Désambiguïsation des mots polysémiques de la ville dans des romans de science-fiction

Sami Guembour, Catherine Dominguès

Désambiguïsation des mots polysémiques de la ville dans des romans de science-fiction

Sami Guembour, Catherine Dominguès

LASTIG, Univ Gustave Eiffel, ENSG, IGN, France –
{sami.guembour, catherine.domingues}@ign.fr

Abstract

This work is part of a project on the representation of the city of the future in a corpus of science fiction novels. The city's words were identified using an existing terminology resource. Some of these words are polysemous or can be used in contexts other than urban description. This article proposes a method to disambiguate them. It uses the CamemBERT language model. Classifiers were fine-tuned to determine the employment context: urban/non-urban. The evaluations on two different corpora show very significant efficiency, which reflects a possibility of distinction in the use of these words.

Keywords: NLP – disambiguation – polysemy – CamemBERT – embedding vector – Fine-Tuning – classification – city – science fiction – corpus

Résumé

Ce travail s'insère dans le contexte d'un projet sur la représentation de ville du futur dans un corpus de romans de science-fiction. Les mots de la ville ont été identifiés grâce à une ressource terminologique existante. Certains de ces mots sont polysémiques ou peuvent être utilisés dans d'autres contextes que la description urbaine. Cet article propose une méthode pour les désambiguïser. Elle utilise le modèle de langue CamemBERT. Des classifieurs sont affinés (fine-tuning) pour déterminer le contexte d'emploi : urbain/non urbain. Les évaluations sur deux corpus différents montrent une efficacité très importante, ce qui traduit une possibilité de distinction dans l'emploi de ces mots.

Mots clés : TAL – désambiguïsation – polysémie – CamemBERT – vecteur de plongement – affinement – classification – ville – science-fiction – corpus

1. Contexte et introduction

PARVIS[1] est un projet scientifique et institutionnel porté par l'université Gustave-Eiffel entre 2019 et 2023. Il visait à analyser les représentations de la ville future afin de mettre en lumière les thèmes et les défis associés aux imaginaires urbains futuristes, notamment en lien avec les enjeux du changement climatique. Il a combiné plusieurs domaines de recherche, tels que la littérature, le traitement automatique des langues, la géographie, l'architecture, la musique et la création littéraire dans une démarche pluridisciplinaire et a développé une démarche de recherche-création qui a produit des œuvres littéraires, sonores, vidéo et théâtrales. Parmi les sources identifiées comme racontant des imaginaires futuristes urbains figurent des romans décrivant des fictions climatiques. Un corpus d'une centaine de romans

[1] PARVIS pour Paroles de villes a tenu un carnet de recherche : https://parvis.hypotheses.org/. Il a été financé par I-SITE FUTURE (http://www.future-isite.fr/accueil/?L=0)

(dorénavant corpus PARVIS) a été constitué et a donné matière à différents travaux. L'un d'eux (Guembour *et al.*, 2023) a proposé de caractériser la ville du futur à partir des éléments urbains (objets et lieux) décrits dans ce corpus, et des fonctions urbaines et sociales associées à ces éléments. Afin d'identifier ces éléments narratifs propres à décrire la ville, une ressource terminologique a été construite comme un sous-ensemble du lexique de l'ouvrage *Les mots de la ville*[2] (Topalov *et al.*, 2010) proposé par une équipe de chercheur.es composée d'urbanistes, architectes, historien.nes. La méthode mise en place dans ce précédent article reposait sur le repérage et l'analyse des romans du corpus dans lesquels la ville constitue un contexte essentiel, ceux-ci étant identifiés comme les romans dans lesquels les mots de la ressource terminologique sont les plus présents. Cependant certains de ces mots sont polysémiques. Le risque est alors que le nombre d'occurrences de ces mots dans le corpus ne soit surestimé à cause de l'addition sous le même lemme des occurrences dont les sens sont différents et ne décrivent pas la ville, ce qui fausserait les résultats de la méthode développée. Par exemple, la forme *cité* peut avoir pour lemme le nom *cité* (qui appartient au lexique de la ville) mais aussi être le participe passé du verbe *citer* qui n'appartient pas au lexique de la ville. En outre l'objectif est aussi d'éliminer les mots qui, sans être réellement polysémiques, ne sont pas employés dans un contexte urbain, comme par exemple *centre de la ville* (contexte urbain) et *centre de la Terre* (contexte non urbain).

Dans cet article, nous proposons une méthode de désambiguïsation de ces mots polysémiques, fondée sur la classification des vecteurs de plongement de phrases contenant un mot polysémique. Le choix de développer une nouvelle méthode plutôt que d'utiliser les méthodes de désambiguïsation existantes est justifié par le fait que ces méthodes permettent de désambiguïser le sens des mots sans offrir la possibilité de différencier leur contexte d'emploi. Ainsi, elles ne permettent pas d'identifier dans quels cas ces mots ont été utilisés spécifiquement pour parler de la ville. De plus, la plupart des méthodes de désambiguïsation existantes ne sont pas adaptées pour le français et/ou reposent sur des étiquettes grammaticales, ce qui ne répond pas à la problématique soulevée dans cet article. La méthode ne peut donc pas se limiter à l'identification de l'étiquette grammaticale du mot polysémique pour décider si son emploi est associé ou non à la ville puisque, par exemple pour le mot *place*, l'étiquette dans l'expression *la place de la ville* et *à la place de* est dans les deux cas *nom*.

Cet article est organisé en six sections. Un état de l'art concernant des travaux déjà réalisés sur la désambiguïsation de mots est présenté dans la section 2. La section 3 détaille le corpus PARVIS et fournit quelques indicateurs textométriques. La méthode proposée pour la désambiguïsation des mots polysémiques est décrite dans la section 4. L'évaluation de la méthode est double : d'abord sur le corpus PARVIS, ensuite afin d'évaluer la robustesse de la méthode, sur d'autres données moins normées ; l'ensemble de ces évaluations ainsi que les résultats de la désambiguïsation sont présentés en section 5. Enfin, la section 6 présente quelques conclusions et perspectives.

2. État de l'art

Différentes méthodes fondées sur les modèles de langues et les vecteurs de plongement de mots qu'ils proposent, ont été élaborées pour identifier les mots polysémiques et les

[2] Ce lexique regroupe 533 mots, majoritairement des noms, qui désignent des éléments de la ville répartis en quatre thématiques : Agglomération, Circulation, Division et Habitation

désambiguïser. Dans (Kågebäck et Salomonsson, 2016), un modèle dédié à la désambiguïsation des mots a été présenté. Ce modèle s'appuie sur les vecteurs de mots générés par Glove (Pennington *et al.*, 2014) et utilise un réseau de neurones bi-LSTM partagé pour tous les mots, facilitant ainsi l'adaptation du modèle à la diversité du vocabulaire. Le modèle est entraîné directement à partir du texte brut, ce qui lui permet de capturer les étiquettes et utiliser efficacement l'ordre des mots dans les phrases. Le modèle ne prédit le sens (dans un ensemble prédéfini) que d'un seul lemme polysémique de la phrase.

Au contraire, (Raganato *et al.*, 2017) prédisent une étiquette sémantique pour chaque mot en entrée, à l'aide d'un modèle LSTM. Ce modèle est enrichi par une couche d'attention et un entraînement multitâche, permettant une annotation plus efficace de tous les mots d'une séquence.

Dans une approche différente, (Du *et al.*, 2019) ont utilisé le modèle BERT (Devlin *et al.*, 2019) fondé sur les transformeurs pour désambiguïser les mots polysémiques d'une phrase. Leur méthode, fondée sur le calcul de vecteurs de plongement des mots et l'utilisation de réseaux de neurones MLP, a rencontré des difficultés avec les mots hors vocabulaire[3], conduisant à l'exploration de stratégies alternatives qui s'appuient sur l'utilisation des mots qui ont une similarité sémantique avec les mots hors vocabulaire.

Concernant la désambiguïsation en langue française, (Norré *et al.*, 2023) ont examiné la traduction de textes français en pictogrammes. Ils ont utilisé plusieurs modèles de langue, dont CamemBERT (Martin *et al.*, 2019), FlauBERT (Le *et al.*, 2020), DrBERT (Labrak *et al.*, 2023), et CamemBERT-bio (Touchent *et al.*, 2023), pour générer les vecteurs de plongement des phrases. Ces vecteurs ont ensuite été comparés aux vecteurs de sens des mots polysémiques de WordNet (Miller, 1994) en utilisant la mesure de similarité cosinus. Cette méthode leur a permis de désambiguïser les mots polysémiques et d'assurer une traduction cohérente en pictogrammes.

Enfin, (Janati *et al.*, 2023) ont proposé une méthode de désambiguïsation fondée sur CamemBERT, exploitant la similarité entre les vecteurs de plongement de mots pour identifier leurs sens.

3. Présentation du corpus PARVIS

Le corpus PARVIS est constitué de 131 romans[4]. Il est le résultat d'une recherche qui croise la littérature critique sur la fiction climatique, les listes de fiction climatique établies par les lecteurs et les bases de données de la Bibliothèque nationale de France[5]. Les romans sélectionnés font référence, du point de vue de leur narration, de leurs thèmes et de leur esthétique, concernant les deux sujets d'étude de PARVIS : le changement climatique et la vie urbaine. Ils décrivent un monde futur qui a subi des changements funestes dus au climat, que le progrès scientifique et technique n'a pu enrayer, conduisant à la mise en place de dystopies violentes, liberticides ou totalitaires.

[3] Il s'agit des mots qui ne font pas partie du vocabulaire d'entraînement. Ils sont donc mal représentés par le modèle de langue.
[4] La construction du corpus a été réalisée par Nadège Perelle ; elle est décrite de manière plus précise dans le carnet de recherche du projet : https ://parvis.hypotheses.org/3400
[5] https ://www.bnf.fr/fr

Ils proviennent principalement de l'aire culturelle nord-ouest, en particulier des pays francophones, britanniques et américains. Le point de départ du corpus est l'année 1961, année de publication de *''The Wind from Nowhere''* de J.G. Ballard, qui est considéré comme le précurseur et l'exemple emblématique du sous-genre de la fiction climatique au sein de la science-fiction. Il s'achève en 2020 afin d'exclure les œuvres créées pendant la période liée à la pandémie de COVID-19. Tous les romans du corpus PARVIS sont en français, soit traduits, soit écrits directement dans cette langue. Les 131 romans composant le corpus totalisent 1 056 287 phrases et 29 038 420 mots (d'après un comptage effectué à l'aide de la bibliothèque NLTK de Python (Bird et al., 2009)).

4. Méthode proposée pour la désambiguïsation

Il s'agit de désambiguïser les mots de la ressource terminologique qui sont polysémiques afin d'identifier, dans le corpus PARVIS, les occurrences d'emploi en relation avec le thème de la ville (comme *cité* dans l'exemple précédent), et celles associées à d'autres thèmes. Parmi les 30 mots du lexique les plus fréquents dans le corpus, 15 sont polysémiques : *allée, base, centre, cité, cœur, cour, enceinte, ferme, lieu, marché, montée, place, porte, tente* et *tour*. La figure 1 classe par nombre d'occurrences décroissant les 30 mots du lexique de la ville les plus fréquents dans le corpus PARVIS. La barre noire donne le nombre d'occurrences des 15 mots polysémiques avant désambiguïsation. *porte* est le mot du lexique de la ville le plus fréquent dans le corpus. Des mots comme *ville* ou *rue*[6] sont considérés comme non polysémiques parce que même dans des emplois figés ou métaphoriques leur caractéristique locative et urbaine est présente de manière essentielle, ce qui n'est pas le cas de *place* (*à la place de*) ou *tour* (*faire le tour de la question*) ; ils sont représentés par une barre gris clair dans la figure.

La méthode proposée vise à donner, pour chaque occurrence de mot polysémique, une réponse booléenne à la question : "le sens de cette occurrence est-il associé à la ville ?" ; ainsi, pour un mot donné, deux catégories sont définies et toutes les occurrences qui ne sont pas associées à la ville sont dans la même catégorie. La méthode repose sur l'annotation des phrases qui contiennent les mots polysémiques (cf. 4.1), et sur la classification des vecteurs de plongement de ces phrases à l'aide de classifieurs entraînés *ad hoc* (cf. 4.2). Ainsi, un classifieur est construit par entraînement pour chaque mot polysémique à désambiguïser. Le classifieur classe chaque phrase contenant un (et un seul) mot polysémique, à l'aide de son vecteur de plongement qui est calculé à partir du vecteur de plongement de chacun de ses composants.

4.1. Annotation des phrases

Des phrases ont été annotées pour construire des jeux de données destinés à l'entraînement des classifieurs, ainsi que des jeux de données pour les évaluer. Ces phrases ont été annotées de manière à attribuer l'étiquette "1" lorsque le mot polysémique de la phrase fait référence à la ville (comme : *Le virus se déplaçait d'un quartier de la cité à l'autre*, où *cité* est le mot à désambiguïser), et "0" dans le cas contraire (comme : *Elle l'a cité comme une sorte de réincarnation*).

[6] La forme rue peut aussi avoir pour lemme le verbe ruer. Après recherche dans le corpus, ce cas s'est avéré suffisamment rare pour que le mot ne soit pas considéré comme polysémique.

Les phrases figées ou métaphoriques dans lesquelles le mot du lexique n'est pas employé dans le contexte de la ville, ont été annotées "0", (comme : *de l'eau plein la tête à la place de la cervelle*, *l'individualisme avait pris place*).

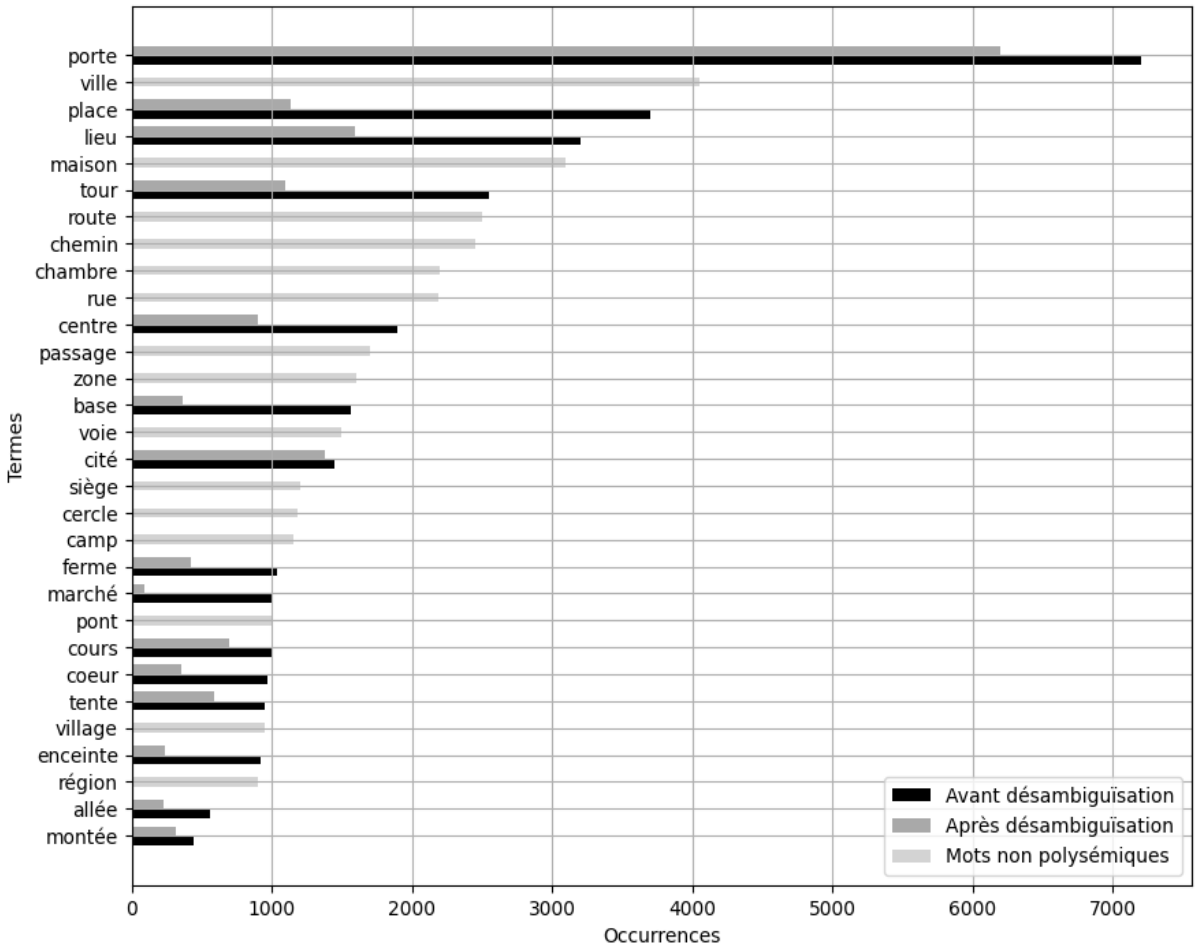

***Figure 1**. Les 30 mots du lexique les plus fréquents dans le corpus avant et après la désambiguïsation des 15 mots polysémiques. Les nombres d'occurrences des mots non polysémiques sont représentés chacun par une barre gris clair.*

4.1.1. Jeux de données de PARVIS

Un jeu de données a été constitué pour chacun des 15 mots polysémiques du lexique les plus fréquents dans le corpus. Chacun d'entre eux regroupe des phrases extraites corpus PARVIS, des exemples dans lesquels les mots polysémiques sont employés pour parler de la ville et aussi d'exemples où ces mots ont d'autres sens. 80% des phrases de ces jeux de données ont été destinées à l'entraînement des classifieurs, et 20 % ont été exclusivement réservées à leur évaluation.

Le tableau 1 indique le nombre total de phrases de chacun des jeux de données ainsi que le nombre de phrases correspondant à chaque étiquette.

Jeu de données PARVIS	# total de phrases	# de phrases avec l'étiquette "0"	# de phrases avec l'étiquette "1"
allée	80	41	39
base	63	35	28
centre	90	54	36
cité	80	27	53
cœur	50	30	20
cour	50	28	22
enceinte	40	18	22
ferme	40	21	19
lieu	56	28	28
marché	50	31	19
montée	60	39	21
place	50	28	22
porte	50	22	28
tente	50	22	28
tour	73	35	38

Table 1. *Description de chaque jeu de données de PARVIS*

4.1.2. Jeux de données du GDN

Les phrases composant les jeux de données de PARVIS proviennent des romans, c'est-à-dire un corpus littéraire dont la typographie est signifiante et cohérente, le vocabulaire approprié et la correction syntaxique fiable. Mais ces caractéristiques ne sont pas partagées par tous les corpus pertinents pour la description de la ville du futur.

Afin d'évaluer la robustesse des classifieurs, ceux-ci ont aussi été testés sur des exemples extraits d'un corpus dont les caractéristiques sont différentes. Il s'agit du Grand Débat National[7] (GDN). Ce corpus est constitué de contributions écrites librement sur une plate-forme par des milliers de contributeurs et contributrices, et il présente moins de garanties en termes de syntaxe et de cohérence des phrases. De plus, le thème de la ville est très présent dans les préoccupations citoyennes même orientées vers d'autres problématiques comme la mobilité ou les politiques publiques, ce qui entraîne des emplois fréquents des mots du lexique de la ville dans des contextes différents, rendant nécessaire la désambiguïsation de leur occurrences. Les phrases des jeux de données de ce corpus ont été annotées de la même manière que celle expliquée précédemment.

4.2. Entraînement des classifieurs

La méthode proposée repose sur la construction de classifieurs de vecteurs de plongement de phrases, calculés à l'aide du modèle de langue *CamemBERT* (Martin *et al.*, 2019). Chaque classifieur a été entraîné sur les phrases des jeux de données de PARVIS en affinant le modèle *CamemBERT* (fine-tuning) à l'aide de la fonction de classification par séquence du modèle (*CamembertForSequenceClassification*). Cela signifie que chaque classifieur est destiné à classifier la phrase contenant le mot polysémique sur lequel il a été entraîné.

[7] Le GDN a été mis en place par le gouvernement français en janvier 2019 en réponse au mouvement des Gilets Jaunes : https ://www.gouvernement.fr/le-grand-debat-national

L'architecture du modèle utilisée est *camembert-base*, et les hyperparamètres du Fine-Tuning ont été définis comme suit : un batch size de *16*, une longueur maximale de *128* pour les séquences, et un maximum de *10* epochs.

5. Évaluations des classifieurs et résultats de la désambiguïsation

L'évaluation des classifieurs est fondée sur deux indicateurs : l'exactitude (accuracy) et la F-mesure. La table 2 présente ces indicateurs sur les deux jeux de données d'évaluation de PARVIS et GDN. Elle met en évidence l'efficacité de la méthode proposée sur le corpus PARVIS avec une moyenne d'accuracy de 96%. Dix des 15 classifieurs ont obtenu une classification correcte de 100 % ; quatre ont atteint des pourcentages d'exactitude de 90 %, ce qui rend les classifications très performants. Le classifieur affichant la performance la moins élevée a été entraîné sur des phrases contenant le mot *centre* et a obtenu une précision de 85%. Cette baisse peut s'expliquer par le fait que *centre* partage un trait sémantique de localisation avec d'autres contexte que celui de la ville, comme dans l'expression *le centre de la terre*. En général, les classifieurs présentent des performances moindres sur le corpus GDN. Une explication possible réside dans la syntaxe des phrases de ce corpus, qui n'est pas toujours correcte au sens où elle ne fournit pas d'indices systématiquement fiables pour la construction du sens, exploitables par les classifieurs. Malgré cette difficulté, les classifications demeurent majoritairement correctes, avec des pourcentages d'exactitude des classifieurs variant entre 70 % et 100 %, et une moyenne de 86 %.

Les évaluations indiquées dans le tableau 2 montrent que la méthode proposée affiche des résultats de classification meilleurs que ceux présentés par les méthodes de l'état de l'art, avec une f-mesure de 0.97, surpassant ainsi les scores de 0.83 pour (Raganato *et al.*, 2017), et 0.78 pour (Du *et al.*, 2019). Ces valeurs représentent les meilleurs résultats de f-mesure obtenus par les méthodes décrites dans l'état de l'art et évaluées sur différentes jeux de données.

La barre en gris foncé dans la figure 1 représente le nombre d'occurrences des 30 mots du lexique les plus fréquents dans le corpus PARVIS après désambiguïsation. On observe que certains mots polysémiques très fréquents ne sont finalement que peu utilisés pour décrire des éléments de la ville, ce qui change le classement des mots les plus fréquents tel qu'il est présenté dans cette figure (voir le nouveau classement dans le tableau 3). Le tableau 3 donne aussi le nombre d'occurrences de ces 15 mots polysémiques après leur désambiguïsation. Le mot du lexique de la ville le plus fréquent dans le corpus avant désambiguïsation, *porte*, le reste après. Ses emplois sont majoritairement associés à l'urbain en tant que fermeture d'un lieu à géométrie variable (une ville, une maison, un placard, etc.). Le mot *enceinte* qui, lorsqu'il désigne une infrastructure de la ville a la même fonction de fermeture, est proportionnellement moins utilisé dans un sens associé à l'urbain (majoritairement, il est employé dans le corpus comme partie de l'expression *femme enceinte*). Cet emploi différencié de deux types de fermeture donne des indications sur l'imaginaire de la ville du futur dans le corpus PARVIS.

Bien que le nombre d'occurrences du mot *cité* ait diminué après la désambiguïsation, son rang a augmenté, en passant du 16^{e} mot le plus fréquent dans le corpus au 12^{e} mot, ce changement s'explique par la réduction significative du nombre d'occurrences d'autres mots polysémiques tels que *base, centre, place, tour* après la désambiguïsation. Ces mots ont été utilisés plus fréquemment dans le corpus que le mot *cité*, cependant, ce dernier est davantage employé dans le contexte de la ville.

Corpus	PARVIS		GDN
Indicateur	**Accuracy**	**F-mesure**	**Accuracy**
allée	1.0	1.0	0.93
base	1.0	1.0	0.93
centre	0.85	0.83	1.0
cité	1.0	1.0	0.73
cœur	1.0	1.0	0.8
cour	0.9	0.93	0.93
enceinte	1.0	1.0	0.7
ferme	1.0	1.0	0.87
lieu	1.0	1.0	0.97
marché	1.0	1.0	0.8
montée	0.9	0.95	0.94
place	0.9	0.93	0.93
porte	0.9	0.93	0.73
tente	1.0	1.0	0.73
tour	1.0	1.0	0.87
Moyenne	**0.96**	**0.97**	**0.86**

Table 2. *Evaluation des classifieurs sur les jeux de données PARVIS et GDN*

Le mot du lexique qui a connu la plus grande diminution en termes d'occurrence après la désambiguïsation est *place*. Cette diminution s'explique par son utilisation multiple et variée en termes de localisation, qui n'est pas toujours spécifique à la ville (comme : *Je fis signe à un interne de prendre ma place à la table*, *Sa place au sein du peuple n'est pas valorisée*), ainsi que par son emploi en tant que lemme du verbe *placer*.

Mot	# occurrences après désambiguïsation	rang avant désambiguïsation	rang après désambiguïsation
allée	226	29	62
base	364	14	48
centre	905	11	22
cité	1381	16	12
cœur	358	24	49
cour	698	23	27
enceinte	233	27	61
ferme	427	20	38
lieu	1596	4	11
marché	93	21	95
montée	317	30	50
place	1089	3	15
porte	6304	1	1
tente	591	25	31
tour	1077	6	16

Table 3. *Nombres d'occurrences des 15 mots polysémiques du lexique dans le corpus avant (tous sens confondus) et après (occurrences du sens associé à la ville) leur désambiguïsation*

6. Conclusions et perspectives

Cet article présente une approche efficace pour désambiguïser des mots polysémiques en distinguant leur contexte d'emploi lié à la ville d'un autre contexte. Cela est réalisé en calculant les vecteurs de contexte des phrases contenant des mots polysémiques avec leurs divers sens, puis en les classifiant en affinant le modèle CamemBERT.

Elle a montré de bonnes performances avec un pourcentage d'accuracy de 96% en moyenne avec une évaluation sur des exemples extraits du même corpus de romans de science-fiction que celui d'entraînement. Les performances ont un peu baissé tout en restant très bonnes (accuracy à 86%) pour une évaluation sur un corpus dont les variations lexicales, syntaxiques et stylistiques sont plus larges et moins prévisibles. La méthode est donc robuste et peu sensible aux variations du corpus.

Malgré ses résultats intéressants sur les deux corpus, cette méthode présente des limitations en termes de temps et de ressources, puisque une annotation de nouvelles phrases et un entraînement de nouveau classifieur sont nécessaires.

Une autre limitation réside dans l'impossibilité à distinguer entre les divers sens d'un mot en dehors du contexte urbain. Elle permet uniquement de déterminer si le mot a été utilisé dans le cadre urbain ou non, sans fournir d'informations sur les autres contextes d'utilisation.

Pour surmonter ces limites, des perspectives de recherche futures incluent l'entraînement de classifieurs sur différentes phrases contenant divers mots polysémiques, avec des annotations spécifiques pour chaque contexte d'utilisation. Ces développements visent à étendre la portée de la méthode et à améliorer sa capacité à traiter un éventail plus large de mots et de sens.

Remerciements

Ce travail s'inscrit dans le cadre du projet PARVIS financé par I-SITE FUTURE (http://www.future-isite.fr) ; il n'aurait pas été possible sans l'aide d'autres chercheurs et stagiaire de l'Université Gustave-Eiffel (UGE) : Olivier Bonin (LVMT) tant pour la construction judicieuse du corpus que pour sa mise en œuvre pratique ; Claude Martineau et Tita Kyriacopoulou (LIGM) pour la chaîne de traitement des données ; Alexandra Li--Combeau-Longuet pour son aide dans la tâche d'annotation des jeux de données du GDN. Qu'ils en soient tous et toutes chaleureusement remercié.es.

Bibliographie

Bird S., Klein E., et Loper E. (2009). Natural language processing with Python : analyzing text with the natural language toolkit. " O'Reilly Media, Inc.".

Devlin J., Chang M.-W., Lee K., et Toutanova K. (2019). BERT : Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran & T. Solorio, Éds., Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers), p. 4171–4186, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1423.

Du J., Qi F., et Sun M. (2019). Using BERT for word sense disambiguation. CoRR, abs/1909.08358.

Guembour S., Dong C., et Dominguès C. (2023). Characterization of the city of the future from a science fiction corpus. In E. Métais, F. Meziane, V. Sugumaran, W. Manning & S. Reiff-Marganiec, Éds., Natural Language Processing and Information Systems - 28th International Conference on Applications of Natural Language to Information Systems, NLDB 2023, Derby,

UK, June 21-23, 2023, Proceedings, volume 13913 de Lecture Notes in Computer Science, p. 313–325 : Springer. DOI : 10.1007/978-3-031-35320-8_22.

Janati B., Ghanimi I., et Ghanimi F. (2023). A disambiguation approach based on the vector representation of words with camembert. p. 020002. DOI : 10.1063/5.0150090.

Kågebäck M. et Salomonsson H. (2016). Word sense disambiguation using a bidirectional LSTM. CoRR, abs/1606.03568.

Labrak Y., Bazoge A., Dufour R., Rouvier M., Morin E., Daille B., et Gourraud P.-A. (2023). Drbert : A robust pre-trained model in french for biomedical and clinical domains.

Le H., Vial L., Frej J., Segonne V., Coavoux M., Lecouteux B., Allauzen A., Crabbé B., Besacier L., et Schwab D. (2020). FlauBERT : Unsupervised language model pre-training for French. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard,

J. Mariani, H. Mazo, A. Moreno, J. Odijk & S. Piperidis, Éds., Proceedings of the Twelfth Language Resources and Evaluation Conference, p. 2479–2490, Marseille, France : European Language Resources Association.

Martin L., Muller B., Suárez P. J. O., Dupont Y., Romary L., Clergerie É. V., Seddah D., et Sagot B. (2019). Camembert : a tasty french language model. arXiv preprint arXiv :1911.03894.

Miller G. A. (1994). WordNet : A lexical database for English. In Human Language Technology : Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994.

Norré M., Cardon R., Vandeghinste V., et François T. (2023). Word sense disambiguation for automatic translation of medical dialogues into pictographs. In R. Mitkov & G. Angelova, Éds., Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing, p. 803–812, Varna, Bulgaria : INCOMA Ltd., Shoumen, Bulgaria.

Pennington J., Socher R., et Manning C. (2014). GloVe : Global vectors for word representation. In

A. Moschitti, B. Pang & W. Daelemans, Éds., Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), p. 1532–1543, Doha, Qatar : Association for Computational Linguistics. DOI : 10.3115/v1/D14-1162.

Raganato A., Delli Bovi C., et Navigli R. (2017). Neural sequence learning models for word sense disambiguation. In M. Palmer, R. Hwa & S. Riedel, Éds., Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, p. 1156–1167, Copenhagen, Denmark : Association for Computational Linguistics. DOI : 10.18653/v1/D17-1120.

Topalov C., Lille L. C., Depaule J.-C., et Marin B. (2010). L'aventure des mots de la ville. Paris, France : Robert Laffont.

Touchent R., Romary L., et Clergerie E. (2023). Camembert-bio : a tasty french language model better for your health.