Construction d'un corpus annoté en genre par apprentissage Zero-Shot.

Nicolas Béchet*, Rémy Kessler**, Gwen Icart***, Gudrun Ledegen***

* UNIV. BRETAGNE-SUD, UMR CNRS 6074 IRISA,
nicolas.bechet@irisa.fr
** UNIV. BRETAGNE -SUD ,
remy.kessler@univ-ubs.fr,,
*** UNIV. RENNES II, PREFICS
gwenicart@outlook.com
gudrun.ledegen@univ-rennes2.fr

Résumé. Afin de s'adapter au mieux au jeune public plus à l'aise avec les nouvelles technologies, une association a développé une application de webchat permettant à toute personne de partager ses angoisses. Plusieurs milliers de conversations anonymes ont ainsi été réunies et forment un corpus inédit de récits sur la détresse humaine, les violences sociales. Nous présentons dans cet article une méthodologie de production d'un modèle d'apprentissage permettant un étiquetage automatique en genre d'un corpus de texte en français. La méthode repose sur l'utilisation d'une combinaison d'un algorithme de classification Zero-Shot, d'une validation humaine et d'un apprentissage supervisé. Nous montrons que cette méthode permet de préannoter efficacement un corpus volumineux en présentant quelques résultats expérimentaux, validé par des experts.

1 Introduction

Depuis les années quatre-vingt-dix, la souffrance sociale est une thématique qui fait l'objet d'une grande attention de la part des pouvoirs publics ainsi que du milieu associatif. Parmi les conséquences, figure l'explosion des lieux d'écoute ou des dispositifs sociotechniques de communication dont les finalités consistent notamment à modérer les diverses formes de souffrance par la libération de la parole dans un but thérapeutique Fassin (2004, 2005). Une association de prévention du suicide a développé une application de webchat afin de répondre à ce besoin. Le webchat est un espace qui permet à toute personne d'exprimer et de partager avec un écoutant bénévole ses préoccupations, sa souffrance et ses angoisses. La principale spécificité de ce dispositif est son caractère non public et anonyme. Protégés par un pseudonyme, les appelants sont invités à confier auprès d'un bénévole les aspects problématiques de leur existence. Plusieurs milliers de conversations anonymes ont ainsi été réunies et forment un corpus inédit de récits sur la détresse humaine. Les travaux présentés dans ce papier s'inscrivent dans le cadre d'un projet de recherche lié à la prévention du suicide. Notre objectif dans ce projet est d'identifier différentes causes de souffrances parfois difficilement décelables, de caractériser ses modalités d'énonciation. Les premiers travaux réalisés ont notamment permis d'identifier

automatiquement la cause de la souffrance des personnes (Kessler et al., 2019). Les travaux présentés dans ce papier visent à proposer une méthodologie permettant la production d'un corpus annoté en genre. L'annotation en genre de notre corpus s'inscrit dans un processus global d'analyse de discours en vue d'une meilleure compréhension des interactions et de leur déroulement. Après une présentation des travaux connexes du domaine, nous présentons plus en détail le corpus au travers de quelques statistiques et d'un exemple. Puis, nous détaillerons dans la section suivante l'approche proposée ainsi que les différents résultats obtenus.

2 Travaux connexes

Dans cette section, nous présenterons brièvement les travaux connexes sur l'apprentissage Zero-Shot. La classification par apprentissage en l'absence de données étiquetées est un problème difficile, et la réalisation des performances meilleures que le hasard nécessite généralement l'introduction de connaissances préalables. Depuis les plongements de mots pré-entraînés (Mikolov et al., 2013; Pennington et al., 2014) jusqu'aux représentations textualisées ((Matthew et al., 2018; Schuster et al., 2019), l'apprentissage par représentation non supervisée a cependant considérablement amélioré l'état de l'art en compréhension de la langue écrite.

En exploitant différents types de représentation, les méthodes d'apprentissage Zero-Shot actuelles peuvent être réparties en deux groupes principaux (Li et al., 2015). Le premier, basé sur les caractéristiques, exploite les attributs partagés entre les catégories de classe (Madapana et Wachs, 2017; Wu et al., 2014), afin de fournir une représentation intermédiaire des étiquettes. Par exemple, des caractéristiques telles que "blanc", "quatre jambes" et "a une queue" peuvent être partagées entre des catégories d'animaux et peuvent fournir une caractéristique significative pour chaque étiquette de classe. Le principal inconvénient de cette approche est qu'elle nécessite une tâche fastidieuse d'annotation manuelle pour les associations classe-caractéristiques. Le second groupe, les méthodes textuelles (Elhoseiny et al., 2013; Rohrbach et al., 2011), extrait une représentation vectorielle intermédiaire à partir de grands corpus textuels tels que WordNet et Wikipédia. L'application de techniques de traitement de la langue afin d'extraire des attributs de manière automatique permet ainsi de considérablement réduire les besoins en annotations manuels. Par exemple, (Elhoseiny et al., 2013; Frome et al., 2013) s'appuient sur des données textuelles issues de l'encyclopédie Wikipédia pour apprendre les relations sémantiques entre étiquettes. (Elhoseiny et al., 2013) utilise la hiérarchie sémantique interne de WordNet pour extraire les caractéristiques de chaque catégorie. Cependant, ces représentations d'étiquettes textuelles sont construites indépendamment de l'entraînement des classificateurs et ne sont pas donc pas optimisées pour une tâche précise.

L'approche par "prompt based learning" issue des agents conversationnels (Madotto et al., 2021) permet d'effectuer un apprentissage avec peu d'exemples, mais nécessiterait de prendre en compte les conversations des écoutants, ce qui n'a pas été le cas dans notre analyse. Proche de notre tâche, la classification en âge et en genre est une tâche récurrente de PAN depuis 2013 afin d'améliorer la détection de style d'auteurs. (Modaresi et al., 2016) propose ainsi l'extraction de certaines caractéristiques stylistiques et lexicales pour la formation d'un modèle de régression logistique.

Nos travaux sont proches des approches classiques à base d'apprentissage Zero-Shot puisque nous utilisons des données d'entraînements issues de larges corpus pour prédire une étiquette et, de manière similaire à (Elhoseiny et al., 2013), nous utilisons uniquement l'information

textuelle comme représentation intermédiaire. L'originalité de ces travaux réside dans le domaine d'application de ces méthodes sur des textes courts, issus de webchat ainsi que dans l'utilisation de l'apprentissage Zero-Shot pour pré-étiqueter un corpus, qui sera ensuite vérifié manuellement. Sans cette étape, l'étiquetage manuel aurait été bien plus chronophage. L'objectif final restant la production d'un modèle d'apprentissage supervisé.

3 Données et statistiques

L'association a fourni à l'équipe de recherche un corpus de conversations entretenues entre les bénévoles et des appelants entre 2005 et 2015. La figure 1 présente un extrait anonymisé de conversation issue de ce corpus. Une des caractéristiques de ce corpus est la présence de phénomènes linguistiques bien particuliers comme des émoticônes, des apocopes (par exemple « ado », « télé », « bi ») des acronymes, des variations (orthographiques, typographiques, mots collés, d'une très grande morphovariabilité). Ces phénomènes doivent leur origine au mode de communication, à la rapidité de composition du message ou aux contraintes technologiques de saisie imposées par le matériel (terminal mobile, tablette, etc.). Le corpus contient 28 422 conversations avec une moyenne de 698 mots par conversation, pour un total de 2 276 973 mots dont 158 361 mots différents.

...
Appelant(12 :06 :36) : J'espère arrivé à m'en sortir un jour . .
Chat-Accueil(12 :06 :59) : je n'arrive pas à savoir si vous êtes un garçon ou une fille
Appelant(12 :06 :59) : Une fille
Appelant(12 :07 :11) : avec un caractére de garçon
Chat-Accueil(12 :07 :37) : oui c'est ce que je pensais.. avec de l'humour en tous cas !
...

FIG. 1 – *Extrait d'une discussion issue du corpus SADSui.*

4 Méthodologie

Principe général de l'approche Le principe général de l'approche est présenté dans la figure 2. La première étape consiste à extraire une graine de 300 conversations qui seront annotées manuellement. Ces dernières vont alors permettre de sélectionner un modèle de classification Zero-Shot. Avec ce modèle, un corpus de 2 396 conversations vont être annotées et vérifiées par des annotateurs humains. Finalement, une étape de classification supervisée est réalisée dans le but d'étiqueter un corpus de 28 422 conversations en genre.

Étiquetage de la graine La première étape est la production d'une graine d'apprentissage de 300 conversations qui ont été étiquetées manuellement en genre. Cet étiquetage a permis d'identifier une liste de structures réunissant les contextes d'identification de genre dans la prise de parole des appelants. Les contextes qui nous apparaissent comme indiquant clairement le genre de la personne qui appelle sont des qualificatifs (adjectifs, noms) dont la forme

Construction d'un corpus annoté en genre par apprentissage Zero-Shot.

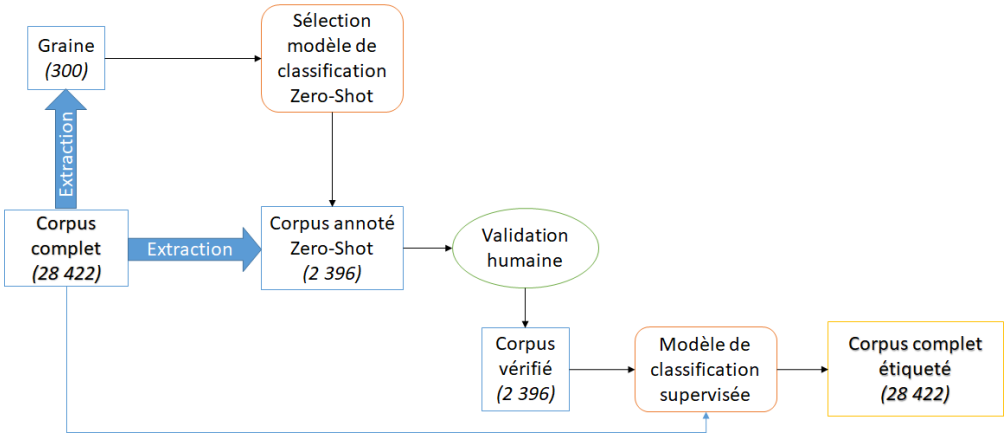

FIG. 2 – *Visualisation de l'approche*

féminine est audible. Blanche-Benveniste (2010) rappelle en effet qu'environ 65% des adjectifs en français ne présentent pas de variation audible et que la morphologie soustractive ('il consiste à former le masculin en soustrayant la consonne finale de la forme longue qui se trouve dans le féminin et dans la base dérivationnelle : blonde / blond') est le "procédé morphologique actuellement le plus important" pour les adjectifs et participes passés en français. Les mêmes observations sont valables pour les formes masculines. Les formes d'adjectifs au féminin qui ne s'entendent pas (énervée, tendue ...) varient souvent fortement en orthographe dans le corpus chat (Lucci et Millet, 1994) et nous les considérons comme des indices fiables uniquement en cas de répétition (a minima 3 / conversation) et en l'absence de variation (masculin et féminin). Une étiquette 'inconnue' est choisie lorsqu'il n'y a pas assez d'éléments pour déterminer le genre ou à l'inverse qu'il existe plusieurs marqueurs contradictoires.

Algorithme de classification Zero-Shot Une fois la graine constituée, nous avons testé différentes méthodes afin de permettre une annotation complète du corpus. La taille de la graine n'étant que de 300 documents, nous n'avons pas été en mesure de produire des résultats intéressants avec une approche classique d'apprentissage supervisé. Ainsi, nous avons exploré les algorithmes de *Zero-Shot Learning*. Ces méthodes ont la particularité de ne nécessiter aucune donnée d'apprentissage et fournissent des résultats de relativement bonne qualité sur des tâches de classification de données textuelles. Le modèle retenu pour la production du corpus annoté en genre est *xlm-roberta-large-xnli* qui a été affiné (fine-tuned) à partir du modèle de connaissance *xlm-roberta-large* (Conneau et al., 2020) sur une tâche d'inférence de langage naturel (NLI) à partir de deux jeux de données XNLI (Conneau et al., 2018) qui comporte 15 langues dont le français. XLNI est un sous-ensemble du jeu de données en anglais MultiNLI [1] (Williams et al., 2018) que les auteurs ont traduit en 14 autres langues, produisant ainsi une ressource permettant l'entraînement multilingue de modèle d'inférence de langage naturel. La particularité de ces jeux de données est de reposer sur la notion de prémisse et d'hypothèse.

1. Multi-Genre Natural Language Inference

Dès lors, l'utilisation de ce modèle pour une tâche de classification repose sur le même principe avec un modèle d'hypothèse à fournir en entrée ainsi qu'une liste d'étiquettes candidates. L'hypothèse contient un masque que le modèle remplacera par les différentes étiquettes fournies. Le tableau 1 présente une partie des hypothèses utilisées dans le cadre de ces travaux. Finalement, le modèle fournira en sortie une probabilité d'appartenance pour chaque étiquette.

Id	Hypothèse	Etiquettes
1	Je suis plutôt {}.	une femme, un homme
2	Cet article est écrit par une personne de sexe {}.	masculin, féminin, inconnu
3	Je suis {}.	une femme, un homme, un inconnu
4	Je suis {}.	une femme, un homme, une fille, un garçon, un inconnu

TAB. 1 – *Extraits des hypothèses utilisées par le modèle de classification*

Généralisation par apprentissage La dernière étape de notre méthodologie est, une fois les données annotées en genre par l'algorithme de Zero-Shot et validées par un annotateur humain, de construire un modèle d'apprentissage supervisé permettant la prédiction du genre de la personne ayant rédigé le texte de la conversation. Ce modèle sera finalement évalué et utilisé pour produire un corpus de 28 422 conversations annotées en genre.

5 Expériences

Résultats de l'apprentissage Zero-Shot sur la graine Nous présentons dans cette section les résultats obtenus sur la graine annotée manuellement qui nous ont permis de sélectionner le meilleur modèle pour la première étape d'annotation. Différents couples (hypothèse, étiquettes) pour l'algorithme de classification Zero-Shot *xlm-roberta-large-xnli* ont été testés comme précisé en section 4. Seuls les meilleurs sont présentés, à savoir ceux obtenus avec le couple numéro 4 du tableau 1. Cette combinaison (hypothèse, étiquettes) comporte cinq étiquettes qui ont été ramenées à 3, une fois l'étiquette prédite, selon les règles suivantes : "*une fille*" devient "*une femme*" ; "*un garçon*" devient "*un homme*". Ces résultats sont de relative-

(genre)	Précision	Rappel	F-score	Support
un inconnu	0.783	0.938	0.854	146
un homme	0.857	0.566	0.682	53
une femme	0.867	0.772	0.817	101
Macro moyenne	0.836	0.759	0.784	300

TAB. 2 – *Résultats obtenus sur la détection de genre avec le couple numéro 4*

ment bonne qualité, en sachant que le modèle n'a jamais été entraîné spécifiquement sur la tâche de prédiction de genre.

Résultats de l'apprentissage supervisé Une fois la combinaison (hypothèse, étiquettes) du modèle de Zero-Shot sélectionnée, nous l'avons utilisé pour étiqueter une partie plus conséquente du jeu de données ce qui a permis d'obtenir un corpus de 2 396 conversations étiqueté

automatiquement. Ces annotations ont alors été vérifiées manuellement ce qui a permis la production d'un corpus de qualité étiqueté en genre de 2 396 conversations. Dès lors, la seconde étape de notre méthodologie consiste à entraîner des modèles d'apprentissage supervisé à partir de ce corpus et d'en vérifier la qualité. Deux modèles d'apprentissage ont été utilisés : un SVM linéaire avec apprentissage par descente de gradient stochastique - SGD (Zhang, 2004) et l'algorithme CamemBERT (Martin et al., 2020), un modèle RoBERTa affiné sur la langue française. Pour vérifier la qualité des modèles produits, nous avons pour SGD divisé aléatoirement [2] le corpus en données d'apprentissage [3] et test avec une répartition de 80-20 et pour CamemBERT en données d'apprentissage, de validation et de test avec une répartition de 60-20-20. Ainsi, le jeu de test comporte 480 conversations. Nous utiliserons comme *baseline* les résultats qui ont été obtenus avec le modèle Zero-Shot *xlm-roberta-large-xnli*. Les macro-moyennes et le taux d'erreur sont présentés dans le tableau 3.

(macro-moyenne)	Précision	Rappel	F-score	Tx Erreur
Baseline	0,806	0,758	0,773	0,194
SGD	0,799	0,759	0,774	0,181
CamemBERT	**0,882**	**0,885**	**0,883**	**0,104**

TAB. 3 – *Macro-moyennes des précisions, rappels et f-scores et taux d'erreur obtenus à partir de l'apprentissage du nouveau corpus*

Ces résultats montrent que l'apprentissage avec CamemBERT produit des résultats de bien meilleure qualité que ceux obtenus avec la baseline (xlm-roberta-large-xnli). Cependant, comme le montre le tableau 4, les résultats obtenus sur le genre "homme" sont encore perfectibles et ne semblent pas encore utilisables sans validation humaine. Nous émettons l'hypothèse que la faible représentativité de la classe dans le corpus d'apprentissage peut en partie expliquer les difficultés rencontrées sur le genre "homme" [4].

(genre "homme")	Précision	Rappel	F-score
Baseline	0,768	0,606	0,677
SGD	0,736	0,549	0,629
CamemBERT	**0,833**	**0,845**	**0,839**

TAB. 4 – *Précisions, rappels et f-scores obtenus pour le genre "Homme"*

6 Conclusion

Nous avons présenté dans cet article une méthodologie permettant la production d'un corpus en français annoté en genre. Notre démarche a l'originalité d'utiliser dans une première étape un algorithme de classification Zero-Shot permettant, avec l'appui d'une validation humaine, la construction d'un corpus d'apprentissage. Dès lors, un modèle d'apprentissage su-

2. nous avons effectué un découpage stratifié

3. dont 10% sont utilisés pour la validation dans l'implémentation de l'algorithme

4. L'association qui nous a fourni le corpus nous a indiqué que près de trois quarts de leur public est féminin, ce qui explique cette faible représentativité

pervisé est utilisé afin d'étiqueter en genre l'ensemble des 28 422 conversations. Nos expérimentations ont montré que le modèle était d'une qualité acceptable avec un taux d'erreur d'un peu plus de 10%. Ainsi, nous envisageons comme futurs travaux de l'étudier plus en profondeur afin d'identifier les causes des erreurs. Par ailleurs, une méthode à base d'apprentissage actif sera aussi étudiée afin de produire un oracle efficace permettant un choix optimum de nouveaux exemples à étiqueter.

Références

Blanche-Benveniste, C. (2010). Le français. Usages de la langue parlée. *Leuven-Paris. Peeters 1*(1), pp. 104–106.

Conneau, A., K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, et V. Stoyanov (2020). Unsupervised cross-lingual representation learning at scale.

Conneau, A., R. Rinott, G. Lample, A. Williams, S. R. Bowman, H. Schwenk, et V. Stoyanov (2018). Xnli : Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Elhoseiny, M., B. Saleh, et A. M. Elgammal (2013). Write a classifier : Zero-shot learning using purely textual descriptions. In *ICCV 2013, Sydney, Australia, December 1-8, 2013*, pp. 2584–2591.

Fassin, D. (2004). Et la souffrance devint sociale. In *Critique, 680(1)*, pp. 16–29. Critique.

Fassin, D. (2005). Souffrir par le social, gouverner par l'écoute. In *Politix,73(1)*, pp. 137–157.

Frome, A., G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. A. Ranzato, et T. Mikolov (2013). Devise : A deep visual-semantic embedding model. In *Advances in Neural Information Processing Systems*, Volume 26. Curran Associates, Inc.

Kessler, R., N. Béchet, G. Ledegen, et F. Pugnière-Saavedra (2019). Word embedding approach to explore a collection of discussions of people in psychological distress. *Proceedings of DDP 2019, London, United Kingdom 1*(1), pp 18–21.

Li, X., Y. Guo, et D. Schuurmans (2015). Semi-supervised zero-shot classification with label representation learning. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 4211–4219.

Lucci, V. et A. Millet (1994). L'orthographe de tous les jours. Enquête sur les pratiques orthographiques des Français. *Paris, Champion. 1*(1), pp. 126–129.

Madapana, N. et J. Wachs (2017). Zsgl : Zero shot gestural learning. In *ICMI*, New York, NY, USA, pp. 331–335. Association for Computing Machinery.

Madotto, A., Z. Lin, G. I. Winata, et P. Fung (2021). Few-shot bot : Prompt-based learning for dialogue systems.

Martin, L., B. Muller, P. J. Ortiz Suárez, Y. Dupont, L. Romary, É. V. de la Clergerie, D. Seddah, et B. Sagot (2020). Camembert : a tasty french language model. In *ACL*.

Matthew, P., N. Mark, I. Mohit, G. Matt, C. Christopher, L. Kenton, et Z. Luke (2018). Deep contextualized word representations. In *Proceedings of NAACL2018*, USA, pp. 2227–2237.

Mikolov, T., I. Sutskever, K. Chen, G. Corrado, et J. Dean (2013). Distributed representations of words and phrases and their compositionality. In *Proceedings of NIPS'13*, USA, pp. 3111–3119. Curran Associates Inc.

Modaresi, P., M. Liebeck, et S. Conrad (2016). Exploring the Effects of Cross-Genre Machine Learning for Author Profiling in PAN 2016. In *CLEF 2016*, Workshop Proceedings.

Pennington, J., R. Socher, et C. Manning (2014). Glove : Global vectors for word representation. In *Proceedings of EMNLP2014*, Qatar, pp. 1532–1543.

Rohrbach, M., M. Stark, et B. Schiele (2011). Evaluating knowledge transfer and zero-shot learning in a large-scale setting. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '11, USA, pp. 1641–1648. IEEE Computer Society.

Schuster, T., O. Ram, R. Barzilay, et A. Globerson (2019). Cross-lingual alignment of contextual word embeddings, with applications to zero-shot dependency parsing. In *Proceedings of the 2019 Conference of the North American Chapter of the ACL : Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota, pp. 1599–1613.

Williams, A., N. Nangia, et S. Bowman (2018). A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the ACL : Human Language Technologies, Volume 1 (Long Papers)*, pp. 1112–1122.

Wu, S., S. Bondugula, F. Luisier, X. Zhuang, et P. Natarajan (2014). Zero-shot event detection using multi-modal fusion of weakly supervised concepts. In *CVPR'14*, Columbus, OH, pp. 2665–2672.

Zhang, T. (2004). Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *Proceedings of ICML*, New York, NY, USA, pp. 116. Association for Computing Machinery.

Summary

In order to best adapt to new technologies, an association has developed a webchat application allowing anyone to express and share their anxieties. Several thousand anonymous conversations have then been brought together and form an unprecedented corpus of stories about human distress and social violence. We present in this paper a methodology to produce a learning model that allows an automatic gender labeling of a corpus of texts in French. The method is based on a combination of a Zero-Shot classification algorithm, human validation, and supervised learning. This method allows us to effectively pre-annotate a large corpus by presenting some experimental results so that an expert can finally more easily validate the annotation produced.