Construction d'une liste de multitermes pour l'étude de la compositionnalité dans les modèles distributionnels

WANG Yizhe

09/12/2019

CLLE (UMR5263)

Plan

Introduction

Introduction - Inspiration du travail

Article: «Évaluation des modèles sémantiques distributionnels: le cas de la dérivation syntaxique» [1]

Objectif de l'article:

- Évaluer la performance du modèle AD et celle de W2V via un jeu de données composé de termes simples reliés par des relations sémantiques
- Analyser l'influence des paramètres sur la performance des modèles.

Modèle AD

- Type de fenêtre de contexte : G+D ou G&D.
- Forme de la fenêtre : rectangulaire ou triangulaire.
- Taille de la fenêtre
- Pondération : log, MI, z-score, etc..

Modèle W2V

- Architecture : CBOW ou skip-gram.
- Sous-échantillonnage : seuil faible / seuil élevé aucun sous-échantillonnage.
- Taille de la fenêtre de contexte
- Dimension des représentations : 100 ou 300.

Introduction

Objectif: Étudier la compositionnalité des multitermes dans les modèles distributionnels via un jeu de données composé de termes complexes reliés par des relations sémantiques.

Jeux de données demandé:

- Une liste de paires de candidats MWTs (CCT_1 et CCT_2) respectant les conditions suivantes:
  - CCT_1 et CCT_1 sont deux candidats composés de deux mots lexicaux.
    $\exists M_1, M_2 \in \mathrm{CCT}_1 \qquad \exists M_3, M_4 \in \mathrm{CCT}_2$
  - $sem(CCT_1, CCT_2)$
  - $M_2 = M_4$
  - $sem(M1, M3)$
    Ex. *agriculture organique* et *agriculture biologique*

Introduction

Si

CCT_1 est compositionnel, alors

$sens(CCT_1) = sens(M_1) \oplus sens(M_2)$

Si

CCT_2 est compositionnel, alors

$sens(CCT_2) = sens(M_3) \oplus sens(M_4)$

On ne prend en compte que les couples: $M_2 = M_4$

Alors

$sens(CCT_1) = sens(M_1) \oplus etc.$

$sens(CCT_2) = sens(M_3) \oplus etc.$

Si $sem(M_1, M_3)$, alors

$sem(sens(M_1), sens(M_3))$

Introduction

Si

$sens(CCT_1) = sens(M_1) \oplus sens(M)$

$\wedge \text{sens}(\text{CCT}_2) = sens(M_2) \oplus sens(M)$

$\wedge \text{sem}(M_1, M_2)$

Hypothèses:

- $sem(CCT_1, CCT_2)$
- $sem(sens(CCT_1), sens(CCT_2))$
- si $sem(V_1, V_2)$, alors $sem(V_1 + A, V_2 + A)$

Introduction

- Reddy et al. (2011): 90 composés nominaux anglais avec un score moyen de 30 jugements de 0 à 5 (considérer la compositionnalité comme la littéralité).
  Ex. *climate change* (4.97)
- Farahmand et al. (2015): 1.042 composés nominaux anglais annotés par quatre experts avec jugement binaire pour non-compositionnalité (le sens ne peut pas être facilement interprétée par les sens de ses constituants) et conventionalité (composés collocatifs dont les constituants coexistent souvent).
  Ex. *flag stop* (non-compositionnel)
  wish list (compositionnel mais conventionnel)
  animal life (compositionnel et non conventionnel)
- Carlos et al. (2016): Composés nominaux de 3x180 en anglais, français et portugais avec scores de compositionnalité pour la tête, l'expansion et l'ensemble de l'expression.
  EX. *climate change*(tête: 4.8, expansion: 4.9, compositionnel: 5.0)

Introduction

Limites:

- La plupart des jeux de données sont en anglais
- La plupart des jeux de données sont des composés nominaux
- Problème de la polysémie
- Difficile de les utiliser pour évaluer les modèles distributionnels sur la compositionnalité dans un cadre contrôlé
- Pas d'informations sur la relation sémantique entre les MWEs.

Compositionnalité et modèle distributionnel

Compositionnalité

Principe de Compositionnalité

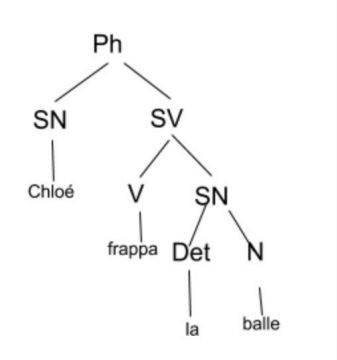
«Le sens d'une expression complexe est déterminé par le sens de ses composants et les règles selon lesquels ils sont combinés.» [9]

Ex. *climat chaud*

Compositionnalité

Compositionnalité aux niveaux différents

- Composition des phrases
  - La représentation syntaxique: l'arbre syntaxique.
    Ex. *Chloé frappa la balle.*



  sens(*Ph*) = *sens*(*Chlo*)⊕sens(frappa)⊕sens(balle)
- Compositionnalité au niveau d'expressions multi-mots (MWEs)
  - MWEs & MWTs
    Ex. *réchauffement climatique*
- Compositionnalité au niveau des mots
    Ex. *disappearance*

Compositionnalité

- **MWEs compositionnelles**
  Ex. *climat chaud*
- **MWEs semi-compositionnelles**: véhicule un sens inhabituel et transformé sur le plan sémantique. Ce sont souvent des collocations [6]
  Ex. *peur bleue*
- **MWEs non-compositionnelles**: aucune analyse décompositionnelle n'est possible
  Ex. *les doigts dans le nez*

Compositionnalité

Problème souvent soulevé dans la pratique: l'ambiguïté [4]

- Le problème de polysémie
  Ex. *air frais: frais ou froid*
- L'ambiguïté de la structure syntaxique
  Ex. *Il a mangé les biscuits sur le canapé*

Hypothèse distributionnelle

Hypothèse distributionnelle:

Les mots utilisés dans les mêmes contextes ont tendance à avoir des significations similaires. [5]

Modèles distributionnels souvent utilisés:

- Fasttext
- Word2vec (Skip-gram & Cbow)
- Glove

Construction de la ressource

Corpus

Projet PANACEA [1]: visant à construire des corpus en plusieurs langues spécialisés dans le domaine de l'environnement.

Corpus utilisé

- Corpus monolingue français construit dans le cadre du projet PANACEA
- **Langue**: français
- **Domain**: environnement
- **Taille**: 50 million mots
- **Format**: XML
- **Prétraitements**: l'extraction de texte à partir de documents XML, normalisation des caractères, minuscule et lemmatisation par TreeTagger

[1] http://www.panacea-lr.eu/en/info-for-researchers/data-sets/monolingual-corpora

DicoEnviro

DicoEnviro: un dictionnaire spécialisé dans le domaine de l'environnement.

écosystème 1 , n. m.

un écosystème : ~ de *zone* 1

Contextes
Liens lexicaux

Explication	Lexie reliée
Voisins	
≈	biome 1 biosphère 1 environnement 1
Autres parties du discours et dérivés	
Dans un é.	dans un ~
Sortes de	
Types d'é.	delta 1 mangrove 1 marais 1 plage 1 terre 1 toundra 1
Combinatoire	
L'é. commence à être différent	l'~ change 1a

Liste de référence

Extrait de la référence

TERM1	POS1	TERM2	POS2	RELATION
infrastructure	NN	arrêt	NN	HYP
absorber	VV	réfléchir	VV	ANTI
apiculteur	NN	éleveur	NN	QSYN
bois	NN	matière	NN	HYP
boisement	NN	plantation	NN	QSYN
boisement	NN	déboisement	NN	ANTI

	Anti	Hyp	Qsyn	Total
Nombre	173	193	689	1055
Pourcentage	16.40%	18.29%	65.31%	100%

Construction de la ressource

Définitions des relations

- QSYN (quasi-synonymes) : les sens voisins (*fragile* et *vulnérable*) et les génériques (ex. *environnement* et *écosystème*)
- ANTI (antonymes) : contraires (*froid* et *chaud*) et contrastifs (ex. *faune* et *flore*)
- HYP (hyperonyme - hyponyme) : relations hiérarchiques (*bois* et *matière*)

TermSuite

- **Développeur**: Béatrice Daille et al. [2]
- **Application**: Extraction et alignement de terminologie
- **Entrée**: corpus / corpus parallèle
- **Langues**: anglais, français, allemand, espagnol, letton, chinois et russe

Sortie de TermSuite

type	pattern	pilot	freq	spec
V[s]	N A A	zone intertropicale humide	6	0,97
V[s]	N P N A	conservation des zones humides	42	1,78
T	N A	parc national	10196	3,85
T	N	polluants	7883	3,74
T	N A	eaux souterraines	3802	3,73
V[m]	N A C A	eaux superficielles et souterraines	87	2,09

Projection sémantique

Hypothèse compositionnelle: si l'on remplace des parties d'un MWT par d'autres ayant la même valeur, le sens du MWT ne change pas.

- On part d'une liste de paires de termes simples (T_1 et T_2) reliés par des relations sémantiques(*sem*):
  sem(T_1,T_2)
  Ex. Anti(*terrestre*, *aquatique*)
- On a construit une liste de paires de candidats bi-termes (CCT_1 et CCT_2) respectant les conditions suivantes:
  - CCT_1 et CCT_1 sont deux candidats termes composés de deux mots lexicaux.
    $\exists T_1, M_1 \in CCT_1 \qquad \exists T_2, M_2 \in CCT_2$
  - Si M_1 et M_2 sont identiques et T_1 et T_2 sont reliés par la relation *sem*, on suppose qu'il y a le même type de relation sémantique entre CCT_1 et CCT_2.
    $M_1 = M_2 \wedge sem(T_1, T_2) \supset sem(CCT_1, CCT_2)$
    Ex. *écologie terrestre* et *écologie aquatique*

Processus

Étapes:

(1) Extraire les candidats MWTs à partir du corpus

(2) Supprimer les candidats qui se présentent moins de 5 fois dans le corpus

(3) Garder uniquement les termes constitués de deux unités lexicales et

(4) Mettre en œuvre la projection sémantique

(5) Supprimer les couples symétriques
Ex. la paire *croissance importante* & *baisse importante* et la paire *baisse importante* & *croissance importante*

(6) Rendre le patron plus détaillé
Ex. *contrôle sur le polluant : NPN — NPDN*

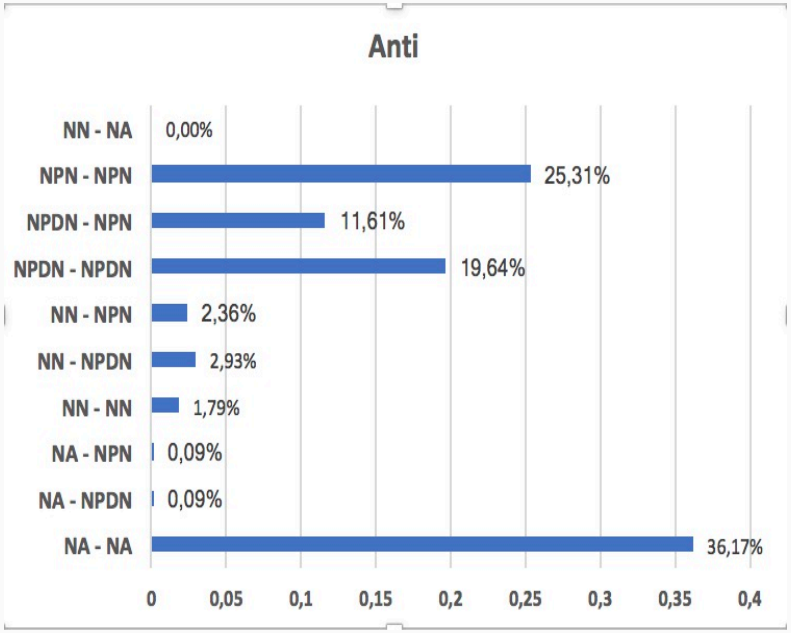
Caractéristiques des données

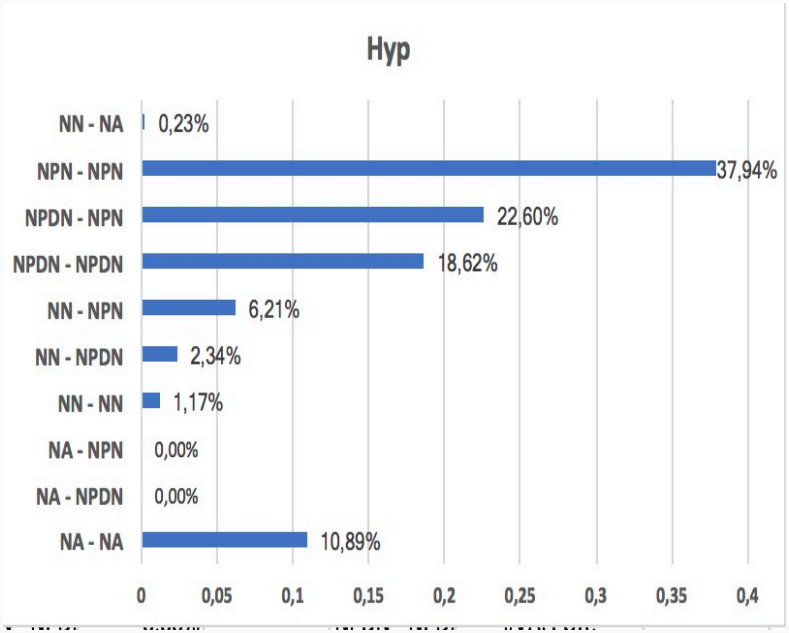
5819 liens entre candidats MWTs sont inférés.

	Anti	Hyp	Qsyn	Total
Number	1059	854	3906	5819
Percentage	18.20%	14.68%	67.12%	100%

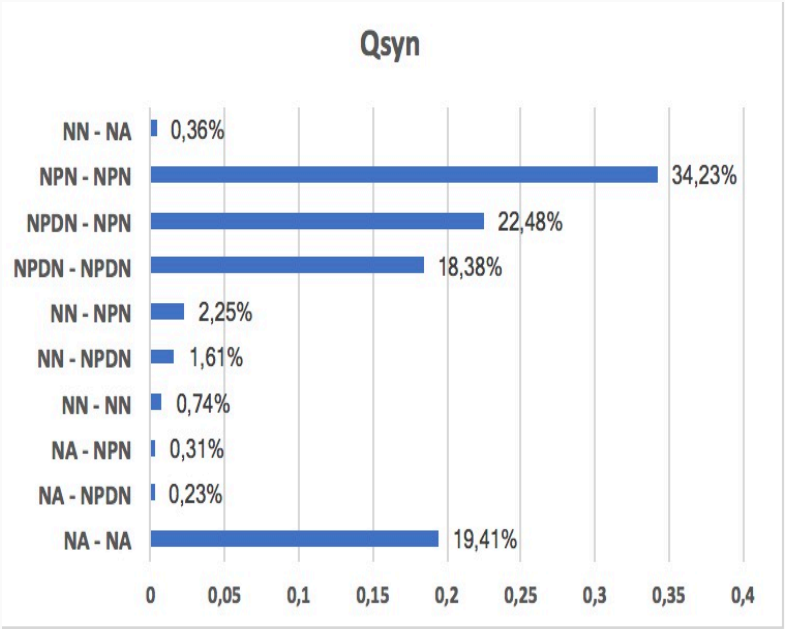
Caractéristiques des données

Analyse du patron de paires de MWTs candidats





Caractéristiques des données



Remarques:

- La plupart des candidats liés par Anti ont le patron **NA**, tandis que le patron de la majorité des candidats ayant du sens hiérarchique ou similaire est **NPN**.
- Les candidats dans la plupart des couples relationnels ont le même schéma.

Évaluation de la préservation des relations

À évaluer:

- Est-ce que les relations inférées sont préservées
  eau de consommation & *alimentation en eau*
  climat humide & *climat sec*

Évaluation de la préservation des relations - Inter-annotation

Évaluation des MWTs candidats:

- Est-ce que les candidats sont syntaxiquement complets
  Ex. *lutte contre le changement* (*lutte contre le changement climatique*)
- Est-ce que les candidats sont liés au domaine de l'environnement
  pouvoir régional

- TERMIUM Plus [2]
- Le Grand Dictionnaire [3]
- IATE (Interactive Terminology for Europe) [4]

	Anti	Hyp	Qsyn	Total
Number	80	42	100	222

[2] https://www.btb.termiumplus.gc.ca/tpv2alpha/alpha-fra.html?lang=fra
[3] http://www.granddictionnaire.com/
[4] https://iate.europa.eu/

Évaluation de la préservation des relations - Inter-annotation

Exemples de bons candidats:

Terme1	Terme2	Relation inférée
faune terrestre	faune aquatique	Anti
flore terrestre	flore aquatique	Anti
gestion agricole	gestion piscicole	Anti
habitat urbain	habitat rural	Anti
agriculture itinérante	culture itinérante	Hyp
puits de carbone	réservoir de carbone	Hyp
gaz de pétrole	gaz combustible	Hyp
consommation de carburant	surconsommation de carburant	Hyp
acidification des sols	contamination des sols	Qsyn
agriculture organique	agriculture biologique	Qsyn
conditionnement des déchets	traitement des déchets	Qsyn
contamination radioactive	pollution radioactive	Qsyn

Évaluation de la préservation des relations - Inter-annotation

- **Désambiguïsation**: Cinq contextes pour chaque candidat
- **Méthode d'évaluation**: kappa de Fleiss

K	**Interprétation**
<0	Pauvre concordance
0.01 - 0.20	Faible concordance
0.21 - 0.40	Légère concordance
0.41 - 0.60	Concordance moyenne
0.61 - 0.80	Concordance importante
0.81 - 1.00	Concordance presque parfaite

Évaluation de la préservation des relations - Inter-annotation

Évaluation de l'inter-annotation

	Anti	Hyp	Qsyn
Fleiss' kappa	0.77	0.68	0.32

Difficulté:

- Interprétations différentes de la définition de quasi-synonyme
  Ex. *dégradation des sols* & *détérioration des sols*
  enfouissement des déchets & *élimination des déchets*

Évaluation de la préservation des relations - automatisation

Entraînement des classifieurs

nécessité d'automatiser le processus d'évaluation des relations inférées:

- Le jugement humain peut provoquer un désaccord considérable
- La grande quantité de données à évaluer (5819 relations)

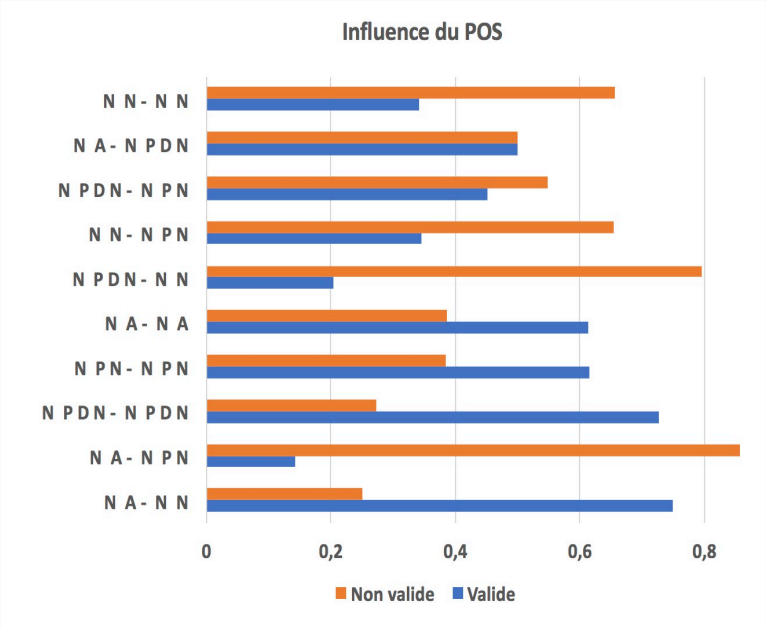
Évaluation de la préservation des relations

Traits linguistiques

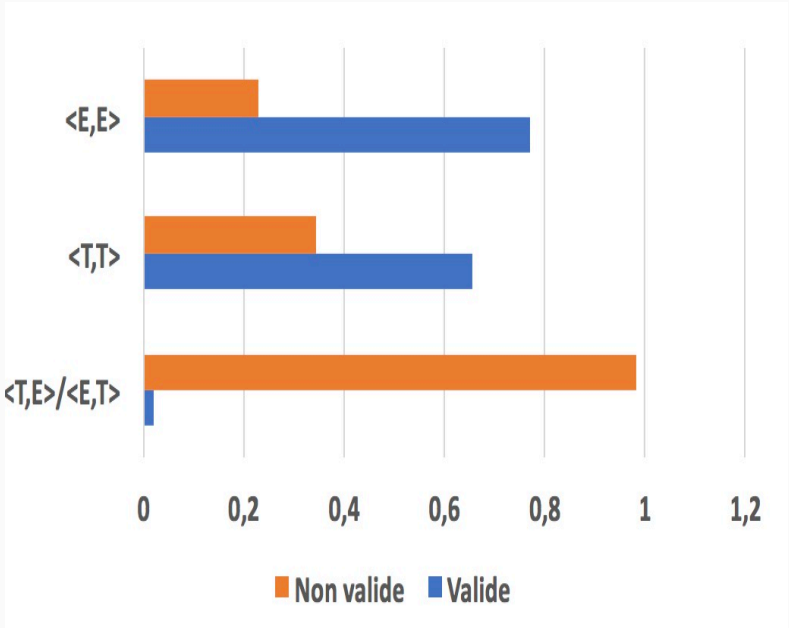
- **Patron POS**: est-ce que le POS de CCT_1 et celui de CCT_2 sont identiques (*écologie terrestre* et *écologie aquatique* vs. *diminution de la demande* et *demande en hausse*).
- **Relation dérivationnelle**: est-ce qu'il existe une relation dérivationnelle entre T_1 et T_2 (*pollution des eaux* et *dépollution des eaux* vs. *électricité éolienne* et *énergie éolienne*).
- **Permutation**: est-ce que T_1 et T_2 se situent à la même position dans CCT_1 et CCT_2 (*monde sans pétrole* et *pétrole sur la planète* vs. *électricité éolienne* et *énergie éolienne*).
- **Tête-Extension**: est-ce que T_1 et T_2 ont le même rôle dans les deux candidats (*environnement global* et *environnement mondial* vs. *population en croissance* et *décroissance de la population*).

Évaluation de la préservation des relations

Environ 2000 paires ont été annotées par une personne pour évaluer l'influence des traits linguistiques sur la préservation des relations.

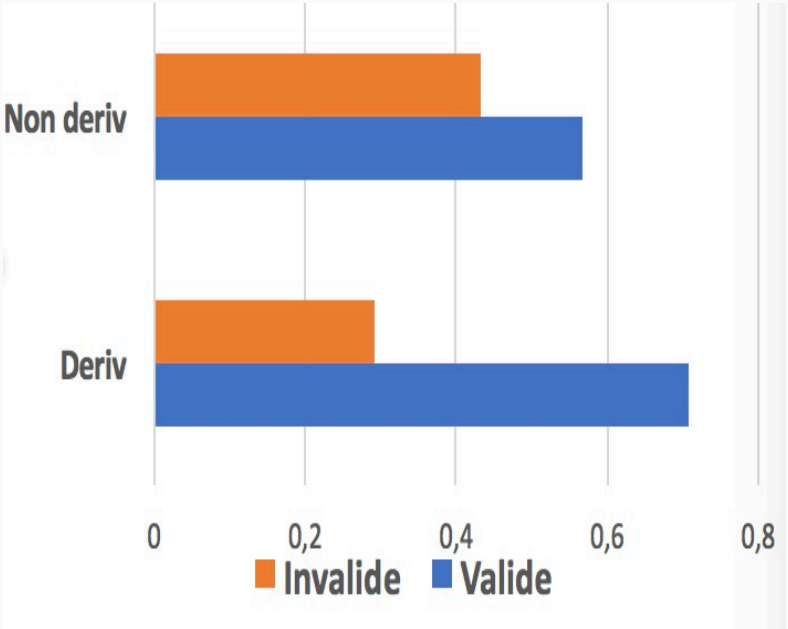


Influence de patron sur la préservation des relations

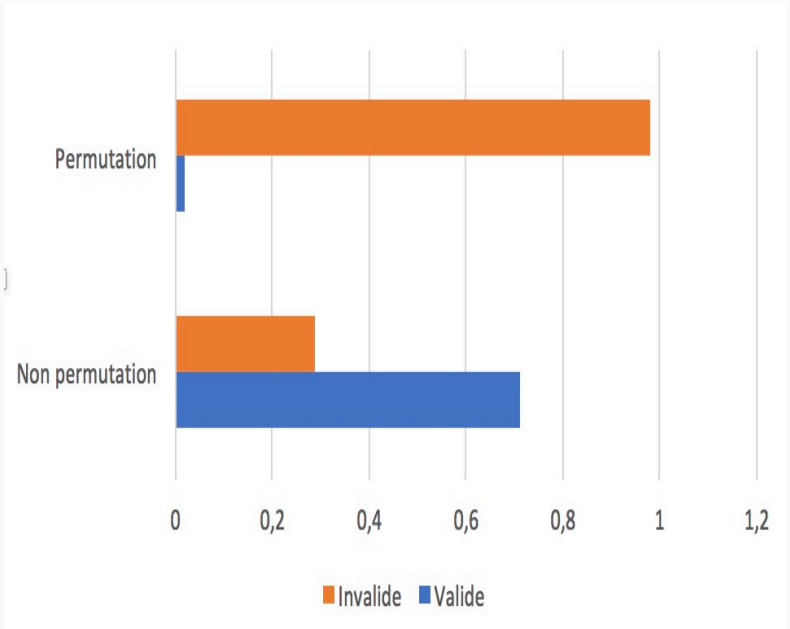


Influence de rôle syntaxique sur la préservation des relations

Évaluation de la préservation des relations



Influence de la relation dérivationnelle sur la préservation des relations



Influence de l'information permutation sur la préservation des relations

Évaluation de la préservation des relations

Résultats des classifieurs

- **Classifieurs choisis**: Random forest & Decision tree
- **Données d'entrée**: 80 paires pour Anti, 42 paires pour Hyp et 100 paires pour Qsyn (un classifieur par relation & décision majoritaire)
- **Hyper-parametres**: par défaut
- **Cross-validation**: GroupKFold
- **Méthode d'évaluation**: précision, rappel, F-mesure

	Anti	Hyp	Qsyn
Precision	88.39%	61.67%	72.71%
Recall	98.75%	72.50%	74.80%
F1	92.33%	63.33%	71.73%

Table 1: Résultats de Decision tree

	Anti	Hyp	Qsyn
Precision	89.17%	65.50%	72.97%
Recall	98.00%	90.00%	96.39%
F1	**92.61%**	**72.46%**	**82.31%**

Évaluation de la préservation des relations

L'analyse de résultats de projection sémantique basée sur l'inter-annotation manuelle

	Anti	Hyp	Qsyn	Total
Liens invalides	13	12	7	32
Liens valides	67	30	93	190

Remarques:

- Plus de paires valides que les invalides
- Les paires invalides ne contiennent pas forcément le MWT non-compositionnel

  Ex. *eau de surface* & *surface de la terre*
- Dans la plupart des cas, ce qui rend le lien invalide c'est la polysémie de terme simple.

  Ex. *route maritime* & *autoroute maritime*

Conclusion et Perspectives

Conclusion et Perspectives

Conclusion:

- On a construit un jeu de données composé de candidats MWTs reliés par les relations lexicales.
  - **Langue**: Français
  - **Volume**: 1059 paires pour Anti; 854 paires pour Hyp; 3906 paires pour Qsyn
- **Volume de données évaluées manuellement**: 80 paires pour Anti; 42 paires pour Hyp; 100 paires pour Qsyn
- **Application**: évaluation des méthodes de composition; comparaison des modèles sémantiques sur la compositionnalité en entraînant les modèles sur un corpus spécialisé; étude sur la représentation vectorielle des candidats MWTs liés sémantiquement.

Conclusion et Perspectives

Perspectives:

- Étudier l'embeddings de MWTs liés sémantiquement à l'aide de tâches spécifiques telles que la reconnaissance des relations.
- Étendre le jeu de données en prenant en compte les MWTs composées de trois unités lexicales afin de mieux couvrir la relation d'hyperonyme.
- Étendre le jeu de données vers d'autres langues ou vers d'autres domaines.

References

[1] Gabriel Bernier-Colborne and Patrick Drouin. Evaluation des modeles sémantiques distributionnels: le cas de la dérivation syntaxique. In *Proceedings the 23rd French Conference on Natural Language Processing (TALN)*, pages 125–138, 2016.

[2] Béatrice Daille, Christine Jacquin, Laura Monceaux, Emmanuel Morin, and Jérôme Rocheteau. Ttc termsuite: une chaîne de traitement pour la fouille terminologique multilingue. In *18ème Conférence francophone sur le Traitement Automatique des Langues Naturelles Conference (TALN 2011).*, 2011.

Références ii

[3] Meghdad Farahmand, Aaron Smith, and Joakim Nivre. A multiword expression data set: Annotating non-compositionality and conventionalization for english noun compounds. In *Proceedings of the 11th Workshop on Multiword Expressions*, pages 29–33, 2015.

[4] Françoise Gayral, Daniel Kayser, and François Lévy. Challenging the principle of compositionality in interpreting natural language texts. *arXiv preprint cs/0609043*, 2006.

[5] Zellig S Harris. Distributional structure. *Word*, 10(2-3):146–162, 1954.

[6] Arnaud Lanoix. *Systèmes à composants synchronisés: contributions à la vérification compositionnelle du raffinement et des propriétés*. PhD thesis, 2005.

Références iii

[7] Carlos Ramisch, Silvio Cordeiro, Leonardo Zilio, Marco Idiart, and Aline Villavicencio. How naked is the naked truth? a multilingual lexicon of nominal compound compositionality. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 156–161, 2016.

[8] Siva Reddy, Diana McCarthy, and Suresh Manandhar. An empirical study on compositionality in compound nouns. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 210–218, 2011.

[9] Zoltán Gendler Szabó. Compositionality. 2004.