Vers une identification automatique de personnes avec schizophrénie dans des conversations contrôlées

Maxime Amblard[1] Chloé Braud[2] Chuyuan Li[1]
Caroline Demily[3] Nicolas Franck[3] Michel Musiol[1,4]

(1) LORIA, UMR 7503, Université de Lorraine, CNRS, Inria, 54000 Nancy, France
(2) IRIT, CNRS, Toulouse
(3) Centre Hospitalier le Vinatier & UMR 5229, CNRS - Univeristé Lyon 1, Lyon, France
(4) ATILF, UMR 7118, Université de Lorraine, CNRS, 54000 Nancy, France

Schizophrénie

- Un trouble mental sévère
- Symptômes : les idées délirantes, les hallucinations, le discours désorganisé

Schizophrénie

- Un trouble mental sévère
- Symptômes : les idées délirantes, les hallucinations, le discours désorganisé
- Enjeu : identification automatique à partir de la production langagière, écrite ou orale
  - aide décisive vers un diagnostic pour les médecins
  - amélioration la compréhension du fonctionnement du langage en général
  - adaptation des systèmes de TAL à des parties de la population affectée

État de l'art

Classification automatique de SCZ[1] fondée sur des données langagières :

[1] SCZ : personnes avec schizophrénie

État de l'art

Classification automatique de SCZ[1] fondée sur des données langagières :

- [Strous et al., 2009] : écrits, traits lexicaux, Acc. = 83,3%
- [Mitchell et al., 2015] : *tweets*, traits lexicaux, Acc. = 82,3%
- [Kayi et al., 2017] : *tweets*, traits morpho-syntaxiques et syntaxiques, F1 = 81,65%
- [Allende-Cid et al., 2019] : textes narratifs, traits morpho-syntaxiques, F1 = 82,8%

[1]SCZ : personnes avec schizophrénie

État de l'art

Classification automatique de SCZ[1] fondée sur des données langagières :

- [Strous et al., 2009] : écrits, traits lexicaux, Acc. = 83,3%
- [Mitchell et al., 2015] : *tweets*, traits lexicaux, Acc. = 82,3%
- [Kayi et al., 2017] : *tweets*, traits morpho-syntaxiques et syntaxiques, F1 = 81,65%
- [Allende-Cid et al., 2019] : textes narratifs, traits morpho-syntaxiques, F1 = 82,8%

⇒ Corpus de nature différente : comparaisons difficiles

[1] SCZ : personnes avec schizophrénie

Table de matières

Approche

Approche

- S'intéresser au dialogue
- 1^e approximation : isoler les tours de parole (TDP) de chaque locuteur :
    1. Extraire les TDP
    2. Concaténer les TDP (cTDP)
    3. Obtenir une instance de classification : cTDP-SCZ ou cTDP-TEM
    4. Classifier les instances dans la classe positive (SCZ) ou négative (TEM)
    5. Obtenir un modèle et analyser

Approche

- S'intéresser au dialogue
- 1^e approximation : isoler les tours de parole (TDP) de chaque locuteur :
  1. Extraire les TDP
  2. Concaténer les TDP (cTDP)
  3. Obtenir une instance de classification : cTDP-SCZ ou cTDP-TEM
  4. Classifier les instances dans la classe positive (SCZ) ou négative (TEM)
  5. Obtenir un modèle et analyser

$\Rightarrow$ Langage plus naturel que les écrits / textes narratifs

$\Rightarrow$ Ignorance les TDP de PSY

Corpus

Corpus SLAM

- Entretiens semi-dirigés entre PSY[2] et SCZ (ou TEM)[3].
- Entretiens enregistrés avec un double système d'*eye-tracker* (données non-utilisées ici)
- Thématique abordée : le quotidien du participant
- PSY non engagé, parole du participant se rapproche d'un monologue.

[2] PSY : psychologue
[3] TEM : témoins

Exemples dialogue

PSY-SCZ

PSY : Et donc là vous avez voir un atelier euh... c'est quoi c'est...

SCZ : Oui donc là je suis allé en atelier thérapeutique euh euhh comment ils appellent ça... pas entretien thérapeutique... j'ai euh...

PSY : Education thérapeutique... c'est ça

PSY-TEM

PSY : Vous voulez faire quoi après

TEM : Euhh je voudrais faire le master de N. de psychopatho de la cognition et des interactions

PSY : Mmh mmh

Recueil des données

- 1 psychologue
- 2 groupes :
  - PSY-SCZ : 18 entretiens
  - PSY-TEM : 23 entretiens (la plupart des étudiants, biais lexicaux)
- 15 hommes dans chaque groupe (biais en termes de genre)

Recueil des données

- 1 psychologue
- 2 groupes :
  - PSY-SCZ : 18 entretiens
  - PSY-TEM : 23 entretiens (la plupart des étudiants, biais lexicaux)
- 15 hommes dans chaque groupe (biais en termes de genre)
- Caractéristiques générales

	TDP/doc	mots/phrase	long. mots	mots gram.
SCZ	~ 200	13,4	4,27	56%
TEM	~ 342	10,5	4,24	51%

Expérience

Contenu de l'expérience

1. Représentation des données : *bag-of-words* (`bow`), *n*-`gram`, `treelet`
2. Sélection de traits
3. Modèles de classification : Naïve-Bayes, Régression logistique, SVM

Représentation des données

- Traits lexicaux
  - *bow*
  - *n-gram* (n=2,3)
- Traits syntaxiques
  - *treelet*
    (parseur syntaxique UDPipe entraîné sur Spoken-French 2.5)

Représentation des données

- Traits lexicaux
  - *bow*
  - *n-gram* (n=2,3)
- Traits syntaxiques
  - *treelet*
    (parseur syntaxique UDPipe entraîné sur Spoken-French 2.5)
- Combinaison de traits (toutes)
  - *bow* + *treelet*
  - *bow* + *n-gram*
  - *n-gram* + *treelet*
  - *bow* + *n-gram* + *treelet*

$\Rightarrow$ 7 combinaisons de traits

Retour sur les `treelet`

- *1-token* ***`treelet`*** : POS tag

NOUN, VERB

Retour sur les `treelet`

- *1-token* ***`treelet`*** : POS tag

  NOUN, VERB

- *2-token* ***`treelet`*** : relation typée entre une tête et un dépendant

  VERB $\xrightarrow{\text{Nsubj}}$ NOUN

Retour sur les `treelet`

- *1-token* **`treelet`** : POS tag

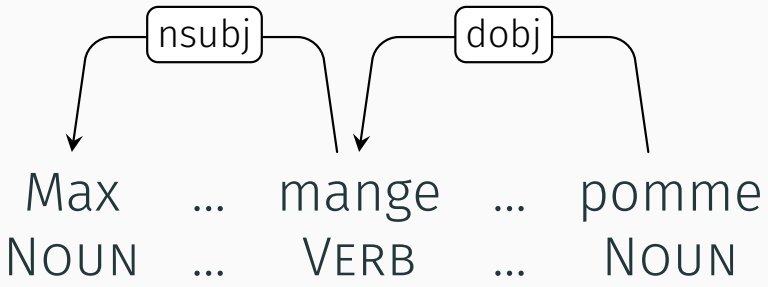
  NOUN, VERB

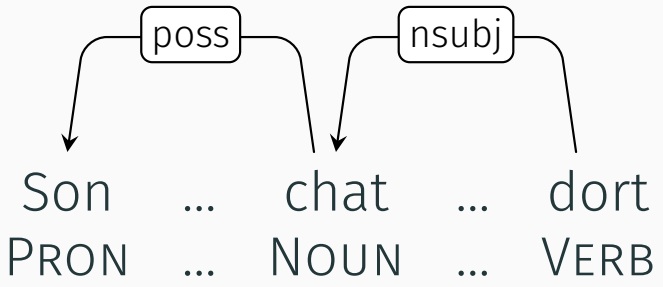
- *2-token* **`treelet`** : relation typée entre une tête et un dépendant

  $\text{VERB} \xrightarrow{\text{Nsubj}} \text{NOUN}$

- *3-token* **`treelet`** : relation une tête et deux dépendants / chaîne de dépendances

  $\text{NOUN} \xleftarrow{\text{Nsubj}} \text{VERB} \xrightarrow{\text{Dobj}} \text{NOUN},$
  $\text{PRON} \xleftarrow{\text{Poss}} \text{NOUN} \xleftarrow{\text{Nsubj}} \text{VERB}$





Sélection de traits

- Problème : peu de données, dimensions très élevées

[4] *https://scikit-learn.org/*

Sélection de traits

- Problème : peu de données, dimensions très élevées
- Sélection : scikit-learn *feature_selection.SelectFromModel*[4]
  - Sans seuil ($1e-5$)
  - 12 seuils : moyenne, médiane, 10 valeurs distribués entre [$1e-5$, 50^e trait le plus important]

⇒ 13 sélections par catégorie de traits
(*bow*, *n-gram*, *treelet*)

[4] *https://scikit-learn.org/*

Sélection de traits

Nombre de traits à l'origine ("#orig.") et sélectionnés ("#sélec") par les classifieurs :

Type de traits	Classifieur	#Orig.	Seuil	#Sélec.	Ratio %
bow	NB	6504	9	6488	99,75
bow	SVM	6504	méd.	3254	50,03
n-gram	SVM	118473	8	98	0,08
treelet	SVM	16865	3	675	4,00
bow + *treelet*	NB	23369	8	11684	49,99
bow + *treelet*	SVM	23369	moy.	3434	14,69
bow + *n-gram*	SVM	124977	4	491	0,39
n-gram + *treelet*	SVM	135338	4	552	0,41
bow + *n-gram* + *treelet*	SVM	141842	5	257	0,18

Classification

- Validation croisée enchâssée
  - À l'extérieur, 1 sous-ensemble parmi N conservé pour l'évaluation
  - À l'intérieur, validation croisée en M sous-ensembles
  - ici $N = M = 5$

Classification

- Validation croisée enchâssée
  - À l'extérieur, 1 sous-ensemble parmi N conservé pour l'évaluation
  - À l'intérieur, validation croisée en M sous-ensembles
  - ici $N = M = 5$
- Classifieurs : scikit-learn
  - Naive Bayes : $\alpha \in \{0.1, 0.01, 0.001\}$
  - Régression logistique : $C \in \{100\}$
  - SVM : $C \in \{5, 100, 1000\}$

$\Rightarrow$ 3 **classifieurs**

Classification

- Validation croisée enchâssée
  - À l'extérieur, 1 sous-ensemble parmi N conservé pour l'évaluation
  - À l'intérieur, validation croisée en M sous-ensembles
  - ici $N = M = 5$
- Classifieurs : scikit-learn
  - Naive Bayes : $\alpha \in \{0.1, 0.01, 0.001\}$
  - Régression logistique : $C \in \{100\}$
  - SVM : $C \in \{5, 100, 1000\}$

$\Rightarrow$ 3 **classifieurs**

Total : 7 combinaisons de traits $\times$
13 sélections de traits $\times$
3 classifieurs
---
273 réalisations

Résultats

Systèmes de référence

- Par la classe majoritaire
- Par longueur de mots
  - taille moyenne des mots
  - taille moyenne des mots au dessus de la taille moyenne de tous les mots
- Ratio *je*/*tu* : *#je* / *#tu* dans chaque document

	Acc.	Prec.	Rec.
Majorité	56,10		
long. mot	49,51	17,21	11,11
$>$ long. moy. mot	52,43	37,43	22,78
ratio *je*/*tu*	72,19	69,87	35,56

Meilleurs systèmes

- Meilleur système : NB avec *bow* (acc. = 93, 66, F_1 = 92, 21)
  - SVM avec *bow* (acc. = 90, 98, F_1 = 90, 38)
  - [Allende-Cid et al., 2019] : SVM avec *bow* (F_1 = 87, 5)

Meilleurs systèmes

- Meilleur système : NB avec *bow* (acc. = 93, 66, F_1 = 92, 21)
  - SVM avec *bow* (acc. = 90, 98, F_1 = 90, 38)
  - [Allende-Cid et al., 2019] : SVM avec *bow* (F_1 = 87, 5)
- Second meilleur système : NB avec *bow+treelet*
  (acc. = 92, 20, F_1 = 90, 38)

Meilleurs systèmes

- Meilleur système : NB avec *bow* (acc. = 93, 66, F_1 = 92, 21)
  - SVM avec *bow* (acc. = 90, 98, F_1 = 90, 38)
  - [Allende-Cid et al., 2019] : SVM avec *bow* (F_1 = 87, 5)
- Second meilleur système : NB avec *bow+treelet* (acc. = 92, 20, F_1 = 90, 38)
- Autres combinaisons de traits : meilleurs scores avec SVM

Différents jeux de traits

Exactitude moyenne avec ou sans sélection ("SVM", "MaxEnt" et "NB") pour chaque combinaison de traits :

Algorithme Sélection	SVM non	SVM oui	MaxEnt oui	NB oui
bow	90,00	90,98	87,07	93,66
n-gram	68,78	81,71	79,76	65,61
treelet	61,46	66,83	58,29	58,05
bow+n-gram	80,49	88,54	86,59	70,49
bow+treelet	87,07	88,78	84,88	92,20
n-gram+treelet	68,54	80,73	77,56	62,20
bow+n-gram+treelet	80,98	85,85	84,15	77,07

NB et SVM, bow et bow+treelet

Comparaison de classifieurs NB et SVM : tests de *Student*

Groupe d'échantillons		*t*-statistique	*p*-value	*d* de Cohen	Taille d'effet
*bow*_nb	*bow*_svm	2,74	0,01	1,23	fort
*bow+treelet*_nb	*bow+treelet*_svm	2,10	0,05	0,94	fort
*bow*_nb	*bow+treelet*_nb	1,21	0,24	0,54	moyen
*bow*_svm	*bow+treelet*_svm	1,49	0,15	0,67	moyen

NB et SVM, bow et bow+treelet

Comparaison de classifieurs NB et SVM : tests de *Student*

Groupe d'échantillons		*t*-statistique	*p*-value	*d* de Cohen	Taille d'effet
*bow*_nb	*bow*_svm	2,74	0,01	1,23	fort
*bow+treelet*_nb	*bow+treelet*_svm	2,10	0,05	0,94	fort
*bow*_nb	*bow+treelet*_nb	1,21	0,24	0,54	moyen
*bow*_svm	*bow+treelet*_svm	1,49	0,15	0,67	moyen

$\Rightarrow$ NB permet des performances significativement supérieures à celles obtenues avec SVM

NB et SVM, bow et bow+treelet

Comparaison de classifieurs NB et SVM : tests de *Student*

Groupe d'échantillons		*t*-statistique	*p*-value	*d* de Cohen	Taille d'effet
*bow*_nb	*bow*_svm	2,74	0,01	1,23	fort
*bow+treelet*_nb	*bow+treelet*_svm	2,10	0,05	0,94	fort
*bow*_nb	*bow+treelet*_nb	1,21	0,24	0,54	moyen
*bow*_svm	*bow+treelet*_svm	1,49	0,15	0,67	moyen

$\Rightarrow$ NB permet des performances significativement supérieures à celles obtenues avec SVM

$\Rightarrow$ La perte de performance avec les *treelet* n'est pas significative.

Analyse des traits

Traits lexicaux

- Test de corrélation de *Spearman* pour évaluer la pré-dominance de certains traits lexicaux

Traits lexicaux

- Test de corrélation de *Spearman* pour évaluer la pré-dominance de certains traits lexicaux
- Exemples avec p-valeur $< 0,05$ et coefficient $|\rho| > 0,3$

Vocabulaire	ρ	p-value
Douleur		
maladie	0,540	$< 1e-3$
hospitalisé	0,509	$< 1e-3$
hallucinations	0,420	0,006
Éducation		
master	$-0,505$	$< 1e-3$
concours	$-0,496$	$< 1e-3$
fac	$-0,490$	0,001

Vocabulaire	ρ	p-value
Psycho		
psychologie	$-0,536$	$< 1e-3$
psychologue	$-0,453$	0,002
Déictique		
j' / je	0,635	$< 1e-5$
mon	0,613	$< 1e-5$
t' / tu	$-0,467$	0,002
nous	$-0,342$	0,028

Traits syntaxiques

- VERB : marqueur fort pour les SCZ
  - VERB $\xrightarrow{\text{Aux}}$ AUX (Ex. : "(j')ai fait", "(c')est (pas) gagné")
  - VERB $\xrightarrow{\text{Nsubj}}$ PRON (Ex. : "ça va", "(je) sais pas")

Traits syntaxiques

- VERB : marqueur fort pour les SCZ
  - VERB $\xrightarrow{\text{Aux}}$ AUX (Ex. : "(j')ai fait", "(c')est (pas) gagné")
  - VERB $\xrightarrow{\text{Nsubj}}$ PRON (Ex. : "ça va", "(je) sais pas")
- NOM : marqueur fort pour les TEM
  - Relation EXPL capture des nominaux explicatifs ou pléonastiques
  - Relation CASE est traitée comme le dépendant du nom

Conclusion

Conclusion

- Premier système identifiant des particularismes dans le discours des SCZ en français
- Tester différentes représentations :
  - Informations lexicales
  - Syntaxiques
- Tester différents classifieurs (*NB*, *SVM* et Régression logistique)
- Biais lexicaux dans les deux groupes
  - SCZ : environnement médical
  - TEM : études et scolarité

Perspectives

- Tester d'autres classifieurs : Random forest, Perceptrons
- Tester d'autres traits :
  - Linguistiques : lexicaux (mots déictiques), sémantiques (connecteurs), *etc.*
  - Extra linguistiques : résultats aux tests neuro-cognitifs
- Introduire le contexte : classification des TDP et pas cTDP

Merci !

Références i

Allende-Cid, H., Zamora, J., Alfaron-Faccio, P., and Alonso, M. (2019).
A machine learning approach for the automatic classification of schizophrenic discourse.
IEEE Access, pages 45544–45554.

Kayi, E. S., Diab, M., Pauselli, L., Compton, M., and Coppersmith, G. (2017).
Predictive linguistic features of schizophrenia.
In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (* SEM 2017)*, pages 241–250.

Références ii

Mitchell, M., Hollingshead, K., and Coppersmith, G. (2015).
Quantifying the language of schizophrenia in social media.
In *Proceedings of the 2nd workshop on Computational linguistics and clinical psychology: From linguistic signal to clinical reality*, pages 11–20.

Strous, R. D., Koppel, M., Fine, J., Nachliel, S., Shaked, G., and Zivotofsky, A. Z. (2009).
Automated characterization and identification of schizophrenia in writing.
The Journal of nervous and mental disease, 197(8):585–588.