Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Vers une analyse automatique de la perception relative à un lieu

Hélène Flamein[1], Iris Eshkol-Taravella[2],
(1) Laboratoire Ligérien de Linguistique (CNRS UMR 7270), Université d'Orléans
(2) MoDyco (CNRS UMR 7114), Université Paris Nanterre
helene.flamein@univ-orleans.fr, ieshkolt@parisnanterre.fr

RESUME

Le travail présenté s'intéresse à la perception qu'ont les habitants de leur ville en se fondant sur un corpus de conversations orales spontanées. La chaîne de traitement conditionnant l'analyse de la perception se décompose en trois étapes : la détection des noms de lieux, l'analyse de la perception identifiée et la visualisation cartographique des informations extraites.

ABSTRACT

Towards an Automatic Analysis of Place Perception.
The work presented deals with the perception that the inhabitants have of their city based on a corpus of spontaneous oral conversations. The processing chain that conditions the analysis of perception is divided into three steps: the detection of place names, the analysis of the identified perception and the cartographic visualization of the extracted information.

MOTS-CLES : Lieux, Subjectivité, Polarité, Oral transcrit, ESLO.
KEYWORDS: Place, Subjectivity, Polarity, Oral transcript, ESLO.

1 Introduction

Le travail présenté s'intéresse à la perception qu'ont les habitants de leur ville en se fondant sur un corpus de conversations orales spontanées tiré d'ESLO (Enquête SocioLinguistique à Orléans)[1]. La perception est donc analysée à travers trois étapes : la détection des noms de lieux mentionnés dans le corpus, l'analyse de la perception dans le contexte des lieux identifiés et les visualisations des informations extraites tout au long du traitement.

2 Détection des lieux dans l'oral transcrit

Le processus de détection automatique de noms de lieux dans le corpus oral transcrit est décrit dans Flamein & Eshkol-Taravella (2020). D'une manière générale, le système s'appuie sur l'exploitation de ressources lexicales Geonames, GEOFLA et Data.gouv.fr qui référencient des noms de lieux normalisés avec leurs coordonnées géographiques. En l'état, ces ressources ne suffisent pas pour l'annotation exhaustive des lieux puisqu'elles ne recensent que les noms officiels, conventionnels des lieux. Pour détecter les noms de lieu non normalisés, le système génère des variantes de noms de

[1] http://eslo.huma-num.fr/

Volume 4 : Démonstrations et résumés d'articles internationaux, pages 26–29.
HAL : hal-02768503.

lieux afin d'enrichir les ressources lexicales. Les règles utilisées pour cela sont établies à partir de l'observation des pratiques des locuteurs du corpus en ce qui concerne la dénomination des lieux. Par exemple, dans les noms de voie, les locuteurs ne conservent que le dernier mot (rue Gauguin au lieu de la rue Paul Gauguin) ou groupe prépositionnel (rue de Sonis au lieu de la rue du Général de Sonis) composant le nom officiel ainsi que le mot caractérisant le type de voie. Dans le cas des noms de villes composés de trois mots ou plus, c'est plutôt le premier terme qui est conservé et les derniers termes qui sont supprimés (La Ferté Saint-Aubin pour La Ferté). De nouvelles entrées sont donc générées et ajoutées au lexique sur la base de ces observations. Le lexique ainsi enrichi est ainsi appliqué sur le corpus et associé à des patrons permettant l'identification des noms de lieux tronqués (*rue de la Rép-* pour la *rue de République*), ou même inventés (*rue de la gare*, *boulangerie du coin*, etc.).

Le module développé est évalué sur un corpus de référence composé de 15 transcriptions, référençant 2292 noms de lieux, parmi lesquels 549 (24%) varient par rapport à la norme. Le module obtient un Rappel de 0,90, une Précision de 0,93 et une F-Mesure de 0,91. Cette évaluation porte sur les frontières du nom du lieu et montre des performances satisfaisantes pour identifier des lieux sous leurs formes officielles et leurs variantes. L'annotation des lieux sert d'ancrage pour la deuxième étape d'analyse de la perception effectuée par apprentissage automatique.

3 Traitement de la subjectivité et de la polarité

La perception est une notion vaste que la linguistique, la psychologie et même les sciences de l'information ont tenté de circonscrire. Du point de vue du TAL, la perception est abordée sous l'angle de l'analyse de sentiment et de la fouille d'opinion (Pak & Paroubek, 2010 ; Marchand, 2015 ; Karaoui et al., 2019). Ces travaux se fondent principalement sur les données issues du Web. Les données orales sont moins facilement disponibles mais présentent tout autant d'enjeux pour la problématique de la fouille d'opinion et de l'analyse de sentiments.

L'analyse de la perception relative à Orléans s'appuie sur la première phase de détection des lieux dans les transcriptions du corpus ESLO et est réalisée grâce à des techniques d'apprentissage automatique supervisé. La classification des énoncés est réalisée par apprentissage automatique sur un corpus de référence annoté manuellement[2] et divisé en trois parties : corpus d'entraînement, de test et d'évaluation. Afin d'entraîner un modèle pour la détection de la subjectivité et de la polarité, différents *features* sont envisagés : la lemmatisation et l'étiquetage morphosyntaxique ainsi que le calcul d'un score de polarité et d'un score d'émotion[3]. Le TF-IDF et Word2Vec sont des méthodes possibles pour créer une représentation vectorielle des segments à analyser. Enfin deux classifieurs, RandomForest et SVM, ont été retenus pour réaliser la classification attendue.

Les expériences mêlant les différents *features*, les méthodes de représentations vectorielles et les classifieurs ont été menées pour sélectionner la combinaison la plus performante. L'utilisation du classifieur SVM, la représentation vectorielle des segments lemmatisés avec le TF-IDF et l'utilisation de tous les features disponibles constituent le modèle le plus efficace. Les expériences révèlent que le même modèle est le plus efficace pour distinguer les énoncés subjectifs des énoncés objectifs mais aussi pour détecter la polarité des énoncés. Pour confirmer les performances de ce modèle, celui-ci est évalué sur le corpus de test. Les résultats de cette évaluation sont présentés dans

[2] Constitué de trente transcriptions du corpus ESLO, annotés en lieux (4519 mentions)

[3] Ces scores sont définis à partir de l'exploitation du lexique FEEL (Abdaoui et al. 2017)

le tableau 1. Pour la tâche de détection de la subjectivité, le modèle obtient une macro-average de 0,77 et une macro-average de 0,76 pour le traitement de la polarité.

Classifieur	Features	Cible	Macro average	Précision	Rappel	F-mesure
TF-IDF + SVM	Score polarité + Score émotions + POS	subj.	0,77	0,69	0,56	0,63
		obj.		0,77	0,89	0,83
		pos.	0,76	0,78	0,91	0,82
		nég.		0,67	0,46	0,54

TABLEAU 1 : Modèle retenu pour la détection de la subjectivité et de la polarité

4 Visualisation de la perception

L'objectif de la dernière étape de notre travail est de représenter visuellement les résultats des analyses réalisées jusqu'ici. Il s'agit de mettre en valeur les informations extraites de notre corpus afin de faciliter leur manipulation et les rendre accessibles. Dans une autre mesure, la visualisation contribue à faire émerger les relations qui unissent les différentes informations détectées. La visualisation de la perception par les Orléanais de leur ville se concrétise donc avec la mise en place d'un Système d'Information Géographique (SIG)[4] et la création de cartes. Pour cela, nous utilisons l'outil ArcGIS Online[5], développé par Esri et disponible en ligne. Plusieurs couches d'informations sont projetées dans le système et permettent de figurer l'ensemble des lieux détectés en fonction des métadonnées qui leurs sont associées au cours du traitement et des déclarations qui leurs sont relatives. La figure 1 présente les lieux situés dans le centre-ville d'Orléans en fonction de la polarité. Un ratio correspondant à la part de déclarations positives et négatives est calculé et corrélé avec le nombre total de déclarations subjectives exprimées à propos du lieu. Plus le nombre de déclarations est important, plus le ratio de la polarité est significatif.

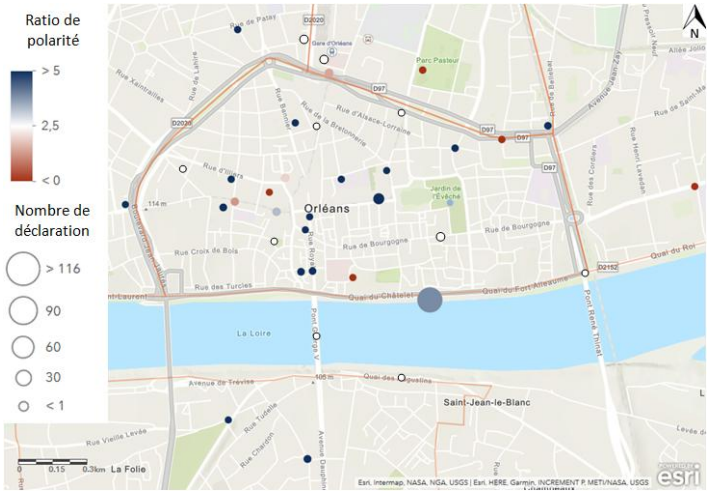

FIGURE 1 : Répartition de la polarité à propos des lieux du centre-ville d'Orléans

4 SIG pour la visualisation de la perception de la ville d'Orléans – http://arcg.is/1v9DaC

5 https://www.arcgis.com/index.html

5 Conclusion et perspectives

Nous avons présenté une méthodologie pour la détection de la perception qu'ont les habitants de leur ville. Afin de procéder à l'analyse de cette perception, nous avons d'abord développé un système de détection des noms de lieux mentionnés tout en tenant compte des spécificités du corpus. A partir des détections réalisées, un modèle pour la détection de la subjectivité et la détermination de la polarité a été entrainé. Finalement, l'ensemble des informations identifiées tout au long du traitement ont été projetées dans un SIG afin de donner corps à l'image que se font les orléanais de leur ville.

Références

ABDAOUI A., AZE J., BRINGAY S. & PONCELET P. (2017). FEEL : A French Expanded Emotion Lexicon. Language Resources and Evaluation, 51(3), pp. 833–855. DOI : https://doi.org/10.1007/s10579-016-9364-5

FLAMEIN H. & ESHKOL-TARAVELLA I. (2020). Noms de lieux dans le corpus de français parlé : Une approche symbolique pour un traitement automatisé. *Le français moderne 2020,* n.1

KARAOUI J. & BENAMARA F. & Véronique Moriceau. (2019) *Détection automatique de l'ironie: Application à la fouille d'opinion dans les microblogs et les médias sociaux*. ISTE Group.

MARCHAND M. (2015). *Domaines et fouille d'opinion: une étude des marqueurs multipolaires au niveau du texte*. Thèse de doctorat, Université Paris Sud - Paris XI. ⟨tel-01157951⟩

NOVAKOVA I. (2019). *Le lexique des émotions*. UGA Éditions.

PAK A. & PAROUBEK P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. *LREc*, 10, pp. 1320–1326. https://www.aclweb.org/anthology/L10-1263/

ZHANG L. (2012). *Analyse automatique d'opinion : Problématique de l'intensité et de la négation pour l'application à un corpus journalistique*. Thèse de doctorat, Université de Caen. ⟨tel-00777603⟩