HAL
open science

Portée de la négation : détection par apprentissage supervisé en français et portugais brésilien

Clément Dalloux, Natalia Grabar, Vincent Claveau, Claudia Moro

To cite this version:

Clément Dalloux, Natalia Grabar, Vincent Claveau, Claudia Moro. Portée de la négation : détection par apprentissage supervisé en français et portugais brésilien. TALN 2018 - 25e conférence sur le Traitement Automatique des Langues Naturelles, May 2018, Rennes, France. 1, Actes de la conférence TALN 2018 - Traitement Automatique de la Langue Naturelle. hal-01937100

HAL Id: hal-01937100

https://hal.science/hal-01937100

Submitted on 27 Nov 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Portée de la négation : détection par apprentissage supervisé en français et portugais brésilien

Clément Dalloux[1], Natalia Grabar[2], Vincent Claveau[1], Claudia Moro[3]

(1) Univ Rennes, Inria, CNRS, IRISA, F-35000 Rennes; (2) UMR 8163 STL CNRS, Université de Lille 3 - France; (3) Pontifícia Universidade Católica do Paraná (PUC-PR),

(1) prenom.nom@irisa.fr, (2) natalia.grabar@univ-lille3.fr, (3) c.moro@pucpr.br

Objectifs : détection de la portée de la négation

- Approche neuronale pour la détection de la portée de la négation
- Deux corpus, en français FR et en portugais brésilien PTBR, annotées avec les informations sur les négations
- Comparaison aux systèmes existants pour l'anglais, analyse des limites et des erreurs

Données : en portugais brésilien et en français

- Deux Corpus créés à partir de **protocoles d'essais cliniques**
- sources FR : hôpital Gustave Roussy, l'Institut National du Cancer
- source PTBR : `http://ensaiosclinicos.gov.br/`
- annotation manuelle de la négation (marqueurs et leur portée)

	FR	PTBR
Mots	134 386	48 204
Vocabulaire	8 133	6 453
Phrases	5 394	3 228
Phrases négatives	820	640
% Phrases négatives	15,20 %	19,83 %

Exemples

1. **absence** [de ganglion métastatique]
2. En cas d'**in**[opérabilité] et/ou **im**[possibilité de réirradier],...
3. ...[des patients éligibles atteints d'une tumeur maligne et porteurs de la mutation BRAFV600] ayant précédemment été inclus et traités dans un protocole antérieur portant sur le Vemurafénib et **n'**[ayant] **pas** [satisfait aux critères du protocole sur la progression de la maladie]...
4. Grupo Controle : **Não** apresentar [DTM].
5. **Ausência de** [evidência clínica de imunossupressão].
6. **não** usuários de [drogas que causem dependência química].

Négation : morphologie (*im-*), lexique (*absence, ausência*), grammaire (*ne pas, não*)

Prétraitements

- FR : étiquetage morpho-syntaxique et lemmatisation avec TreeTagger (Schmid, 1994)
- PTBR : étiquetage morpho-syntaxique avec *RDRPOSTagger* (Nguyen et al., 2015), Snowball

Résultats et analyse d'erreurs

- Résultats proches de ceux obtenus sur le corpus de *SEM-2012 (EN)
- Résultats sont bien moins convainquants sur le corpus brésilien
- Précision privilégiée par le système

	Mots étiquetés			Portées exactes		
Dataset	P	R	F_1	P	R	F_1
FR_{valid}	93,72	86,22	89,81	100	72,48	84,04
FR_{test}	88,29	84,68	86,45	100	53,55	69,75
$PTBR_{valid}$	79,06	69,93	74,22	100	33,71	50,42
$PTBR_{test}$	77,61	64,58	70,50	100	34,29	51,06
EN – Dalloux et al. (2017)	91,24	87,10	**89,12**	100	62,5	76,92
EN – Fancellu et al. (2016)	92,62	85,13	88,72	99,40	63,87	**77,7**

- Dans cet essai, **ni** [le patient,] **ni** [le médecin] **ne** [connaîtront quel traitement] (ProCervix ou placebo) [est administré].
- Dans cet essai, **ni** [le patient], **ni** [le médecin] **ne** [connaîtront quel traitement] (ProCervix ou placebo) est administré.
- Par ailleurs [les patientes du second groupe ayant un risque de rechute potentiellement bas] (grade génomique bas) **ne** [recevront] **pas** [de chimiothérapie].
- Par ailleurs les patientes du second groupe ayant [un risque de rechute potentiellement bas] ([grade génomique bas)] **ne** [recevront] **pas** [de chimiothérapie].
- **Ausencia** de diagnóstico de [doenças neuromusculares], [trauma], [tumores] ou [abscessos raquimedulares], [hemiplegia]/ [paresia], [lesão de plexo] ou [encefalopatia cerebral].
- **Ausencia** de diagnóstico [de doenças neuromusculares], [trauma], tumores ou [abscessos raquimedulares], hemiplegia/paresia, lesão de plexo ou [encefalopatia] cerebral.
- que **não** apresentem outras [doenças neurológicas] ou [ortopédicas diagnosticadas].
- que **não** apresentem [outras doenças neurológicas] ou [ortopédicas] diagnosticadas.

Remerciements

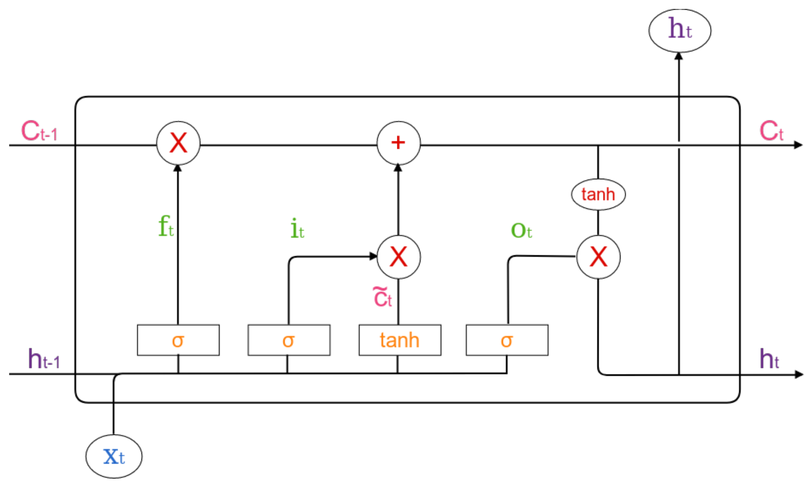
Ce travail a bénéficié d'une aide de l'état attribuée au labex COMIN LABS et gérée par l'Agence Nationale de la Recherche au titre du programme « Investissements d'avenir » portant la référence ANR-10-LABX-07-01. Nous remercions également les relecteurs anonymes pour les remarques constructives qui ont permis d'améliorer la présentation du travail.

Annotation de référence

#Phrase	Position	Forme	Lemme	POS-tag	Marqueur	Portée
5105	0	L'	le	DET :ART	_	_
5105	1	abdomen	abdomen	NOM	_	abdomen
5105	2	est	être	VER :pres	_	_
5105	3	souple	souple	ADJ	_	_
5105	4	et	et	KON	_	_
5105	5	sans	sans	PRP	sans	_
5105	6	défense	défense	NOM	_	défense
5105	7	.	.	SENT	_	_
2247	0	déficit	déficit	NOUN	_	_
2247	1	visual	visual	ADJ	_	_
2247	2	grave	grav	ADJ	_	_
2247	3	sem	sem	ADP	sem	_
2247	4	correção	correçã	NOUN	_	correção
2247	5	.	.	PUNCT	_	_
2916	0	Gestantes	gestant	ADJ	***	
2916	1	.	.	PUNCT	***	

Méthode : réseaux de neurones

Long Short-Term Memory (LSTM) bidirectionnel et Champs Aléatoires Conditionnels (CRF)



Représentation

une instance $I(n, c, t)$, où chaque mot est représenté par

- un vecteur n (*word-embedding*)
- un vecteur c, qui détermine si le mot fait partie d'un marqueur (*cue-embedding*)
- un vecteur t, qui représente l'étiquetage morpho-syntaxique des mots (*postag-embedding*)

Paramètres

- Embeddings de dimension $k = 50$
- Couche cachée avec 2x200 unités (deux couches cachées concaténées pour BiLSTM)
- 50 périodes d'entraînement

Évaluation

- Corpus segmentés en trois parties : 60% (entraînement), 15% (validation), 25% (test)
- Mesures d'évaluation classiques : la précision P, rappel R et la F-mesure F_1

Bilan et Perspectives

Contributions

- mise à disposition des corpus : Été 2018 (`http://people.irisa.fr/Clement.Dalloux/`)
- mise à disposition des outils : `https://allgo.inria.fr/webapps/173`

Points à explorer

- Exploiter différentes architectures de réseaux de neurones récurrents
- Utiliser différents types de *word embeddings*
- Comparer nos résultats avec ceux des systèmes experts disponibles

Références

Dalloux, C., Claveau, V., and Grabar, N. (2017). Détection de la négation : corpus français et apprentissage supervisé. In *SIIM 2017 - Symposium sur l'Ingénierie de l'Information Médicale*, pages 1–8, Toulouse, France.

Fancellu, F., Lopez, A., and Webber, B. (2016). Neural networks for negation scope detection. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, volume 1.

Nguyen, D. Q., Nguyen, D. Q., Pham, D. D., and Pham, S. B. (2015). A robust transformation-based learning approach using ripple down rules for part-of-speech tagging. *AI Communications*, 29(3) :409–422.

Schmid, H. (1994). Probabilistic part-ofispeech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing, Manchester, UK.*