


Le traitement automatique du langage au service des marchés publics.

Présenté par : [Oussama AHMIA](#)

- 
1. Introduction
 2. Extraction d'information
 3. Vectorisation
 - État de l'art
 - LSA+W2V
 - CnHAtt

Introduction

Contexte

OctopusMind est une PME basée à Nantes depuis 2005

- 14 employés, 1/2 IT team, 1/3 experts marchés et analystes

Gère une base de données de plus de **14 000 000** de documents relatifs aux marchés publics. Environ 52 000 documents par an nécessitent une **curation** manuelle par des experts.

Appel d'offres: une procédure qui permet à un commanditaire de faire le choix de l'entreprise la plus à même de réaliser une prestation. Le but est de mettre plusieurs entreprises en concurrence pour fournir un produit ou un service.

Introduction


Problématiques

Fonctionnalités attendues :

- Le **routage** des appels d'offre vers les entreprises concernées,
- La **catégorisation** des appels d'offres similaires et des entreprises positionnées/retenues sur les marchés,
- L'**extraction d'information** à partir de documents non structurés.
- le **suivi des tendances** en matière d'appels d'offres, de marchés remportés, etc

Introduction

Nature des données

– Date et heure limite de remise des plis :	30/03/2017 16:00 (heure de Paris) 
Référence Intitulé :	2017MA0006M13S0000 Classement et reconditionnement d'une partie du fonds d'archives des bulletins d'entree du musee de l'armee
Objet :	Classement et reconditionnement d'une partie du fonds d'archives des bulletins d'entree du musee de l'armee
Entité publique :	Ministère de la Défense
Entité d'Achat :	MINDEF / ESTM / EPA / Musée de l'Armée - Musée de l'Armée
Type d'annonce :	Annonce de consultation
Procédure :	Procédure adaptée
Catégorie principale :	Services
Allotissement :	-
Lieu d'exécution :	(75) Paris, (77) Seine-et-Marne, (78) Yvelines, (91) Essonne, (92) Hauts-de-Seine ...
Code CPV :	92512000 (Code principal)

Introduction

Nature des données

Département : Paris (75)

Date limite des candidatures :

Date limite des offres : 28/06/2019 00:00

Objet de la consultation

Accord cadre alloti multi attributaire à bons de commandes pour la fourniture de denrées alimentaires

Liste des lots

Lot 1 : BOF (beurre, oeufs, fromages)

BOF (beurre, oeufs, fromages)

Lot 10 : POISSON FRAIS

II.2.6) Valeur estimée

Valeur hors TVA : 176 000,00 euros

II.2.7) Durée du marché, de l'accord-cadre ou du système d'acquisition dynamique

Durée en mois : 12

Ce marché peut faire l'objet d'une reconduction : oui

Description des modalités ou du calendrier des reconductions : Les accords-cadres sont reconductibles pour 3 périodes de un an

II.2.9) Informations sur les limites concernant le nombre de candidats invités à participer

Critères objectifs de limitation du nombre de candidats :

II.2.10) Variantes

Des variantes seront prises en considération : non

II.2.11) Information sur les options

Options : non

Extraction d'information

Extraction de surface

- la cité des entreprises (environ 2500 m² de plancher) comprenant des bureaux , salles et espaces spécifiques type living-lab ;
- bureaux (350 m² utiles) répartis en 16 bureaux de 15 m² et un open space
- lot n° 1 : 6000 m² de shon logements + 300 m² pour la réhabilitation ...

Légende:

■ 0/bâtiment

■ 0/surface

■ LINK_SURFACE

■ LINK_SURFACE/surface

■ LINK_SURFACE/bâtiment

(1) la détection d'entités
nommées (surface, bâtiment ou autre)

(2) la recherche de relations de type "est surface de"

Notre approche consiste à effectuer simultanément l'extraction d'entités nommées (surface et bâtiment) et la détection des relations potentiellement existantes entre ces entités

Extraction d'information

Extraction de surface

Pour chaque mot :

- **word.lower()**: mot en minuscule
- **word.isupper()**: Vrai si le mot est en majuscule
- **word.istitle()**: Vrai si le mot commence par une majuscule
- **word.isdigit()**: vrai si le mot est un nombre
- **postag**: "Part of speech", la classe grammaticale du mot
- **postag[:2]**: la seconde partie du postag
- **type (cpt)**: le type de mots (surface, building or other)
- **lemma**: le lemme du mot
- **Caractéristiques à longue portée**: modélisent le contexte, ont démontré leur efficacité en matière d'étiquetage de séquence

Note :

Pour la variables type (cpt), des expressions régulières sont utilisées.

Extraction d'information

Extraction de surface

Example:

– une zone " logements " composée de logements sous forme de maisons en bande de **687_m²** utiles (+ 10_m² de locaux_poubelles poubelles) .

```
{'+1:cpt': '',
 '+1:lemma': 'utile',
 '+1:postag': 'ADJ',
 '+1:postag[:2]': '0',
 '+1:word.istitle()': False,
 '+1:word.isupper()': False,
 '+1:word.lower()': 'utiles',
 '+2:cpt': '',
 '+2:lemma': '(',
 '+2:postag': 'PUN',
 '+2:postag[:2]': '0',
 '+2:word.istitle()': False,
 '+2:word.isupper()': False,
 '+2:word.lower()': '(',
 '-1:cpt': '',
 '-1:lemma': 'de',
 '-1:postag': 'PRP',
 '-1:postag[:2]': '0',
 '-1:word.istitle()': False,
 '-1:word.isupper()': False,
```

```
'-1:word.lower()': 'de',
 '-2:cpt': '',
 '-2:postag': 'NOM',
 '-2:postag[:2]': '0',
 '-2:word.istitle()': False,
 '-2:word.isupper()': False,
 '-2:word.lower()': 'bande',
 '-2lemma': 'bande',
 'bias': 1.0,
 'cpt': 'surface',
 'lemma': '@CARD',
 'postag': 'NUM',
 'postag[:2]': '0',
 'type': ['utiles'],
 'word.isdigit()': False,
 'word.istitle()': False,
 'word.isupper()': False,
 'word.lower()': '687_m²'}
```

Extraction d'information

Extraction de surface

- Les données utilisées pour cette expérimentation sont des annonces extraites de la base de données du **BOAMP**, qui constitue la version électronique du Bulletin officiel des annonces de marchés publics.
- Nous avons indexé les données collectées sur le site BOAMP en utilisant un moteur de recherche (**Lucene**) afin de filtrer les annonces concernant la construction de nouveaux bâtiments.
- La description de chaque annonce est ensuite découpée en phrases puis, à l'aide d'expressions régulières dédiées, les présences de terminologies décrivant une surface sont détectées dans la phrase.
- Corpus final: **2000** séquences étiquetées par des experts.

Extraction d'information

Extraction de surface

80% du jeu de données présenté est utilisé en tant que données d'apprentissage et 20% comme données de test, en procédant à une validation croisée (k=5).

	Précision	Rappel	F1-mesure	exactitude empirique	Écart type
CRF contexte(3)	0.932	0.932	0.932	76.04%	1.10%
CRF contexte(2)	0.926	0.926	0.926	74.65%	1.61%
CRF linéaire	0.899	0.896	0.897	61.75%	2.51%
CRF semi-Markoviens	0.897	0.899	0.897	67.74%	1.69%
HCRF ordre(3)	0.878	0.878	0.877	66.36%	1.36%
HCRF ordre(2)	0.884	0.882	0.882	63.59%	1.38%
Perceptron structuré	0.898	0.897	0.897	64.52%	1.84%
Automate (Regex)	0.884	0.851	0.855	66.89%	N/A
HMM	0.776	0.733	0.667	15.21%	0.28%

Résultats obtenus avec les différents algorithmes.

Extraction d'information

Extraction de surface

	précision	rappel	f1-mesure	support
O	0.954	0.943	0.948	3473
LINK_SURFACE	0.891	0.909	0.900	1764
LINK_SURFACE/bâtiment	0.931	0.940	0.935	315
O/bâtiment	0.807	0.800	0.803	115
LINK_SURFACE/surface	0.973	0.973	0.973	524
O/surface	0.767	0.793	0.780	58

Scores du CRF contexte (3) par étiquette.

Applications OctopusMind

Extraction d'information

Variable	Étiquette	Poids
lemma :m2	LINK_SURFACE/surface	1.087
type :surface	LINK_SURFACE/surface	2.296
type :bâtiment	LINK_SURFACE/bâtiment	4.674
type :surface	O/surface	1.452
word.lower() :m2	LINK_SURFACE/surface	1.079
POS :NOUN	LINK_SURFACE/bâtiment	1.561
type :bâtiment	O/bâtiment	3.551
-1 :type :surface	LINK_SURFACE/surface	0.919
-1 :lemma :mètre	O	-0.53
POS :NUM	O	-0.69
lemma :local	O	-0.81

Extrait des poids associés aux caractéristiques les plus discriminantes, pour le modèle CRF contexte (3)

Extraction d'information

Extraction de budgets

V.4) Informations sur le montant du marché :

Valeur totale finale du marché:

Valeur : 13900 EUR.

Hors TVA.

V.5) LE MARCHÉ EST SUSCEPTIBLE D'ÊTRE SOUS-TRAITÉ :

Non

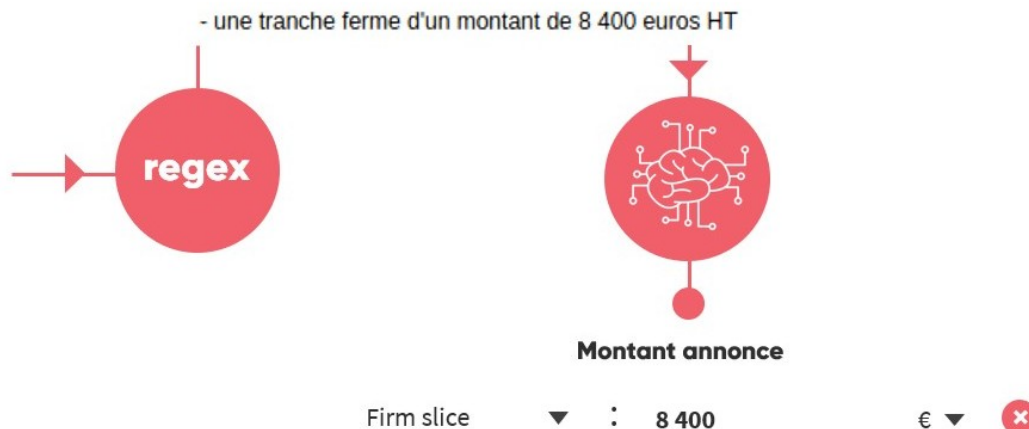
SECTION VI : RENSEIGNEMENTS COMPLÉMENTAIRES

VI.1) LE MARCHÉ S'INSCRIT DANS UN PROJET/PROGRAMME FINANCÉ PAR DES FONDS COMMUNAUTAIRES :

VI.2) AUTRES INFORMATIONS :

Ce marché est décomposé en :

- une tranche ferme d'un montant de 8 400 euros HT
- une tranche conditionnelle d'un montant de 5 500 euros HT.



Vectorisation

Applications

- **Catégorisation** de documents (**classification** des marchés publics par domaine d'activité)
- **Clustering** de documents (découvrir des **tendances** dans les marchés publics)
- **Calcul de similarité** entre documents (**Recommandation** de documents basée sur une distance sémantique)

Vectorisation (État de l'art)

Sac de mots

Bag of Words (Sac de mots) [Harris 1954] consiste à décrire un texte par le nombre d'occurrences des mots (c'est-à-dire leur fréquence) qui le composent.

Exemple « the white cat » :

<i>brown</i>	<i>cat</i>	<i>dog</i>	<i>i</i>	<i>is</i>	<i>love</i>	<i>the</i>	<i>white</i>
0	1	0	0	0	0	1	1

Limites

- Très grande dimensionnalité
- Tous les mots ont le même poids
- Ne prend en considération que la forme du mot : vendre et vendu ont la même distance que vendre et avocat
- Ne capture pas la sémantique

Vectorisation (État de l'art)

Pondération tf-idf

tf-idf (term frequency–inverse document frequency) [Jones 1972] consiste à décrire un texte en pondérant les mots par l'exploitation de leur fréquence d'occurrence.

En donnant plus de poids aux mots moins fréquents.

$$tf_{t,d} = \frac{f_{t,d}}{\sum_k f_{k,d}} \quad idf_i = \log \frac{N}{n_k}$$

Limites

- Très grande dimensionnalité
- Ne capture pas la sémantique

Vectorisation (État de l'art)

Analyse sémantique latente (LSA)

L'analyse sémantique latente (LSA) [Landauer, Foltz, and Laham 1998] est basée sur la décomposition en valeurs singulières de la matrice [termes x documents] qui aboutit à sa factorisation sous la forme :

$$\begin{array}{ccccccc} & X & & U & & \Sigma & & V^T \\ & (\mathbf{d}_j) & & & & & & (\hat{\mathbf{d}}_j) \\ & \downarrow & & & & & & \downarrow \\ (\mathbf{t}_i^T) \rightarrow & \begin{bmatrix} x_{1,1} & \dots & x_{1,j} & \dots & x_{1,n} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{i,1} & \dots & x_{i,j} & \dots & x_{i,n} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{m,1} & \dots & x_{m,j} & \dots & x_{m,n} \end{bmatrix} & = & (\hat{\mathbf{t}}_i^T) \rightarrow & \begin{bmatrix} \begin{bmatrix} \mathbf{u}_1 \end{bmatrix} & \dots & \begin{bmatrix} \mathbf{u}_l \end{bmatrix} \end{bmatrix} & \cdot & \begin{bmatrix} \sigma_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_l \end{bmatrix} & \cdot & \begin{bmatrix} \begin{bmatrix} \mathbf{v}_1 \end{bmatrix} \\ \vdots \\ \begin{bmatrix} \mathbf{v}_l \end{bmatrix} \end{bmatrix} \end{array}$$

Limites

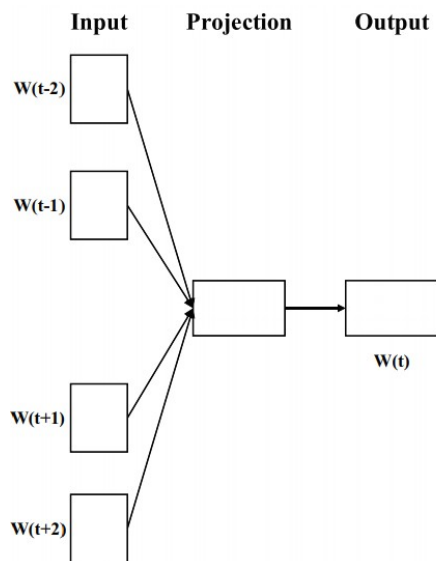
- Gourmand en mémoire
- Perte d'information due à la compression
- Ne capture pas la sémantique locale
- Coûteux algorithmiquement

Vectorisation (État de l'art)

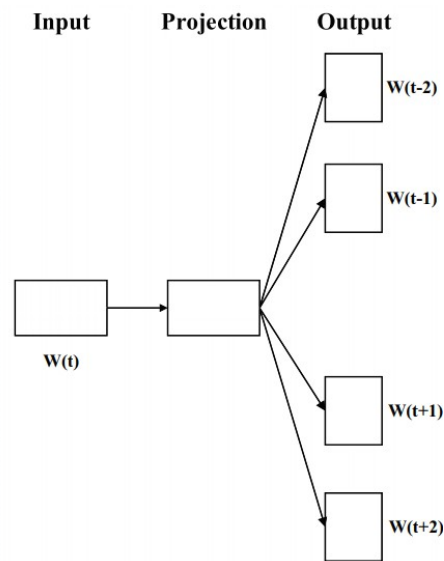
Word2vec

- Constitue une famille de modèles de plongement de mots (**word embedding**) permettant de créer des représentations vectorielles.

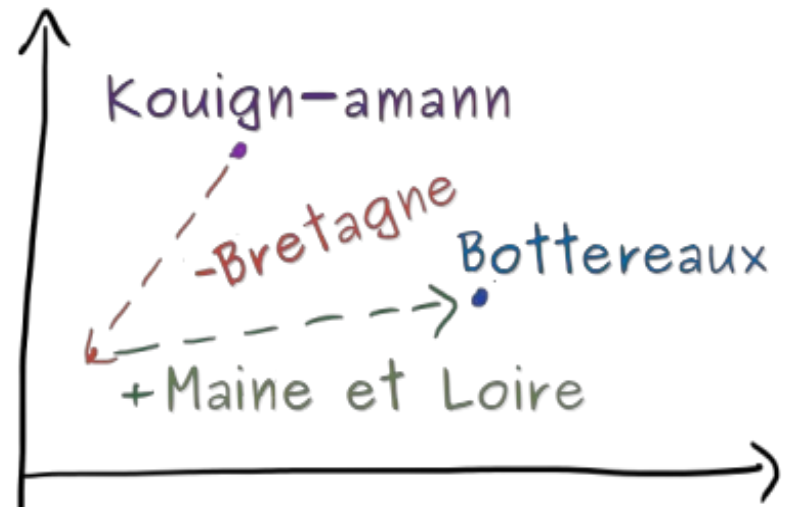
- Word2vec [Mikolov et al. 2013] peut être utilisé via deux architectures différentes : **CBOW** (sac de mots continus) et **Skip-gram** (saut de gramma).



(a) CBOW



(b) Skip-gram

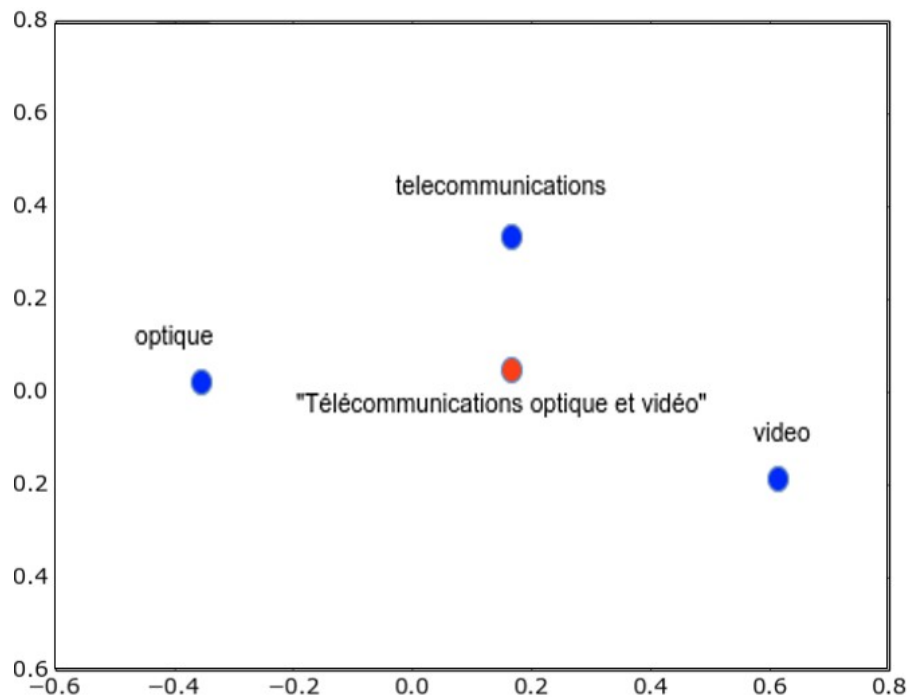


Vectorisation (État de l'art)

Représentation vectorielle (W2V)

En moyennant les représentations vectorielles des mots (word2vec) composant un texte, nous obtenons un vecteur de document de dimension fixe à partir de textes de longueurs variables.

Ainsi, les textes sémantiquement similaires ont des représentations vectorielles similaires.



Limites

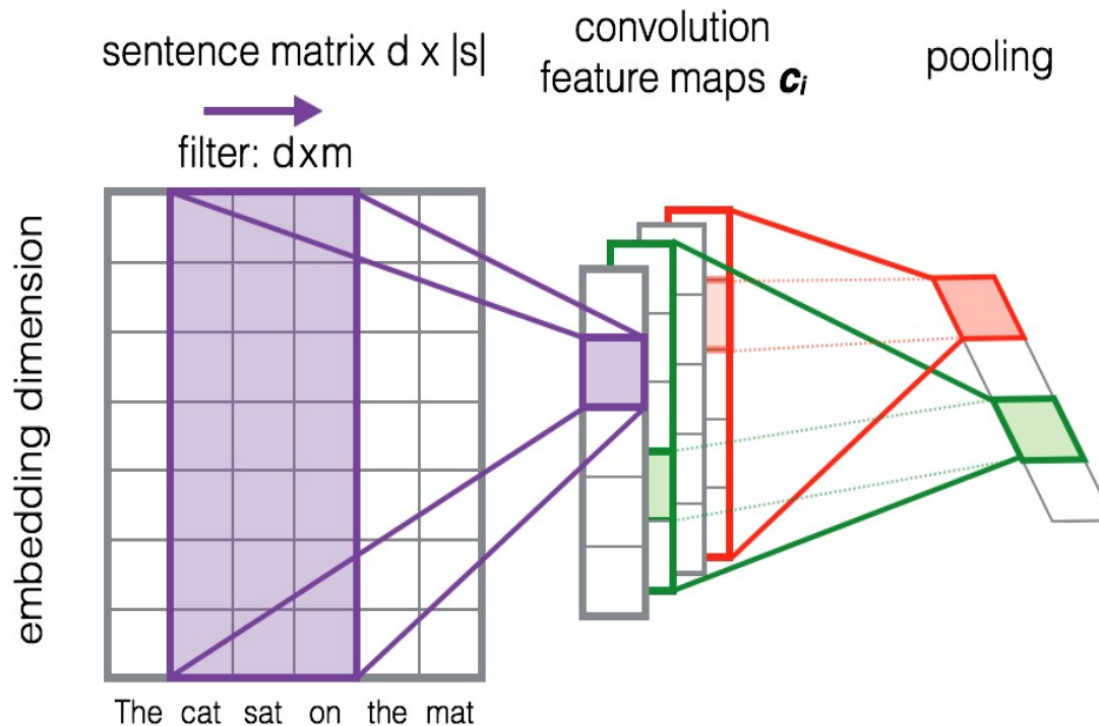
- Pertes d'information due à la moyenne
- Pratiquement impossible de vectoriser de grands documents

Autres approches :

Fast-text [Bojanowski et al. 2016],
Glove [Pennington et al. 2014].

Vectorisation (État de l'art)

Réseaux convolutionnels (CNN)

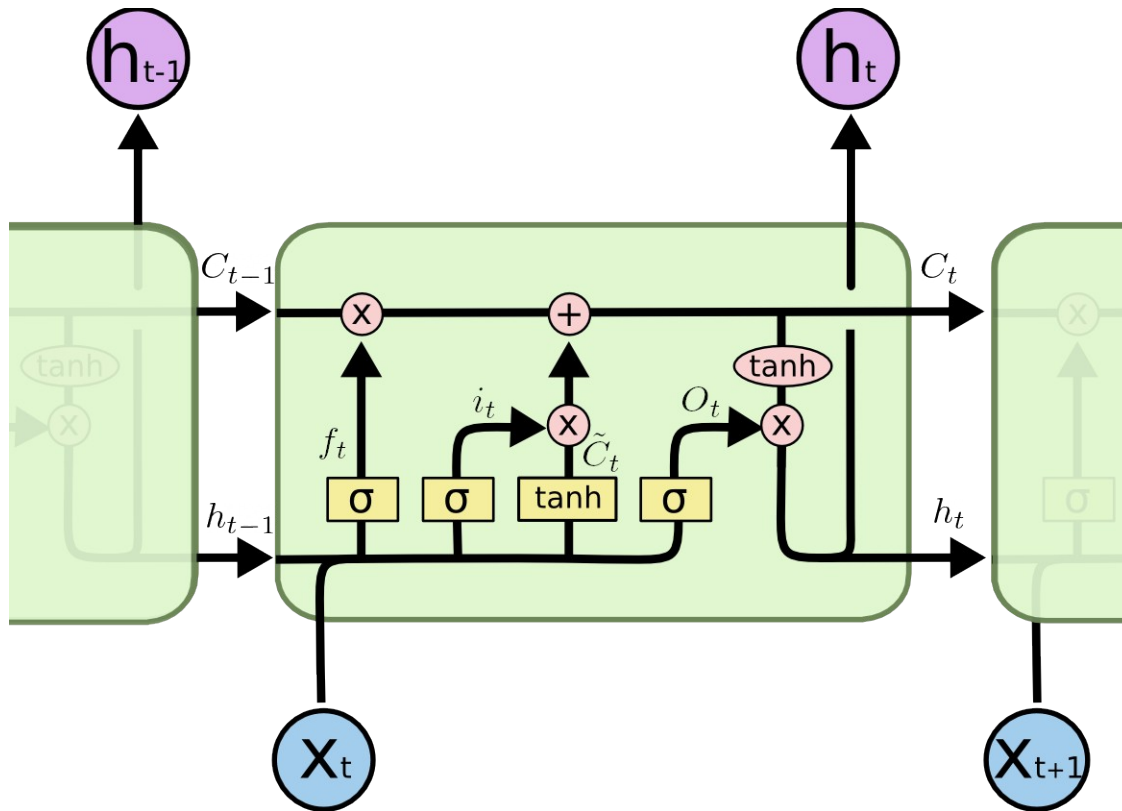


Limites

- Les vecteurs de documents dépendent d'une tâche de classification.
- Sensible aux variations des hyperparamètres.

Vectorisation (État de l'art)

Réseaux Récurrents : LSTM

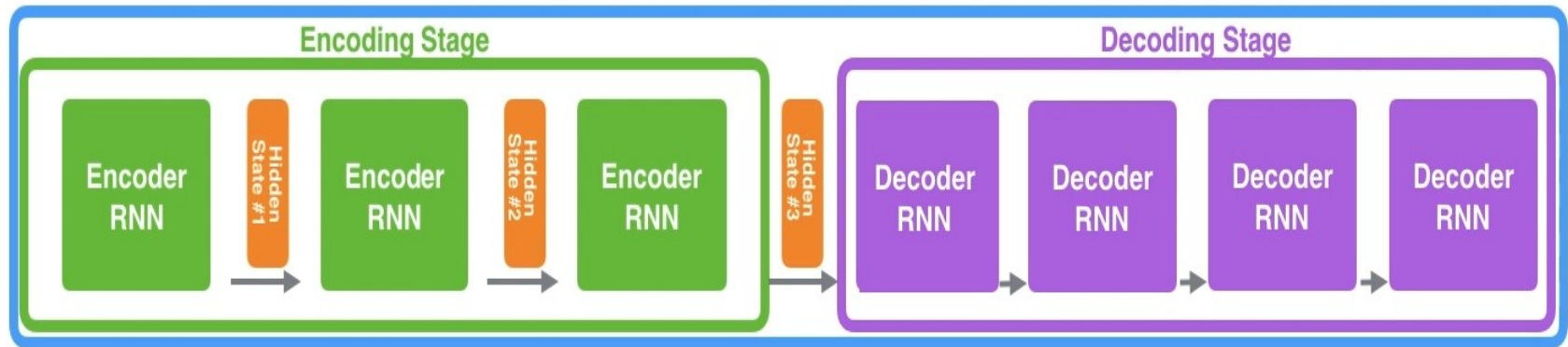


Autres variantes : **GRU** (Gated recurrent unit) [Choi et al. 2014]

LSTM(Long Short-Term Memory) [Hochreiter and Schmidhuber 1997]

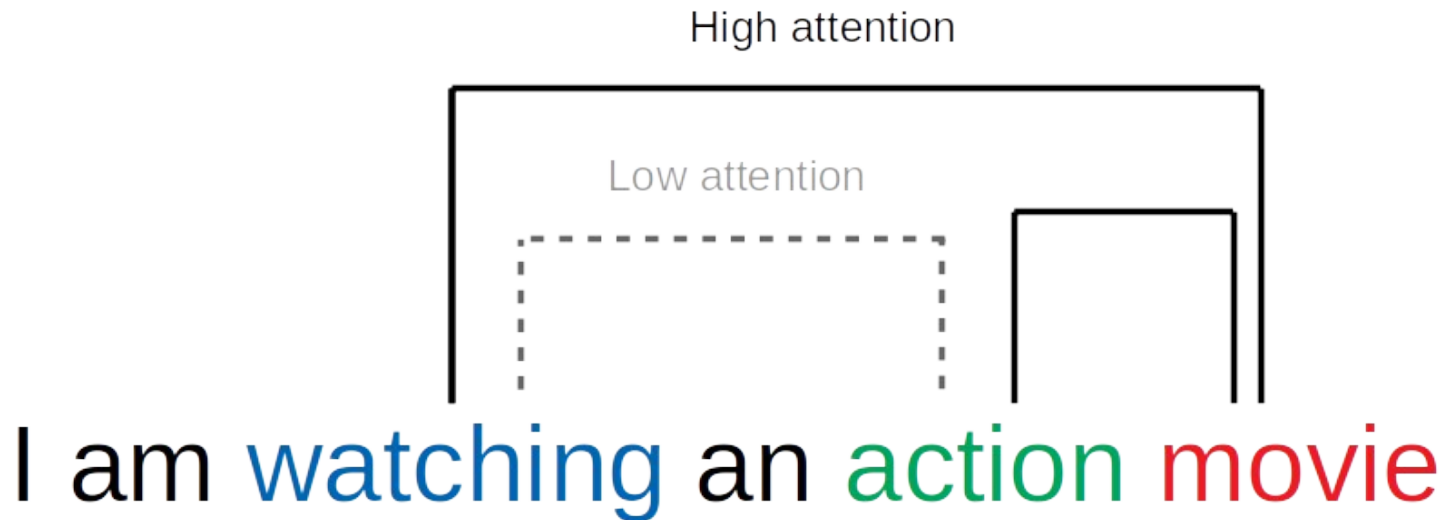
Vectorisation (État de l'art)

Mécanisme d'attention



Vectorisation (État de l'art)

Mécanisme d'attention



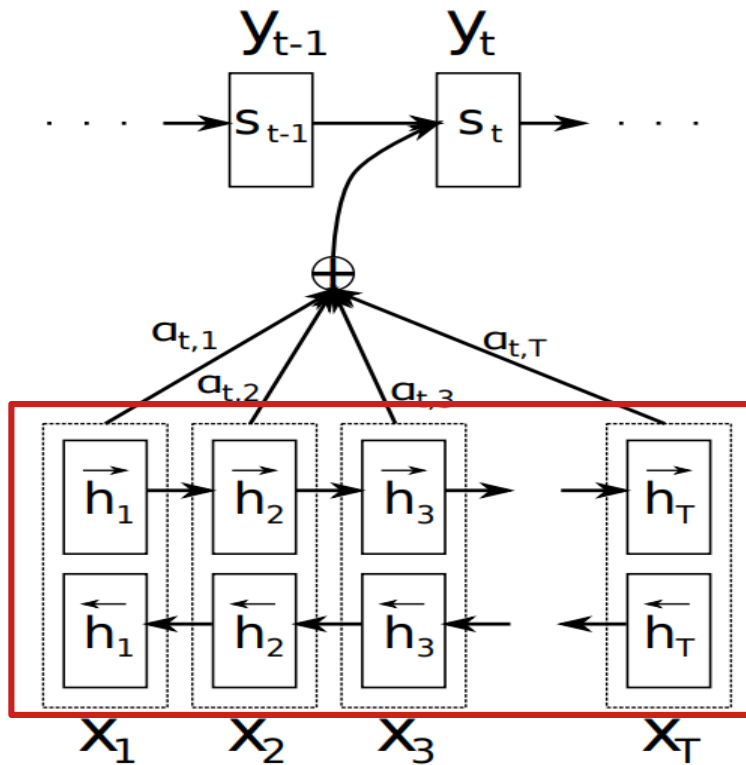
Vectorisation (État de l'art)

Mécanisme d'attention



Vectorisation (État de l'art)

Vanilla attention



$$c_i = \sum_{j=1}^{T_x} a_{ij} h_i$$

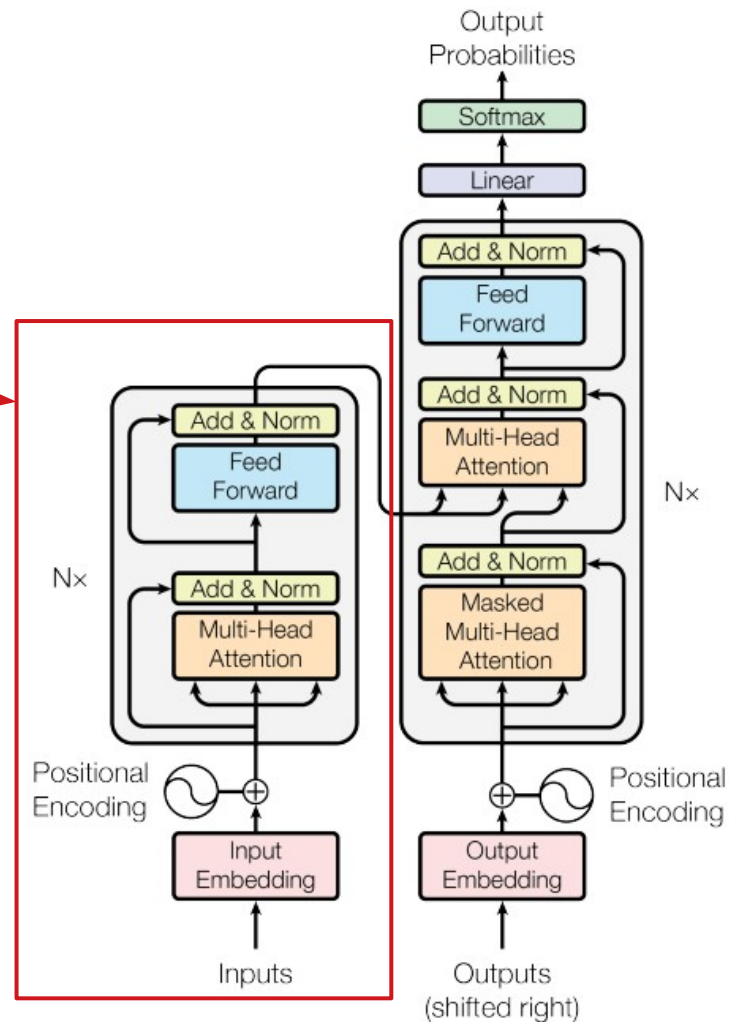
$$a_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})}$$

Bidirectionnal GRU/LSTM

Vectorisation (État de l'art)

Attention (transformer)

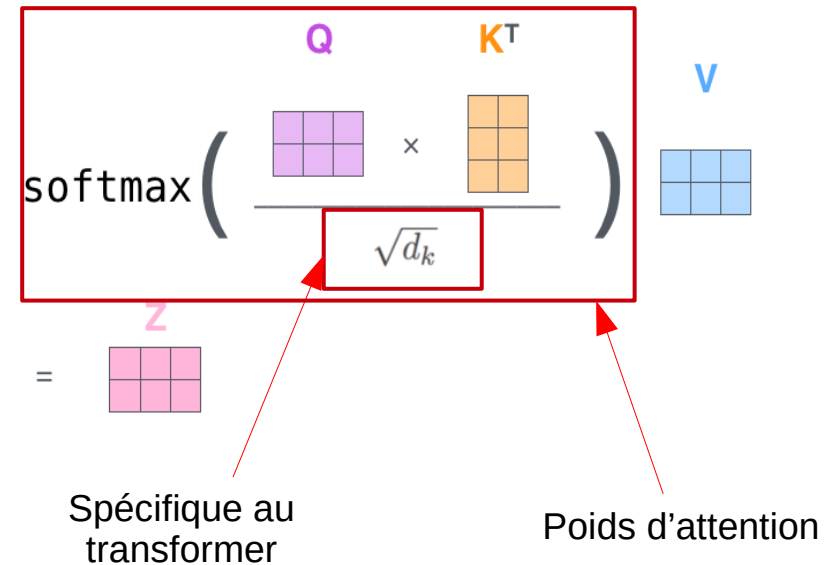
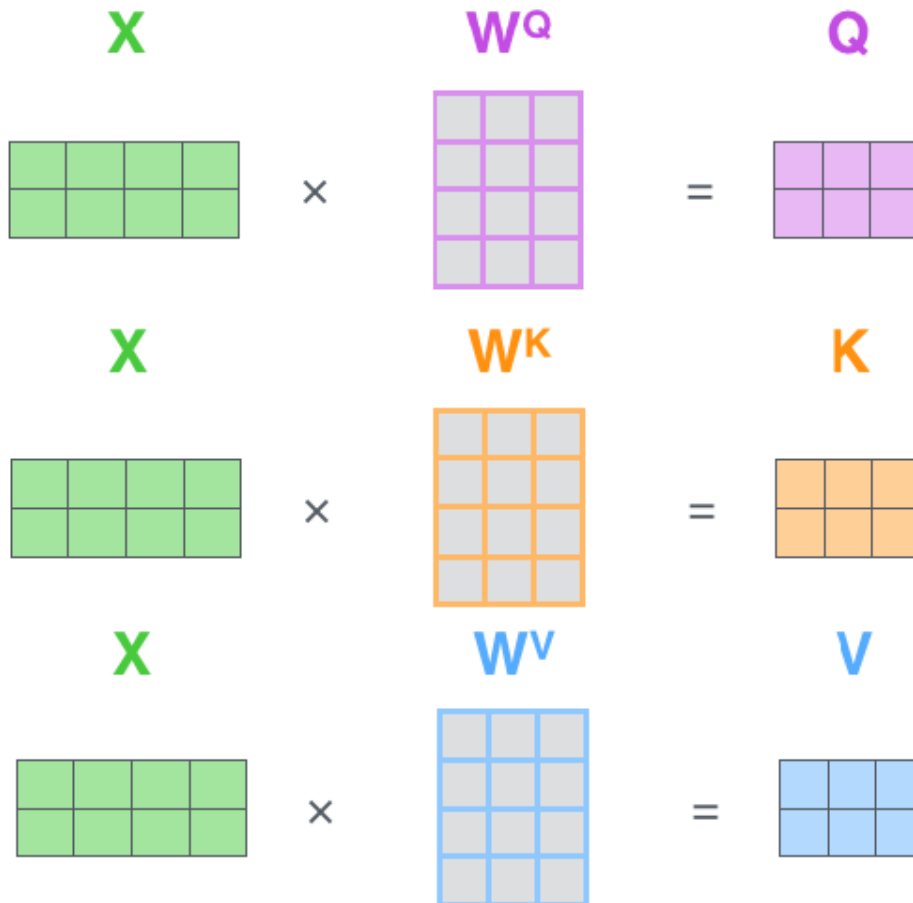
Encodeur



Transformer [Vaswani et al. 2017]

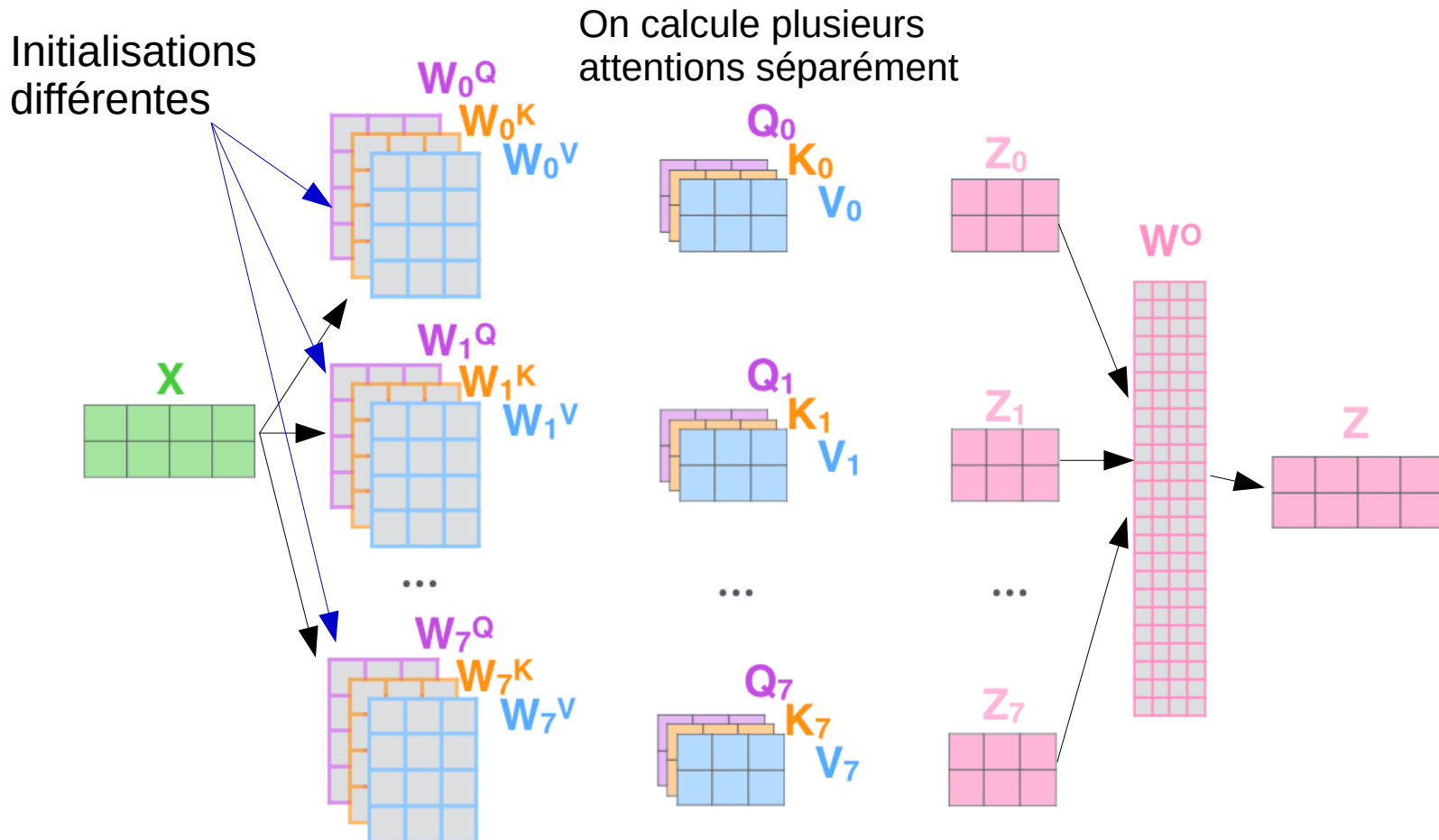
Vectorisation (État de l'art)

Attention (self attention)



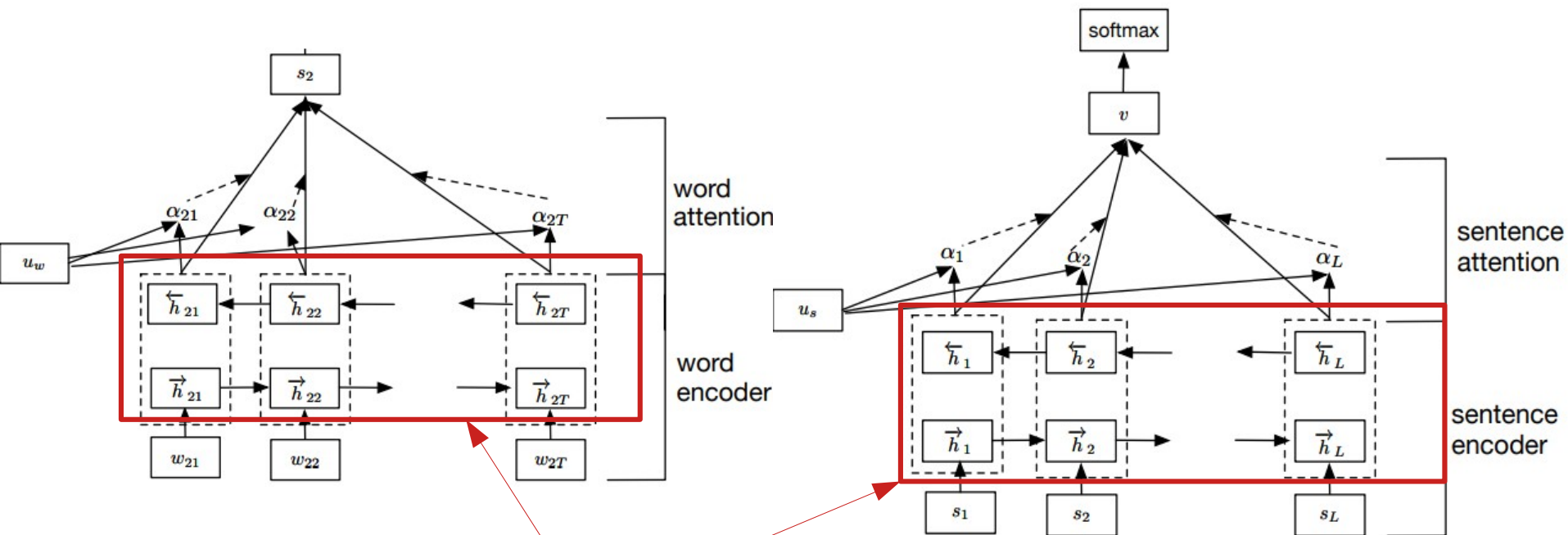
Vectorisation (État de l'art)

Attention (Multi-head attention)



Vectorisation (État de l'art)

Attention hiérarchique (HAT)



Bidirectionnal GRU/LSTM

[Yang et al. 2016]

Vectorisation (Contributions)

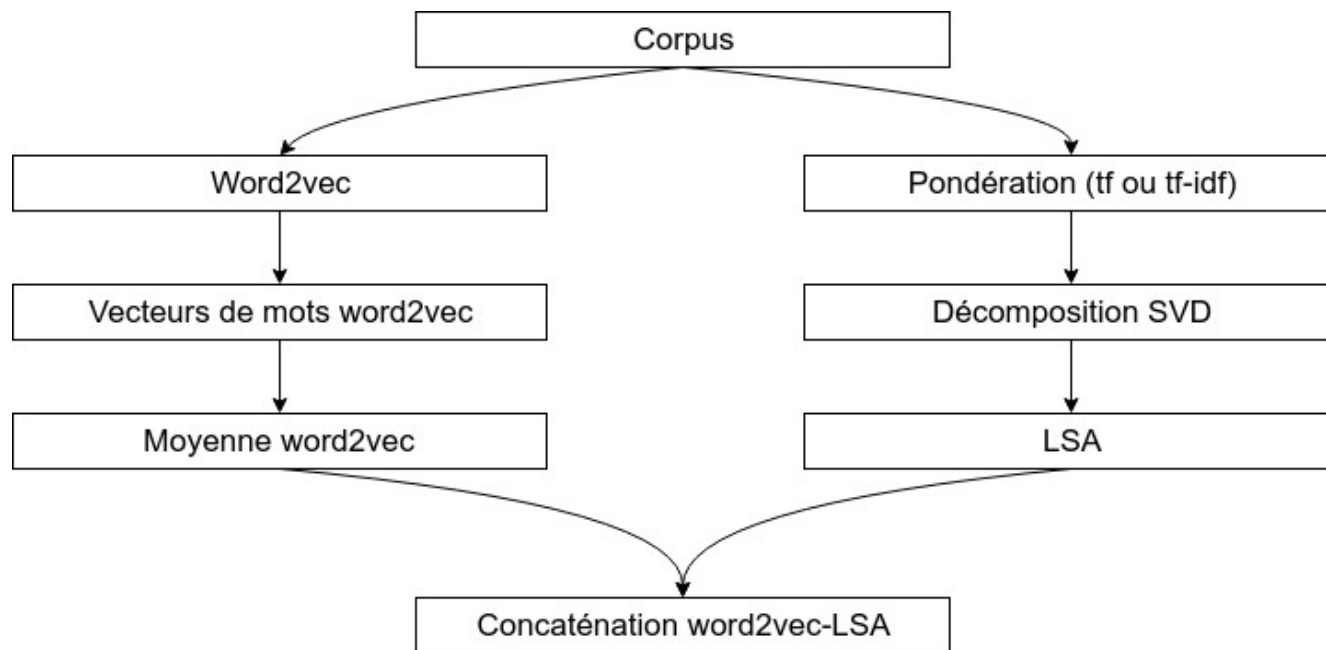
Challenges et limitations

- **Nature des documents** (fautes d'orthographe, contenu juridico- administratif, format différent selon la source)
- **Taille du vecteur** de document
- **Complexité algorithmique** faible (les modèles de vectorisation sont utilisés dans des tâches temps réels)
- **La capacité** à capturer la **sémantique** d'un document

Vectorisation (Contributions)

Contribution LSA+W2V

La concaténation de **W2V** et **LSA** permet d'avoir une représentation vectorielle basse dimension, qui permet de capturer le sens général du texte grâce à **W2V** et aussi de conserver la connaissance des occurrences des mots importants qui caractérisent le texte grâce à **LSA**.



LSA+W2V [Ahmia et al 2019]

Vectorisation (Contributions)

Contribution LSA+W2V

La moyenne **W2V** a tendance à trop généraliser, « **filtre** » le détail des termes présents dans le texte.

	alt.atheism	soc.religion.christian	talk.politics.mideast	talk.religion.misc
talk.religion.misc	23 %	16,25 %	3,4 %	59,45 %

Matrice de confusion 20newsgroups (W2V)

Exemple : les mots « **Mohamed** » et « **Jésus** » ont des représentations très similaires (les deux sont des prophètes). Ce qui peut être problématique si l'on souhaite séparer « **islam** » et « **christianisme** ».

	alt.atheism	soc.religion.christian	talk.politics.mideast	talk.religion.misc
talk.religion.misc	16,43 %	8,58 %	1,69 %	84,5 %

Matrice de confusion 20newsgroups (LSA + W2V)

Vectorisation (Contributions)

Bases de tests utilisées

Cinq jeux de données possédant les caractéristiques suivantes :

- **20NewsGroup** (20NG), **20K multi classes**, **20** catégories.
- **RCV1**, **800K**, **multi-label**, **103** catégories,
- **TED-FR**, **800K**, **multi-label**, **45** catégories.
- **TED-Filtré**, **2000K multi-label**, **45** catégories.
- **Ohsumed**, **22K multi-label**, **23** catégories.

Dans notre expérimentation, nous avons choisi **100** dimensions pour **LSA** et **W2V**, les performances sont évaluées avec différents algorithmes :

- **MLP**, Un perceptron multicouches (**200, 200**)
- **SGD**, Une machine linéaire à vecteurs supports optimisée par descente de gradient stochastique.
- **NB**, Classifieur Bayésien naïf

l'architecture **Word2vec** utilisée est **SkipGram**.

Vectorisation (Contributions)

Pré-traitements

Le prétraitement qui suit est appliqué pour toutes les méthodes utilisées :

- Les mots vides (ou **stop-words**) sont retirés
- Dans le cas spécifique des corpus issus du **fd-TED** les codes **CPV** (système de classification pour les marchés publics) contenus dans le texte sont remplacés par un mot clé neutre (**%digit%**), car ces codes définissent la classe des documents
- Les termes dont le nombre d'occurrences est inférieur à **5** sont supprimés du vocabulaire
- **bi-grammes** (collocation scoring)

Vectorisation (Contributions)

Classification

–Résultats obtenus pour les méthodes testées sur le jeu de données TED-FR :

	Exactitude	Précision	Rappel	F1-score
MLP(LSA+W2V)	55,27	78,2	57	64
MLP(LSA tf-idf)	49,65	76,8	49,4	55,4
SGD(tf)	48,32	78,2	59,6	62,8
MLP(W2V)	45,23	76,6	45	53
MLP(LSA tf)	33,13	64,4	31,8	37

Vectorisation (Contributions)

Classification

– Résultats obtenus pour les méthodes testées sur le jeu de données TED-FILTRE :

	Exactitude	Précision	Rappel	F1-score
MLP(LSA+W2V)	77,47	95,2	75	83,2
MLP(W2V)	75,6	95,2	72,6	81,4
MLP(LSA tf-idf)	69,38	94,4	66,4	76,6
NB(tf-idf)	66,07	88	66,2	74,6
SGD(LSA+W2V)	61,22	91	59,2	69,8

Vectorisation (Contributions)

Classification

Résultats obtenus pour les méthodes testées sur le jeu de données RCV1 :

	Exactitude	Précision	Rappel	F1-score
MLP(LSA+W2V)	58,51	87	81	83,6
MLP(W2V)	57,86	87,4	80	83
MLP(LSA tf-idf)	54,64	86	77	80,4
SGD(tf-idf)	49,67	91,4	69,2	76,2
MLP(tf)	49,41	84,2	72,2	76,6

Vectorisation (Contributions)

Classification

Résultats obtenus pour les méthodes testées sur le jeu de données Ohsumed:

	Exactitude	Précision	Rappel	F1-score
MLP(LSA+W2V)	72,58	89,36	90,05	89,66
SGD(tf)	71,45	87,98	90,25	89,06
MLP (LSA tf-idf)	70,88	88,47	89,50	88,94
MLP (W2V)	69,25	87,48	87,52	87,44
SGD (tf-idf)	64,51	93,28	84,40	88,51

Vectorisation (Contributions)

Classification

Résultats obtenus pour les méthodes testées sur le jeu de données 20NG:

	Exactitude	Précision	Rappel	F1-score
SGD(tf-idf)	92,92	92,8	92,8	92,8
NB(tf-idf)	90,07	90,6	90	89,8
MLP(LSA+W2V)	84,46	84,6	84,6	84,6
NB(tf)	83,96	86,2	83,8	83,4
MLP(LSA tf-idf)	81,86	82	81,8	81,6

Vectorisation (Contributions)

Clustering

Afin de comparer les différentes représentations vectorielles dans une tâche de classification non supervisée, on utilise un **Spherical kmeans** (similarité cosinus) sur la base **20NewsGroup**.

	Adjusted Rand-Index
W2V+LSA	26%
W2V	22%
LSA	18%
TfIdf	9%

Vectorisation (Contributions)

Clustering (LSA+W2V)

En se basant sur une représentation (LSA+W2V), les algorithmes de **clustering** sont capables de regrouper les documents par domaine, ce qui nous permet de mieux comprendre le domaine d'activité et les compétences des clients d'OctopusMind mais aussi le suivi des tendances en matière d'appels d'offres.

CLUSTER N°55, K=10K:

prestations de traiteur et réception pour un déjeuner offert aux personnes âgées de champigny-sur-marne pour une durée de trois ans.

=====
confection et service d'un déjeuner pour les aînés le dimanche 27 octobre 2013
=====

fourniture de prestations de traiteur et de service à table pour deux banquets seniors organisés en 2013.

=====
confection et service d'un déjeuner pour les aînés le dimanche 28 octobre 2012
=====

fourniture, livraison et service d'un dîner pour les seniors de la commune

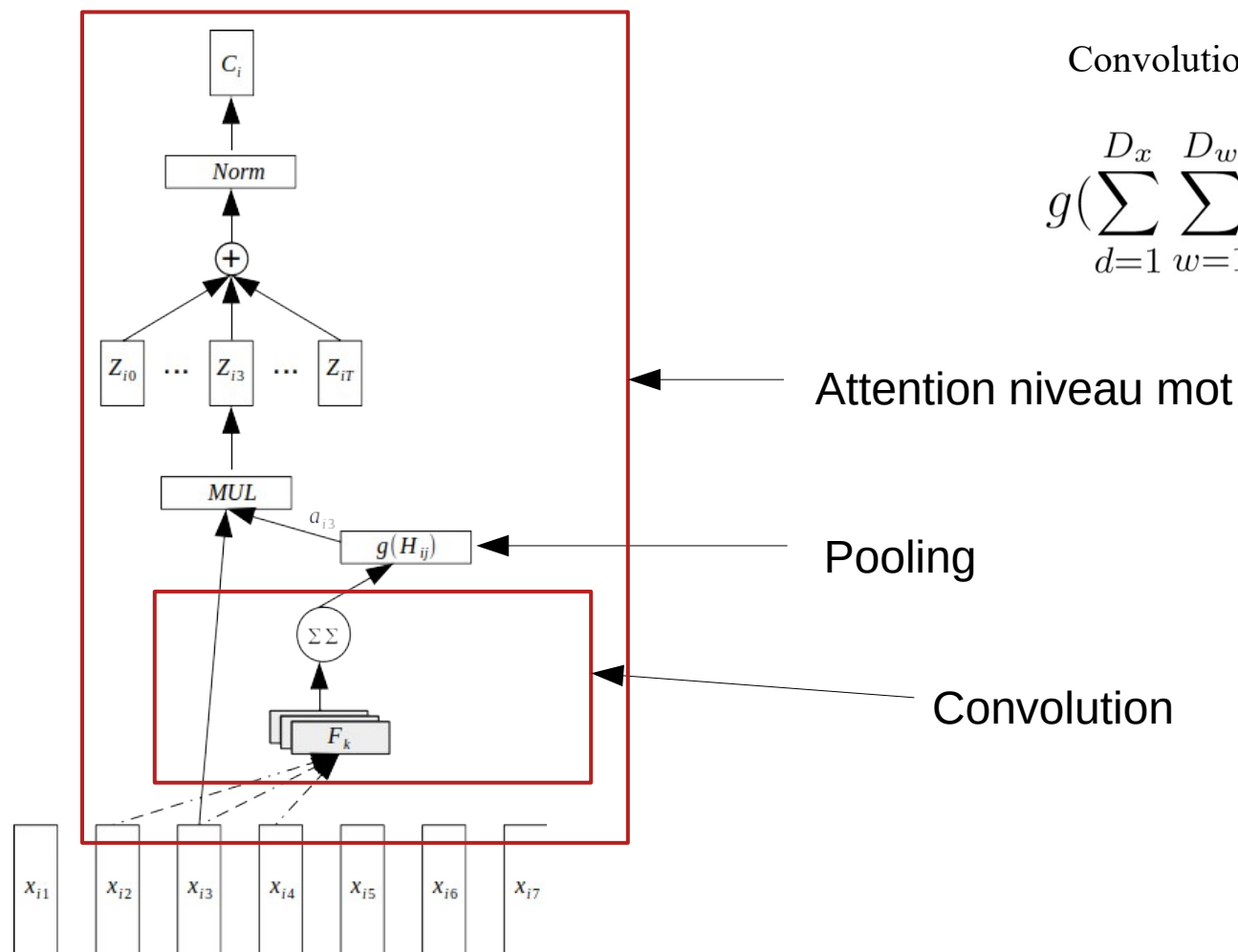
=====
service d'un traiteur le 11 decembre 2011 a l'occasion d'un gouter offert a la rpa roland ricordeau.

=====
confection et service d'un déjeuner pour les aînés le dimanche 30 octobre 2011
=====

...

Vectorisation (Contributions)

Contribution (CnHAtt)

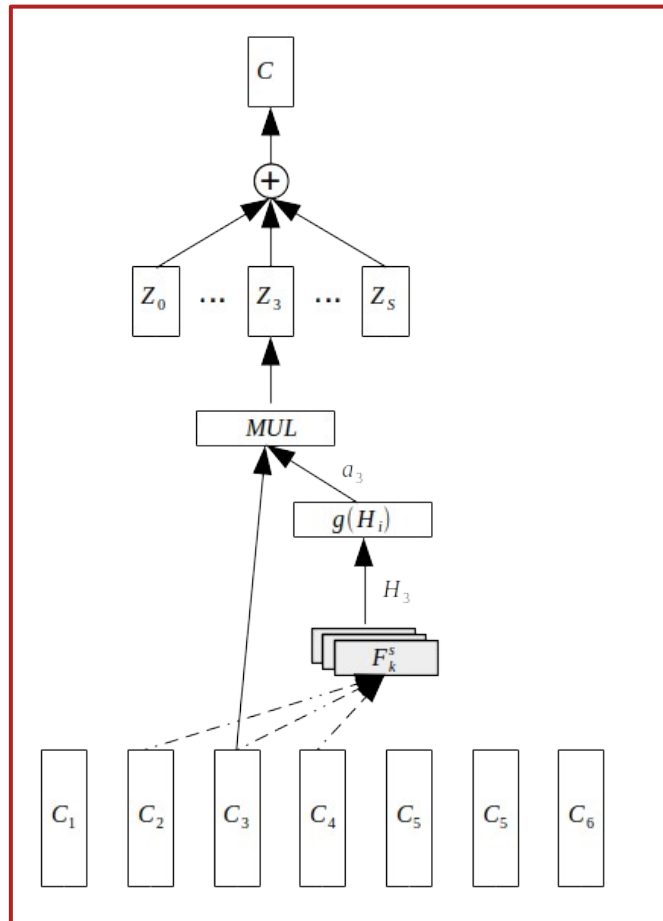


Convolution based Hierarchical Attention

$$g\left(\sum_{d=1}^{D_x} \sum_{w=1}^{D_w} (F_i \odot X_{(sw, sw+D_w)})\right)$$

Vectorisation (Contributions)

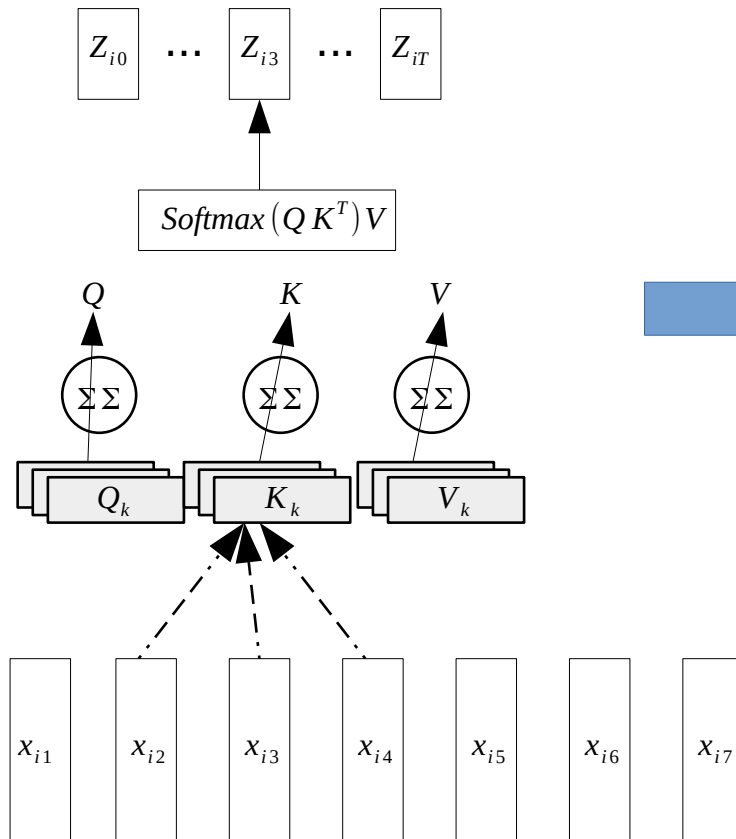
Contribution (CnHAtt)



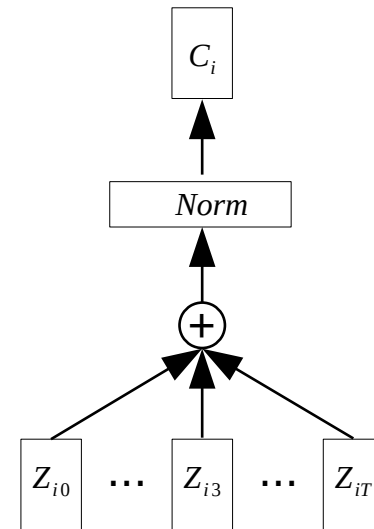
← Attention niveau phrase

Vectorisation (Contributions)

Contribution (CnHAtTr)



Convolution based Hierarchical Attention
Self-attention variant



Vectorisation (Contributions)

Bases de tests utilisées

Quatre jeux de données possédant les caractéristiques suivantes :

- **20NewsGroup** (20NG), **20K multi classes**, **20** catégories.
- **RCV1**, **800K**, **multi-label**, **103** catégories,
- **TED-FR**, **800K**, **multi-label**, **45** catégories.
- **Ohsumed**, **22K multi-label**, **23** catégories.

Le modèle (**CnHAtt**) est comparé avec HAT ainsi que d'autres méthodes de vectorisation (**TF**, **TF-Idf**, **W2V**, **LSA**). Les deux modèles sont connectés à une couche **MLP**.

- **100** dimensions pour **LSA** et **W2V**

les performances sont évaluées avec différents algorithmes :

- **MLP**, Un perceptron multicouches (**200, 200**)
- **SGD**, Une machine linéaire à vecteurs supports optimisée par descente de gradient stochastique.
- **NB**, Classifieur Bayésien naïf

l'architecture **Word2vec** utilisée est **SkipGram**.

Vectorisation (Contributions)

Classification

	Accuracy	Precision	Recall	F1-score
SGD (tf-idf)	92,92	92,8	92,8	92,8
CnHAtt	90,29	96,38	94,19	90,63
NB (tf-idf)	90,07	90,6	90	89,8
HAT	88,24	94,37	92,30	88,72
MLP (LSA+W2V)	84,46	84,6	84,6	84,6
NB (tf)	83,96	86,2	83,8	83,4
MLP (LSA tf-idf)	81,86	82	81,8	81,6
SGD (LSA tf-idf)	78,99	81,2	79	79
NB (LSA tf-idf)	74,1	75,4	74,4	74,4
MLP (W2V)	73,43	73,6	73,6	73,6
SGD (LSA+W2V)	71,49	80,8	71,4	73
SGD (tf)	71,12	83,2	71	74,6
NB (LSA+W2V)	68,46	71,6	68,4	69
MLP (LSA tf)	59,37	59,2	59,4	58,4
SGD (LSA tf)	54,16	68,4	54,2	54,4
NB (W2V)	51,95	55	51,8	52
SGD (W2V)	51,16	65	51,2	52,6
NB (LSA tf)	39,23	51,2	39,2	41,6

Vectorisation (Contributions)

Classification

	Accuracy	Precision	Recall	F1-score
CnHAtt	84,89	99,49	88,95	89,74
HAT	84,40	99,13	82,14	89,25
MLP (LSA+W2V)	55,27	78,2	57	64
MLP (LSA tf-idf)	49,65	76,8	49,4	55,4
SGD (tf)	48,32	78,2	59,6	62,8
MLP (W2V)	45,23	76,6	45	53
MLP (LSA tf)	33,13	64,4	31,8	37
SGD (LSA+W2V)	32,8	66,6	34,4	41
SGD (W2V)	29,9	64,4	31,8	38,4
SGD (LSA tf-idf)	17,63	53	16,6	22
NB (tf)	16,65	34,6	81,2	45,6
SGD (tf-idf)	16,35	73,8	15	20,4
NB (LSA tf-idf)	14,6	32	45,6	35,2
NB (tf-idf)	13,99	68,8	13,4	19,8
NB (LSA tf)	11,41	28,2	41,8	30,4
NB (LSA+W2V)	6,08	26,4	60,8	33,2
SGD (LSA tf)	6,06	30,2	5,8	9
NB (W2V)	5,33	19,8	48,2	25,6

Vectorisation (Contributions)

Classification

	Accuracy	Precision	Recall	F1-score
CnHAtt	78,12	92,99	92,25	92,57
HAT	77,08	92,16	91,52	91,81
MLP (LSA+W2V)	72,58	89,36	90,05	89,66
SGD (tf)	71,45	87,98	90,25	89,06
MLP (LSA tf-idf)	70,88	88,47	89,50	88,94
MLP (W2V)	69,25	87,48	87,52	87,44
SGD (tf-idf)	64,51	93,28	84,40	88,51
MLP (LSA tf)	50,95	84,01	77,13	80,23
NB (tf)	25,80	62,87	90,09	73,31
SGD (LSA+W2V)	19,64	71,96	49,67	56,84
SGD (LSA tf-idf)	19,47	74,52	43,34	53,33
NB (tf-idf)	15,25	88,64	34,20	44,55
NB (LSA tf-idf)	12,94	54,46	53,57	53,33
SGD (W2V)	11,82	63,40	34,42	41,61
SGD (LSA tf)	10,25	69,34	25,13	34,68
NB (LSA tf)	3,29	33,28	47,72	37,13
NB (LSA+W2V)	3,15	38,56	66,54	47,25
NB (W2V)	1,31	32,03	61,63	40,39

Vectorisation (Contributions)

Classification

	Accuracy	Precision	Recall	F1-score
CnHAtt	64,60	89,62	85,30	86,93
HAT	63,15	89,96	83,45	85,37
MLP (LSA+W2V)	58,51	87	81	83,6
MLP (W2V)	57,86	87,4	80	83
MLP (LSA tf-idf)	54,64	86	77	80,4
SGD (tf-idf)	49,67	91,4	69,2	76,2
MLP (tf)	49,41	84,2	72,2	76,6
SGD (tf)	44,84	81,6	75,6	77,8
SGD (LSA+W2V)	38,32	79	71	73,4
SGD (LSA tf-idf)	36,42	82,8	60,6	67,2
SGD (W2V)	32,99	76,6	67,4	70
SGD (tf)	18,81	78,4	44	53
NB (LSA tf-idf)	4,85	40,8	77,2	50,2
NB (LSA+W2V)	2,58	40,6	85,2	50,4
NB (W2V)	1,54	38,6	85,2	48,4
NB (tf)	0,55	27,4	72,8	36,2
NB (tf)	0,29	4,4	0	0
NB (tf-idf)	0,29	0	0	0

Vectorisation (Contributions)

Green computing

Notre modèle « **CnHAtt** » est **10** fois plus rapide comparé à « **HAT** »

Détail de l'expérimentation :

- CPU: 2 x Intel(R) Xeon(R) Gold 6140 CPU @ 2.30GHz, GPU: GeForce RTX 2080 Ti (260 watts).
- Batch de 512 documents.

	CnHAtt	HAT
Training	51 <i>ms</i>	498 <i>ms</i>
Prediction	17 <i>ms</i>	118 <i>ms</i>

- Notre modèle économise **1,87 kwh (90% moins)** sur une durée de **8h**.

Vectorisation (Contributions)

Clustering

Résultats du clustering (adjusted rand-index) :

	CnHAtt	HAT
20NewsGroup	0.50	0.48
RCV1	0.45	0.43
TED-FR	0.46	0.61
ohsumed	0.25	0.16

Kmeans

	CnHAtt	HAT
20NewsGroup	0.38	0.37
RCV1	0.33	0.30
TED-FR	/	/
ohsumed	0.29	0.21

Clustering hiérarchique

Résultats du clustering (adjusted rand-index) en utilisant **Spherical Kmeans**, sur une hiérarchie de classe différente que celle utilisée durant l'entraînement du modèle.

* **0.23** pour **CnHAtt** et **0.17** pour **HAT**

Vectorisation (Contributions)

CnHAtt (Visualisation)

14.25 when trying_to choose a resistor with a tolerance better_than %digit% you
5.95 need a or to screen devices it can_t be made from adding %digit%
4.88 of %digit% value in parallel since the smaller device will have the
2.43 error of %digit% to cope with
2.49 you have %digit% choices
4.58 a live with the error of %digit% tolerance devices for low q circuits or low
3.96 sensitivity designs
8.69 b buy resistors with better_than %digit% tolerance dale
5.51 c use or s select on test

Un exemple de poids d'attention (mots et phrases) sur 20NewsGroup, (classe : sci.electronics)

Vectorisation (Contributions)

CnHAtt (Visualisation)

f paris services d installation materiel radio television audio et video
%digit% s %digit% %digit%
france televisions %digit% esplanade_henri france attn jean_claude france 75907paris
supplement au journal_officiel l union_europeenne %digit% %digit% %digit%
%digit% s %digit% %digit%
objet
cpv
%digit%
services d installation materiel radio television audio et video
procedure incomplete
la procedure passation ete_interrompue
autres informations complementaires
avis annuler pour cause doublon avec le %digit% %digit%

Un exemple de poids d'attention (mots et phrases combinés) sur TED-fr, (class : Installation services of communications equipment)

Applications

Système de recommandation

Bienvenue !

Vous êtes à 2 doigts de consulter des **millions d'offres et d'opportunités commerciales**
Nous aimerions en savoir un peu plus sur vous pour vous proposer les **meilleures offres**

Votre Pays*

France (FR)

Votre entreprise

Secteurs d'activités recherchés*

Activités

Pays

Actes notariés, créances, enchères, courtage, experts judiciaires

Recouvrement de créances, services d'huissiers, actes notariés, services d'enchères, courtage, experts judiciaires, commissaire-priseur, détectives

Action sociale et humanitaire

Insertion sociale, action humanitaire et Droits de l'Homme

Administration des systèmes informatiques

Administration, sécurité, exploitation, maintenance des systèmes informatiques et infrastructures réseau

Aménagement d'intérieur, ergonomie

Aménagement d'espaces, architecture d'intérieur, ergonomie, décoration d'intérieure

✓ Valider et utiliser J360

Applications

Système de recommandation

Consulting Services to conduct UNESCO's IT security audit

📄 SUGGESTIONS DE MARCHÉS

Consulting Services to conduct UNESCO's IT security audit

UNESCO
France - Paris (75)

Multifunctional Devices, Managed Print and Content Services and Records and Information Management – lot 7 : Audit and Consultancy Services

The Minister for the Cabinet Office acting ...
Royaume-Uni

Framework agreement for consultancy services within security (security advisor)

Stortingets Administrasjon (The Norwegia...
Norvège

ICT security and business continuity planning consultancy services

Eurojust
Pays-Bas

ICT security and business continuity planning consultancy services

Eurojust
Pays-Bas

Applications

Analyse de marchés



TED Corpus



Unwatch ▼

5



Unstar

7



Fork

0

<https://github.com/oussamaahmia/TED-dataset>

Shukuria
Tashakkur
bolzin
You
Merci
Gracias
Thank
Biyan
Grazie
Juspaxar
Mehrbani
Arigato
Dankscheen
Komapsumnida
Shukria
Paldies
Hatur
anitha
Make
gozainashita
Fataane
Suzabo
Ekhmet
Nenachalhya
Bala Vasqegatum
Memocher Atto
Guejho
Sukomo
Sikomo
Tavtapuch
Mabaka
elajaj
Shukria
Lah
Dhanyabad
Chaltu
nuhun
Snachalhuya
Mednagge
Merzi
unachheer
Tingki
Wabaja
Sanco
Hil
Makici
Gul
Yaqhanyelay
Efcharisto
Gul