

Représentations sémantiques

Sahar Ghannay

sahar.ghannay@limsi.fr

Introduction

Recherche de sens : de quoi parle-t-on? Qu'est ce que le sens?

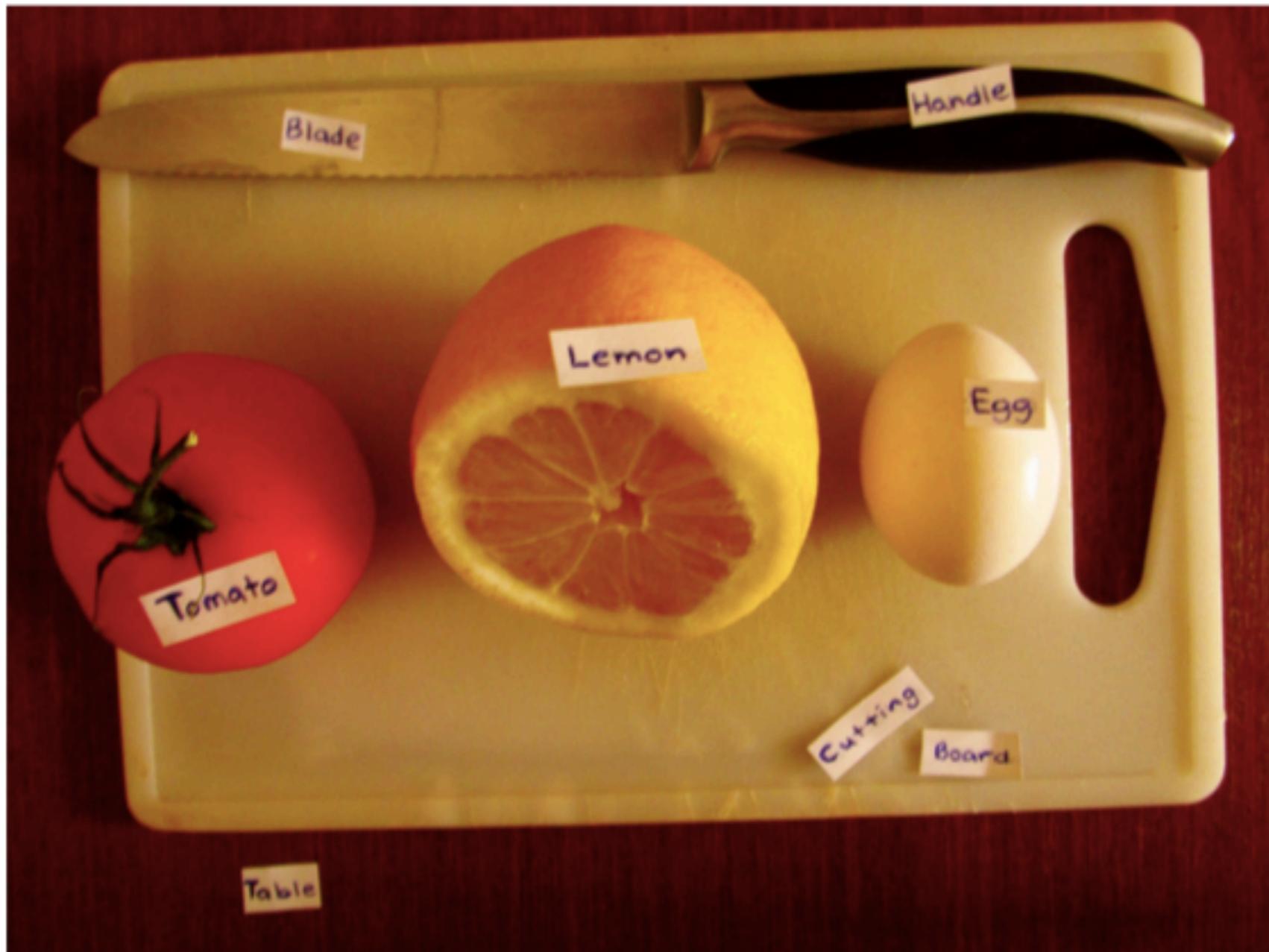
❖ Communément, d'après le Larousse :

- ♦ Ce que quelque chose signifie, ensemble d'idées que représente un signe, un symbole : *Le sens d'une allégorie*.
- ♦ Ce que représente un mot, objet ou état auquel il réfère : *Chercher le sens d'un mot dans le dictionnaire*.

❖ Pour une application informatique :

- ♦ La même chose, mais **restreinte aux concepts nécessaires** au fonctionnement de cette application

Recherche de sens : de quoi parle-t-on? Qu'est ce que le sens?



Analyse sémantique en technologie linguistique

♣ L'extraction de l'information

- ♦ consiste à analyser des textes pour en obtenir des informations en vue d'une application précise :
 - Exemple : la reconnaissance des entités nommées

♣ Compréhension du langage naturel

♣ Similarité sémantique textuelle

♣ *Etc.*

Analyse sémantique en technologie linguistique : Reconnaissance d'entités nommées

❖ La reconnaissance des entités nommées :

- ♦ détecter, repérer des entités nommées (noms de personnes, noms d'organisation, noms de lieux, dates, unités monétaires, ...) dans les flux textuels (on pose les frontières dans le texte)
- ♦ Catégoriser les éléments reconnus selon des catégories sémantiques pré-définies (person, organization, location, date,...)

Analyse sémantique en technologie linguistique : Reconnaissance d'entités nommées

❖ Exemple :

On the invitation of the Festival de Cannes, the Italian actress Monica Belucci has agreed to play the role of Mistress of the Opening and Closing Ceremonies of the 70th festival de Cannes to be held from 17 to 28 May 2017, under the presidency of Spanish filmmaker Pedro Almodovar. [...] Monica Bellucci's friendship with the Festival de Cannes goes back a long way: in 2000, she walked up the steps for the first time to present *Under Suspicion* by Stephen Hopkins.

Analyse sémantique en technologie linguistique : Reconnaissance d'entités nommées

❖ Exemple :

On the invitation of the **Festival de Cannes**, the Italian actress **Monica Belucci** has agreed to play the role of Mistress of the Opening and Closing Ceremonies of the 70th festival de **Cannes** to be held from 17 to 28 **May 2017**, under the presidency of Spanish filmmaker **Pedro Almodovar**. [...] **Monica Bellucci**'s friendship with the **Festival de Cannes** goes back a long way: in **2000**, she walked up the steps for the first time to present *Under Suspicion* by **Stephen Hopkins**.

PERSON, ORGANIZATION, LOCATION, DATE

Analyse sémantique en technologie linguistique

Compréhension du langage naturel

❖ Un système de compréhension du langage naturel (NLU) :

- ♦ peut être considéré comme une machine qui traduit une chaîne de mots en une ou plusieurs actions :
 - I. Associer les mots de la phrase en entrée du système à des messages dans un langage sémantique intermédiaire (souvent appelés concepts).
 - ♦ Un concept est une classe de mots traitant d'un même sujet et partageant des propriétés communes. Par exemple, les mots *hôtel*, *chambre*, *auberge* et *studio* peuvent tous correspondre au concept “hébergement” dans une application touristique.
 2. Traduit les concepts obtenus en actions ou réponses au cours d'une étape d'interprétation de la phrase pour répondre à la requête d'entrée

❖ Le NLU est un module important dans un système de dialogue

- ♦ Fournit une interaction homme-machine directe
- ♦ Permet à l'ordinateur de comprendre les langues humaines

Analyse sémantique en technologie linguistique

Compréhension du langage naturel

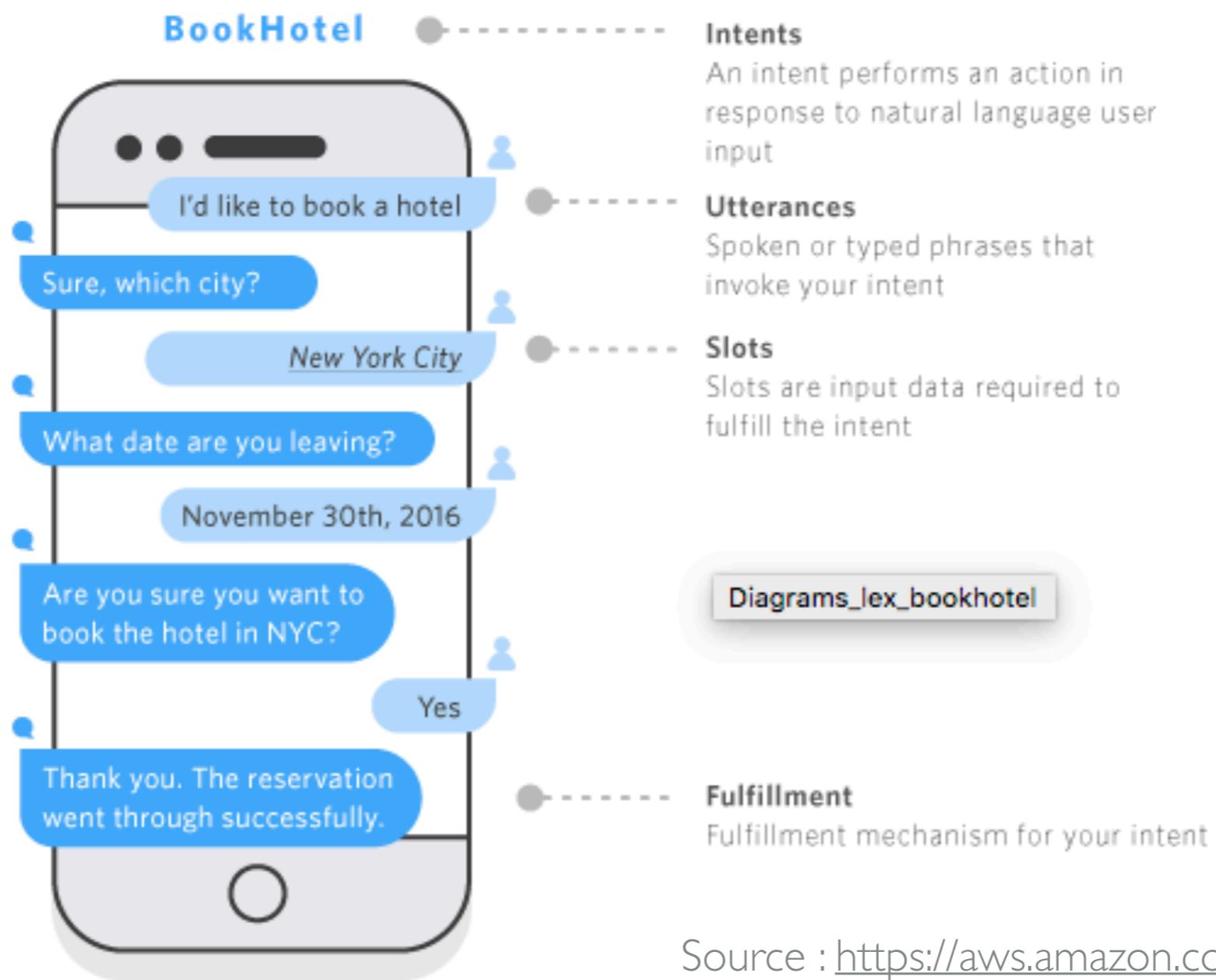
❖ Exemple :

Hyp	je	veux	réserver	une	chambre
Concept		commande		nombre	objet
Label	commande-B	commande-I	commande-I	nombre-B	objet-B
Valeur		réservation		I	chambre

Analyse sémantique en technologie linguistique

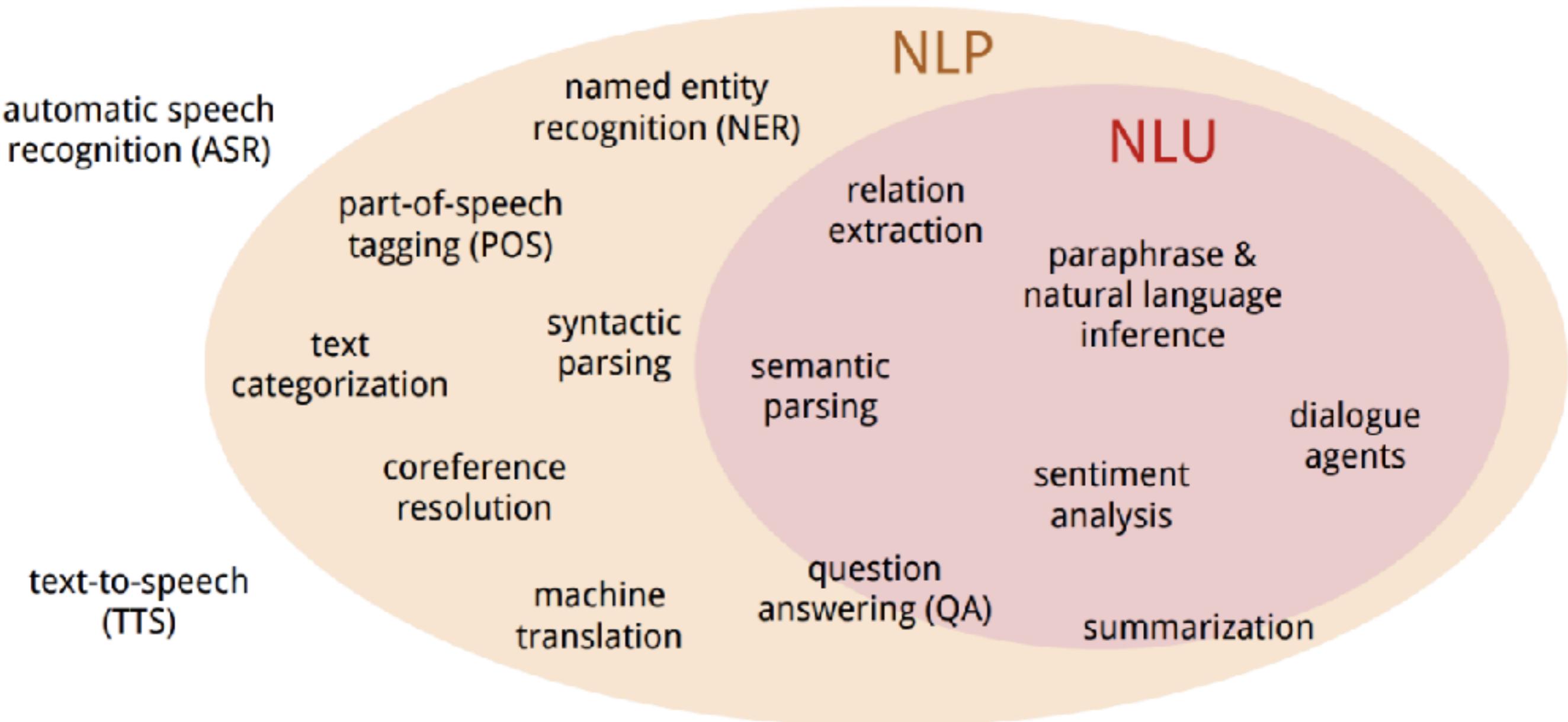
Compréhension du langage naturel

❖ Exemple : Amazon Lex



Source : <https://aws.amazon.com/fr/lex/details/>

Terminology: NLU vs. NLP vs. ASR



Analyse sémantique en technologie linguistique

Similarité sémantique textuel

- ❖ SemEval 2017 tâche 5 (en anglais)

Exemple de paire de phrases

Three men are playing chess.	Two men are playing chess.
A man is playing the cello.	A man seated is playing the cello.
Some men are fighting.	Two men are fighting.



- **Evaluer à quel point ces phrases sont proches sur une échelle de 1 à 5**

Approches récentes

❖ Les approches récentes pour :

- ◆ L'extraction de l'information
 - Détection d'entités nommées
- ◆ Compréhension de langue naturelle
- ◆ Similarité sémantique textuelle
- ◆ *Etc.*

→ s'appuient sur des techniques d'apprentissage profond.

Apprentissage profond (deep Learning)

Deep Learning

- ♣ Deep learning : famille d'algorithmes d'apprentissage automatique dont la stratégie est d'apprendre plusieurs niveaux de représentation pour la réalisation efficace d'une tâche
- ♣ en particulier : réseau de neurones (artificiels) profonds (DNN: Deep Neural Network)

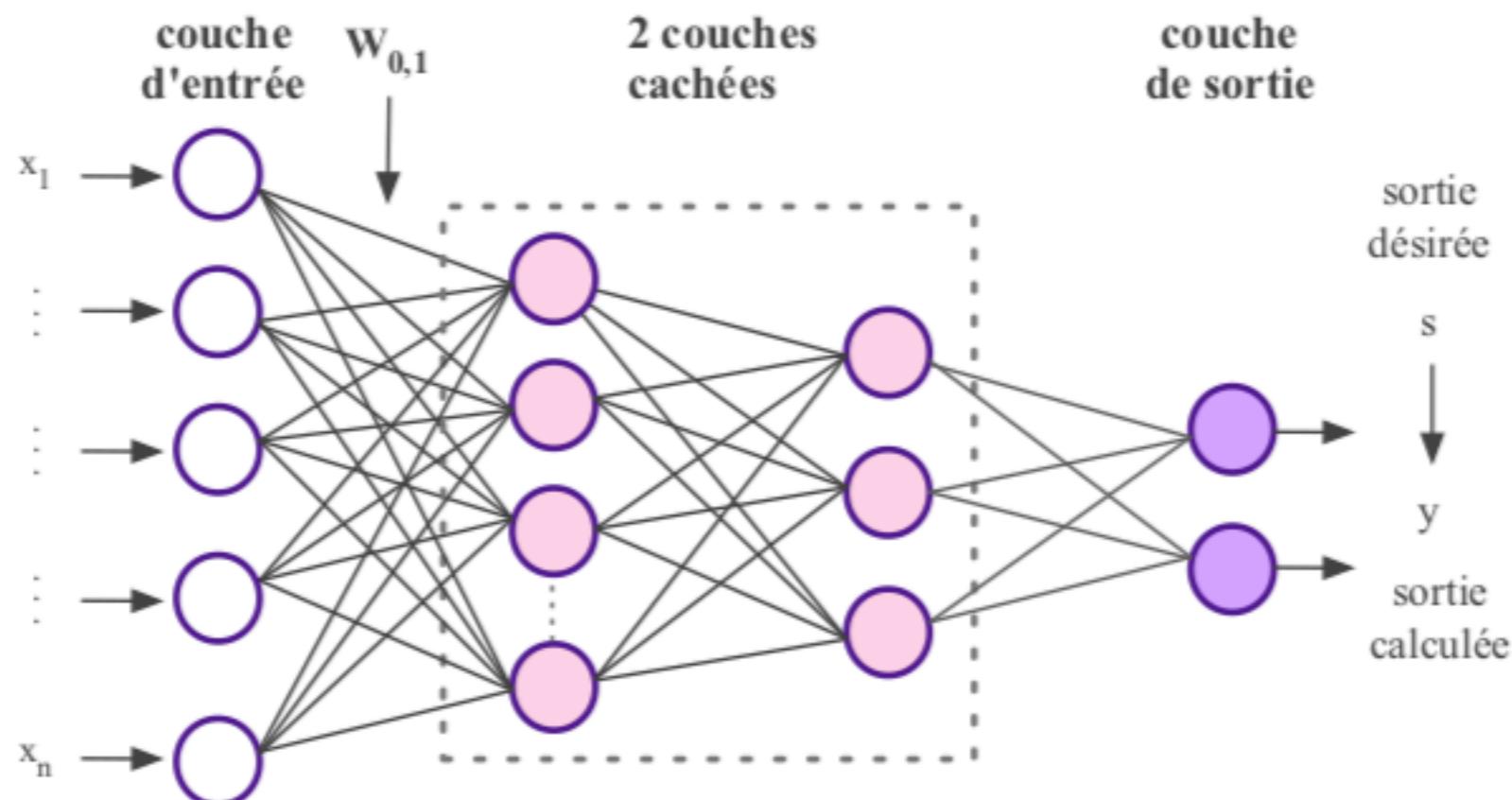
Apprentissage automatique

♣ Apprentissage automatique :

- ♦ Un sous-domaine de l'intelligence artificielle (IA)
- ♦ Se base sur des approches statistiques pour donner aux [ordinateurs](#) la capacité d' « apprendre » à partir de données,
 - Améliorer leurs performances à résoudre des tâches
- ♦ Facilite l'utilisation des ordinateurs dans la construction de modèles à partir de données d'échantillonnage afin d'automatiser les processus de prise de décision en fonction des données saisies.

Réseaux de neurones

- Architecture d'un MLP pour faire la classification :



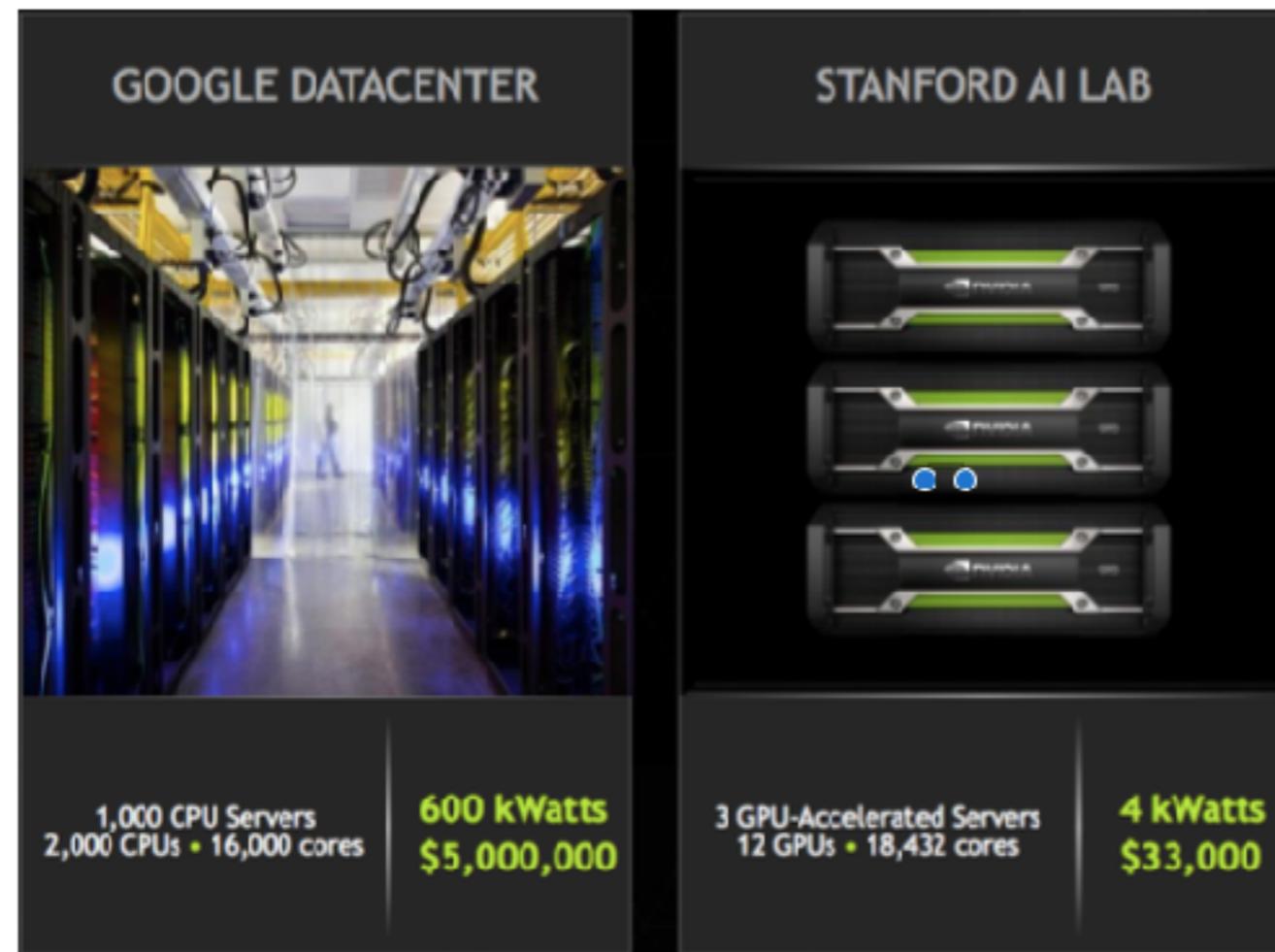
Pourquoi les DNNs
fonctionnent mieux maintenant?

DNN et puissance de calcul

- ♣ Pourquoi les DNN ne fonctionnaient pas si bien auparavant, alors que des réseaux de neurones sont expérimentés depuis les années 80 pour la modélisation acoustique ?
- ♣ Un premier élément de réponse :
 - ♦ les DNN sont particulièrement bien adaptés aux cartes graphiques (GPU) qui permettent d'atteindre une puissance de calcul phénoménale ...

DNN et puissance de calcul

- ❖ ... à un prix très compétitif !
 - Cela rend accessibles des temps d'apprentissage raisonnables et donc les expérimentations réalisables



DNN et les données

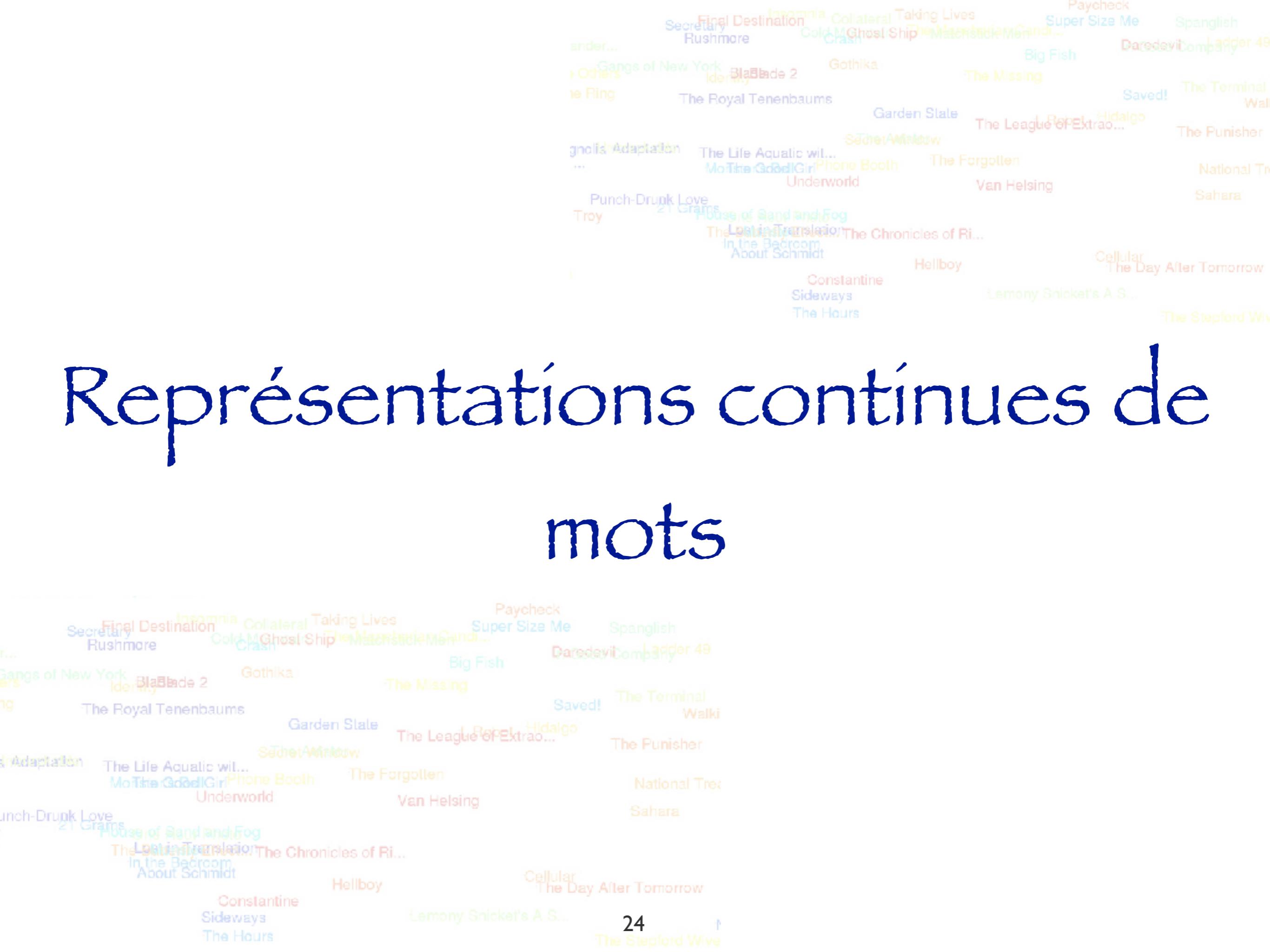
- ❖ Disponibilité de grande quantité de données
 - ❖ i.e. texte, images, audio publiés via sites de news, médias sociaux, plateformes collaboratives, smartphones, capteurs, etc.



DNN et algorithme d'apprentissage

- ❖ Un troisième élément de réponse vient des progrès effectués par les chercheurs pour l'apprentissage des DNN :
 - ♦ meilleurs algorithmes d'apprentissage (pre-training RBM, SDAE, learning rate dynamique,...)
 - ♦ nouvelles architectures (transformateur, mécanisme d'attention, LSTM, GRU...)
 - ♦ Pouvoir utiliser des représentations continues permettant de coder différents types de relations cachées (syntaxiques, sémantiques, contextuel, ...)

Représentations continues de mots



REPRÉSENTATIONS DE MOTS

I. One hot :

- ♦ Exemple : Merci ID=3, vocabulaire=10 mots

Merci :

0	0	1	0	0	0	0	0	0	0
---	---	---	---	---	---	---	---	---	---

- ♦ Difficulté :
 - ne peu pas capter des relations entre les mots
 - Taille du vecteur dépend du taille du vocabulaire

REPRÉSENTATIONS DE MOTS

2. Représentations continue des mots :

- ♦ capter des relations entre les mots
- ♦ Différentes approches ont été étudiées dans la communauté TALN :
 - Clustering
 - représentations distributionnelles
 - Représentations distribuées (neuronales) (connus par : word embeddings, *plongements de mots*, représentations continues de mots)

REPRÉSENTATIONS CONTINUES DE MOTS

❖ Clustering (Brown et al. 1992) :

- ✓ Regroupement de mots en clusters (groupe) en s'appuyant sur leurs contextes (bigrammes)
- ✓ Inconvénient : ne considère pas l'usage des mots dans contexte plus large

❖ Représentations distributionnelles exp. PMI (Pointwise Mutual Information):

- ✓ Connues sous le terme de *Count based models* ou *global matrix factorization*
- ✓ Utilisation de matrice de co-occurrences de mot
- ✓ Le mot est représenté par un vecteur dans lequel chaque entrée est une mesure d'association entre le mot et un contexte particulier
- ✓ Inconvénients : vecteur **sparse** (la plupart des éléments sont nuls) de très **haute dimension** (même taille du vocabulaire)

❖ Représentations distribuée (word embeddings) :

- ✓ Représentation de mot par un vecteur **dense** de **faible dimension** avec des valeurs réelles.

REPRÉSENTATIONS CONTINUES DE MOTS

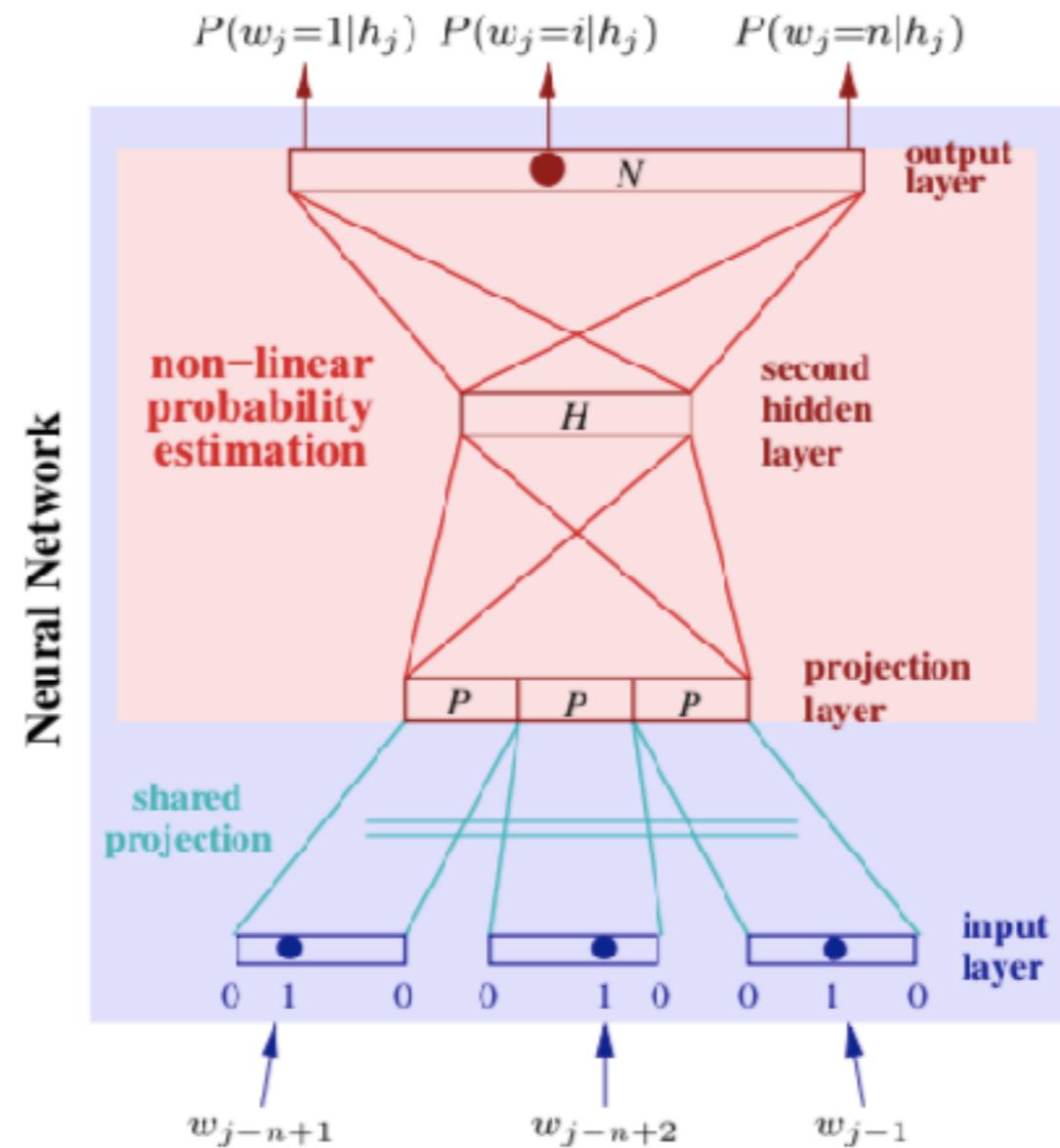
* Pourquoi des vecteurs dense ? :

- ♦ Les vecteurs courts peuvent être facilement utilisés en apprentissage automatique (moins de poids à optimiser)
- ♦ Les vecteurs dense peuvent mieux généraliser
- ♦ etc.

REPRÉSENTATIONS CONTINUES DE MOTS

Les word embeddings :

- Introduites à travers la construction des modèles de langages neuronaux
[Y.Bengio et al. 2003, H.Schwenk et al. 2006]



REPRÉSENTATIONS CONTINUES DE MOTS

Les word embeddings :

- ❖ Introduites à travers la construction des modèles de langages neuronaux [Y.Bengio et al. 2003, H.Schwenk et al. 2006]
- ❖ Constituent une projection des mots de vocabulaire dans un espace de faible dimension de manière à préserver les similarités sémantiques, syntaxiques,...

$$R : Words = \{W_1, \dots, W_n\} \rightarrow Vectors = \{R(W_1), \dots, R(W_n)\} \subset R^d$$

- ❖ Les vecteurs de mots sont proches les uns des autres en terme de distance, les mots doivent être sémantiquement ou syntaxiquement proches.

$$R(W_1) \approx R(W_n) \rightarrow W_1 \approx W_n$$

- ❖ Chaque dimension représente une caractéristique latente du mot, qui peut capter des propriétés syntaxiques et sémantiques.

REPRÉSENTATIONS CONTINUES DE MOTS

Visualisation des word embeddings :



2D t-SNE visualizations of word embeddings. Left:
Number Region; Right: Jobs Region [J.Turian et al . 2010]

What words have embeddings closest to a given word? [R.Collobert et al . 2011]

REPRÉSENTATIONS CONTINUES DE MOTS

Les word embeddings :

❖ Efficace pour de nombreuses tâches [R. Collobert et al 2011], [M. Bansal 2014] et [J.Turian et al 2010] :

- ◆ Reconnaissance d'entités nommées
- ◆ Étiquetage sémantique/grammatical
- ◆ Regroupement en syntagme
- ◆ compréhension du langage naturel
- ◆ etc.

REPRÉSENTATIONS CONTINUES DE MOTS: APPROCHES COURANTES

❖ Embeddings indépendants du contexte :

- Les occurrences du même mot ont la même représentation
- Skipgram, CBOW, Glove, w2vf-deps, fasttext

❖ Embeddings contextuels :

- Chaque occurrence du mot a une représentation calculée en fonction de son contexte
- Exemple : ELMo, BERT



REPRÉSENTATIONS CONTINUES DE MOTS INDÉPENDANTES DE CONTEXTE

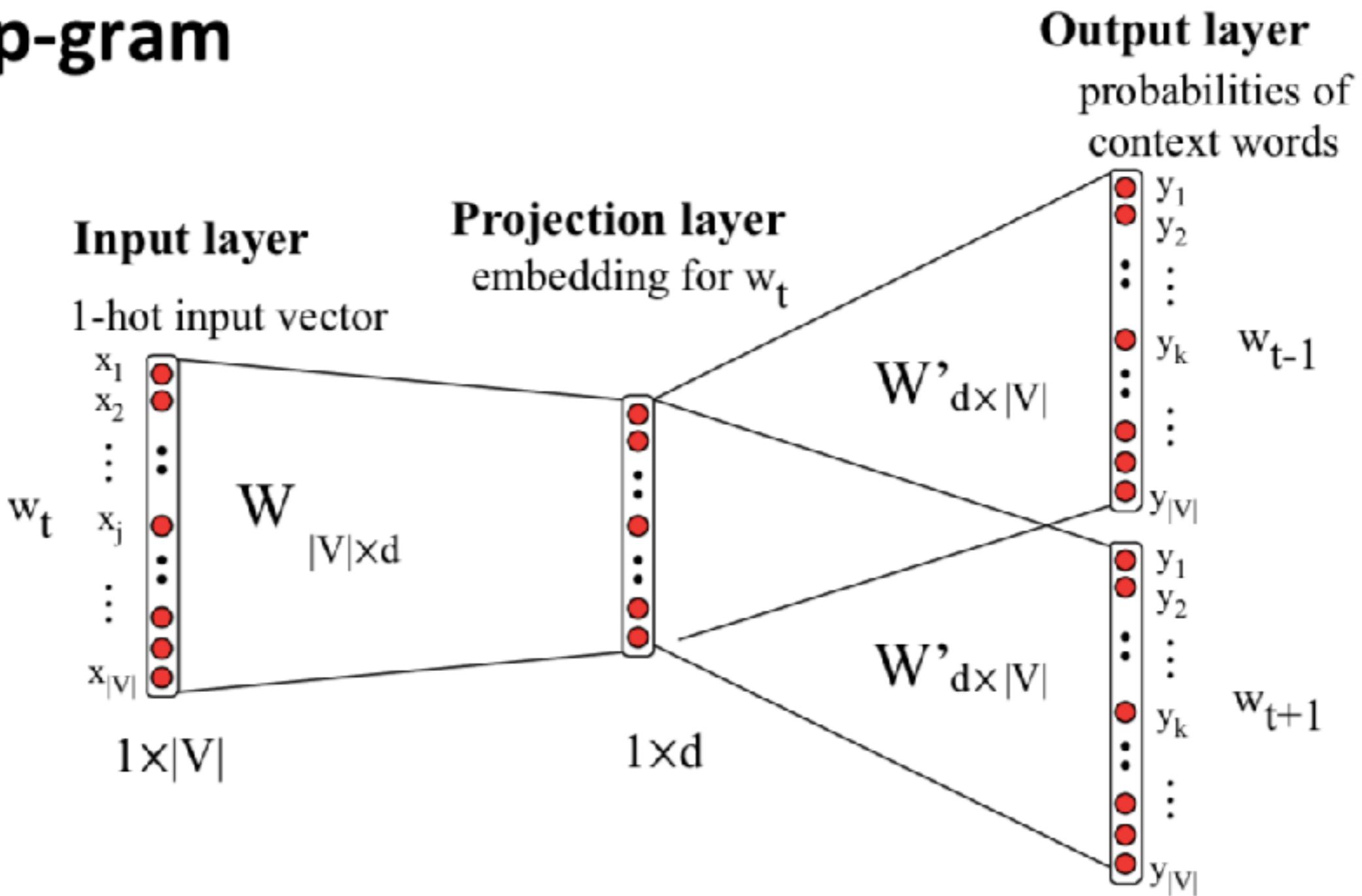
Skip-gram (Mikolov et al. 2013a) et CBOW (Mikolov et al. 2013a)

- ❖ Apprendre des embeddings de mots dans le cadre du processus de prédiction de mots.
 - ◆ Le réseau de neurones doit prédire les mots voisins
 - ◆ Inspiré par les modèles de langage neuronal.

REPRÉSENTATIONS CONTINUES DE MOTS INDÉPENDANTES DE CONTEXTE

REPRÉSEN TATION

Skip-gram

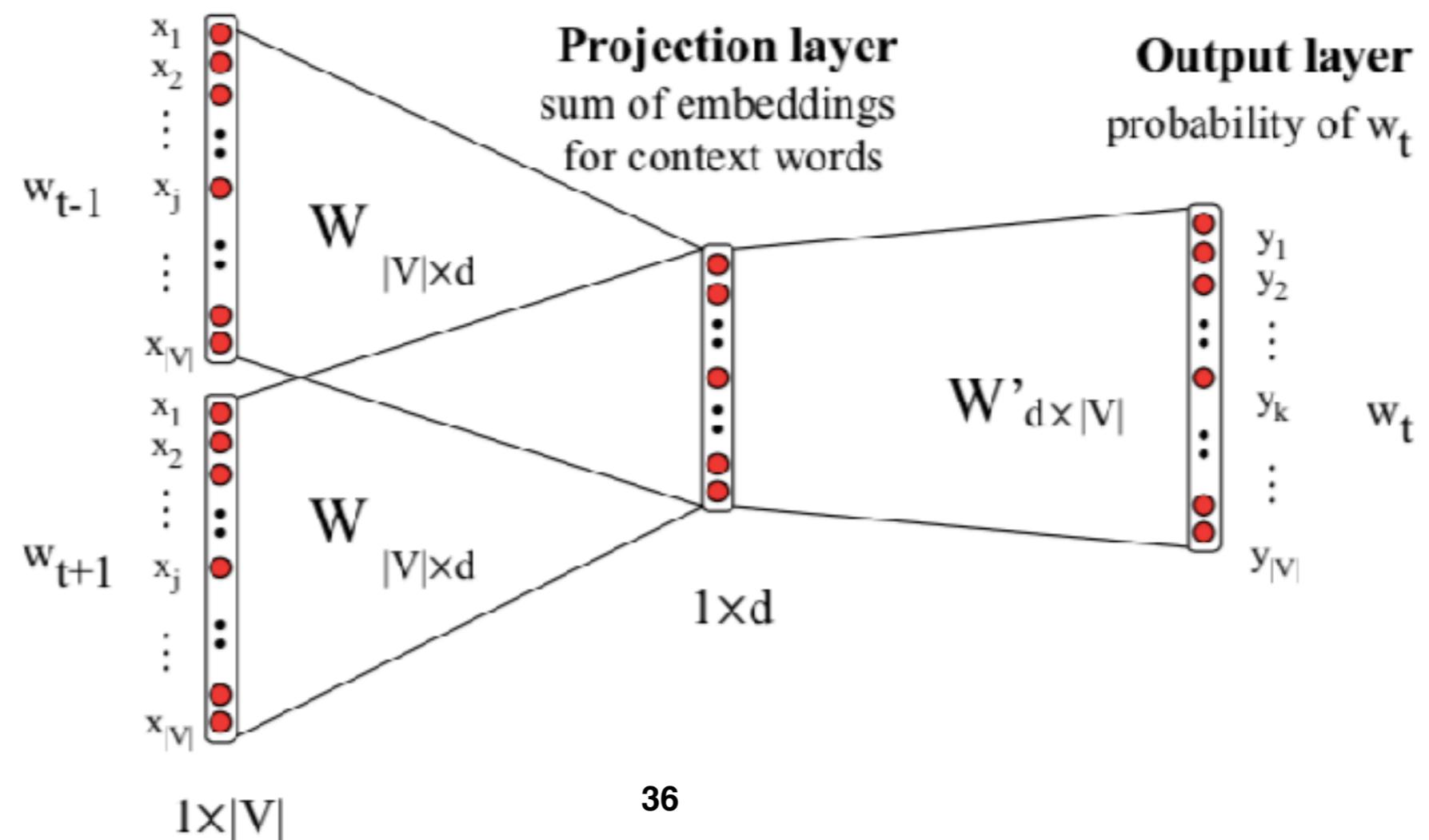


REPRÉSENTATIONS CONTINUES DE MOTS INDÉPENDANTES DE CONTEXTE

CBOW (Continuous Bag of Words)

Input layer

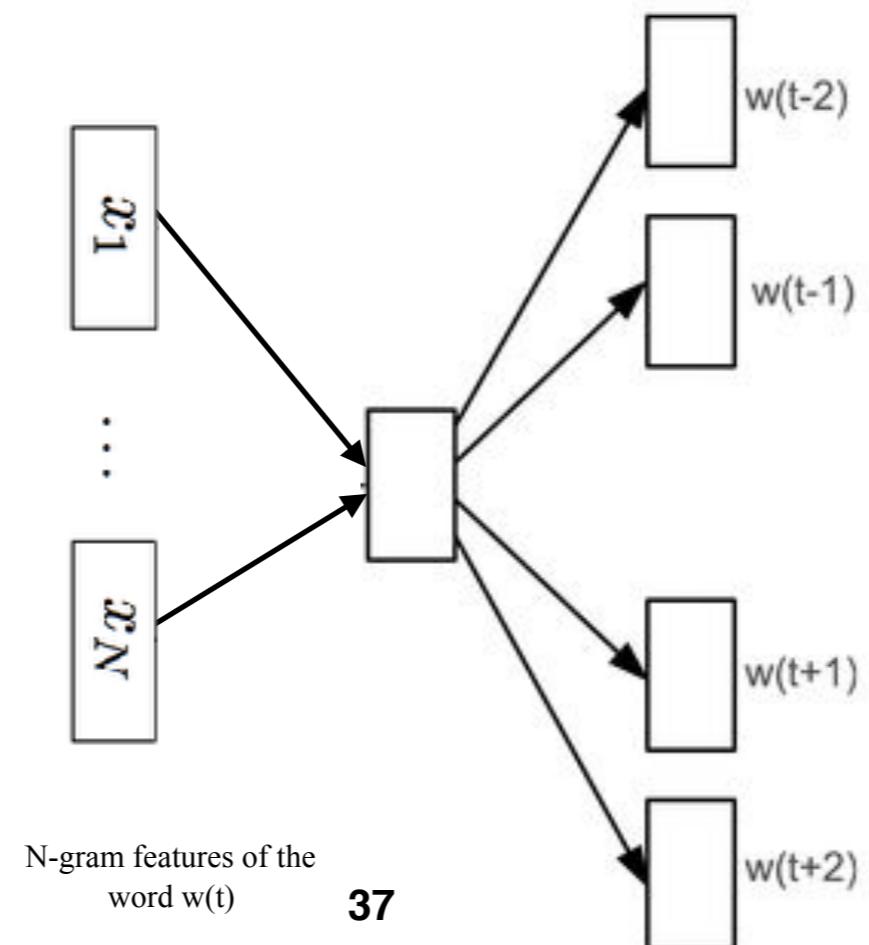
1-hot input vectors
for each context word



REPRÉSENTATIONS CONTINUES DE MOTS INDÉPENDANTES DE CONTEXTE

FastText (P. Bojanowski et al. 2017)

- ❖ Apprendre des embeddings de mots dans le cadre du processus de prédiction de mots basé sur l'architecture skipgram.
 - ◆ Prend en compte l'information morphologique de mot représenté sous forme de ngramme de caractère
 - ◆ Chaque mot est représenté comme la somme de représentations de ses caractères n grammes.
 - ◆ Calcule des représentations de mots pour des mots qui n'apparaissaient pas dans les données d'apprentissage, ce qui n'est pas le cas pour d'autres approches (word2vec)



REPRÉSENTATIONS CONTINUES DE MOTS INDÉPENDANTES DE CONTEXTE

Glove (J. Pennington et al 2014)

- ♣ Analyse des co-occurrences des mots dans le corpus
 - ♦ construction d'une matrice de co-occurrence : utilisation d'une fenêtre contextuelle glissante
 - ♦ estimation des représentations continues des mots

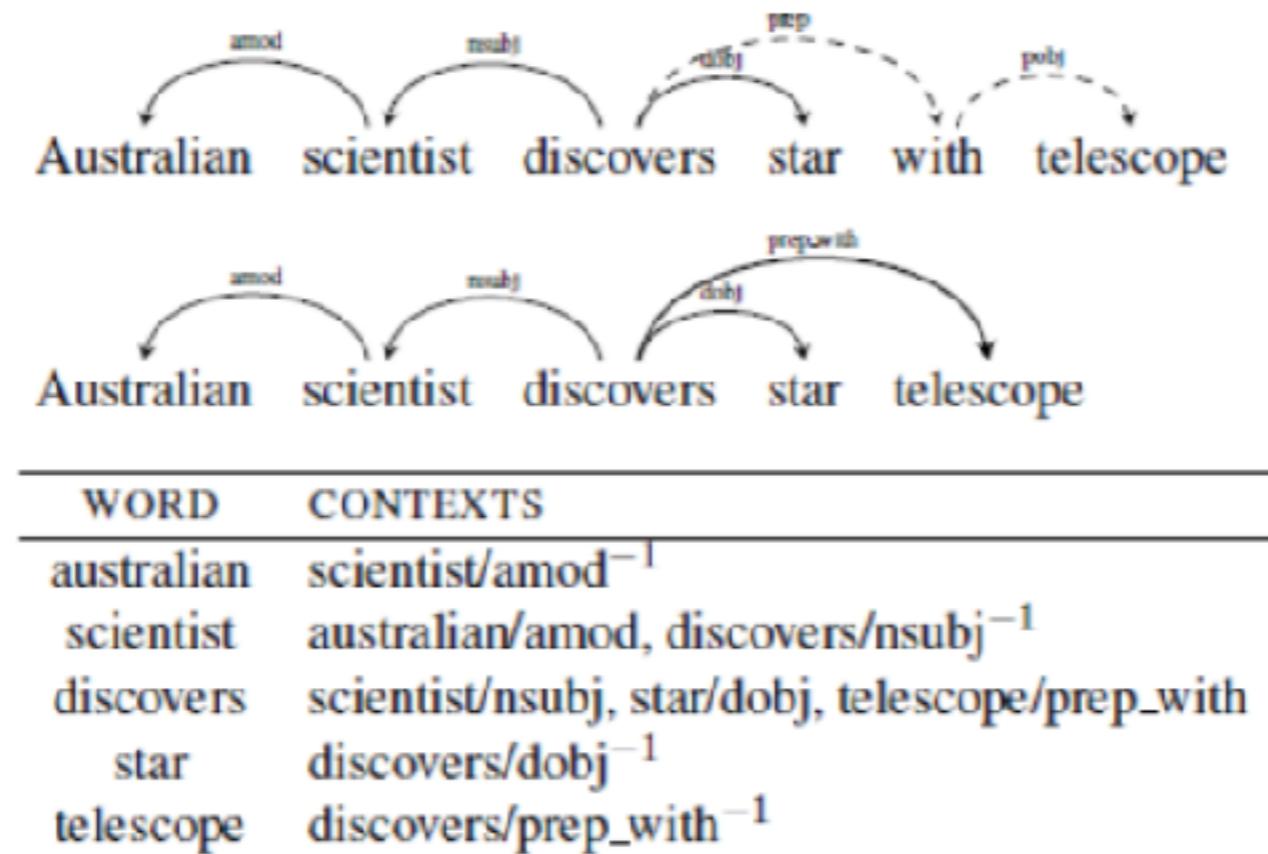
REPRÉSENTATIONS CONTINUES DE MOTS INDÉPENDANTES DE CONTEXTE

w2vf-deps : dependency based word embeddings (O. Levy et al. 2014)

- ❖ Généralisation du modèle skip-gram avec échantillons négatifs
- ❖ S'appuie sur les relations syntaxiques (mots racines, lien de dépendance entre le mot racine et les mots dépendants) pour extraire les mots en contexte
 - ◆ capter des relations entre mots éloignés : impossible de détecter dans une fenêtre de contexte de taille réduite
 - ◆ n'utiliser que les mots de contexte directement liés au mot cible

REPRÉSENTATIONS CONTINUES DE MOTS INDÉPENDANTES DE CONTEXTE

w2vf-deps (O. Levy et al. 2014)

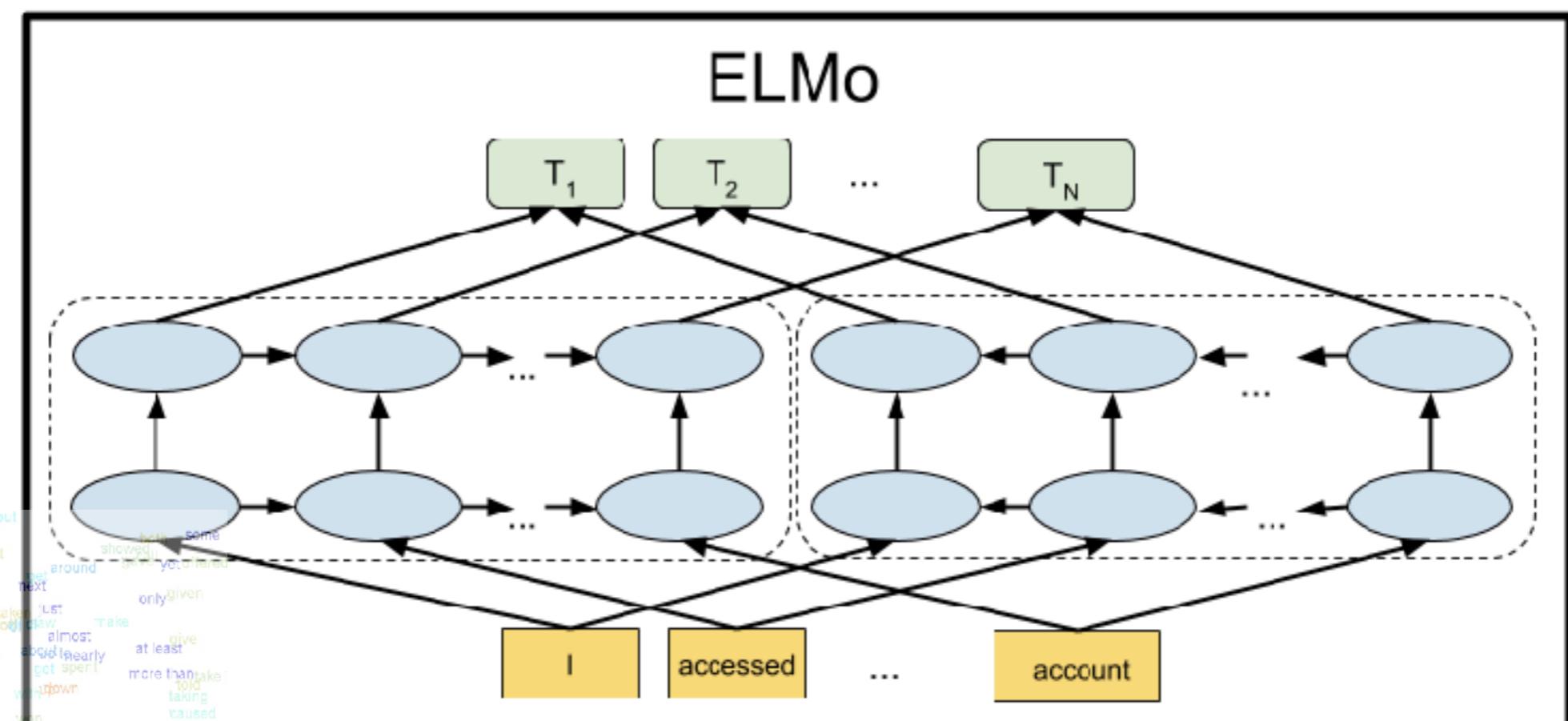


REPRÉSENTATIONS CONTEXTUELLES DE MOTS



ELMo (Embeddings from Language Models) (Peters, et al, 2018)

- ❖ Apprendre les word embeddings en créant des modèles de langage bidirectionnel

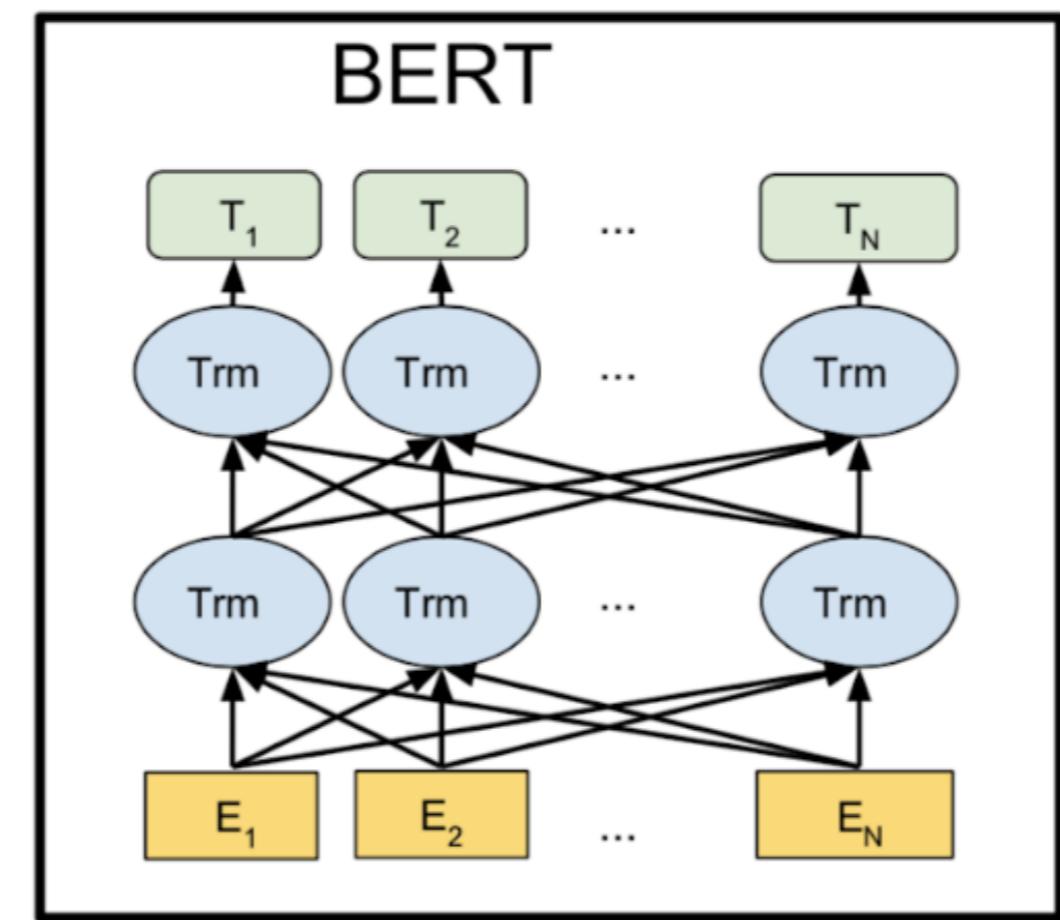


REPRÉSENTATIONS CONTEXTUELLES DE MOTS



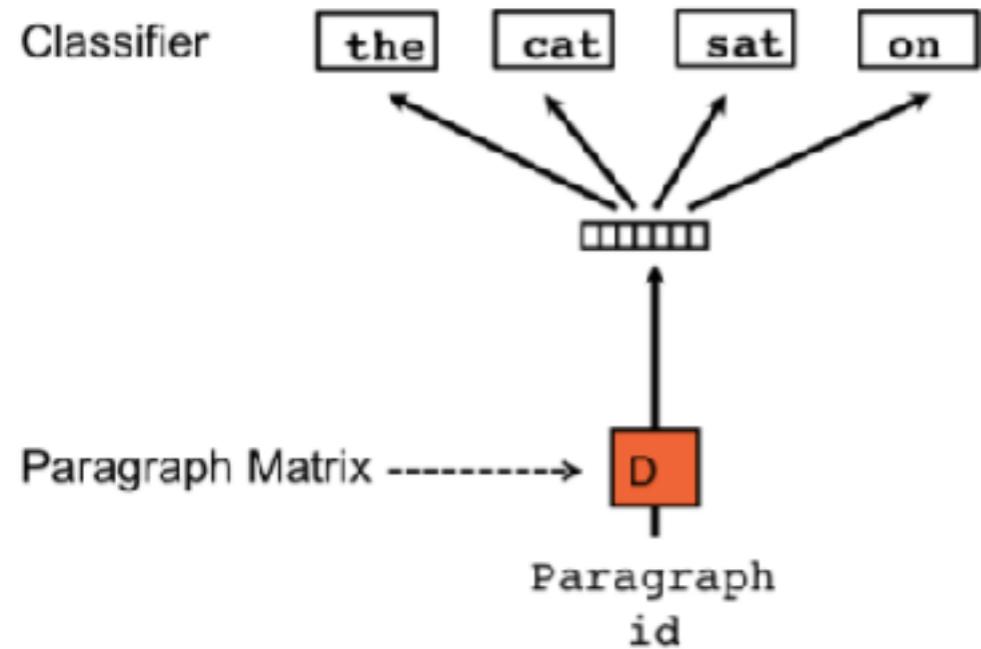
BERT (Bidirectional Encoder Representations from Transformers)(Devlin, et al., 2019)

- ❖ Utilise un transformateur bidirectionnel entraîné conjointement sur une tâche de modélisation de langage masqué et une tâche de prédiction de la phrase suivante.



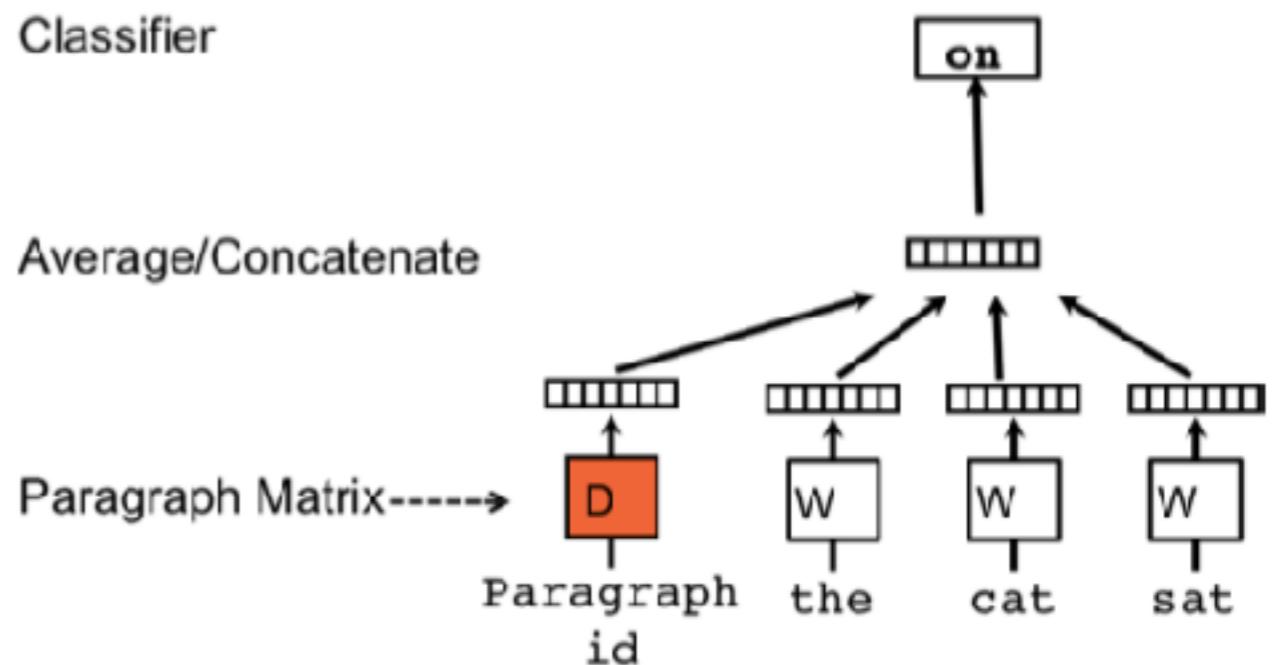
REPRÉSENTATIONS CONTINUES DU DOCUMENT

Distributed bag of words



(Mikolov et al. 2013b)

Distributed Memory model



REPRÉSENTATIONS CONTINUES : TÂCHES D'ÉVALUATIONS

I. Tâches du traitement des langues naturelles (TALN)

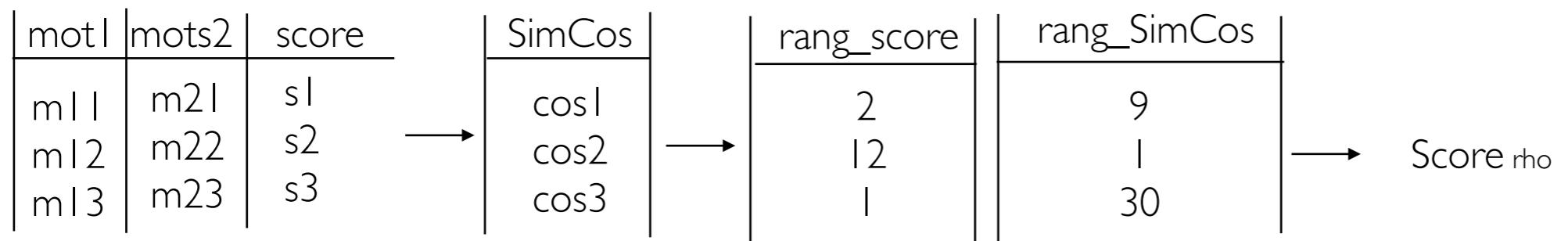
- ♦ Étiquetage morpho-syntaxique (POS) : 48 étiquettes (nom, verbe, etc.)
- ♦ Regroupement en syntagme (CHK) : 22 étiquettes (B-GV, I-GN, etc.)
- ♦ Reconnaissance d'entités nommées (NER) : 8 étiquettes (B-ORG, I-PER, O, etc.)
- ♦ Détection de mention d'entités nommées (MENT) : 3 étiquettes (B, I, O)

REPRÉSENTATIONS CONTINUES : TÂCHES D'ÉVALUATIONS

2. Tâche de similarité (Kiela et al., 2015)

- ♦ **Similarité fonctionnelle** : relations lexicales de synonymie (par exemple “voiture : automobile”) et hyponymie : la signification d’un mot est couverte par un autre terme plus général, comme dans “voiture : véhicule”.
- ♦ **Similarité associative** : tout type d’association lexicale ou fonctionnelle. Exemple : la méronymie (“doigt : main”), antonymie (sens opposés, “chaud : froid”), ou relation fonctionnelle (“pingouin : antarctique”, qui ne sont pas liés par une relation lexicale).

→ Évaluation : score de corrélation de Spearman (rho)



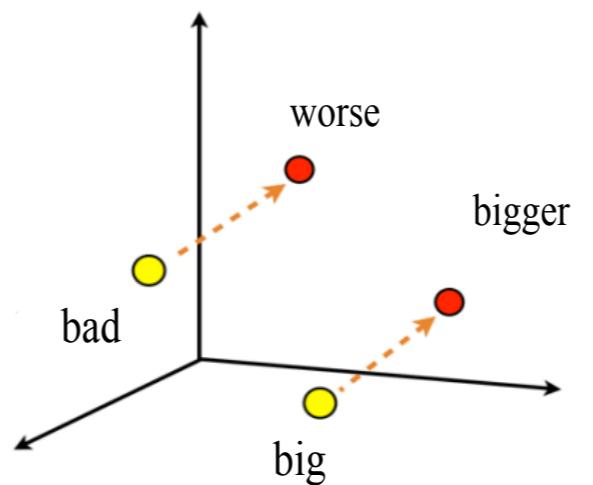
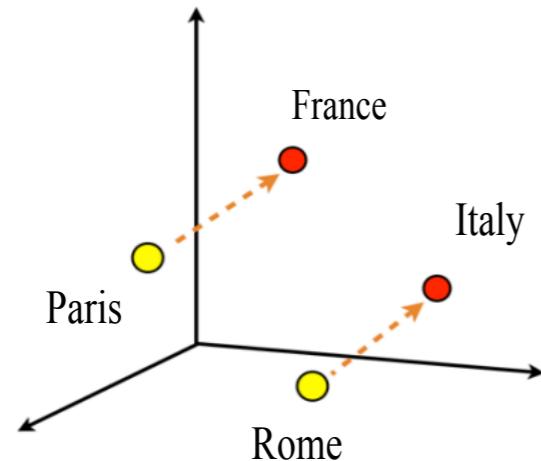
REPRÉSENTATIONS CONTINUES : TÂCHES D'ÉVALUATIONS

3. Tâche d'analogie

- ♦ Évaluer la similarité relationnelle entre les paires de mots
- ♦ Répondre à des questions d'analogie en s'appuyant sur les similarités cosinus
 - Sémantique ou syntaxique

Sémantique Paris:France → Rome:?

Syntaxique bad:worse → big:?

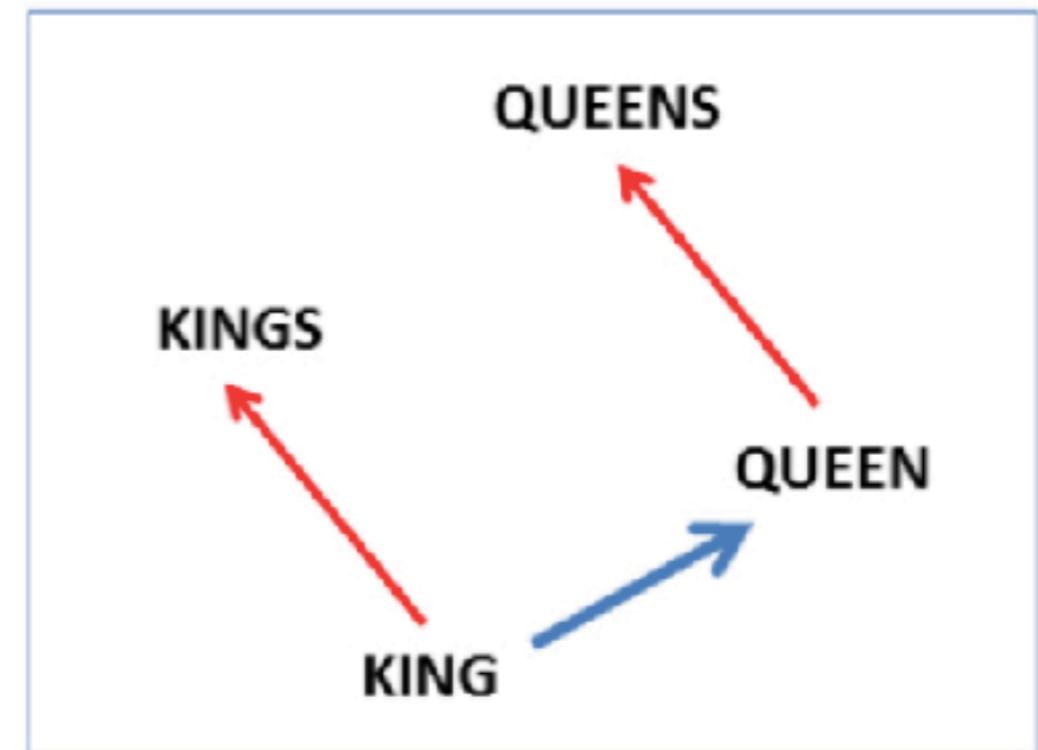
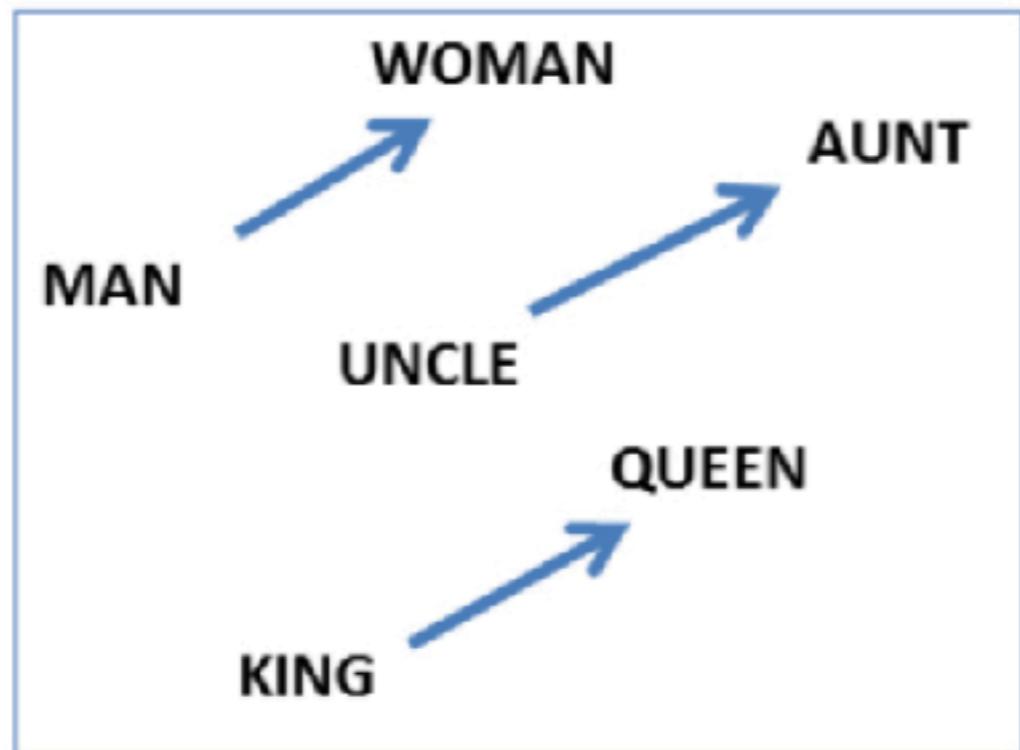


REPRÉSENTATIONS CONTINUES : TÂCHES D'ÉVALUATIONS

3. Tâche d'analogie

$\text{vector('king')} - \text{vector('man')} + \text{vector('woman')} \approx \text{vector('queen')}$

$\text{vector('Paris')} - \text{vector('France')} + \text{vector('Italy')} \approx \text{vector('Rome')}$



ÉVALUATIONS CONTINUES : TÂCHES D'ÉVALUATIONS

Données d'évaluations

- ❖ Word embeddings :
 - ◆ Données : corpus Gigaword (anglais)
 - ◆ Vocabulaire : 239k mots
 - ◆ Paramètres :

Embeddings	Tail.	Dim.	Neg.
CBOW			5
Skip-gram	5		5
GloVe		200	-
w2vf-deps	-		5

- ❖ Tâche d'analogie :
 - ❖ Questions :
 - 8869 sémantiques
 - 10675 syntaxiques

- ❖ Tâche de similarité
 - ◆ Données :
 - WordSim353
 - RW (2034)
 - MEN (3000)

- ❖ Tâches TALN
 - ◆ Données

Tâche	Benchmark	Train	Dev	Test
POS	Penn Treebank	958k	34k	58k
CHK	CoNLL 2000	191k	21k	47k
NER	CoNLL 2003	205k	52k	47k
MENT	Ontonotes	736k	102k	105k

REPRÉSENTATIONS CONTINUES : TÂCHES D'ÉVALUATIONS

Embeddings	Tâches TALN				Tâche de similarité			Tâche d'analogie
	POS	CHK	NER	MENT	WS353	RW	MEN	
	Acc.	FI			Spearman's rank rho			
CBOW	96,01	0,904	0,783	0,554	0,590	0,465	0,609	57,20
Skip-gram	96,43	0,896	0,776	0,578	0,558	0,502	0,662	62,30
w2v-deps	96,66	0,92	0,793	0,580	0,523	0,435	0,557	42,70
GloVe	95,79	0,869	0,764	0,544	0,533	0,410	0,660	65,50

[Ghannay, et al., 2016]

→ Ces embeddings portent des informations différentes

REPRÉSENTATIONS CONTINUES : TÂCHES D'ÉVALUATIONS

DONNÉES DÉPENDANTES DES TÂCHES VS HORS DU DOMAINE

ÉVALUATION QUANTITATIVE

Bench.	task-dependent					Out-of-domain				
	ELMo	FastText	GloVe	Skip-gram	CBOW	ELMo	FastText	GloVe	Skip-gram	CBOW
M2M	88.89	72.13	92.54	88.87	89.39	91.14	93.01	91.77	93.19	92.13
ATIS	94.38	85.72	92.95	90.84	91.87	94.93	95.52	95.35	95.62	95.77
SNIPS	78.68	76.35	87.40	82.10	83.94	90.29	94.85	93.90	94.43	94.05
SNIPS70	53.06	38.19	63.65	47.11	49.76	75.19	79.75	78.68	78.90	80.13
MEDIA	80.26	71.73	82.66	80.01	79.57	86.42	85.30	85.11	85.95	86.06

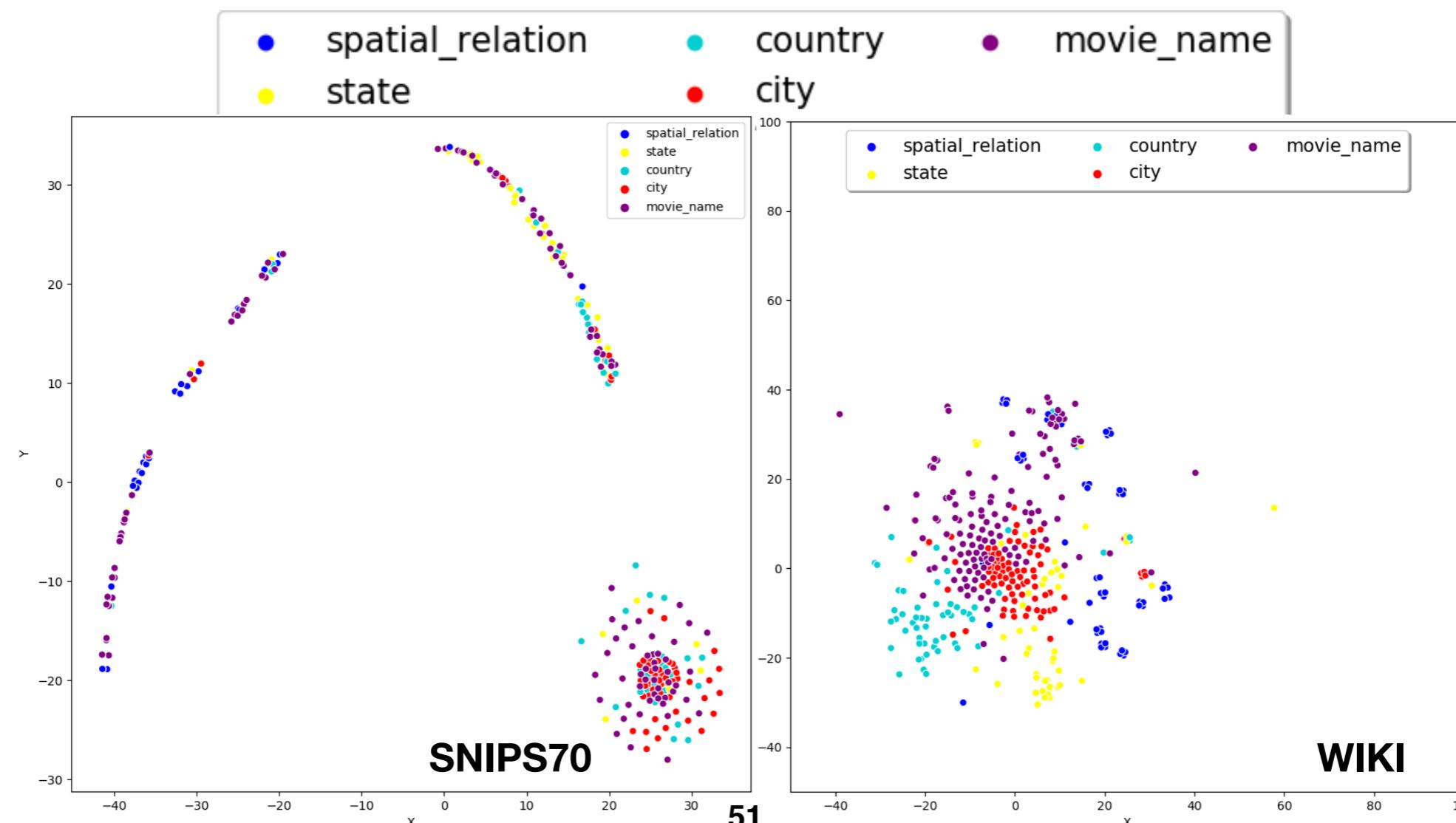
Tagging performance of different word embeddings trained on task-dependent corpus (ATIS, MEDIA, M2M, SNIPS or SNIPS70) and on huge and out of domain corpus (WIKI English or French) on all benchmark corpora in terms of F1 using conlleval scoring script (in %)

[Ghannay, et al., 2020]

REPRÉSENTATIONS CONTINUES : TÂCHES D'ÉVALUATIONS

DONNÉES DÉPENDANTES DES TÂCHES VS HORS DU DOMAINE

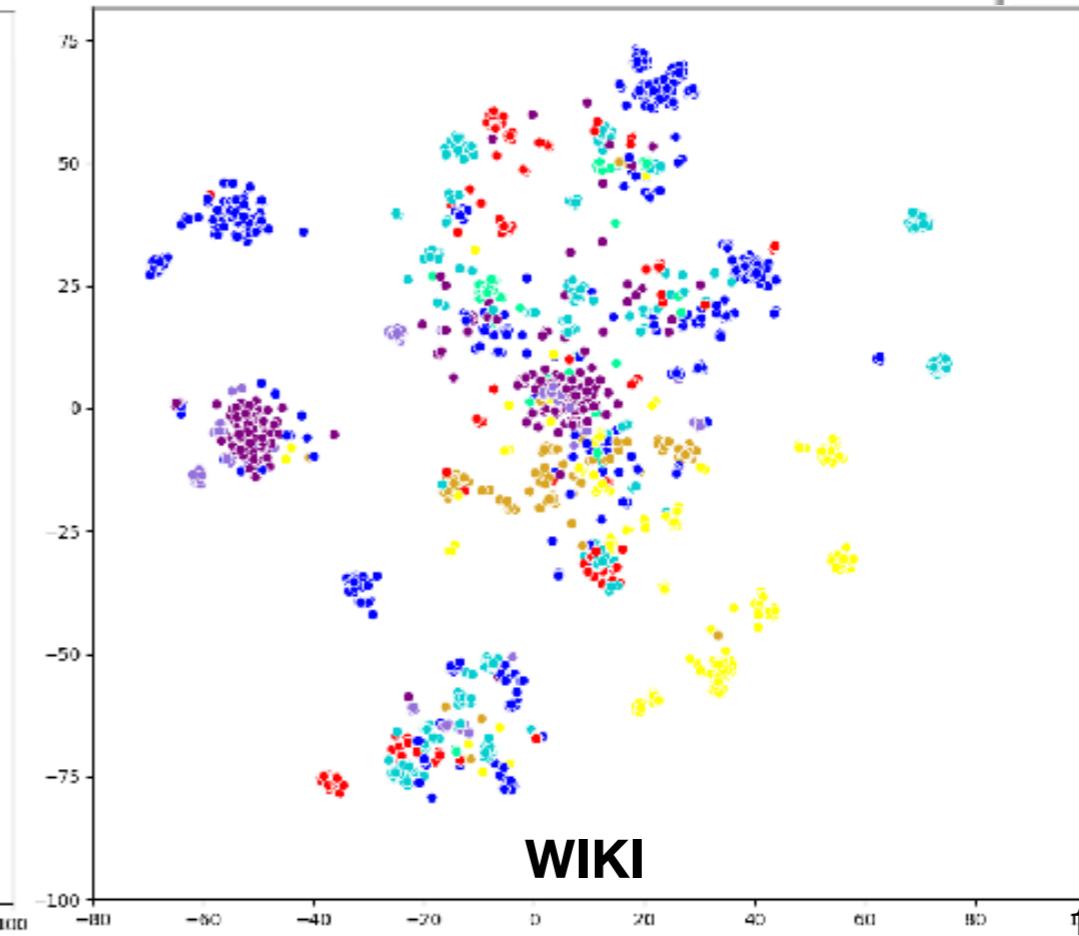
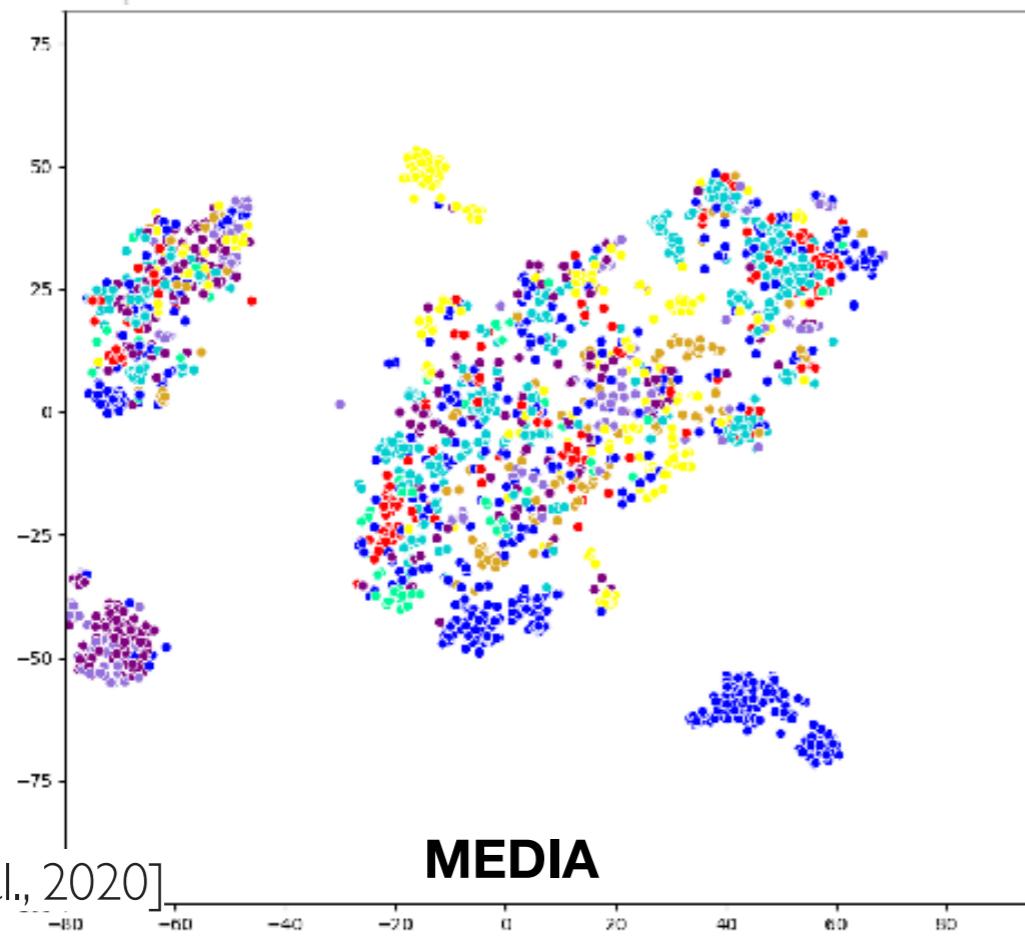
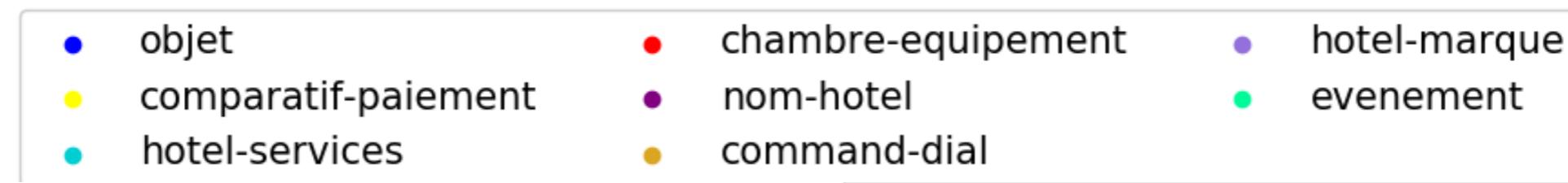
ÉVALUATION QUALITATIVE CBOW:



REPRÉSENTATIONS CONTINUES : TÂCHES D'ÉVALUATIONS

DONNÉES DÉPENDANTES DES TÂCHES VS HORS DU DOMAINE

ÉVALUATION QUALITATIVE ELMO:



REPRÉSENTATIONS CONTINUES : OUTILS

Outils disponibles pour construire les embeddings

- ♦ Word2vec Mikolov : <https://github.com/tmikolov/word2vec>
- ♦ W2vf-deps Levy : <https://levyomer.wordpress.com/2014/04/25/dependency-based-word-embeddings/>
- ♦ Glove Pennington : <https://nlp.stanford.edu/projects/glove/>
- ♦ Fasttext: <https://fasttext.cc/docs/en/support.html>
- ♦ ELMo : allenlp https://github.com/allenai/allennlp/blob/master/tutorials/how_to/elmo.md
- ♦ BERT : <https://pypi.org/project/bert-embedding/>
- ♦ Bibliothèque python : gensim

Sitographie/remerciement

- Cours Yannick Estève, Le Mans université
- Cours Sophie Rosset, LIMSI, CNRS :
 - <https://sophierosset.github.io/docs/eidi-dhm.pdf>
 - <https://bigdataspeech.github.io/EN/>
- Cours Dan Jurafsky:Victor Sementic, Standford university
- Cours en ligne :
 - <https://nlp.stanford.edu/~wcmac/papers/20140716-UNLU.pdf>

Bibliographie

- [Mikolov et al., 2013a] Mikolov, T., Chen, K., Corrado, G. et Dean, J. (2013a). Efficient Estimation of Word Representations in Vector Space. In Proceedings of Workshop at ICLR.
- [Levy et Goldberg, 2014] Levy, O. et Goldberg, Y. (2014). Dependencybased word embeddings. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, volume 2, pages 302–308.
- [Mikolov et al., 2013b] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. et Dean, J. (2013b). Distributed representations of words and phrases and their compositionality.
- [Pennington et al., 2014] Pennington, J., Socher, R. et Manning, C. D. (2014). Glove : Global vectors for word representation. In Proceedings of the Empirical Methods in Natural Language Processing (EMNLP 2014), volume 12.
- [Peters, et al, 2018] Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. arXiv preprint arXiv:1802.05365.
- [Devlin, et al., 2019] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
- [Ghannay, et al., 2016] Ghannay, S., Favre, B., Esteve, Y., & Camelin, N. (2016, May). Word embedding evaluation and combination. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)
- [Kiela et al., 2015] Kiela, D., Hill, F. et Clark, S. (2015). Specializing word embeddings for similarity or relatedness. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 2044–2048, Lisbon, Portugal. Association for Computational Linguistics.
- [Mikolov et al., 2013] Mikolov, T., Yih, W.-t. et Zweig, G. 2013. Linguistic regularities in continuous space word representations. In HLT-NAACL, pages 746–751.
- [P. Bojanowski et al. 2017] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov, “Enriching word vectors with subword information,” Transactions of the Association for Computational Linguistics, vol. 5, 2017.
- [Ghannay, et al., 2020] Ghannay, Sahar, Antoine Neuraz, and Sophie Rosset. "What is best for spoken language understanding: small but task-dependant embeddings or huge but out-of-domain embeddings?" ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020

TP

Cours : https://saharghannay.github.io/files/Cours_MasterISD.pdf

TP : <https://saharghannay.github.io/courses/cours1/example1/>