Extraction d'instances de relations n-Aires issues d'articles scientifiques guidée par une ontologie : comparaison de méthodes structurelles, fréquentistes et sémantiques

IN-OVIVE 2021

Martin Lentschat (Université de Montpellier), Patrice Buche (INRAE UMR IATE), Juliette Dibie-Barthelemy (INRAE MIA Paris), Mathieu Roche (CIRAD UMR TETIS)

21/09/2021

Introduction

Extraction de relations n-Aires dans des publications scientifiques en domaine de spécialité

determined using a calibration curve prepared with gallic acid, and the results reported as mg/l.

3. Results and discussion

3.1. The effect of antioxidants on the OP of edible films containing SA

Fatty acids, such as SA, LA and PA, being edible and having hydrophobic character, are used in coating formulations as water vapour barrier materials. In previous work we had found SA to be more effective than LA and PA for decreasing WVP of cellulose-based edible films (Ayranci & Tunc, 1997, 2001). Therefore, it was of interest to see how the OP of these films was affected by the SA content. The OP values of MC-based edible films, containing varying amounts of SA in their composition, were determined by the method developed in the present work, as described earlier, and are given in Table 2, together with film thickness values.

The general trend is that the OP increases with increasing SA content of the film. This may be attributed to the formation of holes in the crystal structure of edible films as the SA content increases. These holes, which are especially formed above 15 g SA/100 g MC,

Table 2
The SA content, the thickness and the OP values of edible films at 25 °C and 0% RH

SA content g (100 g MC)$^{-1}$	Thickness 10^5 m	OP 10^9 g d^{-1} Pa^{-1} m^{-1}
0.0	1.86±0.00	6.8±0.4
5.0	1.93±0.03	5.2±0.2
15	2.10±0.01	7.7±0.9
25	1.88±0.04	8.6±0.3
40	2.00±0.00	14±1

It is clear from Table 3 that OP values of films decrease with both AA and CA contents. The only exception to this trend is at 16.7 g CA/100 g MC content. The OP values of this film were found to be slightly larger than that of the film with 3.33 g CA/100 g MC. The two antioxidants show similar effects in improving the oxygen barrier property of the films.

3.2. The effects of coating on water loss of fresh foods

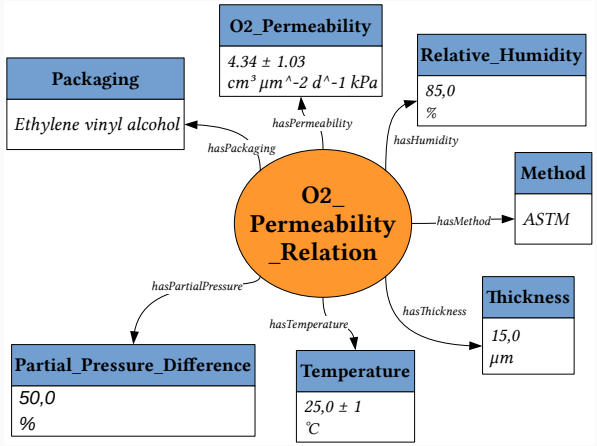
The water loss of mushrooms, with coatings of varying composition, given in Table 1, and of uncoated ones, as a function of time, are shown in Fig. 2. In the coating formulations, an intermediate SA content of 20 g/100 g MC (which is equivalent to 0.6 g/3 g MC) and the highest examined CA or AA content of 16.7 g/100 g MC (which is equivalent to 0.5 g/3 g MC) were maintained according to the results presented above in Section 3.1. The % water losses of uncoated mushrooms are 3.86, 14.7 and 19.7 at the end of first, third and fifth days, respectively. Mushrooms with coatings of varying

Table 3
The antioxidant content, the thickness and the OP values of edible films containing 20 g SA/100 g MC at 25 °C and 0% RH

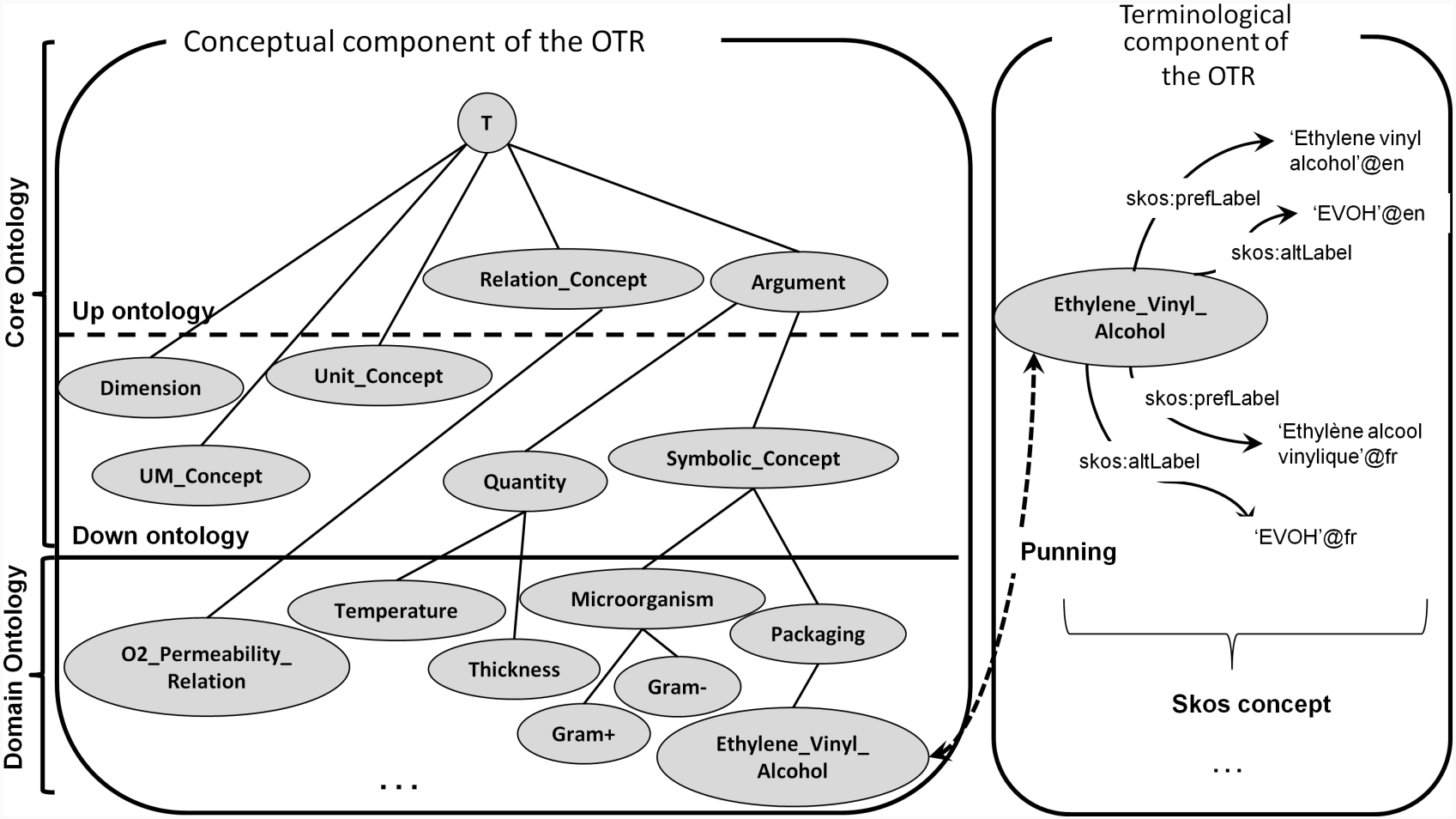
Antioxidant content g (100 g MC)$^{-1}$	Thickness 10^5 m	OP 10^9 g d^{-1} Pa^{-1} m^{-1}
AA		
0.33	1.9±0.2	8.3±0.2
1.67	1.87±0.03	6.5±0.1
3.33	1.8±0.0	5.8±0.2
16.67	1.80±0.02	4.5±0.2
CA		
0.33	1.68±0.0	6.4±0.3
1.67	1.57±0.03	5.39±0.03
3.33	1.49±0.01	3.9±0.2
16.67	1.62±0.02	4.7±0.2

?

?

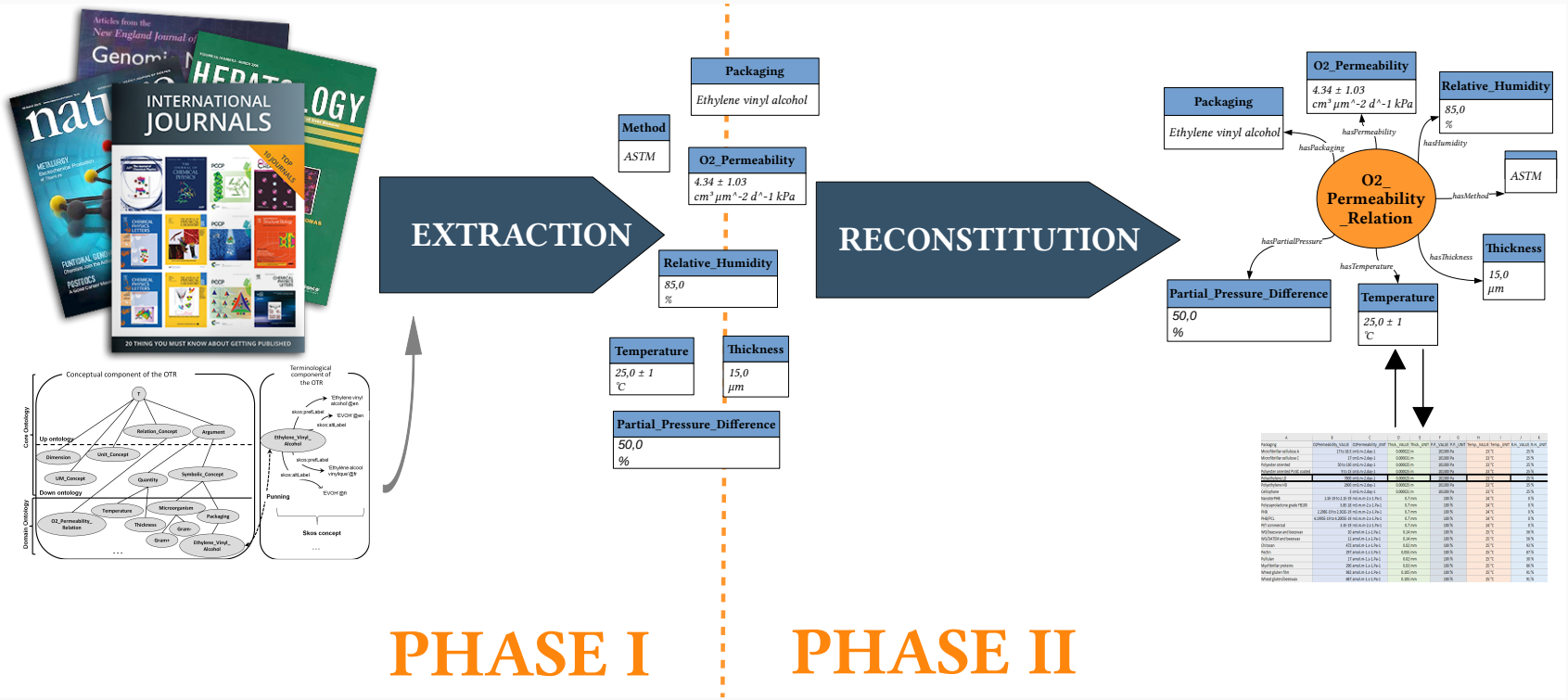


Ressource Termino Ontologique (RTO)



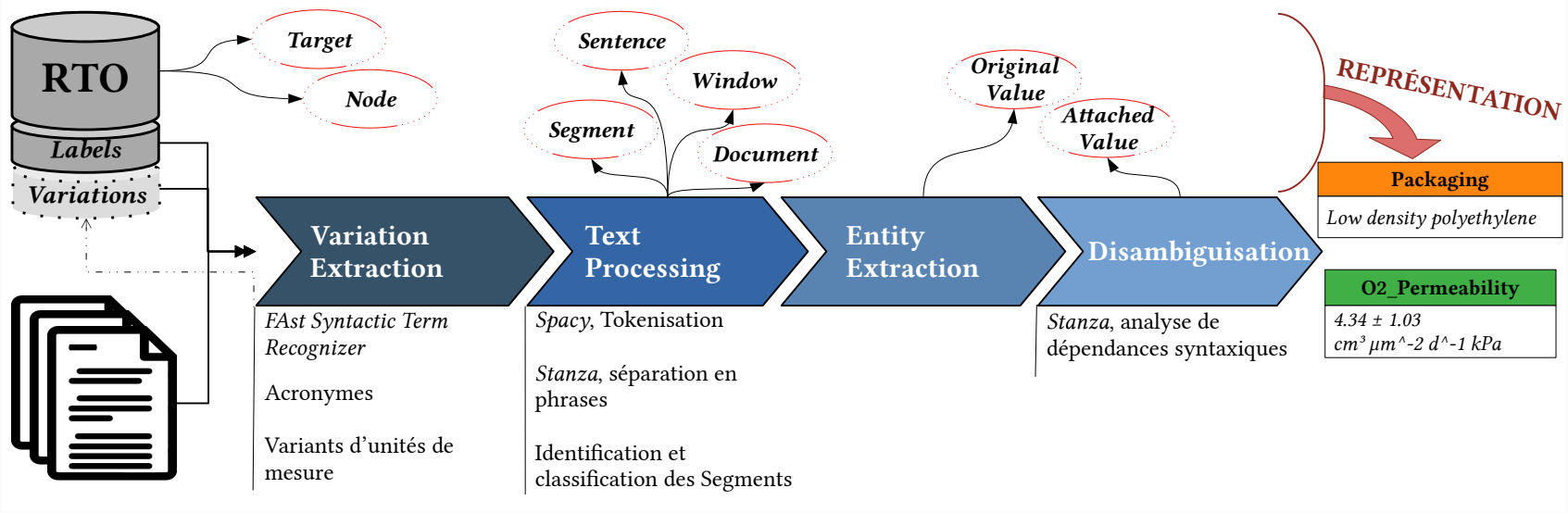
RTO de domaine TRANSMAT `https://ico.iate.inra.fr/atWeb/`

Cadre général



- **Phase I** : extraction des instances d'arguments guidée par une ontologie de domaine
- **Phase II** : reconstitution des instances de relations n-Aires

Scientific Publication Représentation (SciPuRe)



	Feature	Example
LEXICAL ONT.	Target	`Perm.`
	Node	`O2_Perm.`
	OriginalValue	$4.34 * 10^{-3}$
		$cm^3 \mu m m^{-2}$
		$d^{-1} kPa$
	AttachedValue	*permeability*

	Feature	Example
STRUCTURAL	Sentence	*The low ... kPa*
	Window	*Film ...,*
		The ... kPa,
		$\emptyset$
	Segment	*Results*
	Document	*Faro and al.*

Résultats de l'extraction des instances d'arguments

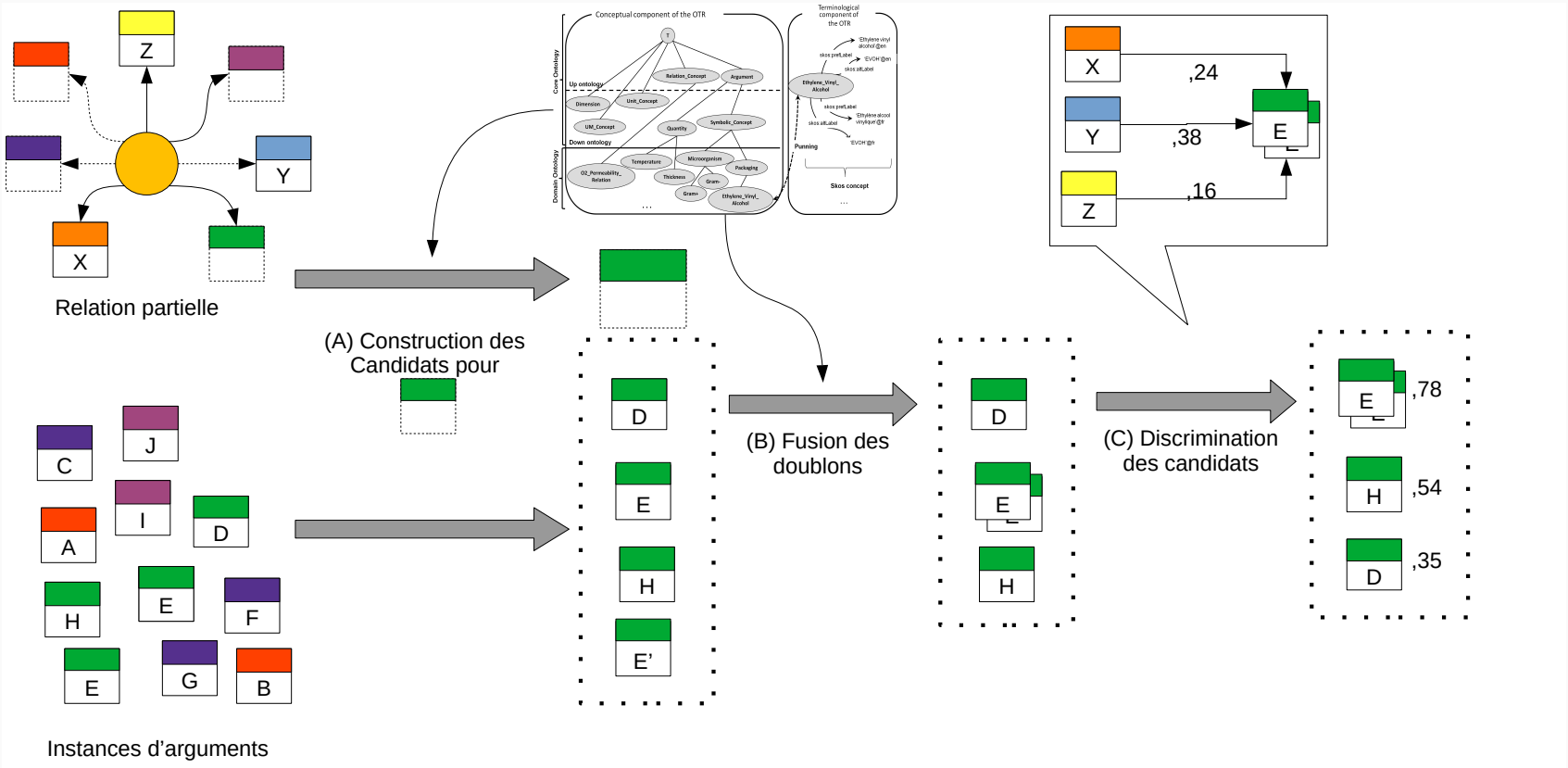
Target	recall (%)	precision (%)	F-score (%)
SYMBOLIC	85	47	61
packaging	86	37	51
component	84	56	67
method	77	16	26
QUANTITATIVE	86	14	24
permeability	83	16	27
relative_humidity	88	28	43
thickness	100	14	24
temperature	83	08	15
GENERAL	**85**	**41**	**55**

Lentschat, M., Buche, P., Dibie-Barthelemy, J., Roche, M., (2021) 'Representation and Relevance Scores of experimental data extracted with an ontological and Terminological Resource', *International Journal of Intelligent Information and Database Systems*. **a paraître**

CIRAD DATAVERSE : "*TRANSMAT Gold Standard*", doi:10.18167/DVN1/U7HK8J

Notre approche pour la reconstitution des relations n-Aires

- Extraction de relations n-Aires partielles dans les tableaux des articles
  Buche, P., Dibie-Barthelemy, J., Ibanescu, L., and Soler, L. (2011). Fuzzy web data tables integration guided by an ontological and terminological resource. *IEEE Transactions on Knowledge and Data Engineering*, 25(4), 805-819.
- Complétion de ces relations avec les instances d'arguments présentes dans le texte :



Scientific Table Representation (STaRe)

	Descripteur	Valeur
ONTOLOGIQUE	Relation	`H2O_Permeability_Relation`
	Result_Argument	`H2O_Permeability`
	Arguments	{*Node* ; *Original_Value* ; *Attached_Value*}
		`Packaging`
		{*Node* ; *Original_Value* ; *Attached_Value*}
		`Method` ∅
		`Relative_Humidity`
		{*Node* ; *Original_Value* ; *Attached_Value*}
		`Temperature` ∅
		`Thickness` ∅
		`Partial_Pressure` ∅
STRUCT.	Table	*Table 3*
	Caption	*Water permeability of tested packaging at* 25°*C*
	Segment	*Results and Discussion*
	Document	*Development of films based on quinoa starch*

Intuitions

Trois approches pour rechercher les instances candidates à la complétion des instances de relations partielles en exploitant les descripteurs de *SciPuRe* et de *STaRe*.

Approche Structurelle

- recherche à proximité des tableaux.
- dans des sections spécifiques des documents selon l'argument.

Approche Fréquentiste

- mesure des cooccurrences fréquentes dans les textes

Approche par Plongements Lexicaux

- calcul de similarité selon un modèle de langage *word-emedding*

Approche Structurelle

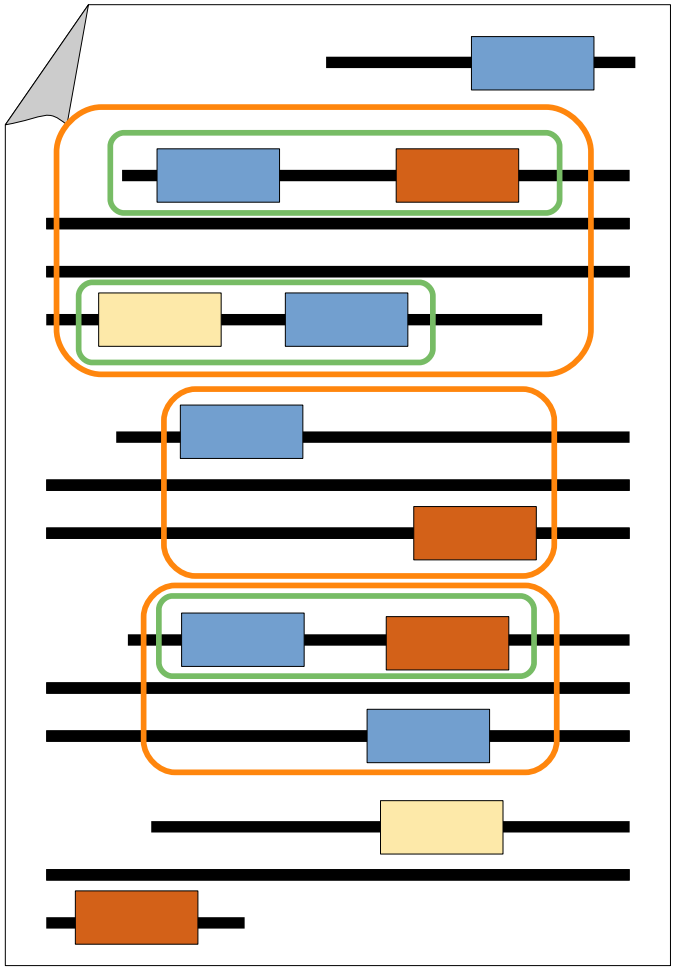
Intuitions

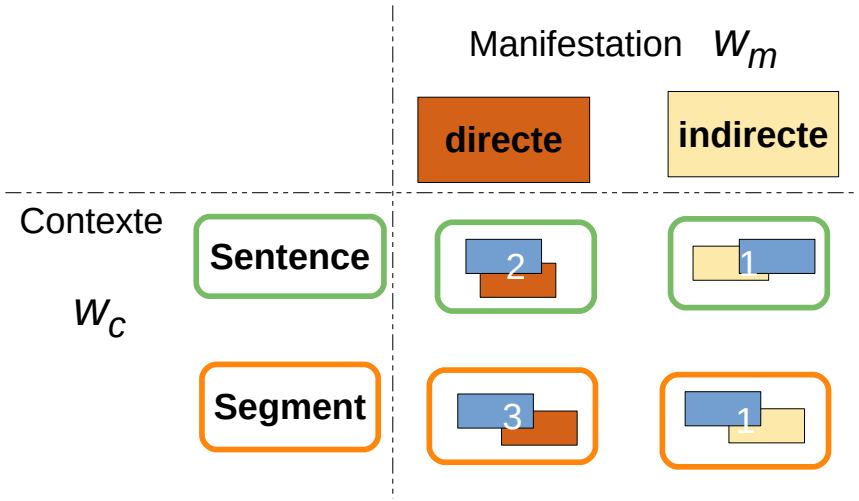
- les instances d'arguments pouvant compléter les relations n-Aires partielles présentes dans les tableaux se situent à proximité de ceux-ci.
- certaines sections sont plus probables de contenir les instances d'arguments à ajouter aux relations

Approche

- Simple : recherche des instances d'arguments les plus proches du tableau
- Guidée : recherche en priorité dans des sections spécifiques (e.g. *Temperature* : *Material_and_Methods* > *Introduction* > *Abstract* > *Results_and_Discussion*)

Approche Fréquentiste





Dice = $\frac{2 * 2}{6 + 4}$ = 0,40

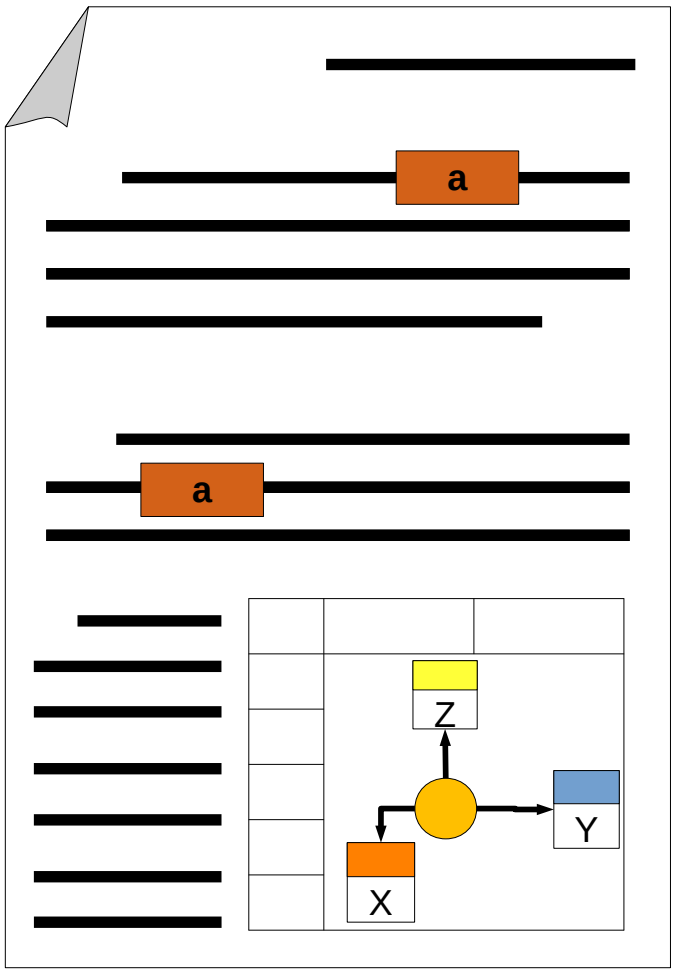
w_m = directe

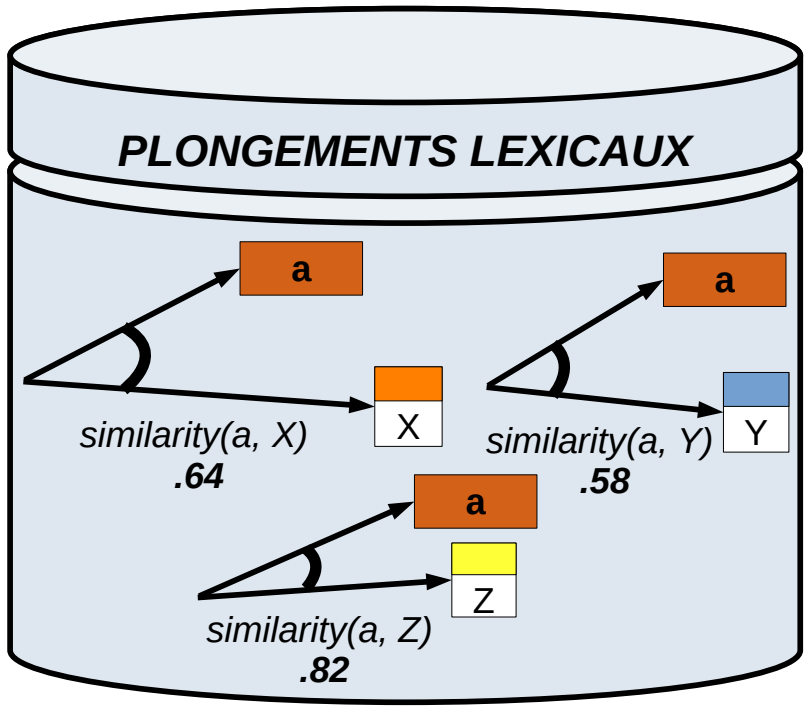
w_c = Sentence

: candidats

Évaluation du calcul de cooccurrence avec les mesures de Dice, Jaccard et Pointwise Mutual Information

Approche par Plongement Lexicaux





Association moyenne = .68

Evaluation des modèles de spaCy, ScipaCy, BERT et BioBERT.

Évaluation

Protocole

- mesures de rappel, précision et f-score sur les instances d'arguments ajoutés aux relations n-Aires.
  $\rightarrow$ trouver l'instance d'argument **ET** l'ajouter à la bonne relation n-Aire

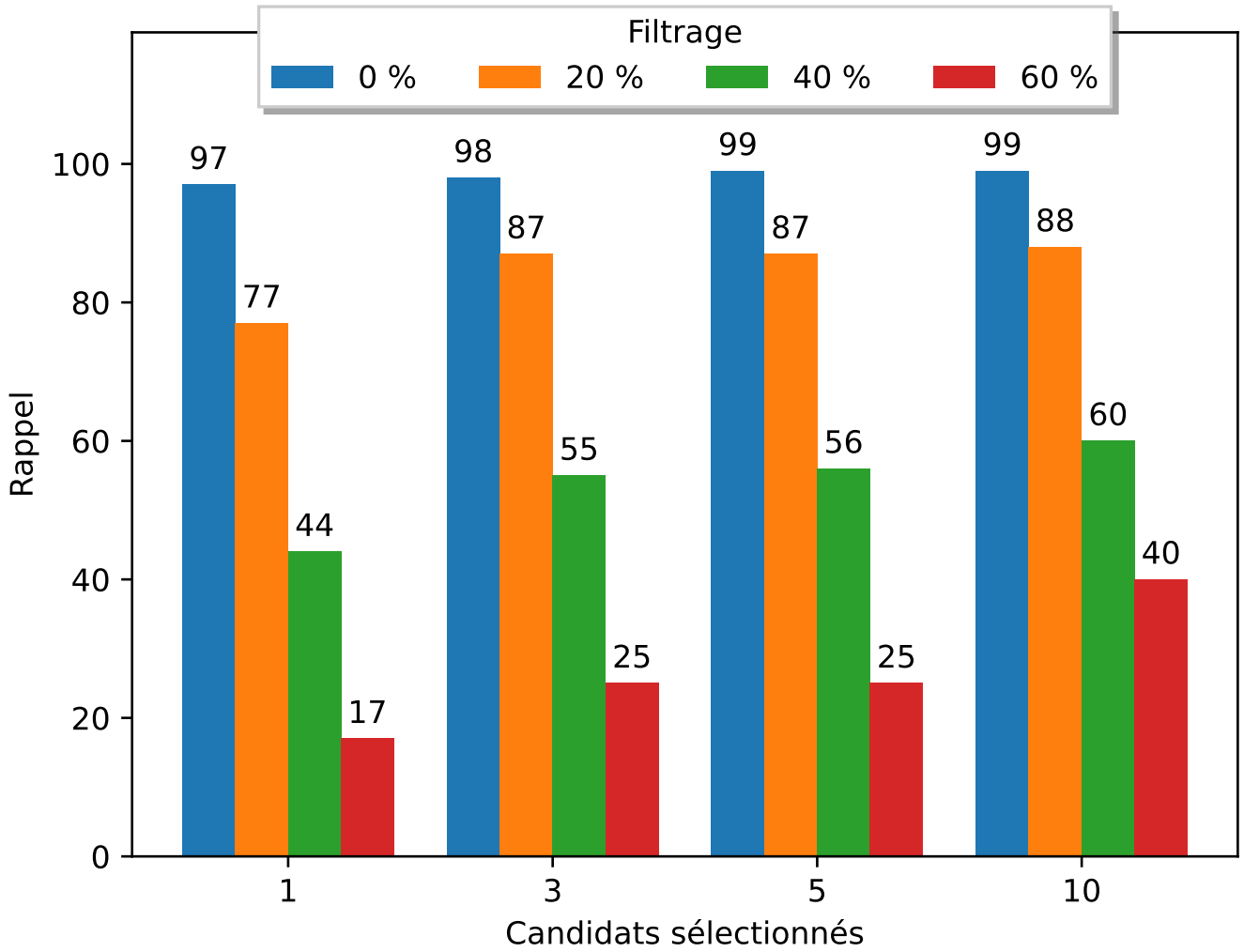
Gold standards : *"TRANSMAT tables data"* (`https://doi.org/10.18167/DVN1/GCZBC9`)
"TRANSMAT relations" (`https://doi.org/10.18167/DVN1/1BBJBQ`)

Paramètres à considérer

fitrage des instances d'argument candidates selon des scores de **pertinence**

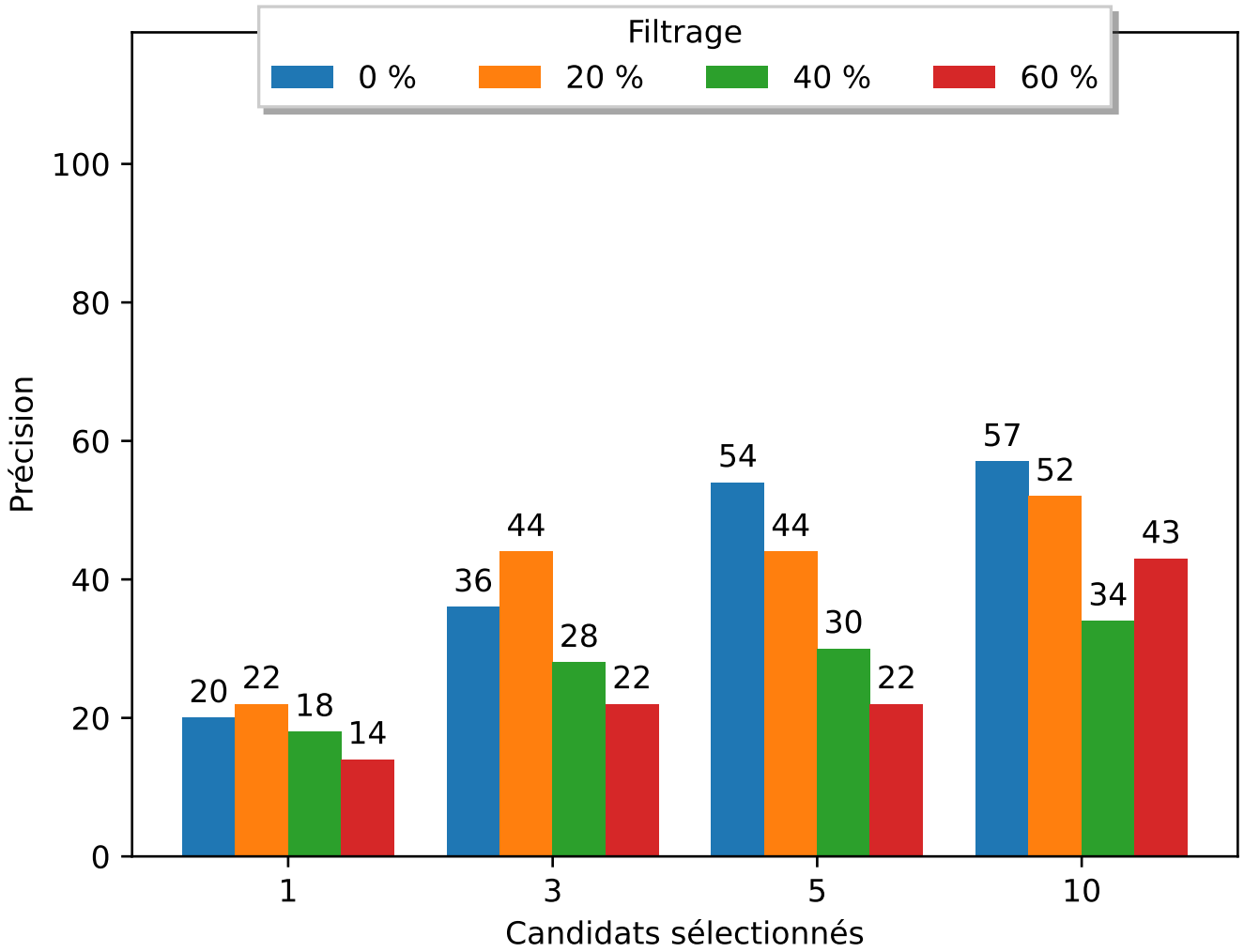
sélection de plusieurs instances d'arguments candidates dans une **démarche d'accompagnement des experts**

Résultats de l'approche Structurelle



Rappel de l'approche Structurelle simple

Résultats de l'approche Structurelle



Précision de l'approche Structurelle simple

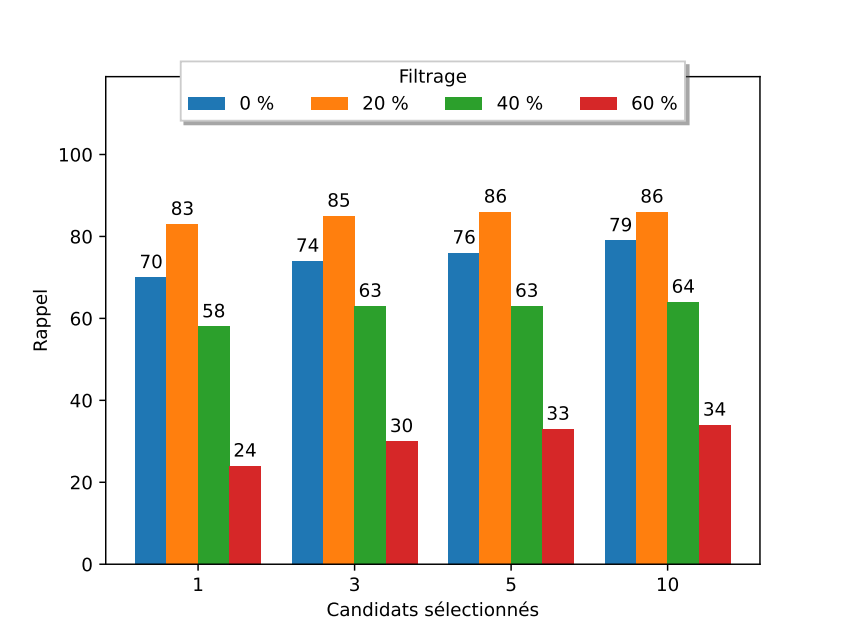
Résultats de l'approche Fréquentiste

Mesure	w_c	w_m	Rappel	Précision	f-score
Jaccard	Document	Attached_Value	.69	.27	.39
Dice	Document	Attached_Value	.71	.28	.40
PMI	Document	Original_Value	.68	.25	.36
...					
Jaccard	Segment	Attached_Value	.50	.12	.19
Dice	Segment	Attached_Value	.52	.13	.20
PMI	Segment	Attached_Value	.50	.12	.19

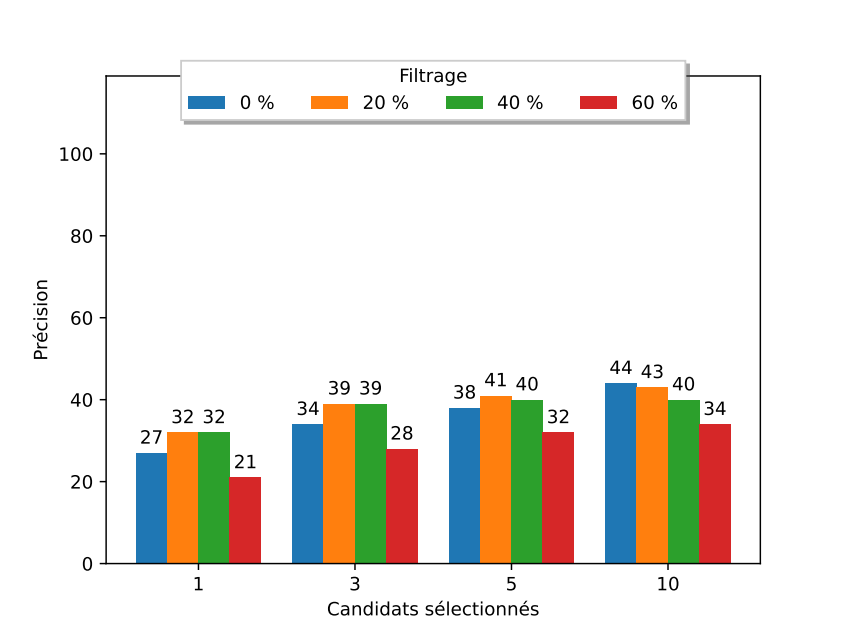
La mesure des cooccurrences au niveau du Document donne les meilleurs résultats.

Jaccard et Dice privilégient les manifestation directes, PMI privilégie les manifestations indirectes.

Résultats de l'approche Fréquentiste



Rappel



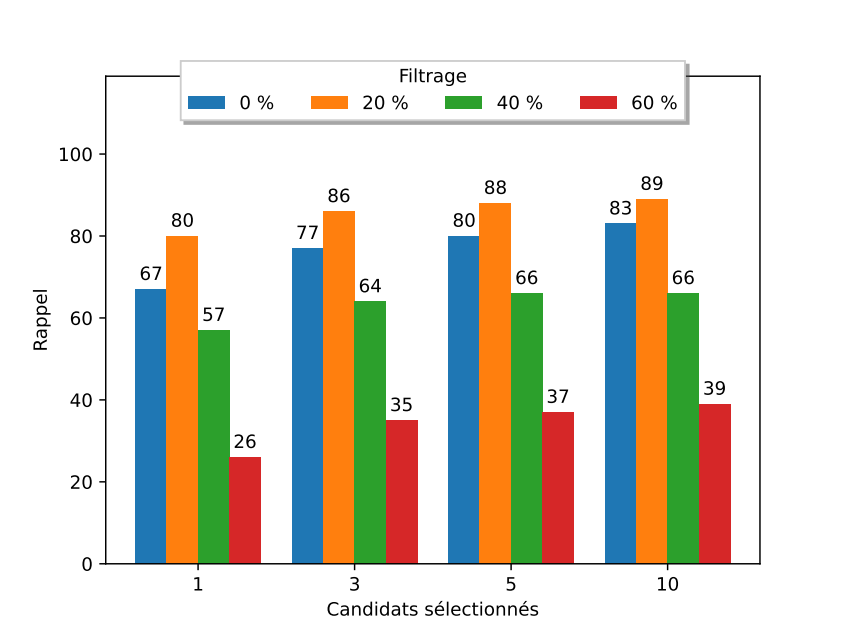
Précision

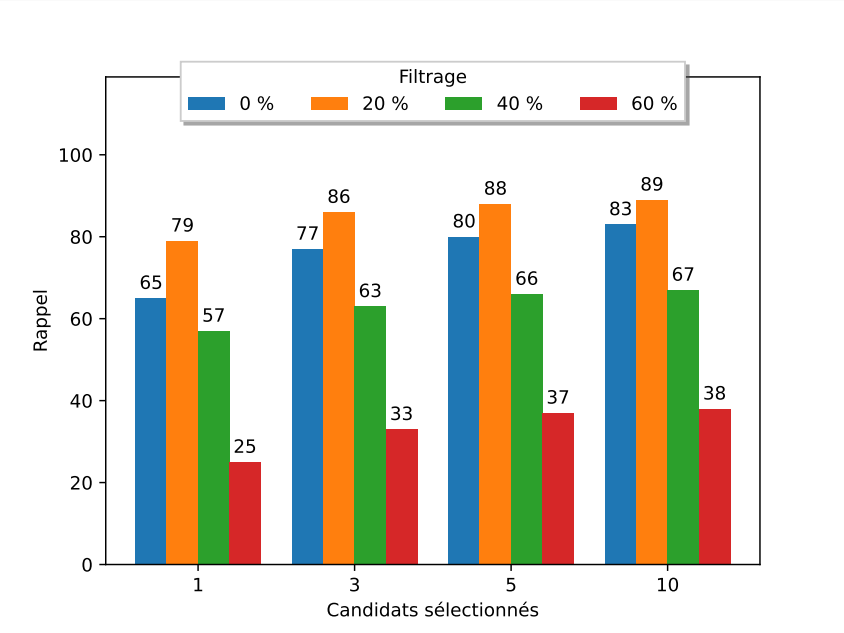
Résultats de l'approche par Plongements Lexicaux

Modèle	Rappel	Précision	f-Score
ner_jnlpba_md*	.50	.12	.19
ner_craft_md*	.49	.11	.18
ner_bionlp13cg_md*	.51	.12	.20
ner_bc5cdr_md*	.49	.11	.19
core_web_lg	.51	.12	.20
core_sci_lg*	.55	.15	.23
core_web_trf	**.67**	**.24**	**.35**
core_sci_scibert*	**.65**	**.22**	**.33**

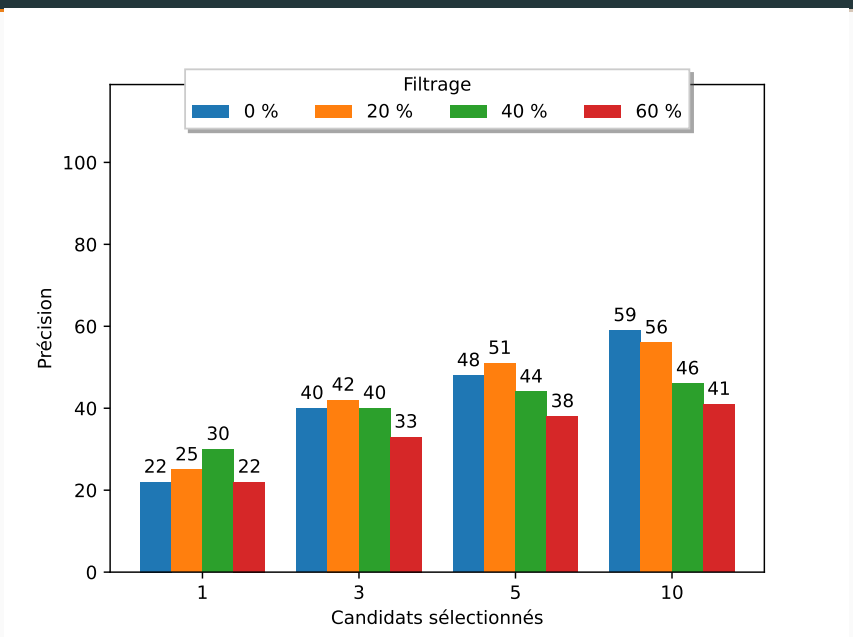
Les modèles suivant la méthode BERT (core_web_trf, core_sci_scibert) donnent les meilleurs résultats, indépendamment du corpus d'entraînement.

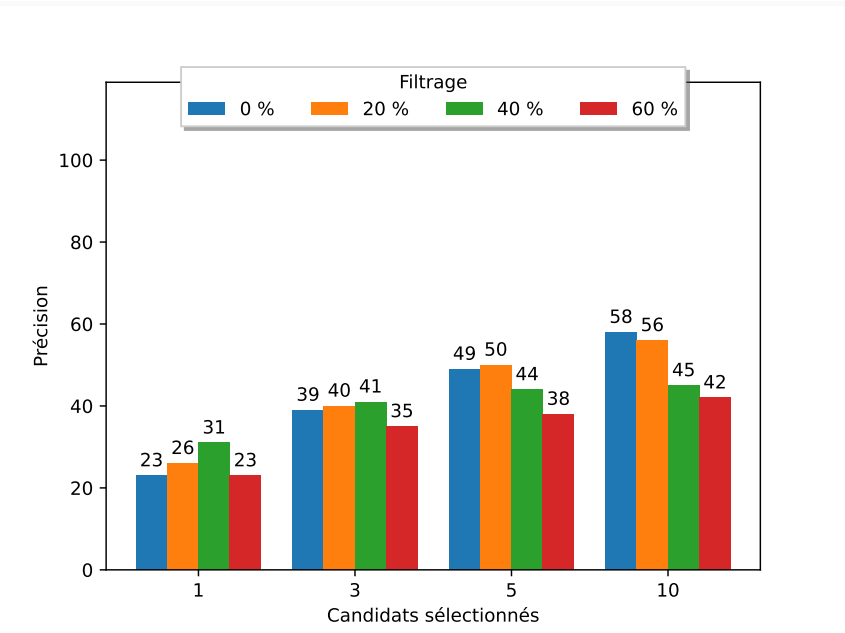
Résultats de l'approche par Plongements Lexicaux





Rappel





Précision

Conclusion

Un rappel haut, une précision selon le nombre de candidats sélectionnés

Approche	Critère	F-SCORE candidats sélectionnés 1	3	5	10
Structurelle	simple	.35	.58	.58	.65
Structurelle	guidée	.45	.56	.61	**.74**
Fréquentiste*	Jaccard	$\mathbf{.48}^{d}_{a}$	$.54^{d}_{a}$	$.61^{p}_{o}$	$.66^{p}_{o}$
Fréquentiste*	Dice	$.46^{d}_{a}$	$.55^{d}_{o}$	$.60^{p}_{o}$	$.66^{p}_{o}$
Fréquentiste*	PMI	$.44^{d}_{o}$	$.53^{d}_{a}$	$.60^{p}_{o}$	$.68^{p}_{o}$
Plongements Lexicaux	core_web_trf	.40	**.59**	.64	.70
Plongements Lexicaux	core_sci_scibert	.39	.57	**.65**	.70

Les meilleurs scores sont obtenus en filtrant 20% des candidats selon leurs scores de pertinence

La méthode de reconstitution des relations n-Aires partielles à choisir varie selon le nombre de candidats sélectionnés

Perspectives

Développements

- alignement des descripteurs de SciPuRe et STaRe avec la RTO
- combinaison des approches de reconstitution
- apprentissage renforcé par la sélection des experts

Extraction synchrone des instances de relations

Raisonnements sur l'**ordre** d'ajout des instances d'arguments, **déductions** au niveaux du document et considération des instances **spécifiques** ou **partagées** entre relations.

Représentation d'un document en graphe

Permet de réduire la dispersion des arguments dans les documents. Les descripteurs de SciPuRe et de STaRe enrichiraient cette approche.

Song, L., Zhang, Y., Wang, Z., and Gildea, D. (2018). N-ary relation extraction using graph-state LSTM. In *Proceedings of the 2018Conference on Empirical Methods in Natural Language Processing*, pages 2226–2235, Brussels, Belgium. Association for Computational Linguistics.

Merci pour votre attention.

martin.lentschat@umontpellier.fr