SENSE-LM : Extraction de références olfactives et auditives dans des textes, basée sur des représentations sensorimotrices et un modèle linguistique.

Cédric Boscher[1], Christine Largeron[2], Veronique Eglin[1], Elöd Egyed-Zsigmond[1]

[1]INSA Lyon, LIRIS, UMR5205, 69621 Villeurbanne, France
[2]Université Jean-Monnet Saint-Etienne, CNRS,
Laboratoire Hubert Curien, UMR5516, 42000 Saint-Etienne, France,
Contact: cedric.boscher@insa-lyon.fr

Note: La contribution suivante a été publiée sous la forme d'un article long dans les Findings de la conférence EACL 2024 (Boscher et al., 2024).

1 Introduction

Bien que l'intelligence artificielle ait fait des avancées considérables dans de nombreux domaines, son application dans la recherche d'informations en sciences humaines et sociales reste peu exploitée. Elle présente pourtant un potentiel considérable pour des applications spécialisées telles que l'étude approfondie de la linguistique sensorielle. Le concept psychophysiologique de sensorialité définit la perception humaine à travers les cinq fonctions sensorielles aristotéliciennes : *visuelle (VIS), gustative (GUS), olfactive (OLF), auditive (AUD) et haptique (HAP).* La linguistique sensorielle examine la relation entre le langage et les expériences sensorielles, avec des applications en sciences cognitives ou en étude du patrimoine. Par exemple, Murphy (2019) a révélé une corrélation entre le type de formulation d'expériences olfactives et le diagnostic de la maladie d'Alzheimer. Pardoen (2019) explore les indices auditifs pour recréer l'atmosphère sonore du Paris du 19ème siècle. Menini et al. (2022) étudie le patrimoine sensoriel des odeurs pour des applications en histoire de l'art et culture.

2 Objectifs et questions de recherche

Nous présentons *SENSE-LM*, un système qui combine modèles de langage, représentations sensorimotrices et techniques de génération lexicale pour détecter les informations évoquant implicitement ou explicitement la sensorialité dans de grands corpus de texte. Nous cherchons ici à mettre en évidence l'intérêt de combiner des modèles de langage, prenant en compte le contexte linguistique des mots, et des représentations basées sur le jugement humain - dites "sensorimotrices" - permettant d'identifier la présence de sensorialités. Contrairement aux approches existantes, notre système prend en compte les cinq sens et a été évalué pour l'ouïe et l'odorat. Faute de jeux de données existants, nous avons créé un ensemble de données artificielles axé sur l'audition, généré avec GPT-4 et annoté manuellement. Le code source et les jeux de données sont disponibles pour la communauté.[1].

3 Méthodologie

Notre système, *SENSE-LM*, extrait des références sensorielles dans de grands corpus, d'abord au niveau des phrases, puis des tokens. La Figure 1 illustre le flux de travail global. L'étape 1 effectue une classification à gros grain pour identifier, dans un ensemble de documents D, les phrases évoquant l'une des cinq fonctions sensorielles. Ensuite, l'étape 2 réalise une classification à grain fin pour extraire les termes reflétant la fonction sensorielle cible dans la phrase.

[1]https://github.com/cfboscher/sense-lm_eacl2024

3.1 Étape 1 - Classification de Phrases Sensorielles

Nous considérons les fonctions sensorielles $\mathbb{M} = \{\text{OLF, GUS, AUD, VIS, HAP, INT}\}$. Un corpus D est composé de phrases, où chaque phrase s a une classe $C(s)$ positive (1) si elle fait référence à une fonction sensorielle m de $\mathbb{M}$, sinon négative (0). Par exemple, pour m = AUD, *"Clocks ticked, marking relentless seconds before thunder growls"* est positive, tandis que *"The cake was delicious, moist, and adorned with colorful frosting"* est négative. L'étape initiale de *SENSE-LM* est de classifier correctement les phrases selon m, en apprenant une fonction ϵ qui associe chaque phrase s à une classe : $\epsilon : D \rightarrow 1, 0$ telle que $\epsilon(s) = C(s)$ pour tout $s \in D$.

Nous définissons ici le concept de représentation sensorimotrice, basé sur les normes sensorimotrices de Lancaster (Lynott et al., 2020). Cette ressource modélise 40 000 lemmes anglais, évalués par des annotateurs humains selon leur appariement sémantique avec 11 fonctions, soit 6 fonctions sensorielles humaines (les cinq sens aristotéliciens et l'interoception) et 5 fonctions motrices (utilisation des parties du corps). Chaque lemme est représenté par un vecteur à 11 dimensions avec des valeurs entre 0 et 5. La représentation sensorimotrice de la phrase s est égale à la somme des représentations de chaque mot. Nous détaillons la méthode de calcul dans l'Annexe A.

La première étape de *SENSE-LM* combine le plongement sémantique contextuel de la phrase, produit par BERT, ainsi qu'une représentation sensorimotrice de cette même phrase. *SENSE-LM* prend en entrée une phrase s. La première branche utilise BERT pour extraire un plongement de dimension 768, noté s_B. La seconde branche convertit s en une représentation sensorimotrice s_{SN} de dimension 11. Ces deux représentations, s_B et s_{SN}, sont ensuite fusionnées par concaténation et passées à une couche Fully Connected (dimension 779) pour prédire si la phrase s est sensorielle (valeur 1) ou non (valeur 0) pour la fonction sensorielle m. Nous détaillons l'architecture du modèle précédent dans l'Annexe B.

3.2 Étape 2 - Extraction de Termes Sensoriels

La deuxième étape de *SENSE-LM* vise à extraire les tokenss associés à un sens spécifique $m \in \mathbb{M}$ dans une phrase s, parmi les phrases classées positivement à l'étape 1. Pour chaque sens m et chaque phrase s de D_{pos} (phrases classées positivement), s est divisée en une séquence de tokens $t(s)$. Par exemple, pour $m = AUD$ et $s =$ *"Clocks ticked, marking relentless seconds before thunder growls."*, on note $t(s) =$ (**Clocks, ticked**, marking, relentless, seconds, before, **thunder, growls**, ..., <PAD>), où les tokens en gras sont positifs pour m. L'objectif est d'apprendre la fonction γ qui, dans cet exemple, produit :

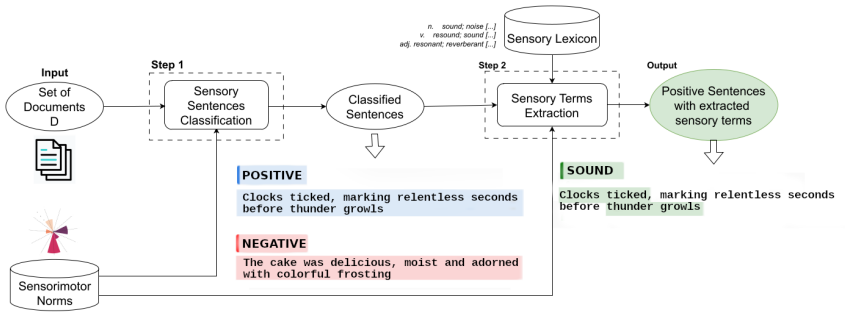

Figure 1: Flux de travail global de *SENSE-LM*, suivant l'exemple de la tâche de recherche d'information Auditive.

$$\gamma(t(s), m) = (\mathbf{1}, \mathbf{1}, 0, 0, 0, 0, \mathbf{1}, \mathbf{1}, \ldots, < PAD >)$$

Pour réaliser cette tâche, nous proposons une approche combinatoire basée sur trois étapes successives (détaillées en Annexe C :

1. Une première phase de classification basée sur RoBERTa
2. Une seconde phase d'extension de mots positifs, basée sur des ressources lexicales
3. Une troisième phase de repêchage de faux négatifs basée sur une fonction heuristique combinant les plongements sémantiques et représentations sensorimotrices.

4 Travaux Connexes

Approches Basées sur des Ressources Lexicales

Les approches lexicales visent à créer automatiquement une liste de termes ou une taxonomie liée à un domaine sensoriel à partir d'un échantillon de termes de référence. Lexifield (Mpouli et al., 2020), un système d'expansion sémantique, a été appliqué pour identifier des termes liés aux fonctions sensorielles, auditives ou olfactives dans les œuvres littéraires, surpassant les méthodes comme Empath (Fast et al., 2016). Il enrichit un ensemble de termes de départ via la similarité de plongements sémantiques et des ressources externes multilingues. Cette méthode a été utilisée pour détecter des descriptions sonores (Mpouli et al., 2019), malgré certains cas limites dus à la polysémie, et des résultats perfectibles dus à la formulation d'hypothèses naïves sur les plongements sémantiques.

Approches Basées sur des Modèles de Langage

Certains travaux préliminaires ont ouvert les premières contributions de l'exploration de l'information sensorielle basée sur des modèles de langage. Nous pouvons formuler la tâche de classification binaire de phrase de la façon suivante : *"La phrase s contient-elle une référence à l'olfaction ?"*, en s'appuyant sur MacBERTh (Manjavacas and Fonteyn, 2021), une variante de BERT pré-entraînée sur des textes historiques (1450–1950). Comme l'efficacité de ces solutions dépend fortement de la qualité des annotations de vérité terrain et ont un comportement difficilement explicable, elles sont difficiles à exploiter par les non-spécialistes.

5 Expériences et Analyses

Nos expériences sont réalisées sur deux jeux de données :

Odeuropa : English Benchmark. (Menini et al., 2022) Ce jeu de données se focalise sur les expériences olfactives décrites par des sources documentaires allant du 17ème au 20ème siècle. Il contient 2176 phrases et 5530 occurrences de termes liés à l'odeur.

	Odeuropa			Jeu de données Artificiel Auditif		
Méthode	**Précision**	**Rappel**	**F1-Score**	**Precision**	**Recall**	**F1-Score**
BERT	91.51	90.12	90.80	96.03	96.14	96.08
LR(s_{SN})	82.25	72.33	76.97	87.64	87.04	87.23
GPT-4	91.59	89.42	90.4	N/A*	N/A*	N/A*
SENSE-LM	**94.09**	**92.26**	**93.16**	**97.01**	**97.22**	**97.12**

Table 1: Évaluation comparative de la tâche de classification binaire de phrases.

	Odeuropa			Jeu de données Artificiel Auditif		
Méthode	**Précision**	**Rappel**	**F1-Score**	**Précision**	**Rappel**	**F1-Score**
Lexifield ($\mathbb{L}_m$)	77.3	43.53	55.69	43.25	16.32	23.69
GPT-4	52.90	70.99	60.62	N/A*	N/A*	N/A*
SENSE-LM (2.1)	80.01	66.32	72.52	91.51	89.25	90.36
SENSE-LM (2.1 ∪ 2.2)	81.5	72.7	76.84	**91.75**	92.49	92.11
SENSE-LM (2.1 ∪ 2.3)	80.48	70.21	74.99	91.19	92.32	91.75
SENSE-LM	**82.01**	**73.62**	**77.58**	91.65	**93.01**	**92.32**

Table 2: Évaluation comparative de la tâche d'extraction des termes sensoriels.

Jeu de données Artificiel Auditif. En raison du manque de jeux de données disponibles pour d'autres fonctions sensorielles, nous avons créé un ensemble de données artificiel avec 1000 phrases synthétiques générées par GPT-4, de structure et longueur variable, contenant des références à des sons. Le protocole de génération est détaillé dans notre dépôt[2].

Évaluation de l'Étape 1 — Classification de Phrases Sensorielles

Nous comparons *SENSE-LM* avec BERT, une régression logistique entrainée sur la représentation sensorimotrice (11 caractéristiques), notée LR(s_{SN}), et GPT-4. Les protocoles sont détaillés dans notre dépôt. Les résultats du Table 1 montrent que *SENSE-LM* surpasse les modèles de référence en précision, rappel et score F1 sur les deux jeux de données en intégrant une représentation conceptuelle basée sur le jugement humain, contrairement à BERT et GPT-4.

Évaluation de l'Étape 2 - Extraction de Termes Sensoriels

La Table 2 présente les résultats fournis par les approches de référence (en haut) et par *SENSE-LM*, avec une évaluation ablative de chaque composante (en bas). *SENSE-LM* affiche les meilleures performances globales. Notre raisonnement sur les limites de performance de GPT-4 est valable dans ce nouveau cas, et les hypothèses même renforcées par l'échantillon de données encore plus petit utilisé pour la tâche d'entraînement (600 phrases). Dans un second temps, l'évaluation ablative de *SENSE-LM* met en évidence l'intérêt de combiner successivement ses 3 étapes, donnant les meilleurs résultats.

[2] https://github.com/cfboscher/sense-lm/tree/main/gpt4_prompts

References

Cédric Boscher, Christine Largeron, Véronique Eglin, and Elöd Egyed-Zsigmond. 2024. SENSE-LM : A synergy between a language model and sensorimotor representations for auditory and olfactory information extraction. In Findings of the Association for Computational Linguistics: EACL 2024, pages 1695–1711, St. Julian's, Malta. Association for Computational Linguistics.

Dash. 2021. Extract the right Phrase From Sentence. Medium, Analytics Vidhya.

Ethan Fast, Binbin Chen, and Michael Bernstein. 2016. Empath: Understanding Topic Signals in Large-Scale Text. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, pages 4647–4657. ArXiv:1602.06979 [cs].

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

Dermot Lynott, Louise Connell, Marc Brysbaert, James Brand, and James Carney. 2020. The Lancaster Sensorimotor Norms: multidimensional measures of perceptual and action strength for 40,000 English words. Behavior Research Methods, 52(3):1271–1291.

Enrique Manjavacas and Lauren Fonteyn. 2021. Macberth: Development and evaluation of a historically pre-trained language model for english (1450-1950). pages 23–36.

Stefano Menini, Teresa Paccosi, Serra Sinem Tekiroglu, and Sara Tonelli. 2022. Building a multilingual taxonomy of olfactory terms with timestamps. Proceedings of the Thirteenth Language Resources and Evaluation Conference, pages 4030–4039.

George A. Miller. 1995. Wordnet: A lexical database for english. Commun. ACM, 38(11):39–41.

Suzanne Mpouli, Michel Beigbeder, and Christine Largeron. 2020. Lexifield: a system for the automatic building of lexicons by semantic expansion of short word lists. Knowledge and Information Systems, 62(8):3181–3201.

Suzanne Mpouli, Christine Largeron, and Michel Beigbeder. 2019. Identifying sound descriptions in written documents. In 2019 13th International Conference on Research Challenges in Information Science (RCIS), pages 01–06. IEEE.

Claire Murphy. 2019. Olfactory and other sensory impairments in alzheimer disease. Nature Reviews Neurology, 15(1):11–24.

Mylène Pardoen. 2019. Projet Bretez: une pincée de son dans l'Histoire. Digital Studies/Le champ numérique, 9(1):11.

A Algorithme de calcul de la représentation sensorimotrice

Algorithm 1 Méthode de calcul de la représentation sensorimotrice

Entrée: Phrase s, Normes Sensorimotrices L_{SN}
Sortie: Représentation sensorimotrice s_{SN}

$s_{SN} \leftarrow (0, 0 \ldots 0)$
$s \leftarrow RemoveStopWords(s)$
for $w \in s$ **do**
  if $lemma(w) \in L_{SN}$ **then**
    $v \leftarrow lemma(w)_{SN}$
  else
    $v \leftarrow (0, 0 \ldots 0)$
    for $i \in$ Synsets(w) **do**
      if $lemma(i) \in L_{SN}$ **then**
        $v \leftarrow lemma(i)_{SN}$
        $break$
      end if
    end for
    $s_{SN} \leftarrow s_{SN} + v$
  end if
end for
  return s_{SN}

L'Algorithme 1 détaille la méthode de calcul proposée pour obtenir la représentation sensorielle d'une phrase s, illustrée par la Figure 2.

Nous désignons par L_{SN} un dictionnaire correspondant aux mots disponibles dans les normes sensorimotrices de Lancaster : il cartographie chaque mot w en s avec sa représentation sensorimotrice w_{SN} comme un vecteur à 11 dimensions $w_{SN} = (w_{SN}(j), j = 1, \ldots, 11)$, où $w_{SN}(m)$ correspond à la composante de w_{SN} associée à la fonction sensorielle $m \in \mathbb{M}$.

La représentation sensorimotrice w_{SN} de w est égale à $lemma(w)_{SN}$ si le lemme associé à w existe dans L_SN. Dans le cas où ce lemme n'est pas inclus dans L_{SN}, nous considérons que le premier élément appartient à l'ensemble $Synsets(w)$ des synsets WordNet de w tels que définis par Miller (1995), c'est-à-dire les mots synonymes. Enfin, s'il n'y a pas non plus de synset de w inclus dans L_{SN}, la représentation sensorimotrice de w équivaut à un vecteur à 11 dimensions avec des valeurs nulles.

Comme détaillé dans l'algorithme, après avoir déterminé cette représentation sensorimotrice pour chaque mot $w \in s$, la représentation sensorimotrice de phrase $s_{SN} = (s_{SN}(j), j = 1, \ldots, 11)$ de s est obtenue en additionnant ces vecteurs de mots.

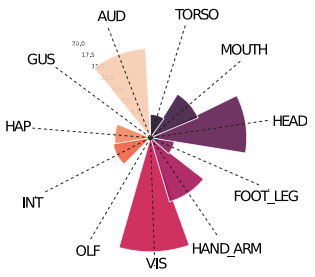

Figure 2: Représentation sensorimotrice de la phrase *"Clocks ticked, marking relentless seconds before thunder growls"*, représentant les 6 attributs sensoriels et les 5 attributs moteurs.

B Architecture du modèle de Classification Binaire de Phrase Sensorielles (Étape 1)

Comme le montre la Figure 3, *SENSE-LM* prend une phrase s en entrée. Sa première branche met en œuvre les étapes successives de BERT : les couches Embedding, Transformers et Pooler, qui permet d'extraire un plongement de s de dimension 768, noté s_B. La seconde branche du modèle transforme la phrase s en sa représentation sensorimotrice s_{SN}, produisant un vecteur de taille 11. Enfin, le modèle concatène s_B et s_{SN} en une représentation commune, fournie à une couche Fully Connected en sortie (dimension = 779) qui renvoie soit la valeur 1 si la phrase s est considéré comme sensorielle par rapport à la fonction sensorielle m, soit 0 sinon.

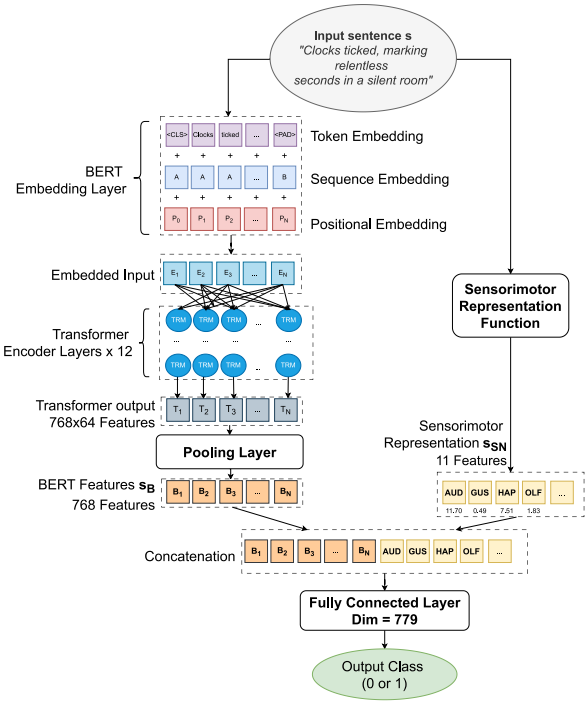

Figure 3: Architecture du modèle utilisé par *SENSE-LM* pour l'étape 1 (classification de phrases sensorielles)

C Description de la Méthode de Classification de Mots Sensoriels

Pour réaliser cette tâche, nous proposons une approche combinatoire basée sur trois étapes successives :

Étape 2.1. Classification avec RoBERTa.

Tout d'abord, nous proposons d'affiner un modèle de langage sur la tâche d'extraction de sous-phrases dans des phrases qui expriment la présence d'une sensorialité donnée, en suivant l'intuition de Dash (2021) qui s'est précédemment penché sur la tâche d'identifier les termes qui reflètent le mieux le sentiment principal (Positif, Neutre ou Négatif) exprimé par tweets[3]. Par analogie, nous utilisons un principe similaire pour détecter les mots qui reflètent le mieux la présence de la sensorialité cible m.

Nous utilisons une architecture BERT, avec les paramètres pré-entraînés RoBERTa set (Liu et al., 2019), qui montre

[3] https://www.kaggle.com/competitions/tweet-sentiment-extraction/leaderboard

empiriquement des performances améliorées sur la tâche de classification des tokens sensoriels dans un contexte de phrase.

Notre entrée est la phrase tokenisée t(s), et la sortie prédite est un vecteur noté $V(t(s), m)$, avec des1 pour les termes prédits positivement, et des 0 pour les négatifs. Ainsi, cette première étape permet d'extraire un premier ensemble de mots, classés comme positifs dans le contexte par RoBERTa. $P_{pos}(s, m)$ désigne l'ensemble des mots dans $t(s)$ qui associent les mots classés positivement dans $V(t(s), m)$, et $P_{neg}(s, m)$ les mots négatifs.

Étape 2.2. Extension avec des Ressources Lexiciales

Deuxièmement, nous utilisons une ressource lexicale, telle que Lexifield (Mpouli et al., 2020) dans le but d'élargir la liste des jetons sensoriels extraits au préalable à l'étape 2.1. Ce lexique noté $\mathbb{L}_m$ contient un ensemble de mots appartenant au champ lexical de la fonction sensorielle cible m. Par exemple, nous pouvons considérer $\mathbb{L}_{\text{OLF}} = \{$odeur (nom), odeur (verbe),...$\}$ si $m =$ OLF.

Pour chaque mot $w \in P_{neg}(s, m)$, on change la valeur correspondante en $V(t(s), m)$ à 1 si $w \in \mathbb{L}_m$.

Étape 2.3. Heuristique basée sur le langage et le jugement humain.

Enfin, dans le but de récupérer les mots faux négatifs omis par la première étape de classification, et en même temps, d'éviter d'introduire de manière significative des exemples faux positifs, nous établissons une heuristique qui considère à la fois la représentation sensorimotrice des termes candidats et leur proximité sémantique avec les exemples positifs. Nous désignons par $\mathbb{E}$ un ensemble d'espaces d'encastrement sémantique, et $\text{CosSim}_e(a, b)$ la mesure de similarité cosinus entre les mots a et b dans un espace d'encastrement $e \in \mathbb{E}$. Pour chaque mot $w \in P_{neg}(s, m)$, on change la valeur correspondante en $V(t(s), m)$ à 1 s'il combine les deux conditions suivantes :

1. $w_{SN}(m) > T$
2. $\exists e \in \mathbb{E}$, et $\exists x \in P_{pos}(s, m)$, t.q. $\text{CosSim}_e(w, x) > U$

où $w_{SN}(m)$ désigne, dans la représentation sensorimotrice du mot w, la dimension correspondant à la fonction sensorielle m.

La condition 1 permet d'abord de s'assurer que le terme candidat est, par essence, cohérent avec la fonction sensorielle cible m. T définit la valeur seuil minimale de $w_{SN}(m)$, avec $T \in [0, 5]$. Ensuite, la condition 2 garantit que la classification de w comme positif a un sens dans le contexte, car elle est sémantiquement proche d'au moins un des termes positifs. $U \in [0, 1]$ définit la valeur minimale de similarité cosinus entre un terme candidat et au moins un des termes positifs. T et U sont réglés manuellement sur la base d'analyses empiriques, bien qu'ils puissent être déterminés par une recherche par grille. À la fin de cette étape, le système renvoie la sortie $\gamma(t(s), m) = V(t(s), m)$.