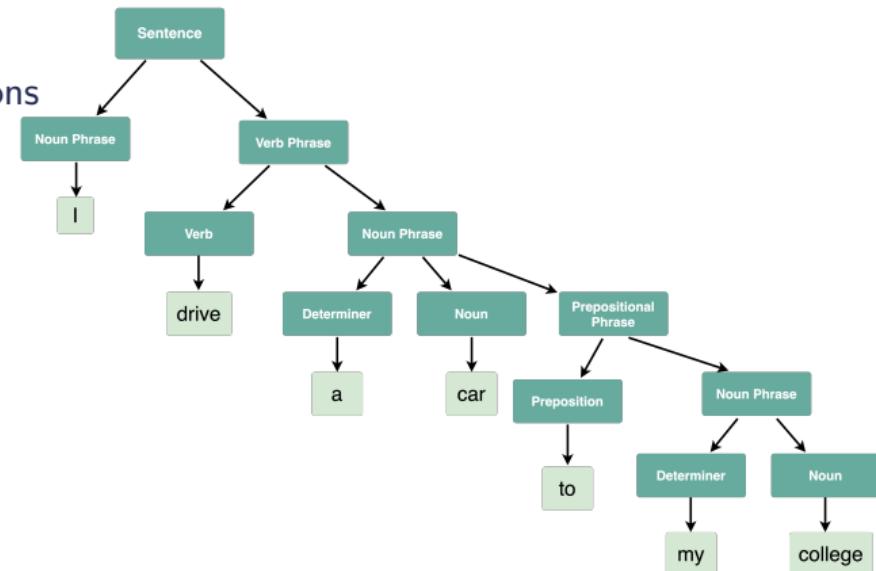
- Traduction automatique
- Extraction d'information
 - NER, Key/value, event, relations

■ Sous-tâches requises

- Analyse linguistique
 - Part-of-speech
 - Syntax Tree Parsing
 - Semantic Role Labeling

■ Tâche plus récentes

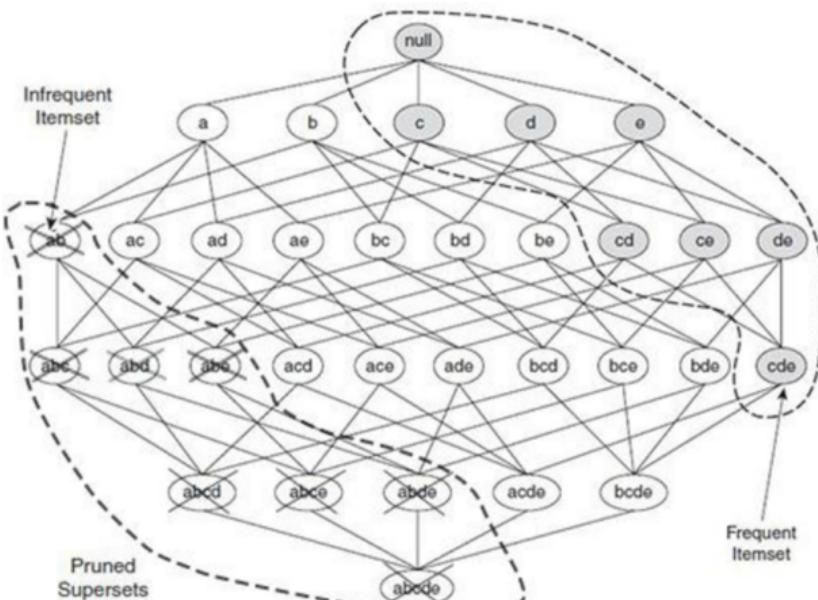
- Analyse thématique (sup/non-sup)
- Opinion mining
- Question Answering
- ...



Premiers modèles de NLP

(80'-90')

- Pattern matching / regex
 - = rules implementation
- Frequent Itemsets
 - = rule discovery
- + Linguistic resources (Synonyms, rules extension)
- + rely on subtask (e.g. PoS)
- High precision / (Low recall)
- Naive Bayes
 - Indexation / classification
 - Feature engineering
 - Scale ++ (SQL)



Specific Data/Tools/Scale/resources ⇒ NLP ≠ ML community

SVM / HMM: rapprochement ML-NLP

(90'-2000')

- **SVM** (linéaires, régularisés) ⇒ Nouvelles performances sur des jeux de données plus gros
 - Niveau document (classification)
 - (Niveau mot, *String kernel*)
- **HMM** ⇒ Nouvelles performances en analyse de séquence
 - PoS, Tree Parsing, Semantic Role Labeling, **NER**
 - Mais aussi : analyse fine des opinions

↗ mémoire, ↗ puissance ⇒ Nouvelles opportunités algo.

High precision / low recall

L'industrie du NLP
ne prend pas le virage

Low(er) Precision / high(er) recall

Les besoins en ML de la
communauté académique NLP
augmentent

2000'-2010': Transition

- Multiplication des **outils & nouvelles pratiques**
 - SVMlight, Liblinear
 - Torch+NLP (Collobert 08: Senna)
 - Modèles pré-entraînés
- Multiplication des **tâches**
 - SVM à sortie structurée (Structured output)
 - Analyse de documents structurés (& HTML)
 - Question Answering
 - Machine Translation
- Multiplication des **jeux de données** (de + en + libres)
 - CoNLL 02' (+03', 04'), ACE 05', NYTimes 08'
 - Wikipedia (?01'), DBPedia/wikidata (?07')
 - User-Generated-Content / Réseaux Sociaux

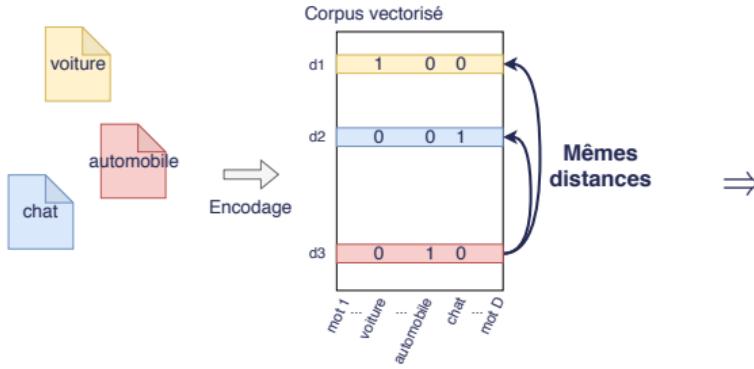
⇒ **Emergence des nouveaux acteurs** industrie+recherche

- GAFAM (mais pas seulement)
- Créneau = NLP + ML

2010'-2020': basculement = ↗↗ performances en ML

- 1 Apprentissage de représentation des mots (auto-supervisé)
 - Synonymie, encodage des caractéristiques grammaticales & sémantiques
- 2 Fonction d'agrégation efficace (CNN, RNN, Transformer)
 - Encodage du sens des phrases
- 3 Ouverture vers le transfert (comme en image)
- 4 Optimisations diverses (passage aux sous-mots, transformer, ...)

Fossé sémantique



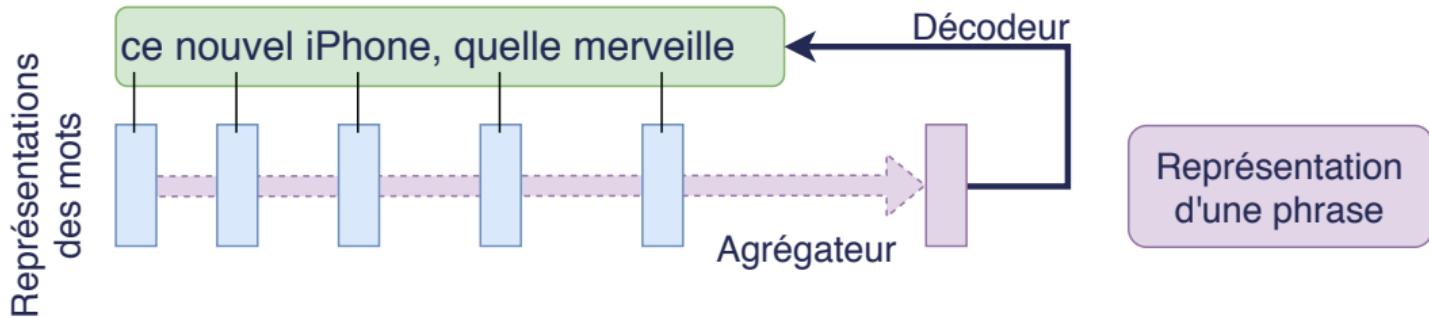
Représentation continue

(1) Initialisation aléatoire des positions des mots



2010'-2020': basculement = ↗↗ performances en ML

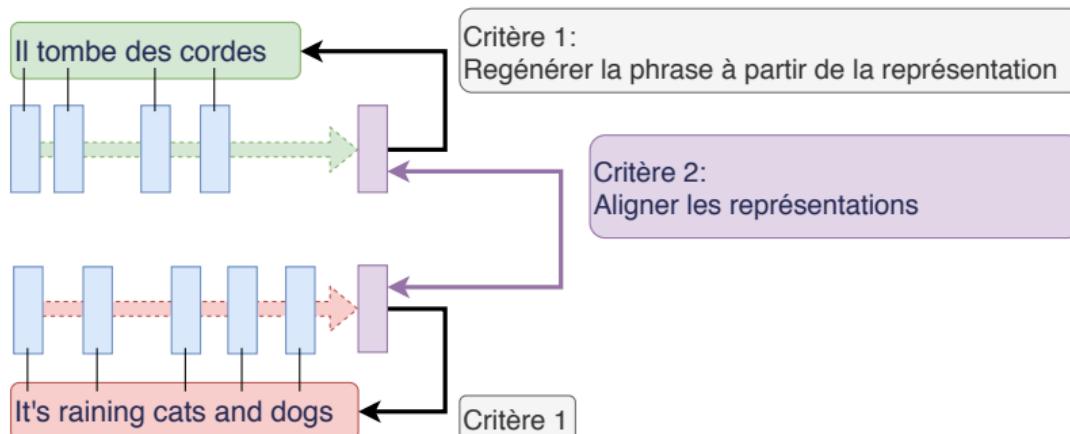
- 1 Apprentissage de représentation des mots (auto-supervisé)
 - Synonymie, encodage des caractéristiques grammaticales & sémantiques
- 2 Fonction d'agrégation efficace (CNN, RNN, Transformer)
 - Encodage du sens des phrases
- 3 Ouverture vers le transfert (comme en image)
- 4 Optimisations diverses (passage aux sous-mots, transformer, ...)



Typiquement un espace vectoriel de taille 768

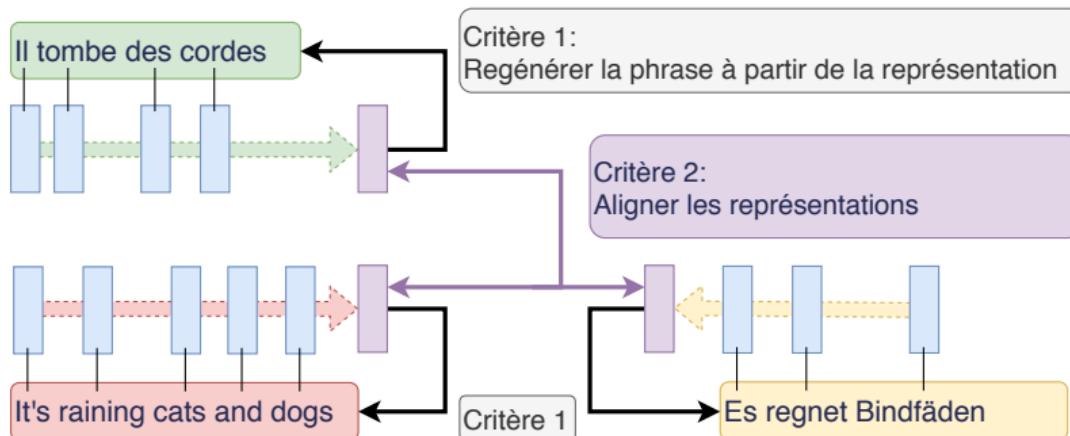
2010'-2020': basculement = ↗↗ performances en ML

- 1 Apprentissage de représentation des mots (auto-supervisé)
 - Synonymie, encodage des caractéristiques grammaticales & sémantiques
- 2 Fonction d'agrégation efficace (CNN, RNN, Transformer)
 - Encodage du sens des phrases
- 3 Ouverture vers le transfert (comme en image)
- 4 Optimisations diverses (passage aux sous-mots, transformer, ...)



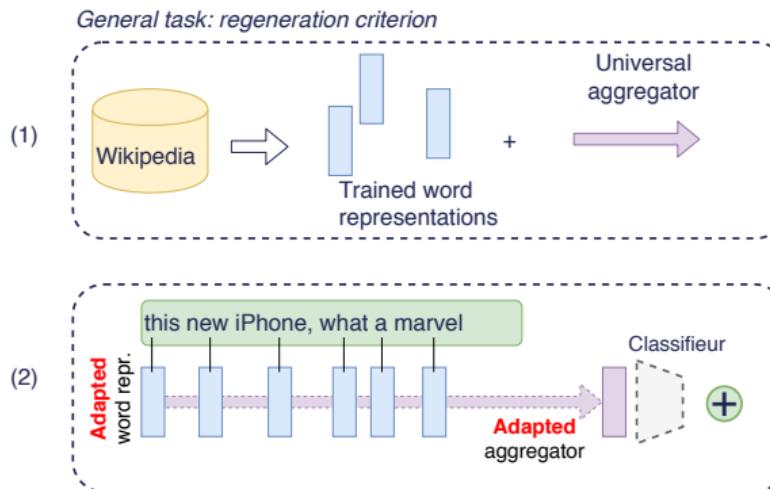
2010'-2020': basculement = ↗↗ performances en ML

- 1 Apprentissage de représentation des mots (auto-supervisé)
 - Synonymie, encodage des caractéristiques grammaticales & sémantiques
- 2 Fonction d'agrégation efficace (CNN, RNN, Transformer)
 - Encodage du sens des phrases
- 3 Ouverture vers le transfert (comme en image)
- 4 Optimisations diverses (passage aux sous-mots, transformer, ...)



2010'-2020': basculement = ↗↗ performances en ML

- 1 Apprentissage de représentation des mots (auto-supervisé)
 - Synonymie, encodage des caractéristiques grammaticales & sémantiques
- 2 Fonction d'agrégation efficace (CNN, RNN, Transformer)
 - Encodage du sens des phrases
- 3 Ouverture vers le transfert (comme en image)
- 4 Optimisations diverses (passage aux sous-mots, transformer, ...)



Synthèse

1990 Matrix factorization :

- PCA is a historical way to learn representations
 - Criterion = reconstruction
- In NLP ⇒ SVD / PCA

[Deerwester, 1990]

2003 Recurrent neural architecture

[Bengio, 2003]

2005 Simplified neural architecture for text representation

- an alternative to PLSA

[Keller, 2005]

2008 Convolutional Neural architecture for text

- precursor of modern architectures
- multi-tasks

[Collobert, 2008]

2012 The Word2Vec wave

- Qualitative & cheap word embeddings

[Mikolov, 2012]

2013 Manifest for representation learning

[Bengio, 2013]

2014 Seq2seq paradigm

[Sutskever, 2015]

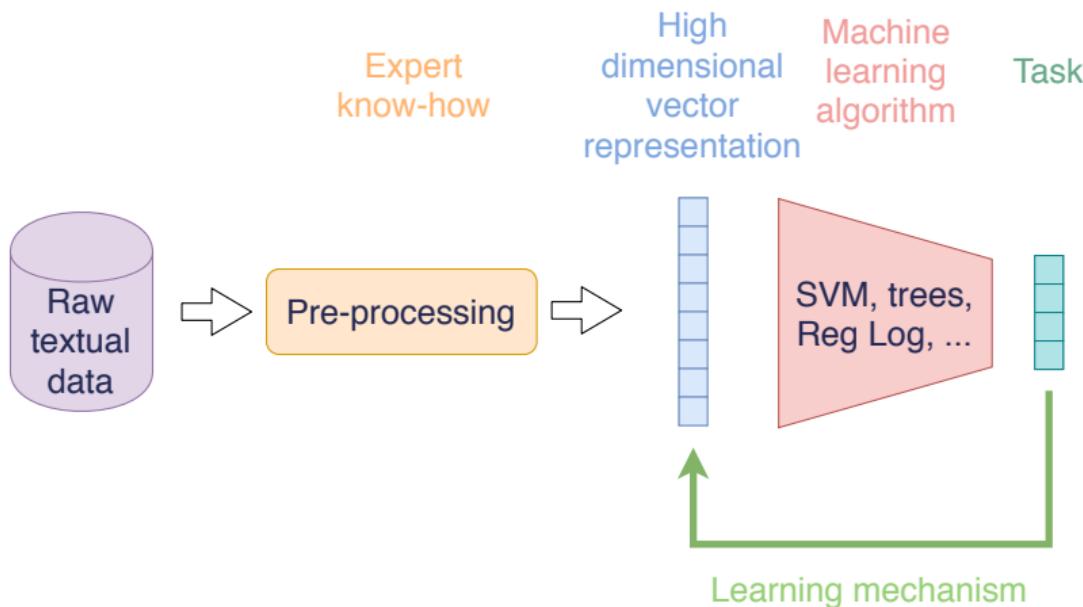
⇒ Also an approximate schedule of the presentation

Synthèse

General philosophy : building a virtuous cycle

- Tackling directly raw data
- Learn a representation that can be useful for several tasks...

Important keyword = **embedding**

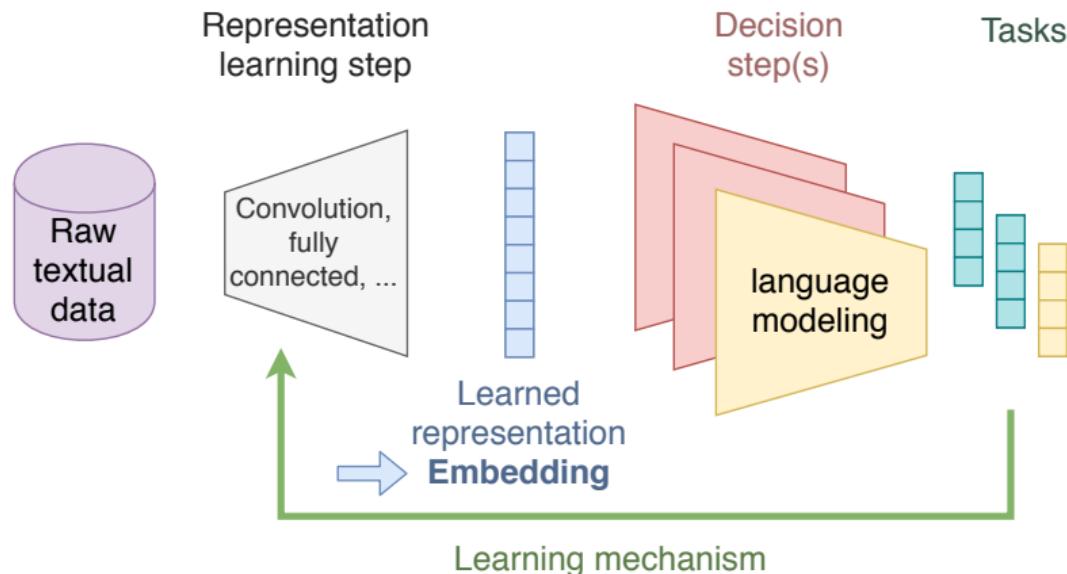


Synthèse

General philosophy : building a virtuous cycle

- Tackling directly raw data
- Learn a representation that can be useful for several tasks...

Important keyword = **embedding**



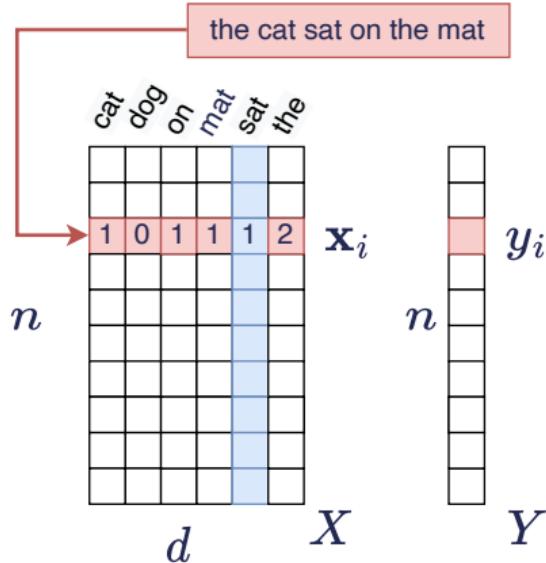
MODÉLISATIONS EN ML / NLP

RUPTURE:
ML \Rightarrow REPRESENTATION
LEARNING



De Naive Bayes à SVM

the cat sat on the mat



Fonction de classification

$$f(\mathbf{x}_i) = \sum_{j=1}^d w_j x_{ij}, \quad \log p(\mathbf{x}_i | c) = \sum_{j=1}^{|D|} x_{ij} \underbrace{\log P(\text{mot}_j | c)}_{\theta_j}$$

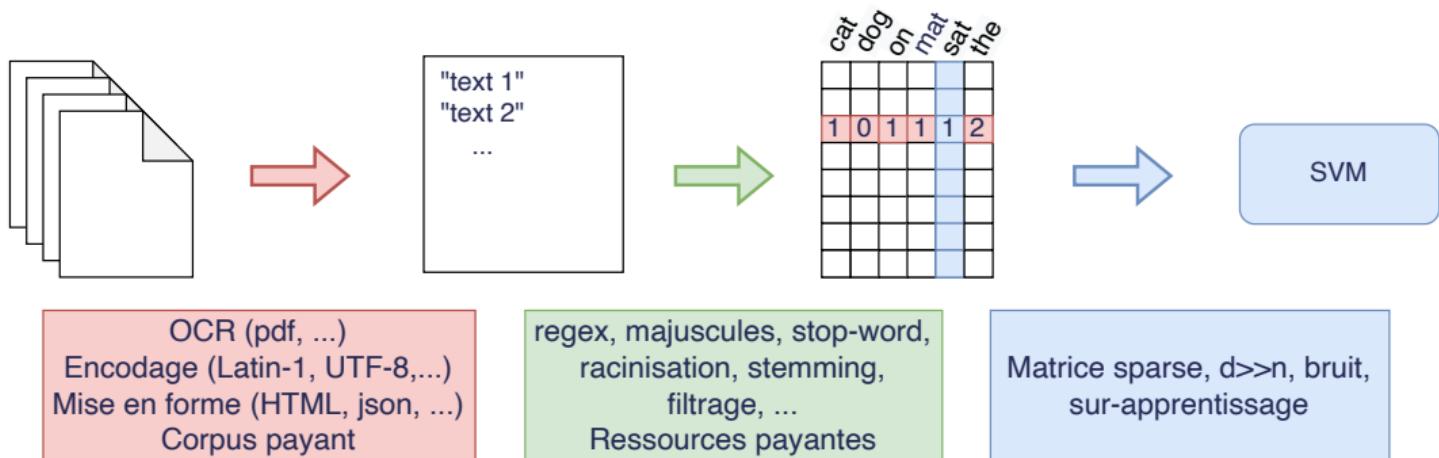
- Problème spécifique: $d \gg n$
- SVM = bonne implémentation + régularisation
- Interprétation \Rightarrow analyse w_j/θ_j

Classification au niveau paragraphe/document

- Analyse thématique, opinion, sondage, suivi temporel dans les réseaux sociaux

Difficultés/spécificités NLP

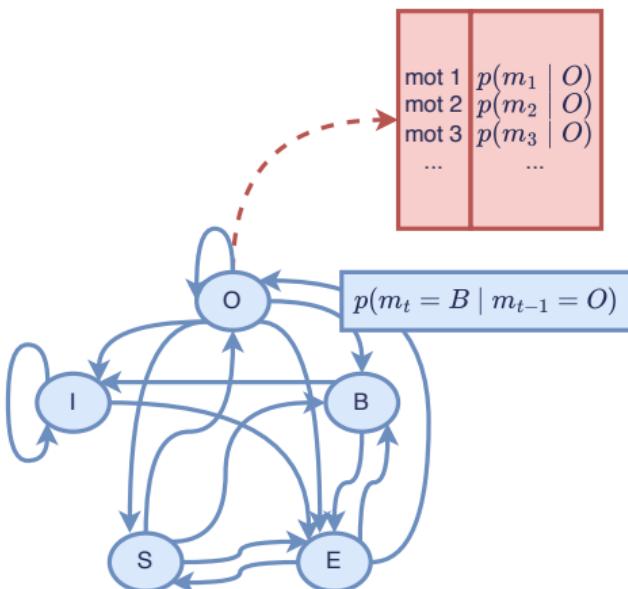
- Savoir faire sur les traitements
- Outils+ressources spécifiques ⇒ communauté spécifique





Modélisation historique des séquences

- HMM = Etats cachés (classes), transition (entre classes), émission (classes \Rightarrow observations)



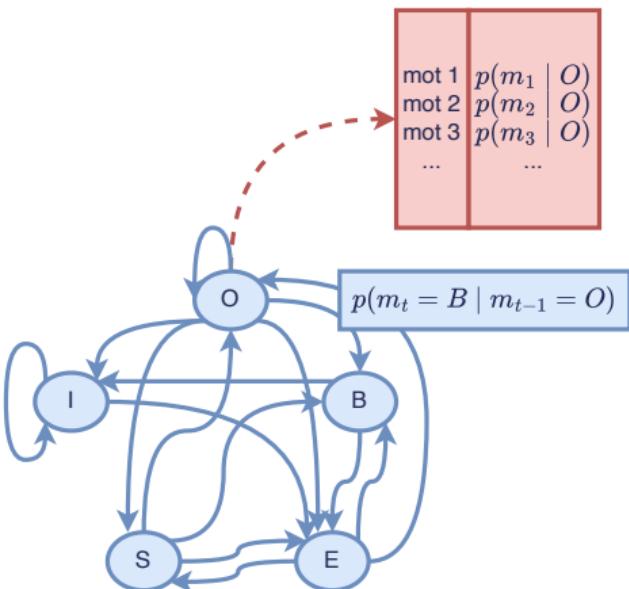
Words	BIOES Label
Jane	B-PER
Villanueva	E-PER
of	O
United	B-ORG
Airlines	I-ORG
Holding	E-ORG
discussed	O
the	O
Chicago	S-LOC
route	O
.	O

'B' : Beginning of named entity
 'I' : Inside of named entity
 'O' : Outside of named entity
 'E' : End of named entity
 'S' : Single named entity



Modélisation historique des séquences

- HMM = Etats cachés (classes), transition (entre classes), émission (classes \Rightarrow observations)



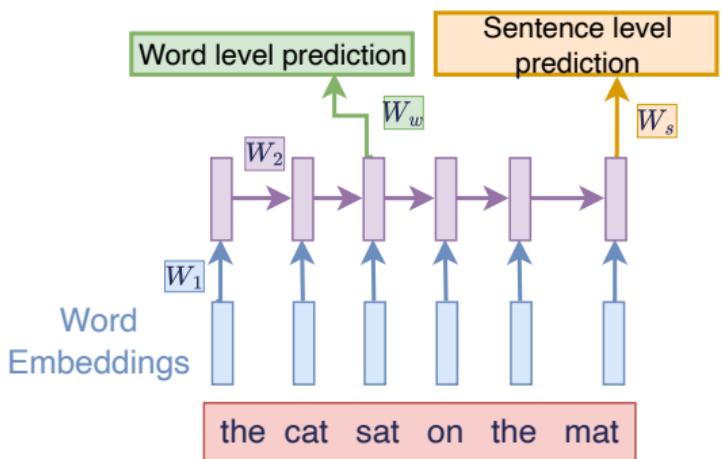
- Apprentissage (par comptage)
 - Données étiquetées requises
- Inférence Viterbi
 - Mémoire + calcul
- Performance ++
- Evolutions algo. \Rightarrow CRF

Bcp de pré-traitements
+ astuces pour limiter les paramètres
 \Rightarrow Communauté NLP



Introduction des modèles de langue

[Bengio 2003]



Représentation continue de concepts discrets:

$$w_i = \text{"cat"} \Rightarrow \mathbf{x}_i \in \mathbb{R}^d$$

- \mathbf{x}_i = embedding
- $\{\mathbf{x}_i\}_{i=1,\dots,D}$ = lookup table
- Distance : $s(w_i, w_j) = \mathbf{x}_i \cdot \mathbf{x}_j$

Idée

Distance/similarité = encodage du sens+grammaire



Bengio et al., JMLR, 2003
A Neural probabilistic language model

Representation learning in text

[Keller, 2005]

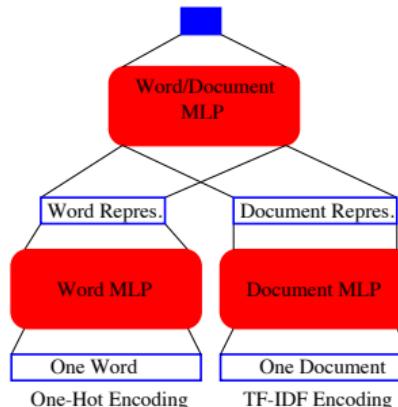
Task =

Is the word w present in document d ?

\Rightarrow Self-supervision

Close to LSA / PLSA paradigm: compressing the original data matrix through embeddings

\Rightarrow Learning a language model



Conclusion:

Words & documents representations are more efficient than PLSA ones for several TREC information retrieval tasks



M. Keller, S. Bengio, ICANN 2005
A neural network for text representation

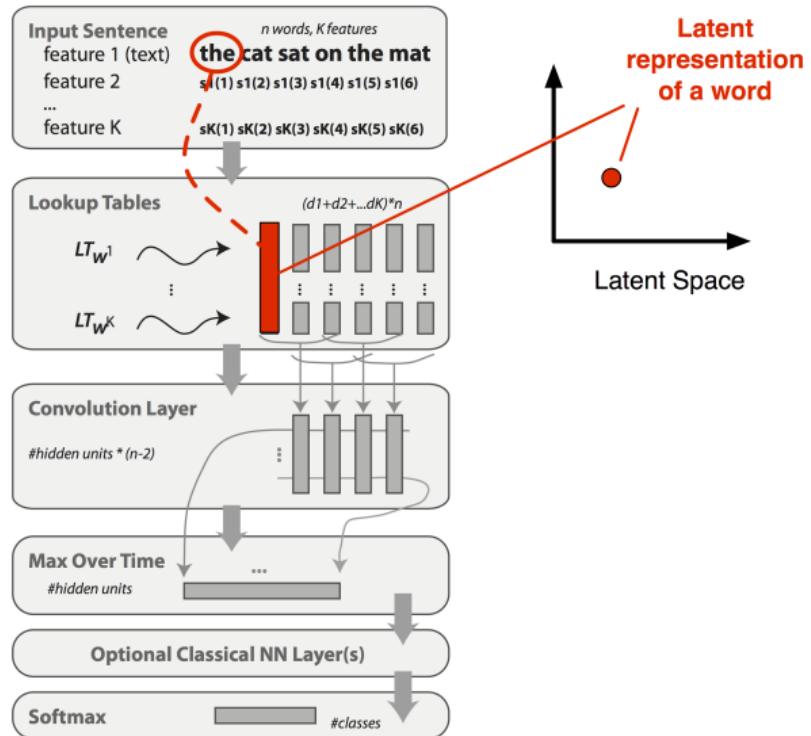
Convolutional Neural Architecture

[Collobert, 2008]

Multi-task representation learning

- State of the art on :
POS, NER, SRL
- Quite difficult to set...
- ... But open source +
open embeddings
- Based on torch...

By Collobert



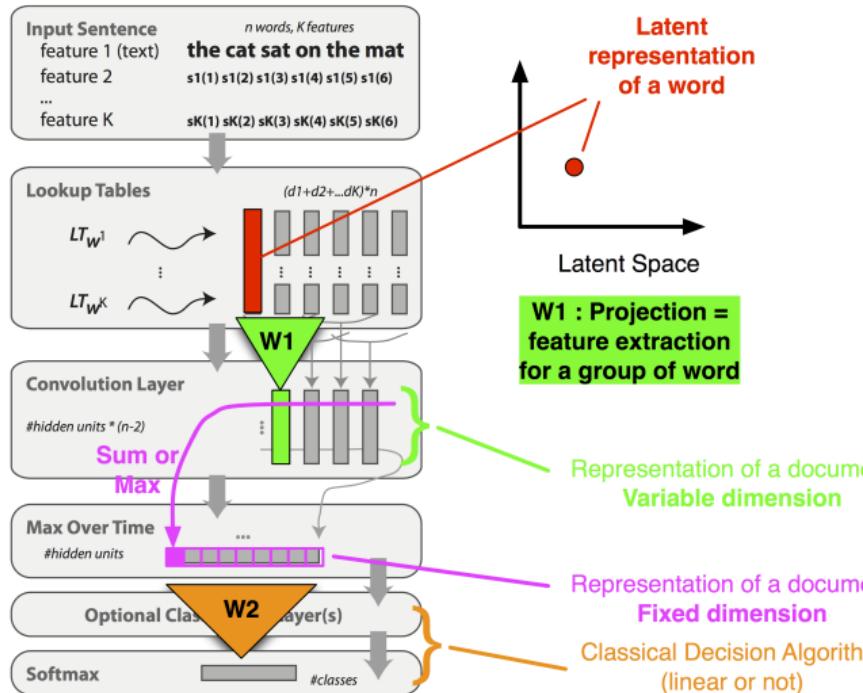
Convolutional Neural Architecture

[Collobert, 2008]

Multi-task representation learning

- State of the art on :
POS, NER, SRL
- Quite difficult to set...
- ... But open source +
open embeddings

By Collobert



R. Collobert, J. Weston ICML 2008

A unified architecture for natural language processing: Deep neural networks with multitask learning

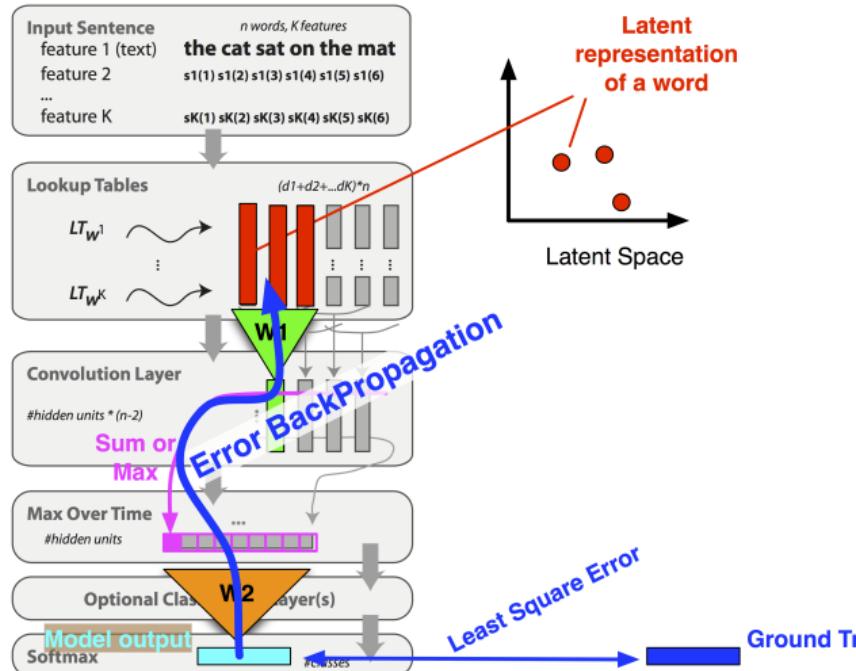
Convolutional Neural Architecture

[Collobert, 2008]

Multi-task representation learning

- State of the art on :
POS, NER, SRL
- Quite difficult to set...
- ... But open source +
open embeddings
- Based on torch...

By Collobert



R. Collobert, J. Weston ICML 2008

A unified architecture for natural language processing: Deep neural networks with multitask learning

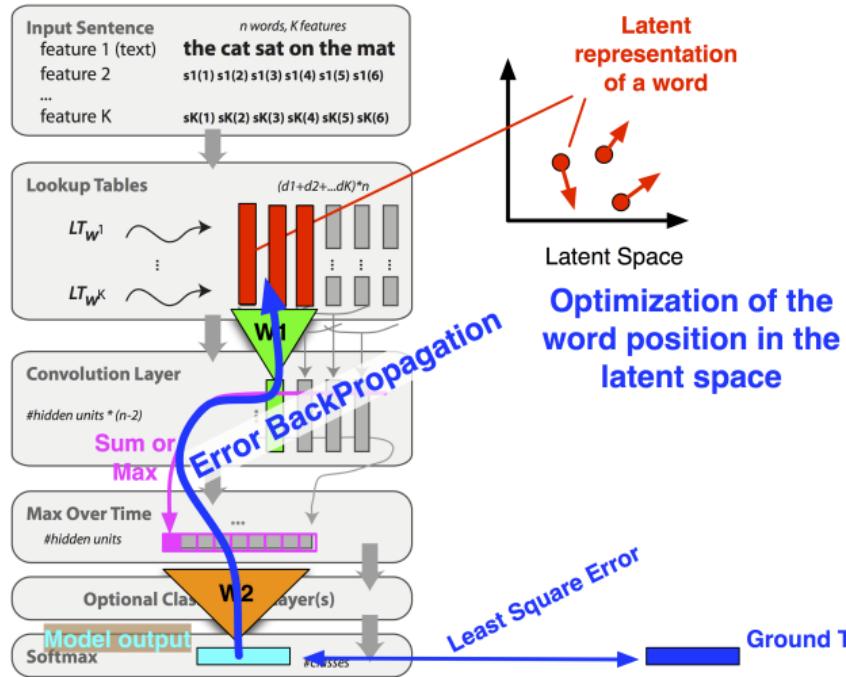
Convolutional Neural Architecture

[Collobert, 2008]

Multi-task representation learning

- State of the art on :
POS, NER, SRL
- Quite difficult to set...
- ... But open source +
open embeddings
- Based on torch...

By Collobert



R. Collobert, J. Weston ICML 2008

A unified architecture for natural language processing: Deep neural networks with multitask learning

Convolutional Neural Architecture (2)

[Collobert, 2008]

Several important informations

- Embedding are learned **keeping the sentence structure**
- Embeddings benefit from **multi-tasks**

Our embeddings have been trained for about 2 months, over Wikipedia.

<https://ronan.collobert.com/senna/>

- Learning is slow... ...But inference is fast & efficient

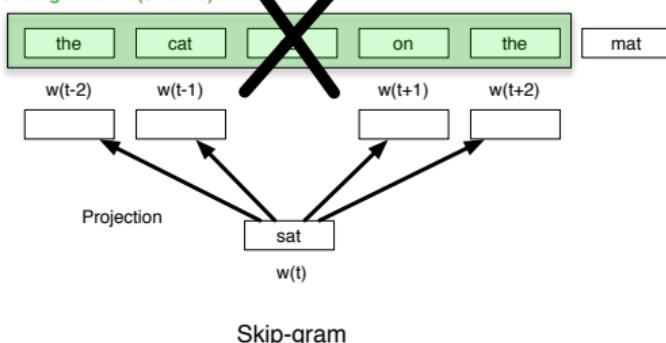
Task	Benchmark	Performance	Timing (s)
Part of Speech (POS)	Toutanova et al, 2003	(Accuracy) 97.29%	3
Chunking (CHK)	CoNLL 2000	(F1) 94.32%	2
Name Entity Recognition (NER)	CoNLL 2003	(F1) 89.59%	2
Semantic Role Labeling (SRL)	CoNLL 2005	(F1) 75.49%	36
Syntactic Parsing (PSG)	Penn Treebank	(F1) 87.92%	74

- + Costly **embeddings available** to the community

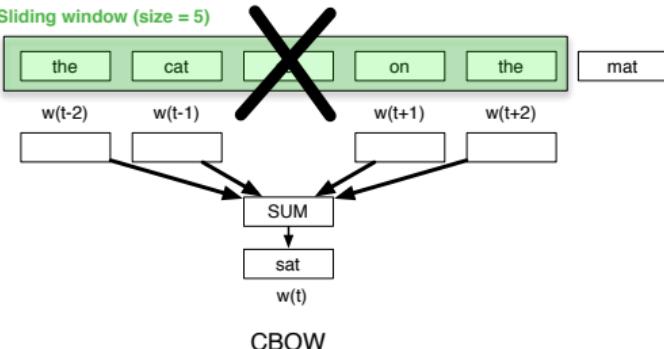
Word2Vec : extracting local semantics

[Mikolov, 2013]

Sliding window (size = 5)



Sliding window (size = 5)



Skip-Gram (& *negative sampling* implementation) : easy learning

- Local analysis
- *predictive* criterion : estimating missing value
- Sliding window = local context C of word w



Mikolov, Sutskever, Chen, Corrado, Dean, NIPS 2013 (arXiv 2012)

Distributed representations of words and phrases and their compositionality



Word2Vec : extracting local semantics

[Mikolov, 2013]

he curtains open and the moon shining in on the barely
 ars and the cold , close moon " . And neither of the w
 rough the night with the moon shining so brightly , it
 made in the light of the moon . It all boils down , wr
 surely under a crescent moon , thrilled by ice-white
 sun , the seasons of the moon ? Home , alone , Jay pla
 m is dazzling snow , the moon has risen full and cold
 un and the temple of the moon , driving out of the hug
 in the dark and now the moon rises , full and amber a
 bird on the shape of the moon over the trees in front
 But I could n't see the moon or the stars , only the
 rning , with a sliver of moon hanging among the stars
 they love the sun , the moon and the stars . None of
 the light of an enormous moon . Theplash of flowing w
 man 's first step on the moon ; various exhibits , aer
 the inevitable piece of moon rock . Housing The Airsh
 oud obscured part of the moon . The Allied guns behind

(1) Initialisation aléatoire des positions des mots

	happy	temple
sofa		
night	cold	car
stars		city
	moon	shape
toy		
kitchen	shining	river
		toy

Skip-Gram (& negative sampling implementation) : easy learning

- Local analysis
- predictive criterion : estimating missing value
- Sliding window = local context C of word w



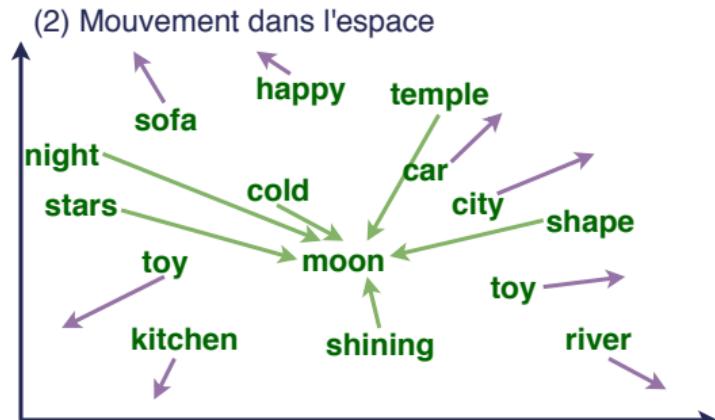
Mikolov, Sutskever, Chen, Corrado, Dean, NIPS 2013 (arXiv 2012)

Distributed representations of words and phrases and their compositionality

Word2Vec : extracting local semantics

[Mikolov, 2013]

he curtains open and the moon shining in on the barely
 ars and the cold , close moon " . And neither of the w
 rough the night with the moon shining so brightly , it
 made in the light of the moon . It all boils down , wr
 surely under a crescent moon , thrilled by ice-white
 sun , the seasons of the moon ? Home , alone , Jay pla
 m is dazzling snow , the moon has risen full and cold
 un and the temple of the moon , driving out of the hug
 in the dark and now the moon rises , full and amber a
 bird on the shape of the moon over the trees in front
 But I could n't see the moon or the stars , only the
 rning , with a sliver of moon hanging among the stars
 they love the sun , the moon and the stars . None of
 the light of an enormous moon . Theplash of flowing w
 man 's first step on the moon ; various exhibits , aer
 the inevitable piece of moon rock . Housing The Airsh
 oud obscured part of the moon . The Allied guns behind



Skip-Gram (& negative sampling implementation) : easy learning

- Local analysis
- predictive criterion : estimating missing value
- Sliding window = local context C of word w



Mikolov, Sutskever, Chen, Corrado, Dean, NIPS 2013 (arXiv 2012)

Distributed representations of words and phrases and their compositionality

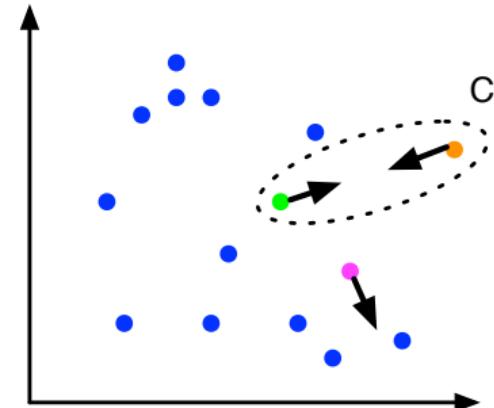


W2V : detailed explanation

- Given word w and local contexts C :

$$\text{Idée SG: } \arg \max_{\theta} \prod_C \prod_{w \in C} p(C|w; \theta)$$

- $p(D = 1|w_i, w_j; \theta) \Rightarrow$ proba. that w_i and w_j occur in the same context



$$\arg \max_{\theta} \underbrace{\prod_{i,j \in C} p(D = 1|w_i, w_j; \theta)}_{\text{Negative Sampling}} + \underbrace{\prod_{i,j \in \bar{C}} p(D = 0|w_i, w_j; \theta)}$$

Goldberg, Levy, arXiv 2014
word2vec Explained: Deriving Mikolov et al.'s Negative-Sampling Word-Embedding Method

Hammer, NN 2002
Generalized Relevance Learning Vector Quantization

W2V : Skip-gram & negative sampling

$$\arg \max_{\theta} \prod_{i,j \in C} p(D = 1 | w_i, w_j; \theta) + \underbrace{\prod_{i,j \in \bar{C}} p(D = 0 | w_i, w_j; \theta)}_{\text{Negative Sampling}}$$

- Using logistic function : $p(D = 1 | w_i, w_j) = \frac{1}{1 + \exp(-\mathbf{z}_i \cdot \mathbf{z}_j)}$
- Global log-likelihood : $\arg \max_{\mathbf{z}} \left(\sum_{i,j \in C} \log \sigma(\mathbf{z}_i \cdot \mathbf{z}_j) + \sum_{i,j \in \bar{C}} \log \sigma(-\mathbf{z}_i \cdot \mathbf{z}_j) \right)$

σ : fct sigmoide, C : Set of Cooccurrences, \bar{C} : Set of Non-Cooc

- Stochastic Gradient Descent + triplet loss
- Frequent word subsampling trick : picking words with

$$p(w_i) = 1 - \sqrt{\frac{t}{\text{freq}(w_i)}}, \quad t = 10^{-5}$$



W2V: Parametrization & results

- Choosing a large latent space: $z \in [500, 1000]$
 $(vs \text{ PLSA/LDA } z \in [10, 100])$
- Operate on large & small corpora...
- ... very fast !

Model (training time)	Redmond	Havel	ninjutsu	graffiti	capitulate
Collobert (50d) (2 months)	conyers lubbock keene	plauen dzerzhinsky osterreich	reiki kohana karate	cheesecake gossip dioramas	abdicate accede rearm
Turian (200d) (few weeks)	McCarthy Alston Cousins	Jewell Arzu Ovitz	- - -	gunfire emotion impunity	- - -
Mnih (100d) (7 days)	Podhurst Harlang Agarwal	Pontiff Pinochet Rodionov	- - -	anaesthetics monkeys Jews	Mavericks planning hesitated
Skip-Phrase (1000d, 1 day)	Redmond Wash. Redmond Washington Microsoft	Vaclav Havel president Vaclav Havel Velvet Revolution	ninja martial arts swordsmanship	spray paint grafitti taggers	capitulation capitulated capitulating

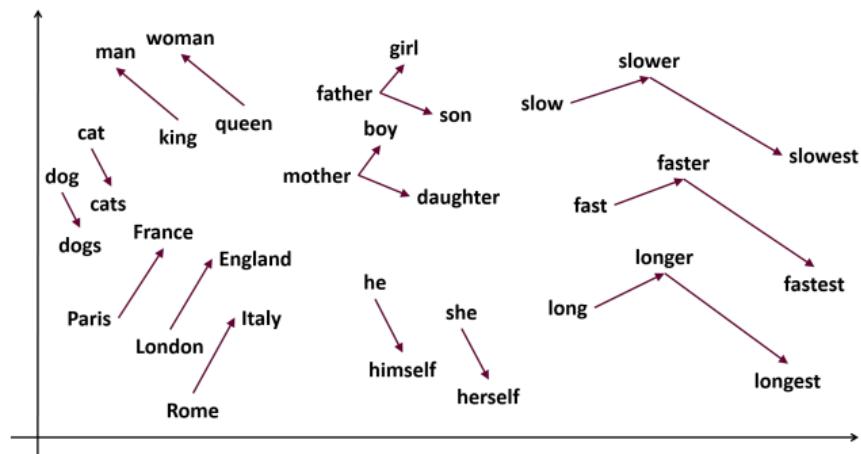
W2V: Playing in the latent space

a is to *b* what *c* is to ???

↔

$$\mathbf{z}_b - \mathbf{z}_a + \mathbf{z}_c$$

Syntactical property (1):



Query:

$$\mathbf{z}_{woman} - \mathbf{z}_{man} + \mathbf{z}_{king} = \mathbf{z}_{req}$$

Nearest neighbor:

$$\arg \min_i \|\mathbf{z}_{req} - \mathbf{z}_i\| = \text{queen}$$

$$\mathbf{z}_{woman} - \mathbf{z}_{man} \approx \mathbf{z}_{queen} - \mathbf{z}_{king}$$

$$\mathbf{z}_{kings} - \mathbf{z}_{king} \approx \mathbf{z}_{queens} - \mathbf{z}_{kings}$$



W2V: Playing in the latent space

$$a \text{ is to } b \text{ what } c \text{ is to } ??? \Leftrightarrow \mathbf{z}_b - \mathbf{z}_a + \mathbf{z}_c$$

Syntactical property (2):

Query:

$$\mathbf{z}_{easy} - \mathbf{z}_{easiest} + \mathbf{z}_{luckiest} = \mathbf{z}_{req}$$

Nearest neighbor:

$$\arg \min_i \|\mathbf{z}_{req} - \mathbf{z}_i\| = \text{lucky}$$



W2V: Playing in the latent space

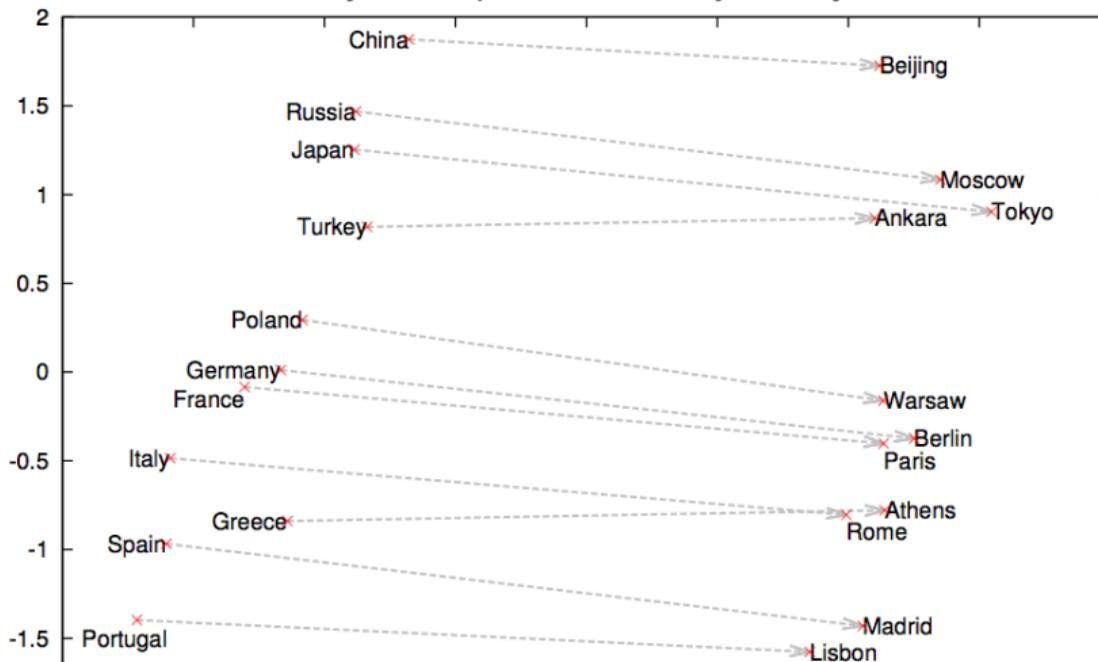
a is to *b* what *c* is to ???

↔

$$\mathbf{z}_b - \mathbf{z}_a + \mathbf{z}_c$$

Semantic Property (1):

Country and Capital Vectors Projected by PCA





W2V: Playing in the latent space

$$a \text{ is to } b \text{ what } c \text{ is to } ??? \Leftrightarrow \mathbf{z}_b - \mathbf{z}_a + \mathbf{z}_c$$

Semantic Property (2)

Czech + currency	Vietnam + capital	German + airlines	Russian + river	French + actress
koruna	Hanoi	airline Lufthansa	Moscow	Juliette Binoche
Check crown	Ho Chi Minh City	carrier Lufthansa	Volga River	Vanessa Paradis
Polish zolty	Viet Nam	flag carrier Lufthansa	upriver	Charlotte Gainsbourg
CTK	Vietnamese	Lufthansa	Russia	Cecile De

Table 5: Vector compositionality using element-wise addition. Four closest tokens to the sum of two vectors are shown, using the best Skip-gram model.



fastText : syntactic robustness

[Bojanovski, 2016]

How to deal with unknown/rare words ?

- Word2Vec : creating a specific embedding
to keep the global structure of the sentence
- fastText :
 - Word embeddings
 - **Subword embeddings = N-grams of characters**
 - Impl.: special bounding characters: <> + skip-gram & negative sampling

Word where / $N = 3 \Rightarrow \langle \text{where} \rangle + \langle \text{wh}, \text{ whe}, \text{ her}, \text{ ere}, \text{ re} \rangle$

- ⇒ **Rare/unknown words** can be represented by their N-grams of characters
- ⇒ **Preprocessings** are strongly reduced / dictionary has a fixed size
- ⇒ Evolution: N-grams of characters ⇒ **byte pair encoding** (variable N),
focus on freq. patterns

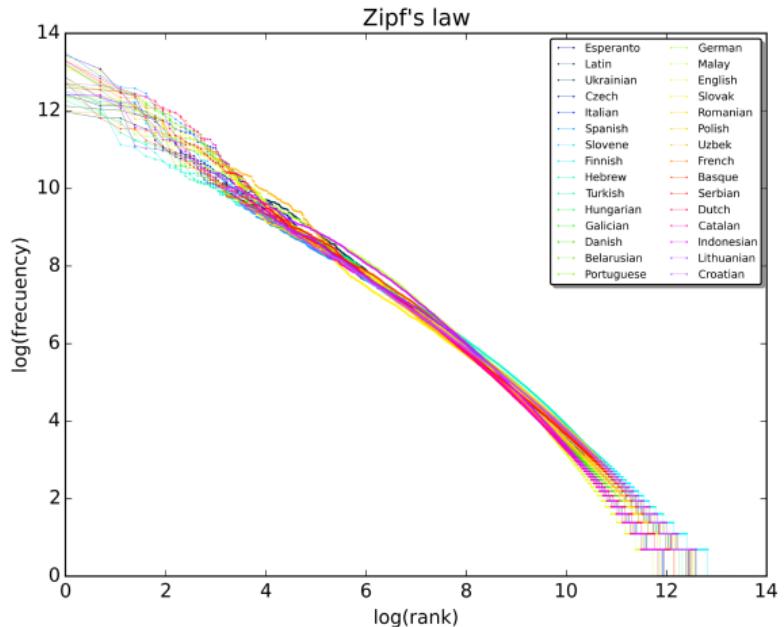




fastText : syntactic robustness

[Bojanovski, 2016]

How to deal with unknown/rare words ?



Few words appear a lot...
many words appear rarely

Great issue !



P. Bojanowski, E. Grave, A. Joulin and T. Mikolov, Trans. ACL 2017 (arXiv 2016)
Enriching Word Vectors with Subword Information

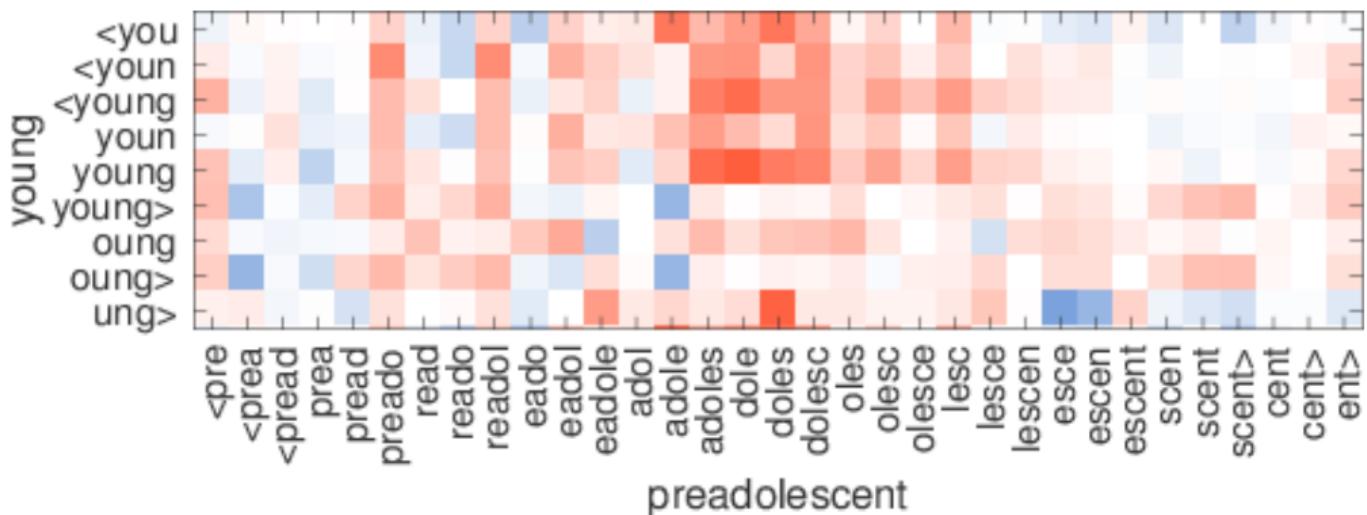


fastText : syntactic robustness

[Bojanovski, 2016]

How to deal with unknown/rare words ?

Typical result with fastText:



Cherry on the cake: fastText... is fast

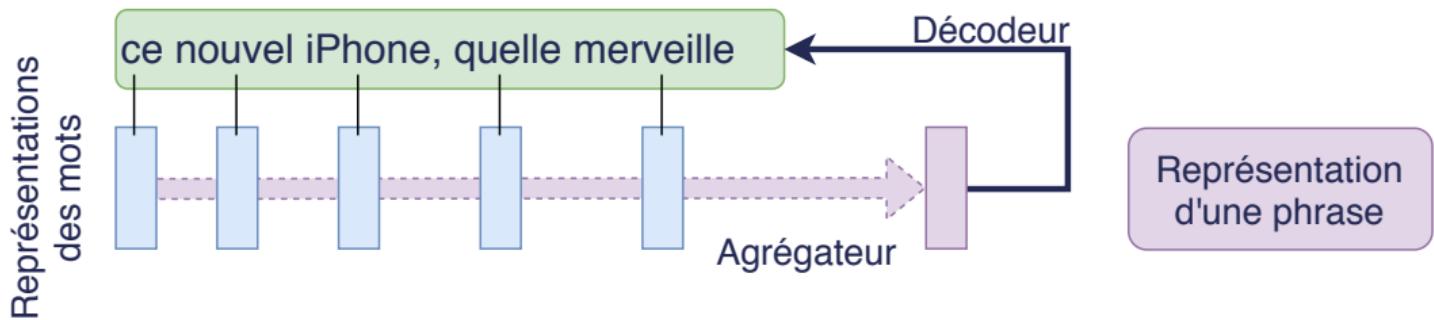


P. Bojanowski, E. Grave, A. Joulin and T. Mikolov, Trans. ACL 2017 (arXiv 2016)
Enriching Word Vectors with Subword Information

NLP: AGRÉGATION DES REPRÉSENTATIONS DE MOTS

(Rappel) Enjeux de l'agrégation des représentations de mots

- Très peu de tâche au niveau des **mots**
- **Phrase:** Sentiment, thématique, décomposition...
- Classification de mots... Au niveau de phrase
 - e.g. : la détection d'entités/PoS nécessite l'analyse de la phrase



Moyenne & max: des limites très vite atteintes [Le, 2014]

W2V= powerful semantics at the **word level**...

How scaling to the **sentence or document level**?

Simple averaging (or max) of word embeddings:

- + great results on small word groups
- poor results on larger groups
 - quickly converge to a central abstract point of the latent space

Moyenne & max: des limites très vite atteintes [Le, 2014]

W2V= powerful semantics at the **word level**...

How scaling to the **sentence** or **document level**?

- Aggregate multiple words associated to a single entity

- Pointwise Mutual Information* threshold:

$$\text{score}(w_i, w_j) = \frac{\text{count}(w_i w_j) - \delta}{\text{count}(w_i) \times \text{count}(w_j)}.$$

- Include new terms in the dictionary before running word2vec

Newspapers			
New York	New York Times	Baltimore	Baltimore Sun
San Jose	San Jose Mercury News	Cincinnati	Cincinnati Enquirer
NHL Teams			
Boston	Boston Bruins	Montreal	Montreal Canadiens
Phoenix	Phoenix Coyotes	Nashville	Nashville Predators
NBA Teams			
Detroit	Detroit Pistons	Toronto	Toronto Raptors
Oakland	Golden State Warriors	Memphis	Memphis Grizzlies
Airlines			
Austria	Austrian Airlines	Spain	Spanair
Belgium	Brussels Airlines	Greece	Aegean Airlines
Company executives			
Steve Ballmer	Microsoft	Larry Page	Google
Samuel J. Palmisano	IBM	Werner Vogels	Amazon

Moyenne & max: des limites très vite atteintes [Le, 2014]

W2V= powerful semantics at the **word level**...

How scaling to the **sentence** or **document level**?

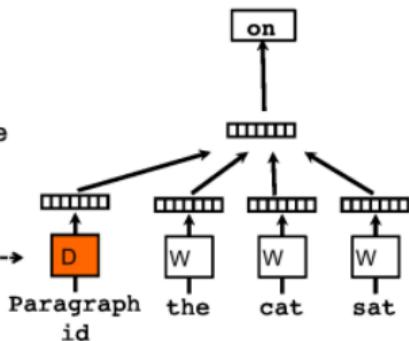
Classifier

on

Average/Concatenate

.....

Paragraph Matrix ----->



Classifier

the

cat

sat

on

Paragraph Matrix ----->

D

Paragraph
id



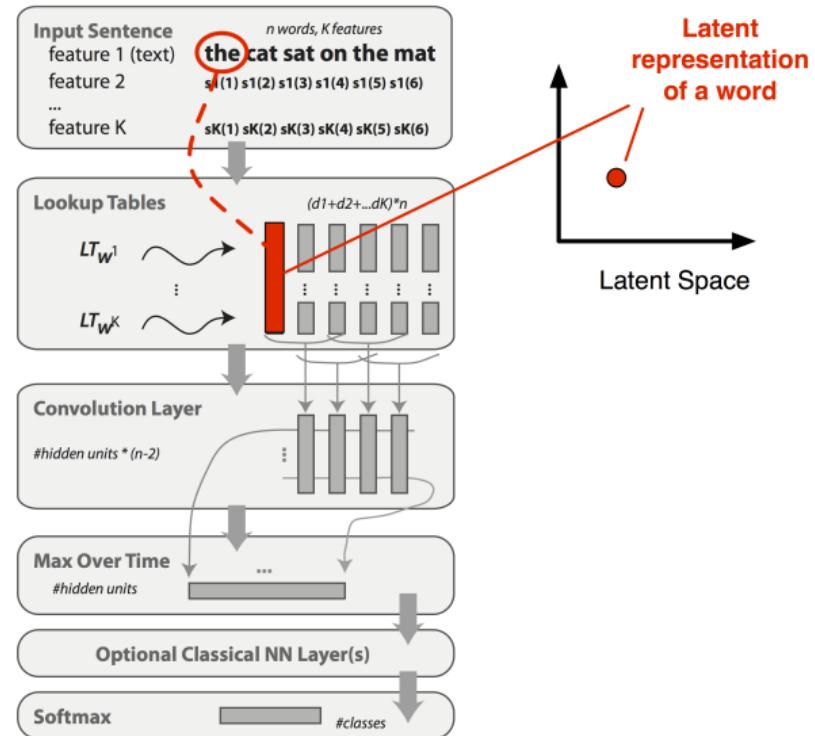
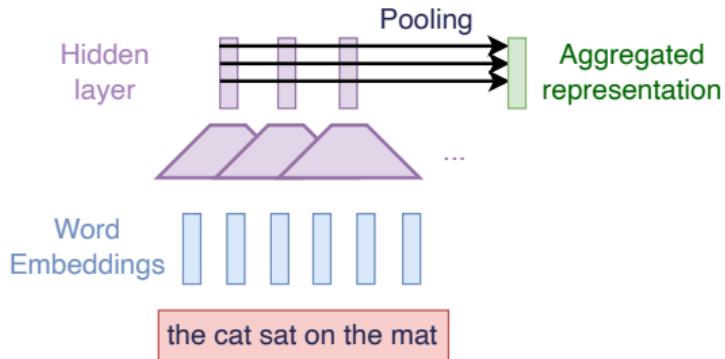
Q. Le, T. Mikolov, ICML 2014

Distributed representations of sentences and documents

Convolutional Neural Network (CNN)

[Collobert 08]

Principe:



- Convolution = Agrégation locale
- Agrégation globale
= pooling / attention
- Multiplier les couches
- Parallélisable



Recurrent Neural Network & LSTM

[Bengio 03]

The next step:

providing a better modeling of the link between words

⇒ Learning a latent representation of a **specific word group**

- ... but sentences have different lengths
- ⇒ using a **recurrent aggregator** ... ie a RNN

General RNN:

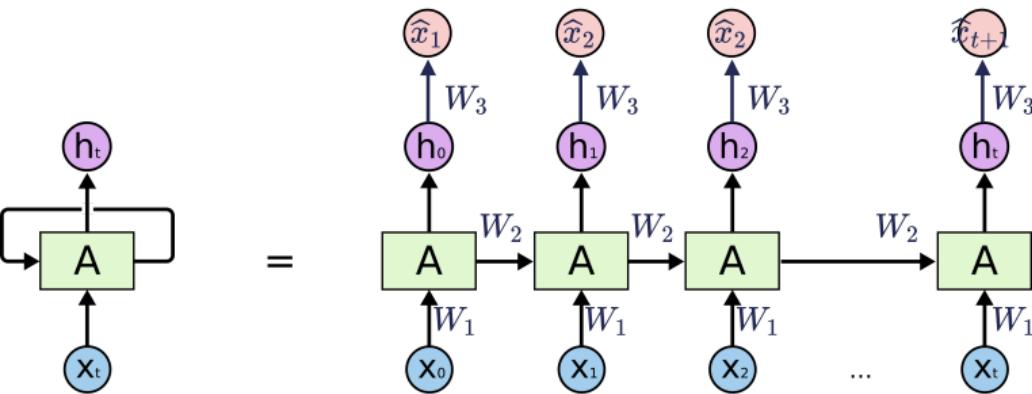
$$\mathbf{h}_t = W_1 \mathbf{x}_t + W_2 \mathbf{h}_{t-1}$$

Specific task :

$$\tilde{y}_t = W_3 \mathbf{h}_t$$

Loss example :

$$\mathcal{L}_t = (\tilde{y}_t - y_t)^2$$





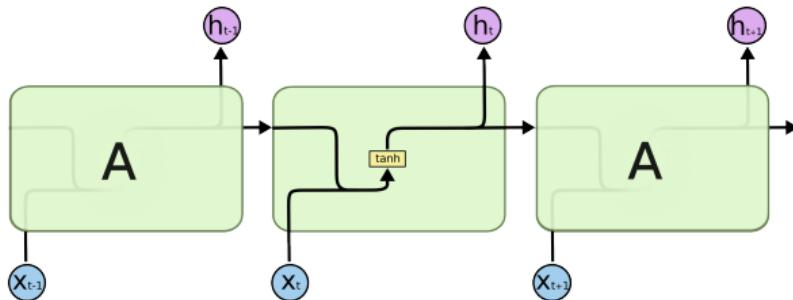
Recurrent Neural Network & LSTM

[Bengio 03]

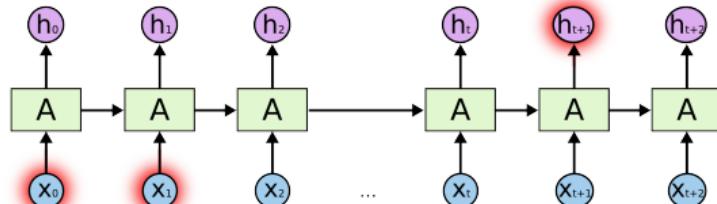
The next step:

providing a better modeling of the link between words

Classical RNN :



Gradient vanishes & long term dependancies are not modeled....



Chris Olah's blog <http://colah.github.io/posts/2015-08-Understanding-LSTMs/>



Recurrent Neural Network & LSTM

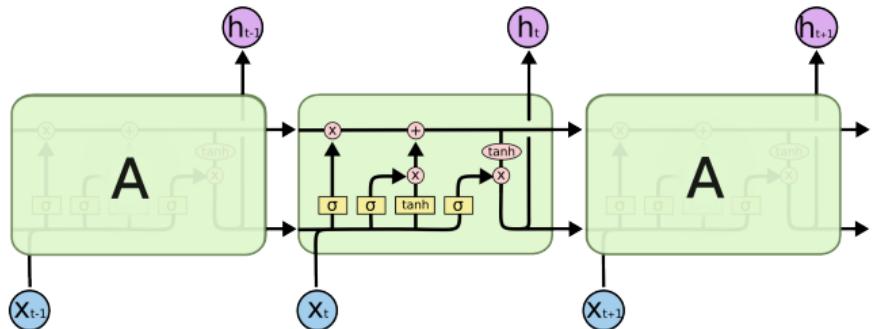
[Bengio 03]

The next step:

providing a better modeling of the link between words

The phenomenon has been understood & (partially) overcome:

Neurons **learn** what should be **kept in memory** and what should be **forgotten**



Gated architecture



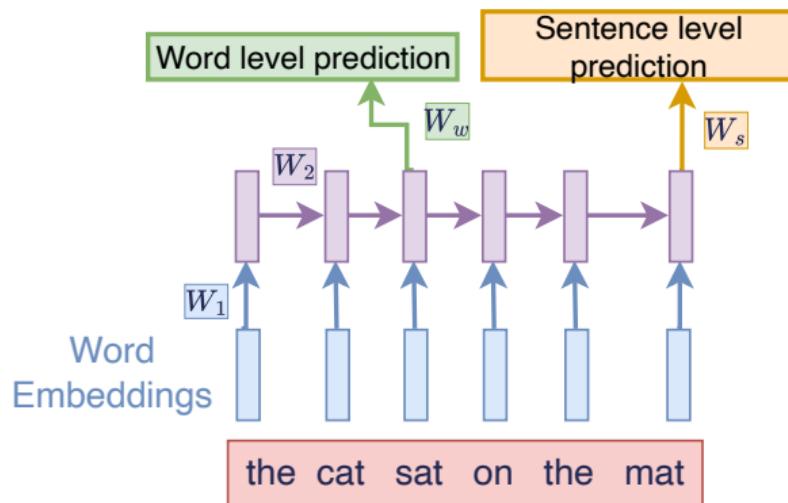
S. Hochreiter, J. Schmidhuber, Neural computation 1997
Long short-term memory

Chris Olah's blog <http://colah.github.io/posts/2015-08-Understanding-LSTMs/>



Vers l'apprentissage de bout en bout

- **Supervision classique:** Un tâche au niveau des mots ou des



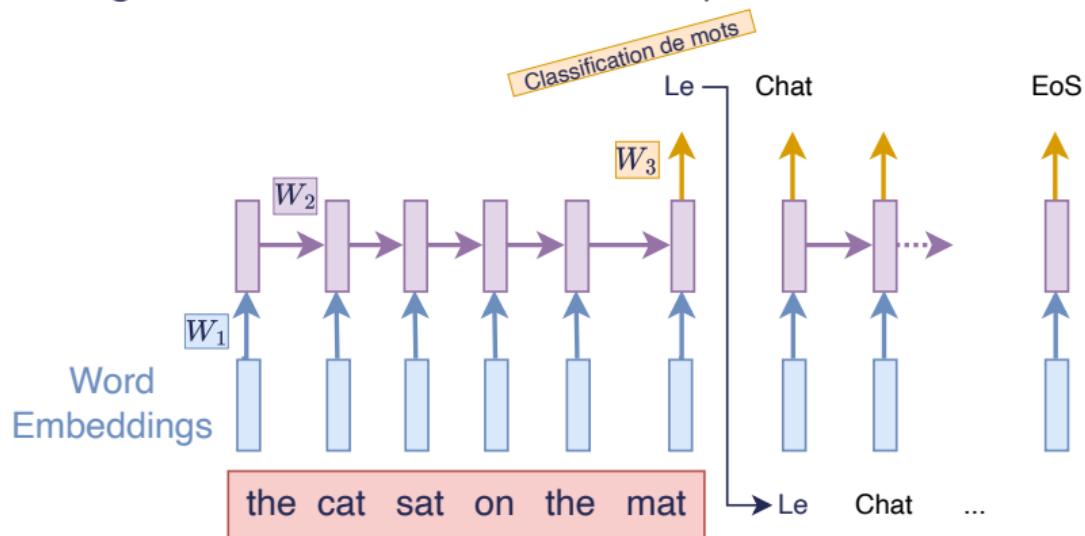
Bengio et al., JMLR, 2003
A Neural probabilistic language model



Sutskever et al., NeurIPS, 2014
Sequence to Sequence Learning with Neural Networks

Vers l'apprentissage de bout en bout

- **Supervision classique:** Un tâche au niveau des mots ou des
- **End-2-end:** Supervision de bout en bout (en génération)
 - Nouvelle génération de traducteurs automatiques



Bengio et al., JMLR, 2003

A Neural probabilistic language model



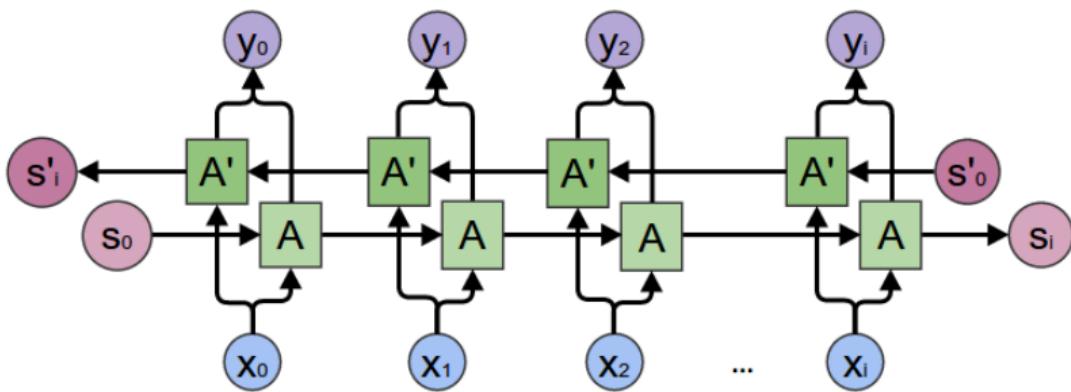
Sutskever et al., NeurIPS, 2014

Sequence to Sequence Learning with Neural Networks

Limites & amélioration des RNN : Bi-LSTM [SOTA 2019]

LSTM

- + Sequential modeling
- Sequential dependencies ! = partial modeling



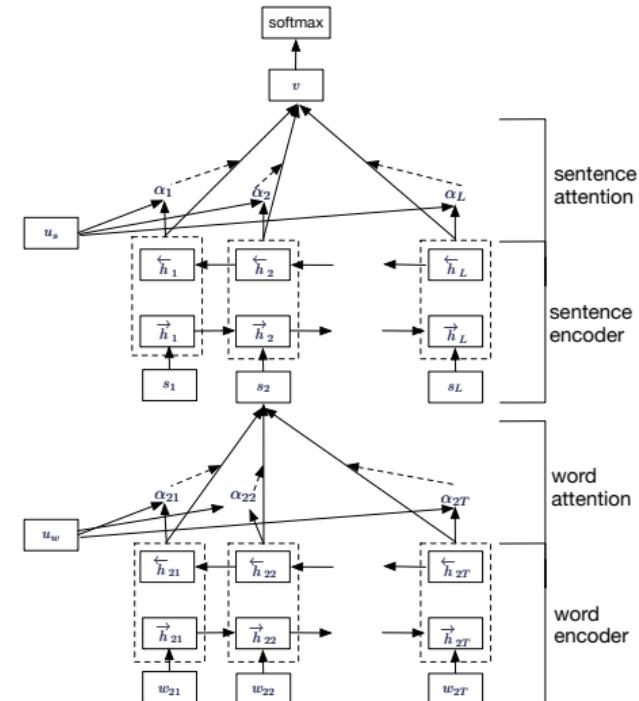
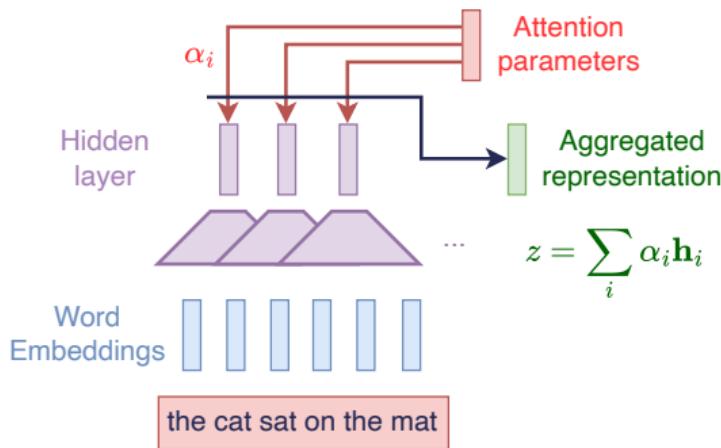
Bi-dimensional representation $[S_1, S'_1]$ is more powerful representation of the sentence S than each single representation.

Classical notation: $\mathbf{s} = [\overrightarrow{\mathbf{s}}, \overleftarrow{\mathbf{s}}]$



Agrégation & attention (+ hierarchie)

Principe (Sur CNN ou RNN):



- Agrégation = sélection d'information = attention

$$\alpha_i = \frac{\exp(\mathbf{w}_i \cdot h_i)}{\sum_j \exp(\mathbf{w}_j \cdot h_j)} = \text{softmax}$$



Yang et al., NAACL-HLT 2016
Hierarchical Attention Networks for Document Classification

Agrégation & attention (+ hierarchie)

Principe (Sur CNN ou RNN):

- Agrégation = sélection
d'information = attention

$$\alpha_i = \frac{\exp(\mathbf{w}_i \cdot h_i)}{\sum_j \exp(\mathbf{w}_j \cdot h_j)} = \text{softmax}$$

GT: 4 Prediction: 4

pork belly = delicious .
scallops ?
i do n't .
even .
like .
scallops , and these were a-m-a-z-i-n-g .
fun and tasty cocktails .
next time i 'm in phoenix , i will go
back here .
highly recommend .



Yang et al., NAACL-HLT 2016

Hierarchical Attention Networks for Document Classification



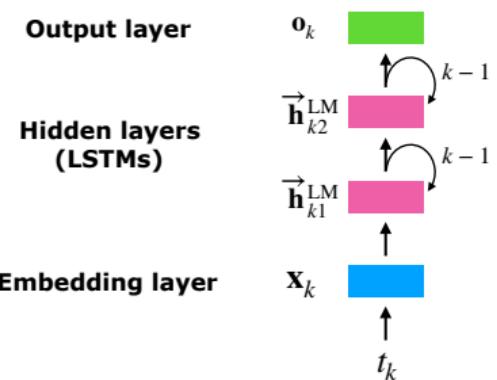
ELMo : Deep contextualized word representations

Static word embeddings

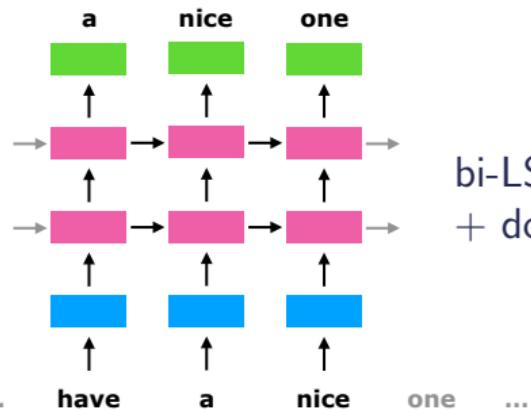


Mapping function = dynamic embeddings

The forward LM architecture



Expanded in the forward direction of k



bi-LSTM architecture
+ double hidden layer



M. E. Peters et al., arXiv 2018
Deep contextualized word representations

ELMo : Deep contextualized word representations

Static word embeddings



Mapping function = dynamic embeddings

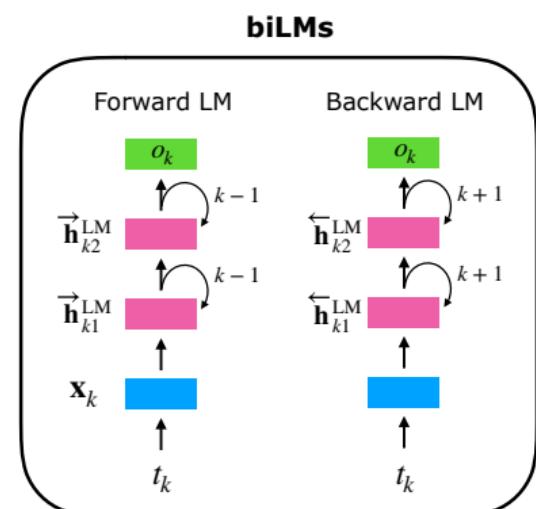


ELMo is a task specific representation. A down-stream task learns weighting parameters

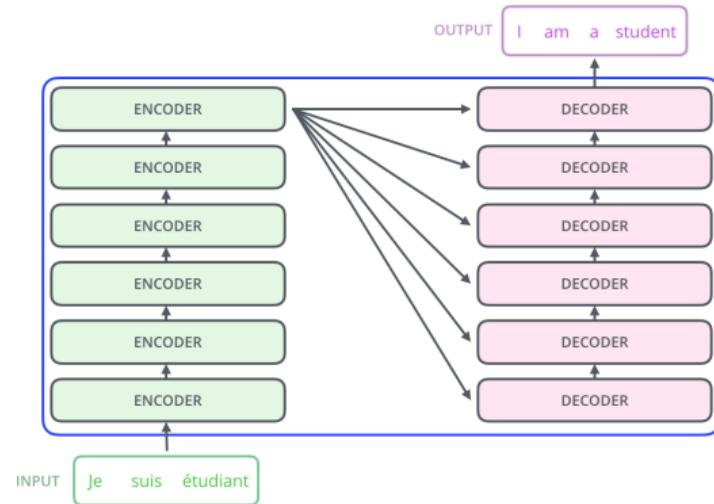
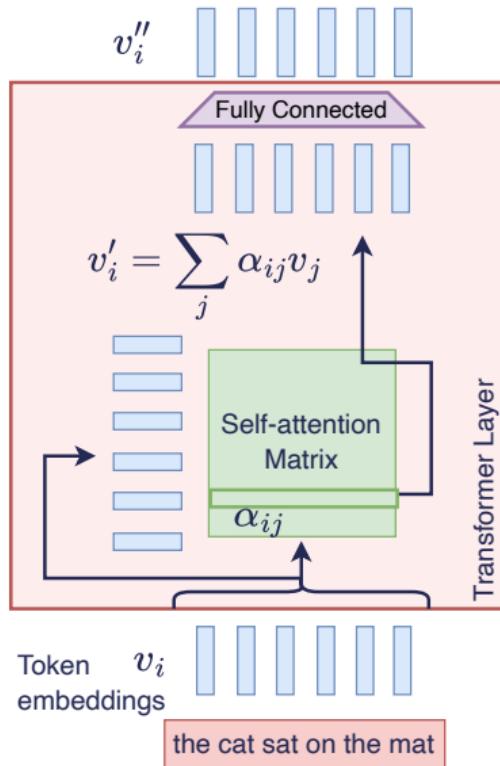
$$\text{ELMo}_{k, \text{task}} = \gamma^{\text{task}} \times \sum \left\{ \begin{array}{l} s_2^{\text{task}} \times \mathbf{h}_{k2}^{\text{LM}} \\ s_1^{\text{task}} \times \mathbf{h}_{k1}^{\text{LM}} \\ s_0^{\text{task}} \times \mathbf{h}_{k0}^{\text{LM}} \\ ([\mathbf{x}_k; \mathbf{x}_k]) \end{array} \right. \xrightarrow{\text{Concatenate hidden layers}} [\mathbf{\vec{h}}_{kj}^{\text{LM}}; \mathbf{\hat{h}}_{kj}^{\text{LM}}]$$

Unlike usual word embeddings, ELMo is assigned to every *token* instead of a type

ELMo represents a word t_k as a linear combination of corresponding hidden layers (inc. its embedding)



Transformer: removing the recurrent architecture



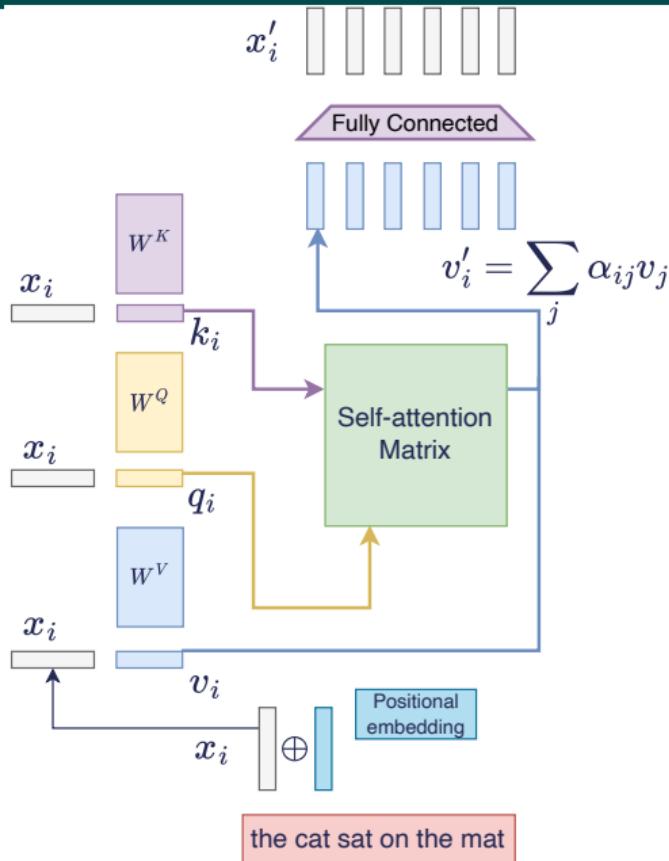
- 5 to **12** layers
- 5 to **12** heads
- No more computational bottleneck

Vaswani et al., NIPS 2017
Attention Is All You Need

Devlin et al., arXiv 2018
BERT: Pre-training of Deep Bidirectional Trans-
formers for Language Understanding

Jay Alammar, Blog 2018
[http://jalammar.github.io/
illustrated-transformer/](http://jalammar.github.io/illustrated-transformer/) 30/45

Transformer: removing the recurrent architecture



Paramètres :

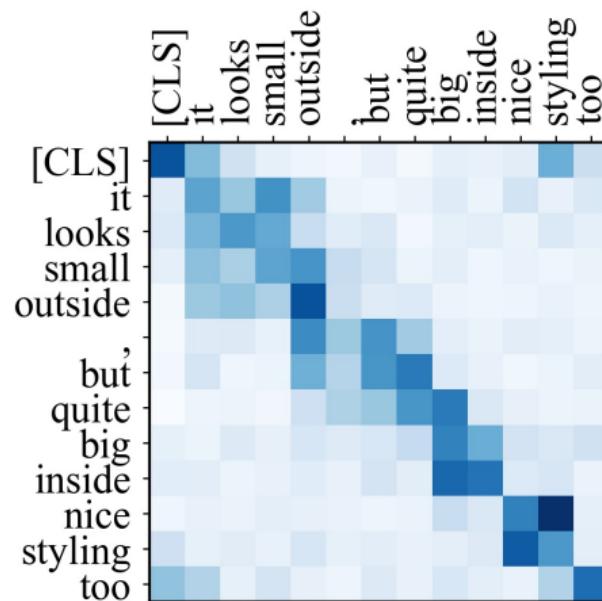
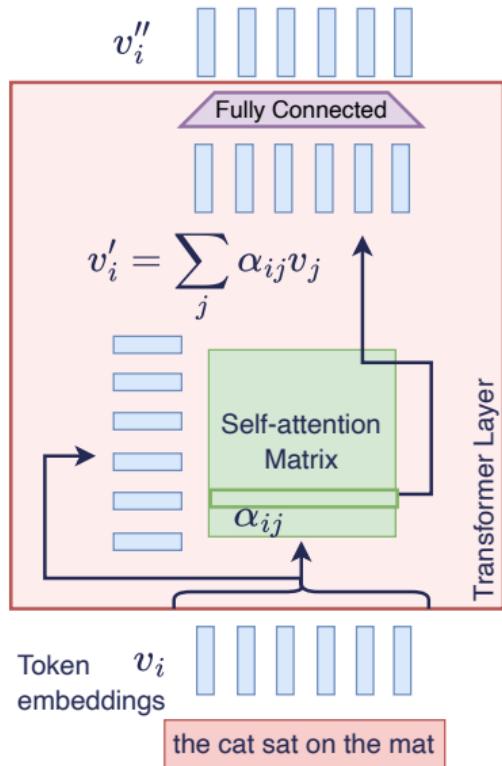
- Token = sub-word ($\sim 30k$)
- Embeddings (dim=500-1000)
- Projections (W^Q, W^K, W^V)
 \Rightarrow dim= $\infty 64 \times 12$ heads
- Positional embeddings
- Fully connected

Self-attention :

lignes normalisées en soft-max

$$\alpha_{ij} = \frac{\exp(k_i \cdot q_j)}{\sum_{j'} \exp(k_i \cdot q_{j'})}$$

Transformer: removing the recurrent architecture

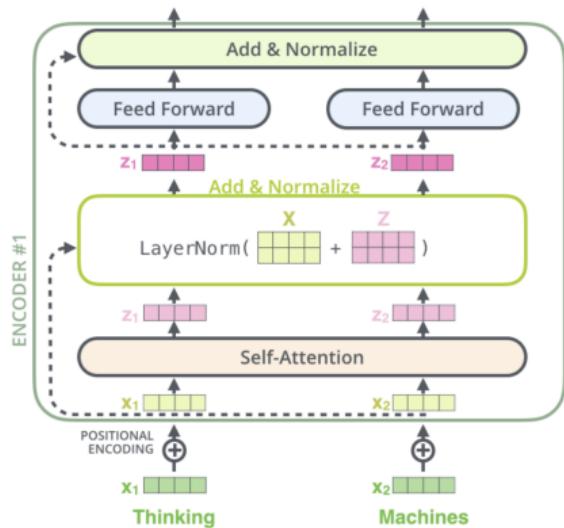


Vaswani et al., NIPS 2017
Attention Is All You Need

Devlin et al., arXiv 2018
BERT: Pre-training of Deep Bidirectional Trans-
formers for Language Understanding

Jay Alammar, Blog 2018
[http://jalammar.github.io/
illustrated-transformer/](http://jalammar.github.io/illustrated-transformer/) 30/45

Transformer: removing the recurrent architecture



■ skip-connection = conserver l'information centrale + stabiliser le gradient



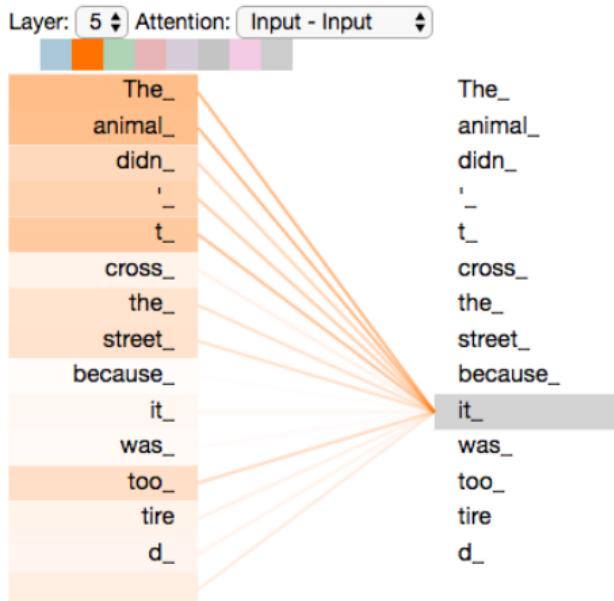
Vaswani et al., NIPS 2017
Attention Is All You Need

Devlin et al., arXiv 2018

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Jay Alammar, Blog 2018
<http://jalammar.github.io/illustrated-transformer/>

Transformer: removing the recurrent architecture



Vaswani et al., NIPS 2017
Attention Is All You Need

Devlin et al., arXiv 2018

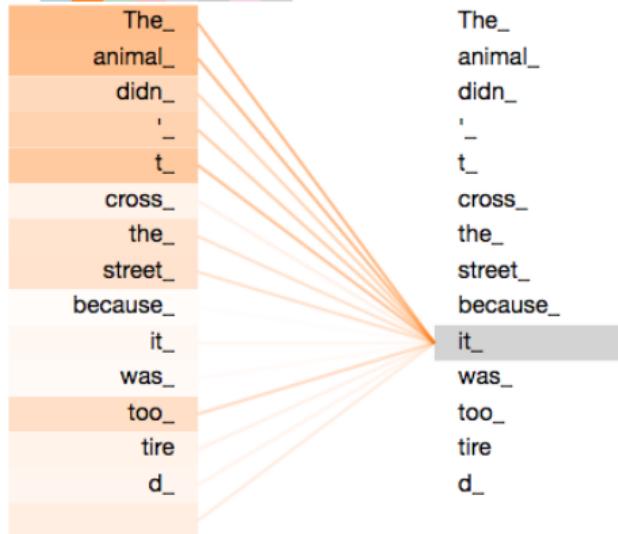
BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Jay Alammar, Blog 2018

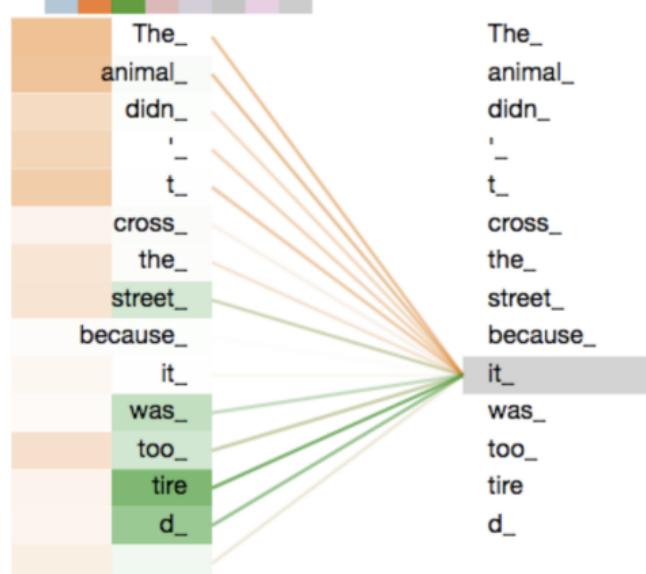
<http://jalammar.github.io/illustrated-transformer/>

Transformer: removing the recurrent architecture

Layer: 5 ⚡ Attention: Input - Input ⚡



Layer: 5 ⚡ Attention: Input - Input ⚡



Vaswani et al., NIPS 2017
Attention Is All You Need

Devlin et al., arXiv 2018

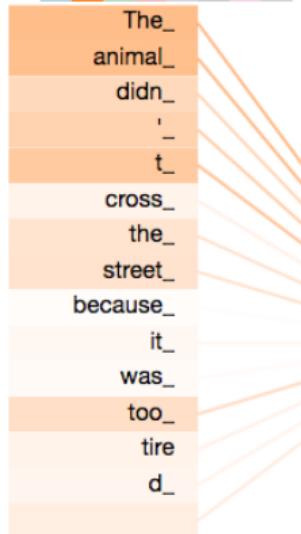
BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Jay Alammar, Blog 2018

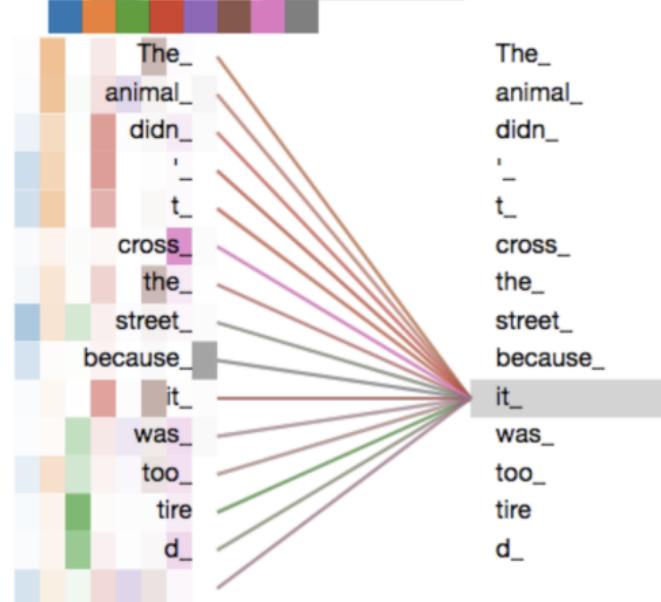
<http://jalammar.github.io/illustrated-transformer/>

Transformer: removing the recurrent architecture

Layer: 5 ⚡ Attention: Input - Input ⚡



Layer: 5 ⚡ Attention: Input - Input ⚡



Vaswani et al., NIPS 2017
Attention Is All You Need

Devlin et al., arXiv 2018

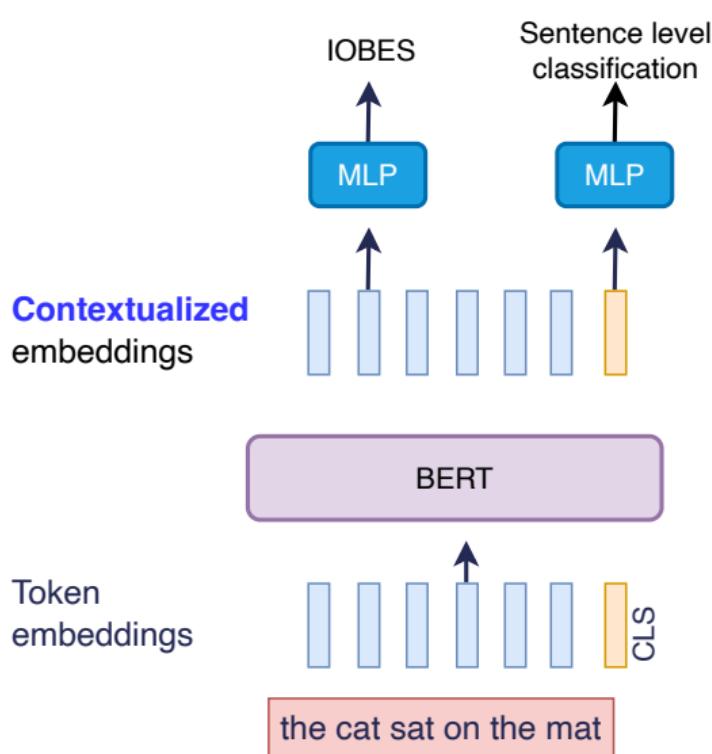
BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Jay Alammar, Blog 2018

<http://jalammar.github.io/illustrated-transformer/>



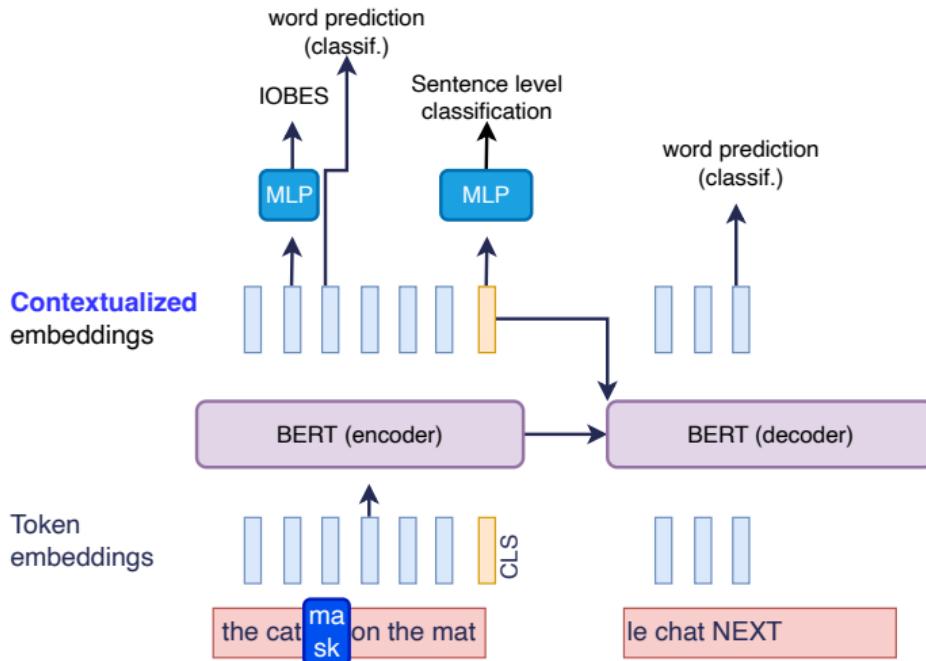
BERT Implementation & use



Words	BIOES Label
Jane	B-PER
Villanueva	E-PER
of	O
United	B-ORG
Airlines	I-ORG
Holding	E-ORG
discussed	O
the	O
Chicago	S-LOC
route	O
.	O

- Nouvel état de l'art en NER
- Efficace sur l'extraction de relation
- End-to-end Relation Extraction
- Traduction automatique
- Question Answering ...

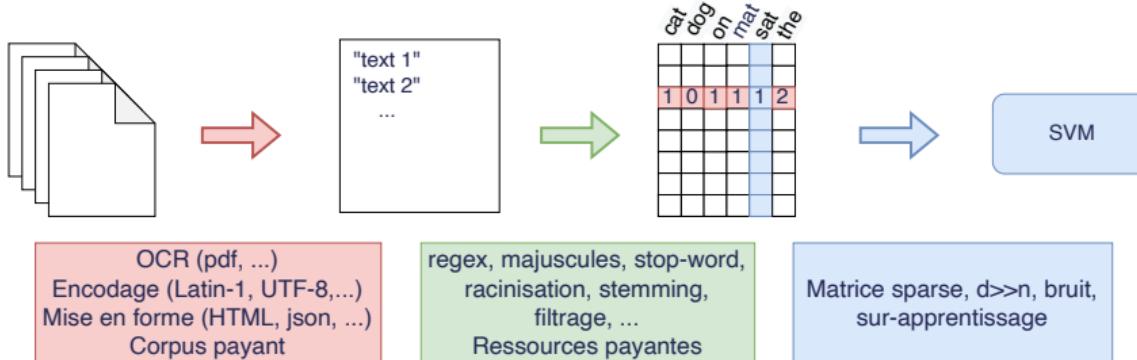
Apprentissage sur des tâches multiples



- Très performant en texte...
- ... En image
- ... En analyse de séries temporelles

- Problème de complexité:
une couche = complexité en $\mathcal{O}(N^2)$

Bilan



- + Corpus largement disponibles
- + Ressources linguistiques difficiles d'accès desulettes
- + Quasiment plus de pré-traitements [byte pair encoding, dico ~ 30k]
- + Plus de matrices sparses
- + Modèles pré-entraînés directement efficace sur de nombreuses tâches
- + Robuste aux nouveaux mots
- + Sur-apprentissage limité par la masse de données
- Puissance & mémoire requises
- Etat de l'art en mouvement perpetuel

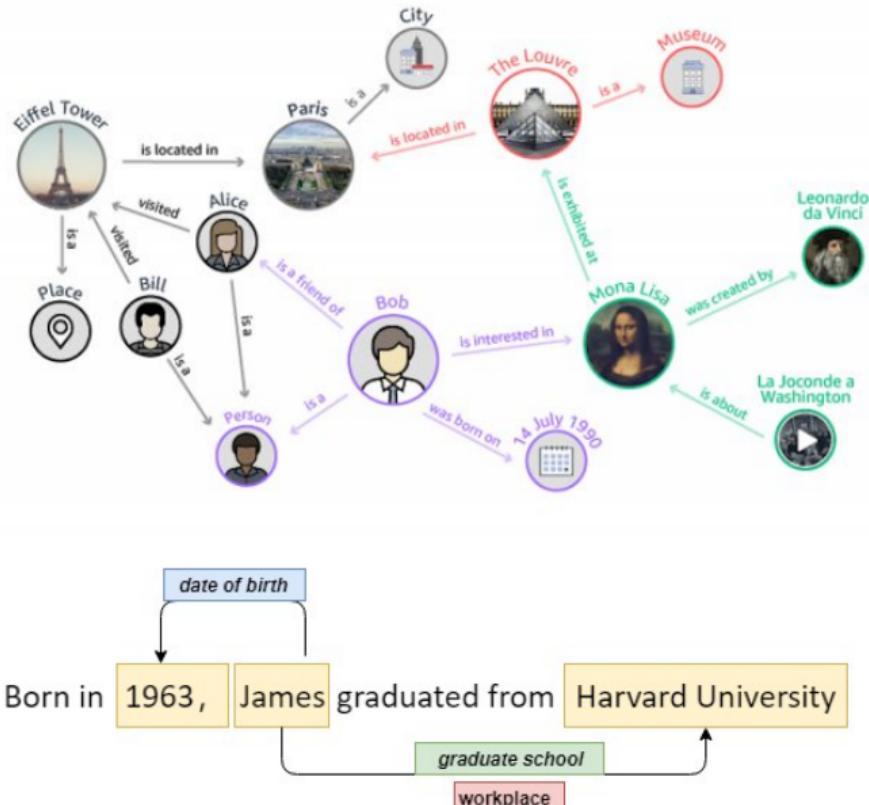
EVOLUTION DES TECHNIQUES DE NLP VERS LA CONNAISSANCE ET LE RAISONNEMENT

Knowledge Base building = Information Extraction

- Base exploitable par les ordinateurs
 - RDF / SparQL
 - Google Knowledge Graph

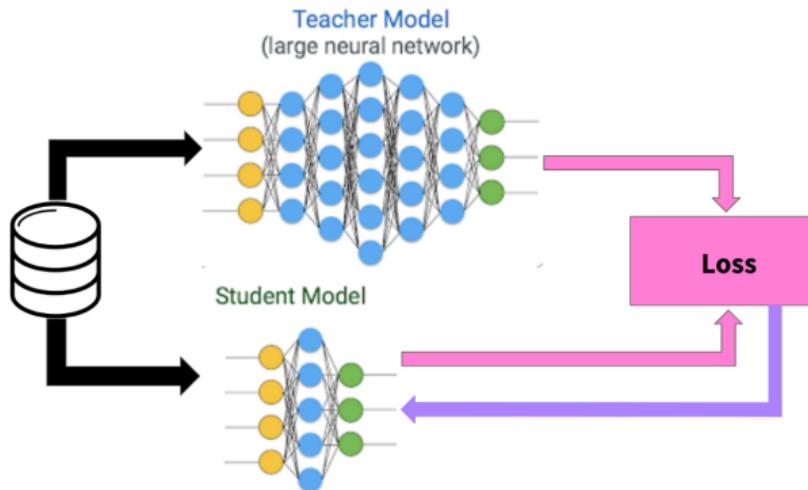
Idée:

- Construire une base exploitable automatiquement
- A partir de textes bruts
⇒ NER, Relation Extr., End to End Relation Extr.



Precision/recall ⇒ un débat lié aux appli.

- Tom Mitchell, NELL (Never Ending Language Learning)
⇒ ++ Precision (utiliser les informations extraites comme vérité terrain)
- Distillation
 - Utilisation des étiquettes prédites comme vérité terrain
 - Risque d'effondrement (archi. teacher-student)
- Indexation de données juridiques
 - ++ Rappel (interdiction de rater une jurisprudence)





KB & Language Model

Modèle de langue ≈ compréhension et complément...

Exemples avec GPT:

A rabbit is sitting on |

the ledge and you know it's not going to go down unless you

a box.

the grass, holding a flower in his mouth.



Roberts et al., arXiv 2020

How much knowledge can you pack into the parameters of a language model?



KB & Language Model

Modèle de langue ≈ compréhension et complément...

Exemples avec GPT:

JFK died in 1963. JFK was shot dead in Dallas, TX on 5 November 1963.

John Fitzgerald Kennedy had been president of the United

The event was called "A day of reckoning

President John F Kennedy had been assassinated by Lee



Roberts et al., arXiv 2020

How much knowledge can you pack into the parameters of a language model?

Language Model & hallucination

Mais évidemment, tout n'est pas présent... Et surtout, on manque de confiance.

Charles De Gaulle , who succeeded Churchill as Prime Minister in

1940.

1940, was not a man

France, who was a passionate

- Question de complexité? de contrôle? de contexte?

KB vs Question Answering

[SQuAD v2, 2018]

Les modèles de langue savent répondre aux questions... Dans un certain contexte :)

Passage Sentence

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity.

Question

What causes precipitation to fall?

- Est-il possible/facile d'extraire des passages avec un moteur de recherche?
- Est-ce une alternative aux KB?

Answer Candidate

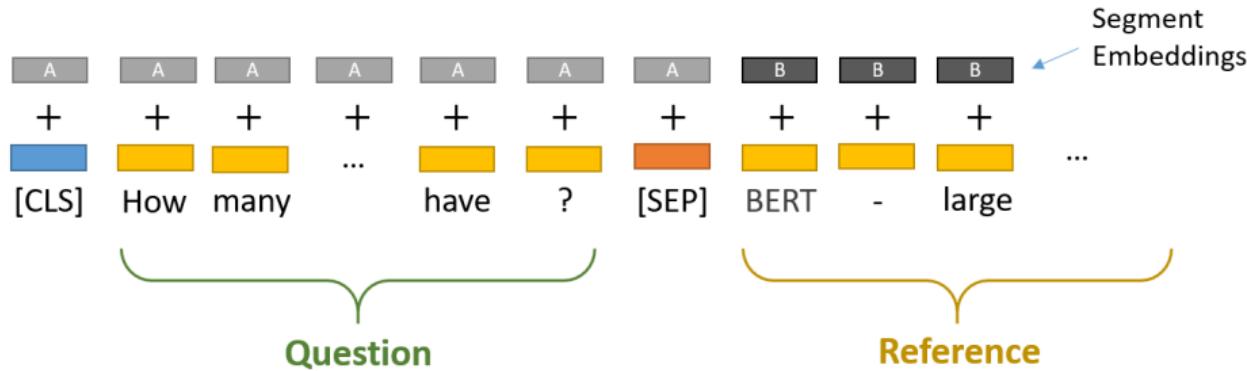
gravity



Rajpurkar et al., ACL 2018

Know What You Don't Know: Unanswerable Questions for SQuAD

QA: une approche par classification



Question: How many parameters does BERT-large have?

Reference Text: BERT-large is really big... it has 24 layers and an embedding size of 1,024, for a total of 340M parameters! Altogether it is 1.34GB, so expect it to take a couple minutes to download to your Colab instance.

- Classification de *span* sur la référence
- Intérêt de l'attention pour la tâche

QA: une approche par classification

Des problèmes de plus en plus difficiles, impliquant du raisonnement

Multi-hop QA: mixer plusieurs proposition pour répondre:

Where was Facebook launched? (A) Cambridge (B) Silicon Valley

H_c : Facebook was launched in Cambridge.

P1: Facebook was launched at Harvard University.

P2: Facebook headquarters was set up in Silicon Valley.

P3: Harvard University is at Cambridge, Massachusetts.

P4: Harvard is only a few miles from Boston.

Relevance
Model
→
Sentence-wise

P1: 0.4
P2: 0.1
P3: 0.4
P4: 0.1



QA: une approche par classification

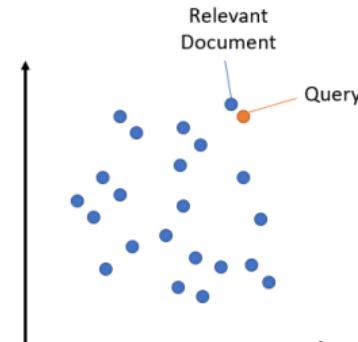
Reasoning	Passage (some parts shortened)	Question	Answer	BiDAF
Subtraction (28.8%)	That year, his Untitled (1981) , a painting of a haloed, black-headed man with a bright red skeletal body, depicted amid the artists signature scrawls, was sold by Robert Lehrman for \$16.3 million, well above its \$12 million high estimate.	How many more dollars was the Untitled (1981) painting sold for than the 12 million dollar estimation?	4300000	\$16.3 million
Comparison (18.2%)	In 1517, the seventeen-year-old King sailed to Castile . There, his Flemish court In May 1518, Charles traveled to Barcelona in Aragon.	Where did Charles travel to first, Castile or Barcelona?	Castile	Aragon
Selection (19.4%)	In 1970, to commemorate the 100th anniversary of the founding of Baldwin City, Baker University professor and playwright Don Mueller and Phyllis E. Braun, Business Manager, produced a musical play entitled The Ballad Of Black Jack to tell the story of the events that led up to the battle.	Who was the University professor that helped produce The Ballad Of Black Jack, Ivan Boyd or Don Mueller?	Don Mueller	Baker



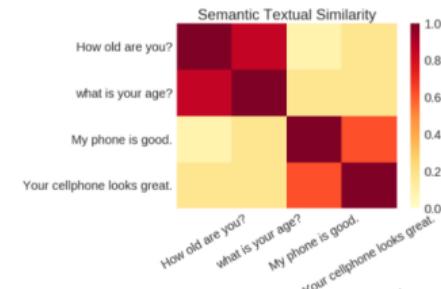
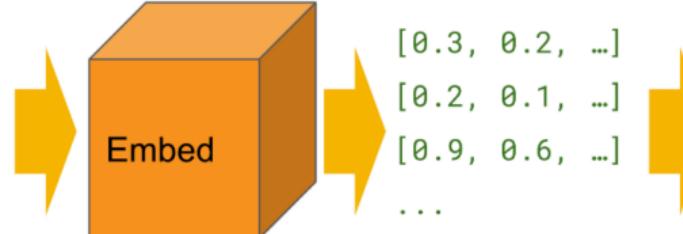
Passage à l'échelle, indexation

Comment passer à l'échelle? ⇒ Indexer les embeddings de phrases
 Index de représentations vectorielles continues + recherche de voisin

- Google indexe déjà les images à partir de représentation vectorielles continues
- Elastic search a un module *dense-vector*



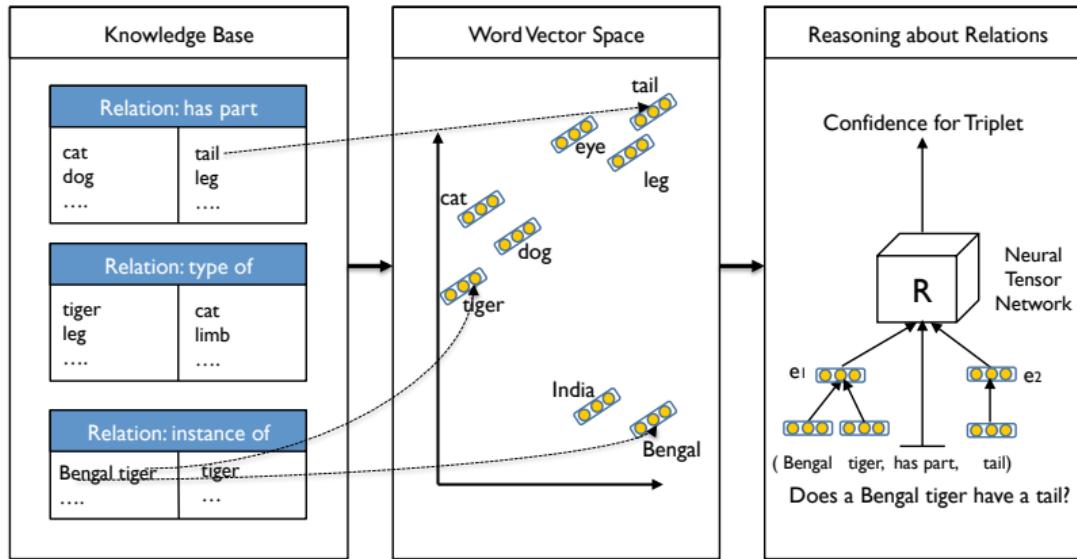
"How old are you?"
 "What is your age?"
 "My phone is good."
 ...





KB Extension

Algo. de representation learning = outil pour la complétiōn de KB



Bordes et al., AAAI, 2011

Learning structured embeddings of knowledge bases



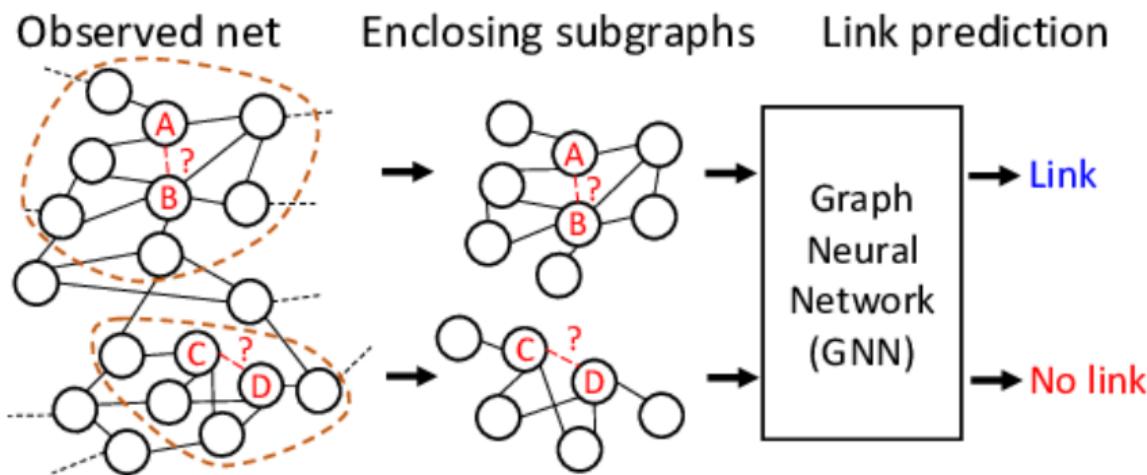
Socher et al., NeurIPS, 2013

Reasoning With Neural Tensor Networks for Knowledge Base Completion

KB Extension

Algo. de representation learning = outil pour la complétiōn de KB

Exploitation des GCNN: convolution sur des graphes



Les Graph-CNN sur la *knowledge completion* font-ils du Multi-hop QA?



NGuyen et al., NAACL-HLT, 2018

A Novel Embedding Model for Knowledge Base Completion Based on Convolutional Neural Network

MODÈLES GÉNÉRATIFS: MODE OU TENDANCE DURABLE?

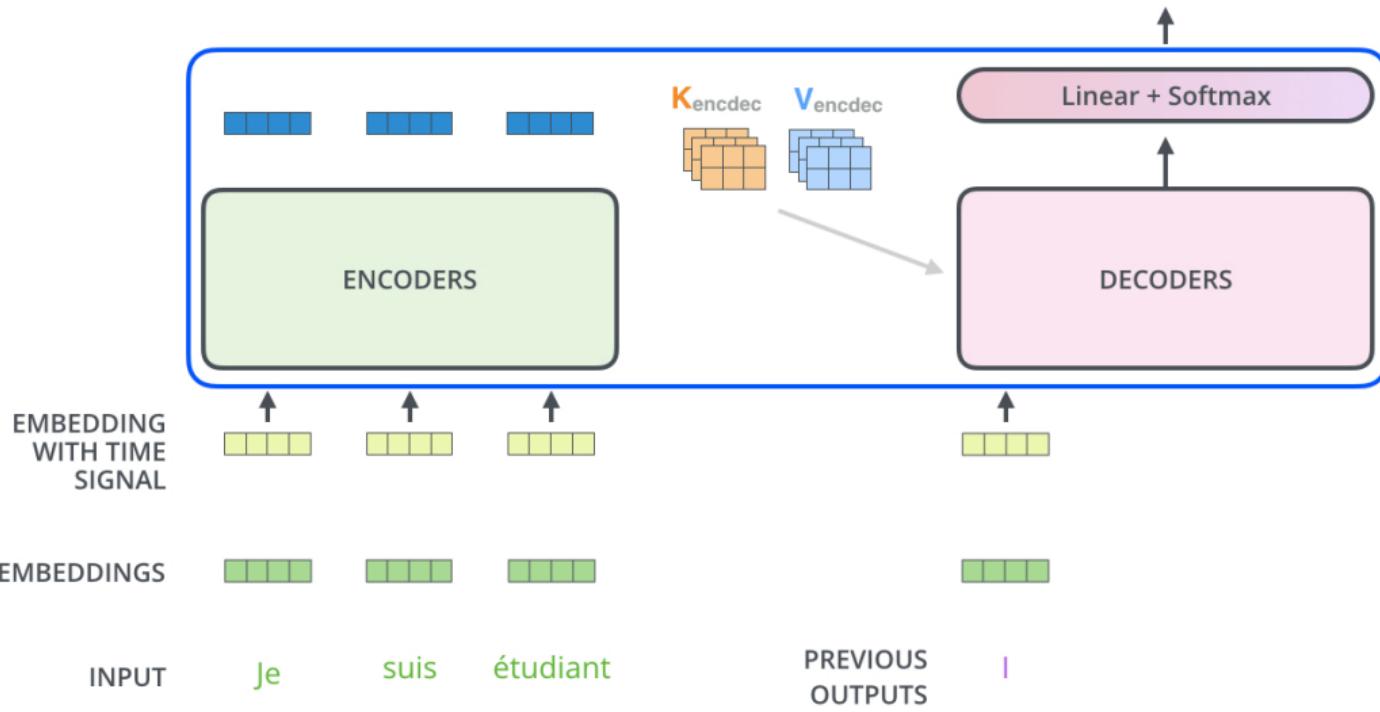


Coût des modèles génératifs

Decoding time step: 1 2 3 4 5 6

OUTPUT

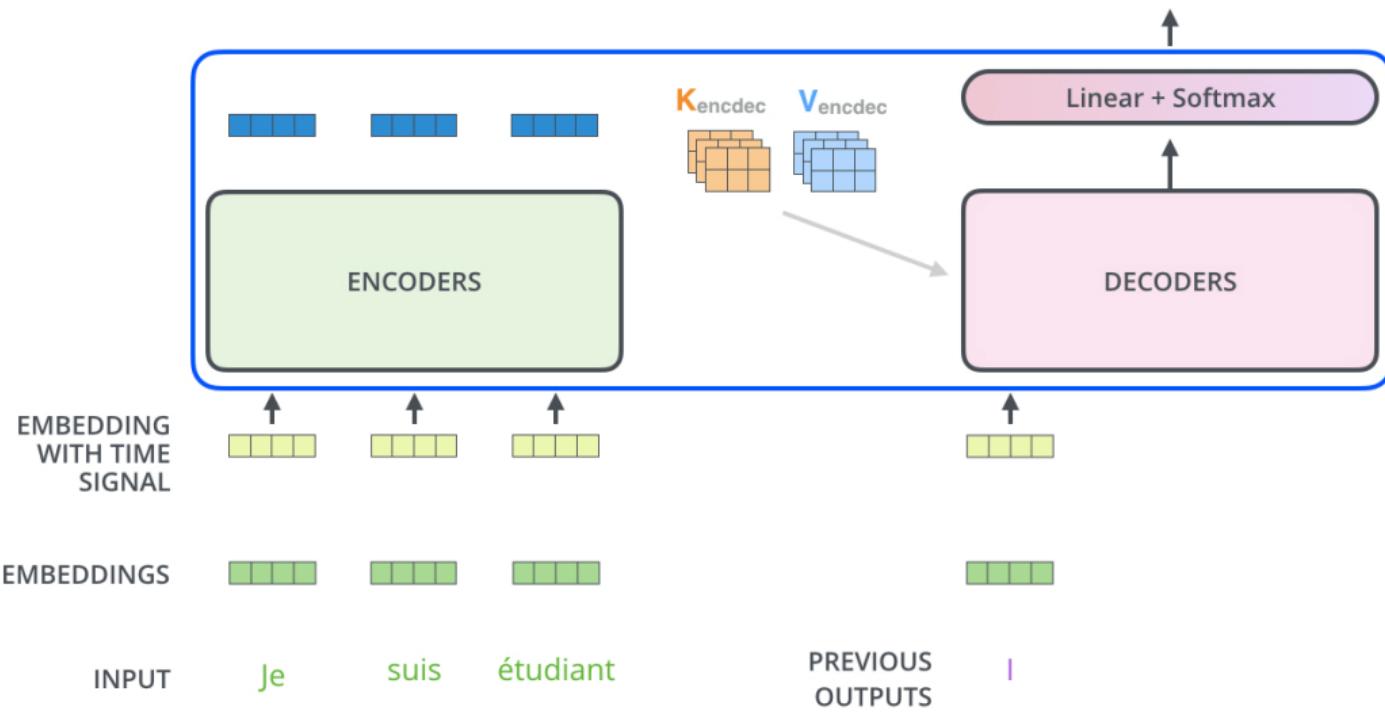
|



Coût des modèles génératifs

Decoding time step: 1 2 3 4 5 6

OUTPUT | am

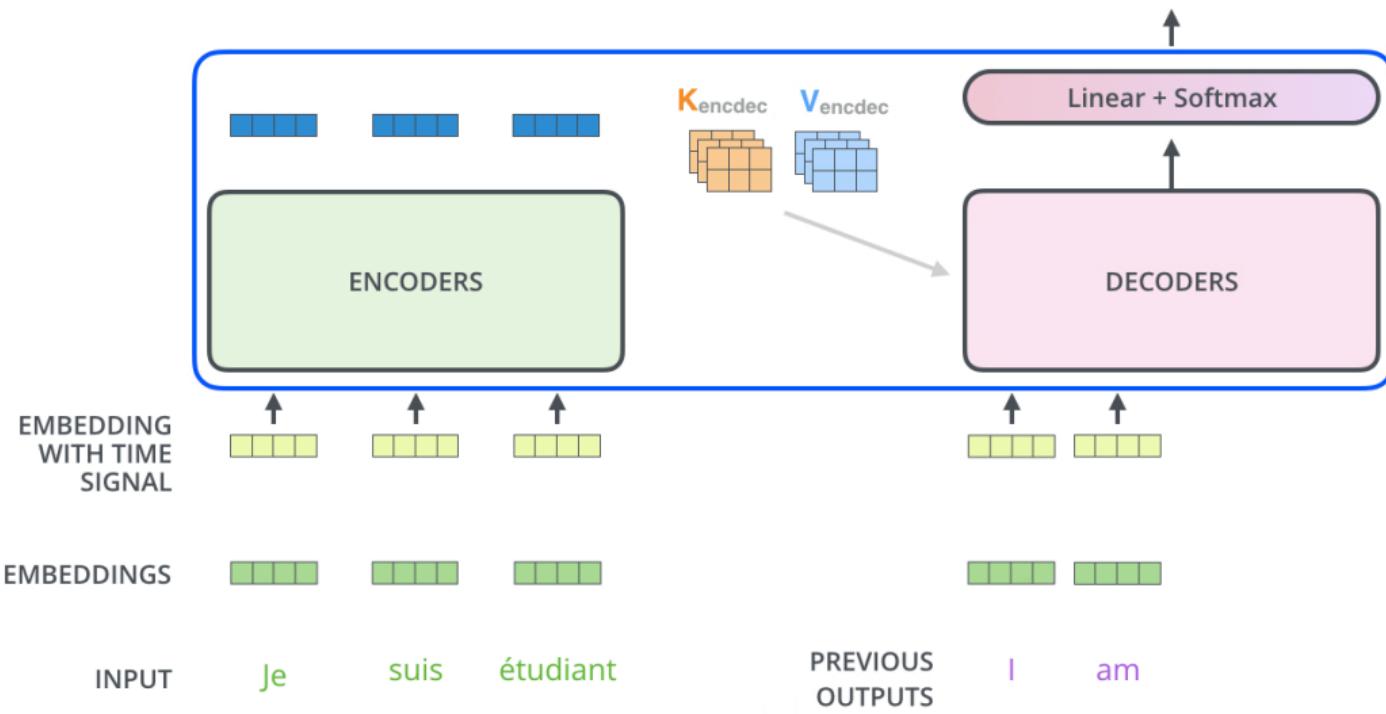


Coût des modèles génératifs

Decoding time step: 1 2 3 4 5 6

OUTPUT

I am a





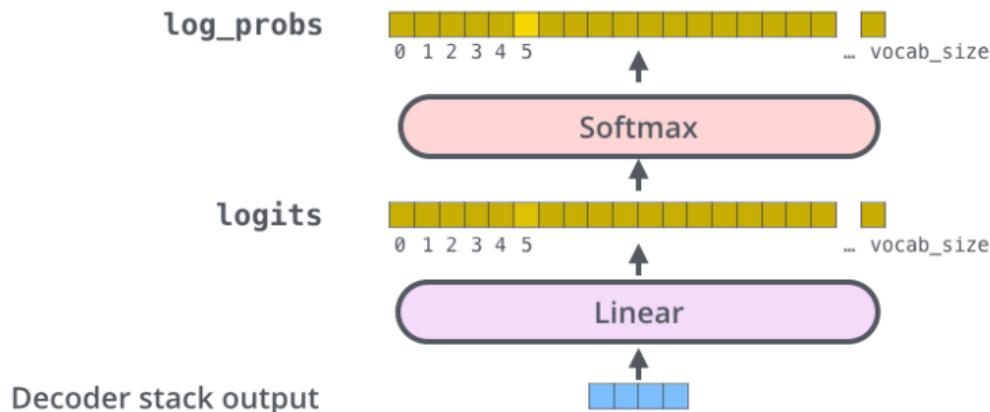
Coût des modèles génératifs

Which word in our vocabulary
is associated with this index?

am

Get the index of the cell
with the highest value
(argmax)

5

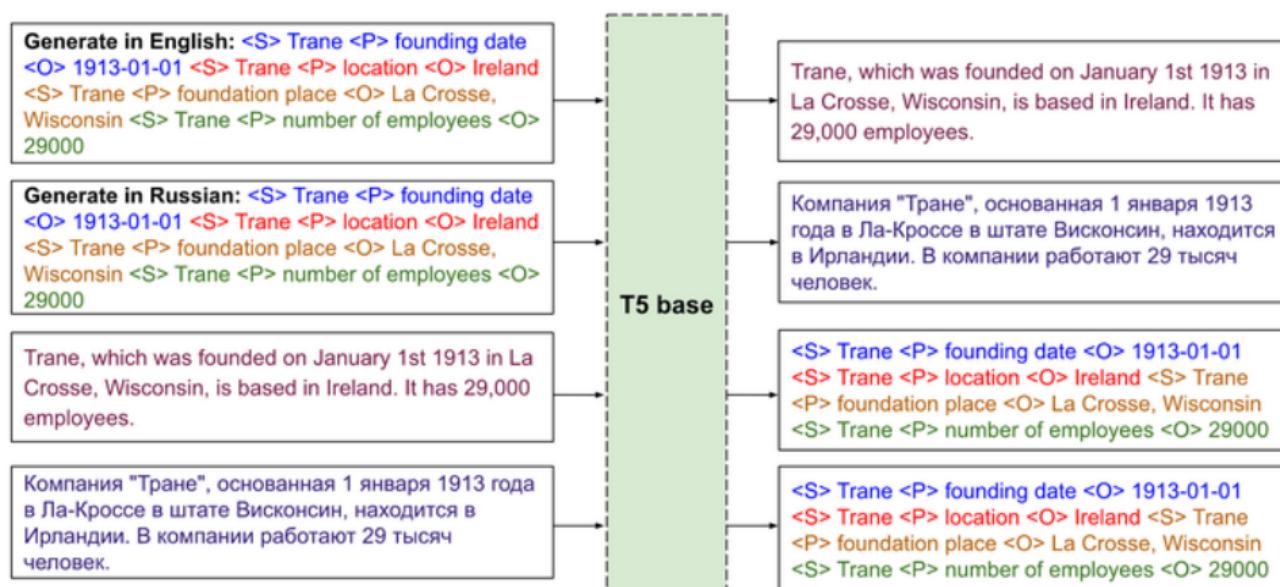


+ beam search pour certains algos



Modèle Génératifs

Tendance depuis GPT (OpenAI, 2018):
redéfinir les tâches NLP en génération de texte (Modèle *prompt*)



Gestion des connaissances = traduction humain/machine?



Modèle Génératifs

Corpus Web-NLG

Triplets:

- (Alan Bean, nationality, United States)
- (Alan Bean, birthDate, 1932-03-15)
- (Alan Bean, almaMater, UT Austin, B.S. 1955)
- (Alan Bean, birthPlace, Wheeler, Texas)
- (Alan Bean, selection, 1963)

Text:

Alan Bean was an American astronaut, born on March 15, 1932 in Wheeler, Texas. He received a Bachelor of Science degree at the University of Texas at Austin in 1955 and was chosen by NASA in 1963.



Gardent et al., ACL 2017
Creating Training Corpora for NLG Micro-Planners

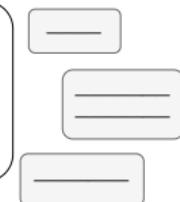


Guo et al., INLG 2020
CycleGT: Unsupervised Graph-to-Text and Text-to-Graph Generation via Cycle Training

Apprentissage cyclique
(pour des données alignées... Ou pas)

Text Corpus (No Matched Graph)

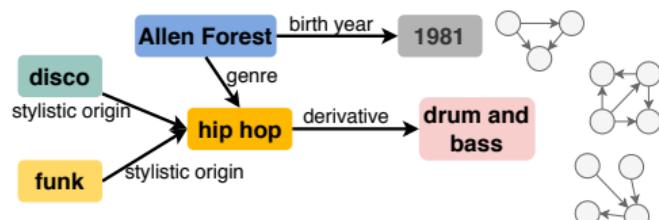
Allen Forest, a **hip hop** musician, was born in the year **1981**. The music genre **hip hop** gets its origins from **disco** and **funk** music, and it is also which **drum and bass** is derived from.



CycleGT



Graph Dataset (No Matched Text)



QA en approche générative & raisonnement

Task: Basic Math

Problem: Before December, customers buy 1346 ear muffs from the mall. During December, they buy 6444, and there are none. In all, how many ear muffs do the customers buy?

Predicted Answer: 1346.0 X

Generated Program:

```
answer = 1346.0 + 6444.0
print(answer)
# Result ==> 7790.0
```

Gold Answer: 7790.0 ✓

Task: Muldiv

Problem: Tickets to the school play cost 6 for students and 8 for adults. If 20 students and 12 adults bought tickets, how many dollars' worth of tickets were sold?

Predicted Answer: 48 X

Generated Program:

```
a=20*6
b=12*8
c=a+b
answer=c
print(answer)
# Result ==> 216.0
```

Gold Answer: 216 ✓



Mishra et al., arXiv 2022

Lila: A Unified Benchmark for Mathematical Reasoning

⇒ On cherche déjà à faire plus que de l'exploitation de connaissances.

Résumé temporel

80-90' Corpus à accès restreint +
Ressources spécifiques +
savoir faire particulier

2000' Montée du ML +
accès aux outils et ressources

2015' Changement complet de modèles (deep) + ressources (modèle pré-entraîné)
+ savoir-faire (descripteur=mot)

2020' Changement de modèle (RNN⇒Transformer) +
accès ressources pour tous (HuggingFace) +
descripteur de plus en plus simples (lettres)

- Prix du ticket d'entrée = accès deep-learning
- Multiplication des tâches ⇒ amener des idées neuves
- Tendance: ~~NLP/nouvelle KB~~ ⇒ NLP contourne les KB