L’IA à travers des travaux de recherche en traitement automatique des langues (TAL)

Mathieu Constant
Université de Lorraine, CNRS, ATILF (UMR 7118)

Objectifs de l’exposé

- Montrer le fonctionnement de l’IA **de manière simplifiée** à partir de trois travaux de recherche

- Participer à la démystification de l’IA

Traitement automatique des langues (TAL)

- **Manipulation par la machine de productions en langue naturelle** (ex. texte)

- **Deux paradigmes**
  - **Analyse** : production en langue naturelle → représentation abstraite
  - **Génération** : représentation abstraite → production en langue naturelle

- **Exemple de la traduction automatique**
  - phrase en anglais → représentation abstraite (cachée) → phrase en français

Quelques applications classiques du TAL

- Traduction automatique
- Questions-réponses
- Résumé automatique de textes
- Recherche d’information, fouille de textes
- Analyse de sentiments
- Correction orthographique/grammaticale

Le TAL actuellement

- Les récentes avancées de l'IA (générative) ont profondément bousculé le domaine du TAL

- Les frontières traditionnelles du domaine sont aussi bouleversées avec la multimodalité notamment : ex. génération d'une video à partir d'une description textuelle

- Une compétition féroce avec les grands groupes Tech notamment
- Difficile pour un chercheur académique de se faire une place

Trois travaux de recherche de l’équipe TAL de l’ATILF

1) Prédire la complexité d’un mot dans un contexte donné dans le cadre de l’apprentissage d’une langue étrangère

2) Identifier des expressions idiomatiques dans des textes

3) Reformuler/expliquer des termes médicaux pour des non-initiés

Prédiction de la complexité des mots
(Thèse d'Abdelhak Kelious)

Le problème à résoudre

- **Entrée** : un mot dans un contexte
- **Sortie** : un nombre entre 0 et 1 indiquant la complexité du mot dans le contexte donné

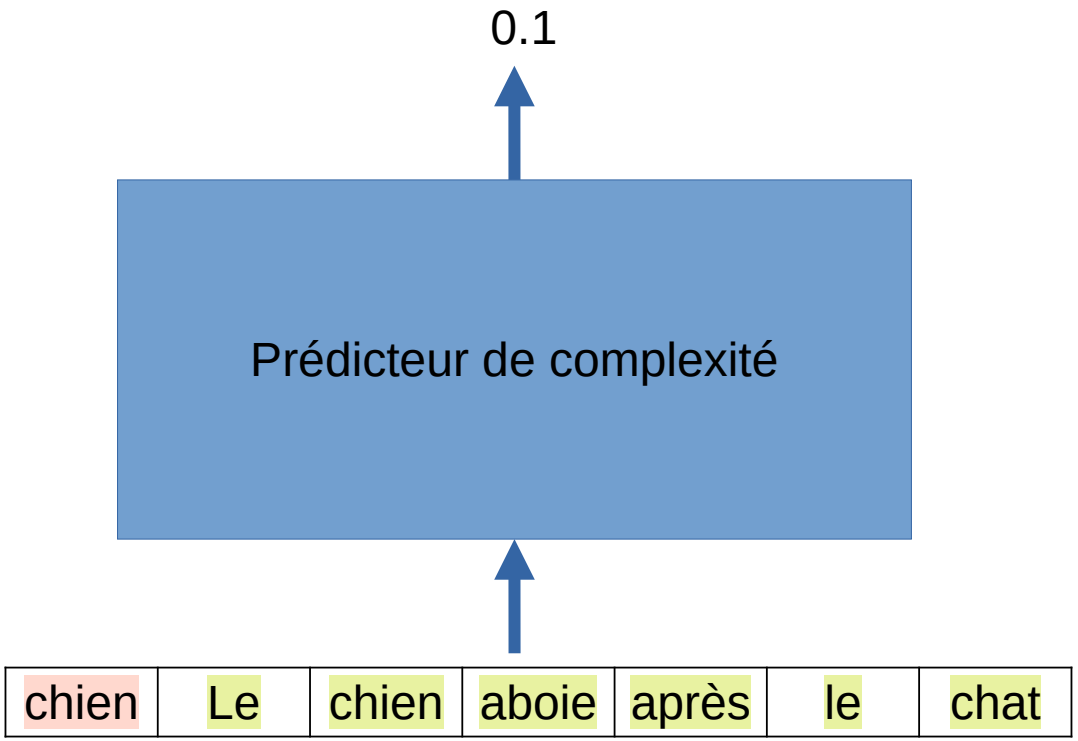
Exemples

Le **chien** aboie après le chat → 0.1 (facile)

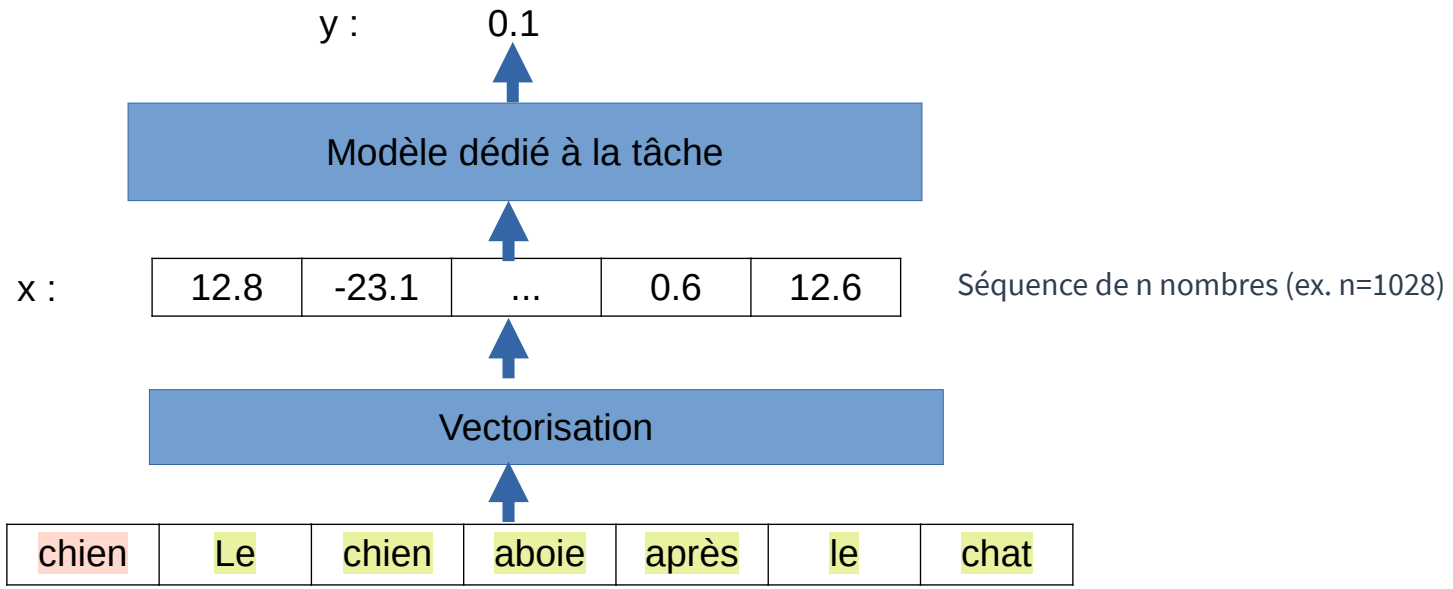
Le **chien** de mon pistolet est cassé → 0.9 (très difficile)

Motivations : identifier les mots difficiles dans un texte pour des apprenants d'une langue et pouvoir les aider pour la lecture

Prédiction de la complexité des mots (2)



Prédiction de la complexité des mots (3)



Prédiction de la complexité des mots (4)

Les modèles récents s'appuient sur des réseaux de neurones

- Un réseau de neurones est l'équivalent d'une fonction mathématique f qui permet de passer d'une entrée vectorisée x à une sortie y [y = f(x)]

- Exemple totalement fictif

  f(x) = f(12.8,-23.1,…,0.6,12.6) = (12.8-23.1+...+0.6+12.6) /1028 = 0.1 = y

La fonction f calcule la moyenne des nombres de la séquence x

Prédiction de la complexité des mots (5)

Apprentissage d'un réseau de neurones

- Un réseau de neurones est appris sur un ensemble de données pour lesquels on connaît déjà le résultat (généralement issu d'une annotation humaine)

  Le **chat** miaule → 0.1
  La **sirène** des pompiers hurle dans la nuit → 0.5
  ...
  J'ai un **chat** dans la gorge → 0.8
  [des milliers de données annotées]

- Le réseau de neurones est conçu par un développeur avec des millions de paramètres ; ces paramètres sont optimisés de telle manière que l'erreur de prédiction soit minimale par rapport à ce qui est attendu (annotations humaines)

Prédiction de la complexité des mots (6)

Vectorisation de l'entrée

- L'entrée est une séquence de mots (des symboles sans valeurs numériques)
- Chaque mot est transformé en une séquence de nombres
- Peut s'appuyer sur des grands modèles de langue appris sur de gigantesques volumes de textes
- Un modèle de langue repose souvent sur un réseau de neurones (ex. type Transformer) appris à l'aide de la tâche de prédiction d'un mot masqué

(je ne rentre pas dans les détails techniques)

Prédiction de la complexité des mots (7)

Résultats expérimentaux (Kelious et al. 2024a)

- Données en anglais annotées LCP 2021 (Shardlow et al., 2021)
- Découpage des données :
  - 7662 instances (mot + contexte → complexité) pour l'apprentissage
  - 917 instances pour l'évaluation

- Des performances honorables : une correlation [Pearson] de 0.8 entre les complexités attendus et prédites [0 : aucune correlation ; 1 : correlation parfaite]
- Des expériences pour l'espagnol et l'allemand montrent des performances similaires (Kelious et al. 2024b)
- Pour info, aussi des expériences avec des modèles génératifs

Identification d’expressions multi-mots (1)
(Projet ANR PARSEME-FR)

Les expressions multi-mots

- Des expressions de plusieurs mots dont le sens ne peut être prédit à partir du sens de ses mots
- Exemples : *faire face*, *mettre les voiles*, *cordon bleu, pomme de terre*

Problème à résoudre

- **Entrée :** une phrase (i.e. une séquence de mots)
- **Sortie :** les occurrences des expressions

Identification d’expressions multi-mots (2)

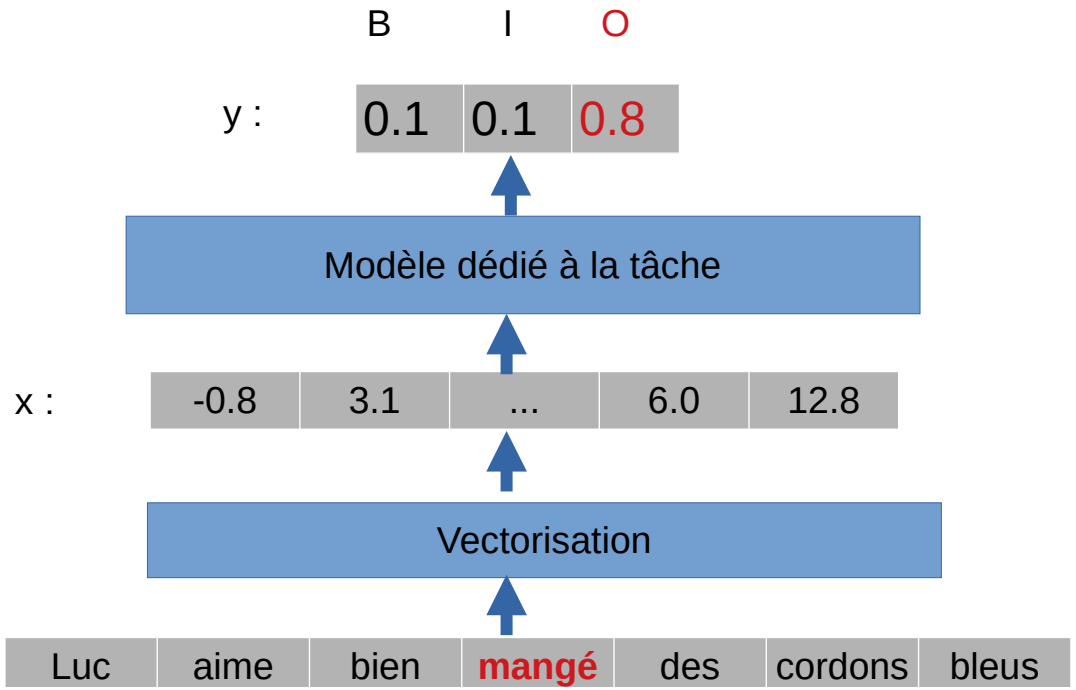
Formalisation du problème en pratique : étiquetage de chaque mot de la phrase avec une catégorie de l’ensemble {B, I, O}

- B : le mot est au début d’une expression multi-mots
- I : le mot appartient à une expressions multi-mots dans une position non-initiale
- O : le mot n’appartient pas à une expression multi-mots

Luc	aime	les	cordons	bleus
O	O	O	B	I

Identification d’expressions multi-mots (3)

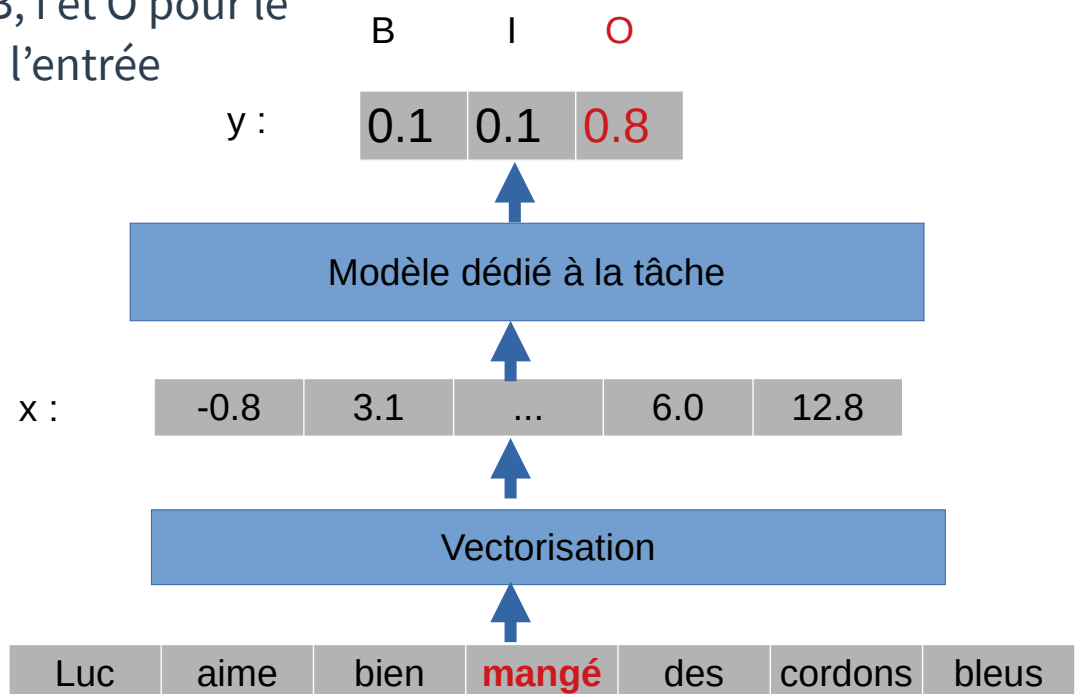
Tâche de classification multi-classes



Identification d’expressions multi-mots (3)

y : probabilités des étiquettes B, I et O pour le mot cible en fonction de l’entrée

Tâche de classification multi-classes

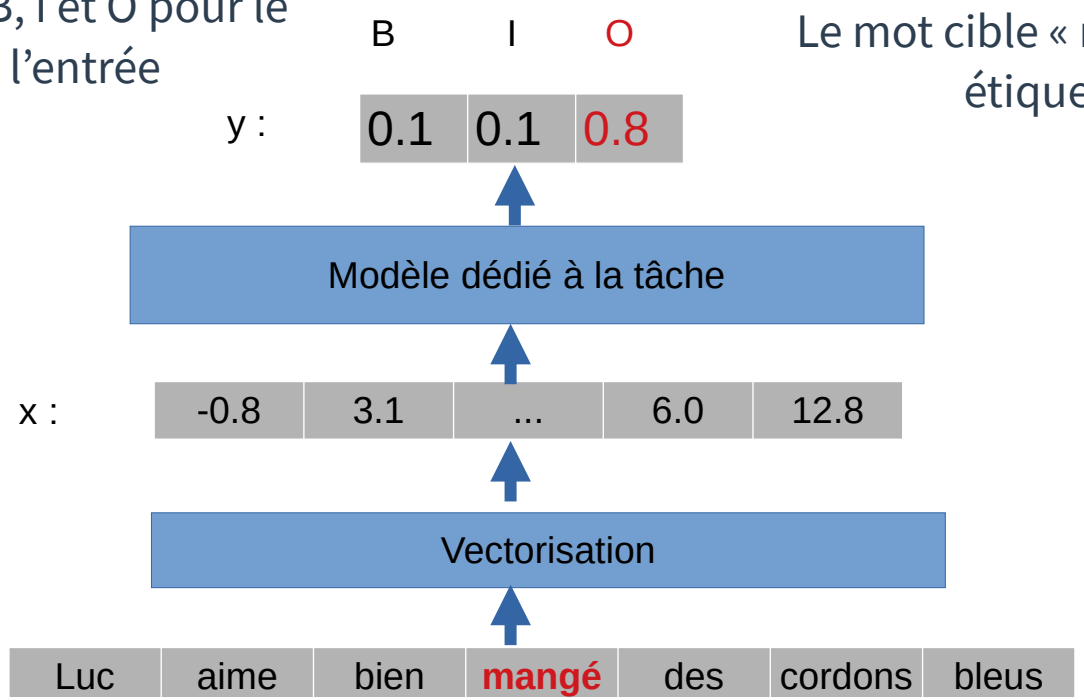


Identification d’expressions multi-mots (3)

y : probabilités des étiquettes B, I et O pour le mot cible en fonction de l’entrée

Le mot cible « mangé » sera étiqueté O

Tâche de classification multi-classes



Identification d’expressions multi-mots (4)

Résultats expérimentaux pour les expressions verbales

- Données PARSEME 1.2 (Ramisch et al. 2020) pour 14 langues
- Annotées pour les expressions verbales
- Découpage en données d’apprentissage et d’évaluation

- Le meilleur système MTLB-STRUCT (Taslimipoor et al. 2020) fondé sur un réseau de neurones et un grand modèle de langue obtient
  - Une exactitude globale de 70 % pour l’identification des expressions verbales
  - Mais seulement 40 % pour les expressions non vues lors de l’apprentissage

Expliquer/reformuler des termes médicaux
(post-doc de Ioana Buhnila)

Contexte

- Les termes médicaux sont difficiles à comprendre pour des non-initiés
- Des outils d'explication ou reformulation adaptés peuvent faciliter la compréhension des patients (vs. practiciens)

Exemples pris de Buhnila et al. (2024)

- hypopnée → respiration partiellement bloquée
- myasthénie grave → trouble qui entraîne une faiblesse musculaire et une fatigue musculaire excessive
- céphalée d'allure commune → migraine
- akathisie → agitation intérieure et incapacité à rester assis

Expliquer/reformuler des termes médicaux (2)

Approche utilisée (sur le français) [Buhnila et al. 2024]

- Exploitation de modèles de langue génératifs librement disponibles : ex. BIOMISTRAL [Labrak et al. 2024, modèle spécialisé au médical]
- Interrogation à l'aide de requêtes textuelles (prompts)

  ex. *qu'est-ce qu'un gliome ?*

Principe des modèles génératifs

Le modèle va générer la réponse de la manière suivante : chaque mot est généré consécutivement en fonction de la requête (séquence de mots) et des mots déjà générés, à partir d'un réseau de neurones

Quel est le mot le plus probable étant donné la requête et les mots déjà générés ?

Expliquer/reformuler des termes médicaux (3)

Résultats (Buhnila et al. 2024)

- Le modèle BIOMISTRAL produit des erreurs falgrantes dans 5 % des cas environ sur la base de données Refomed (Buhnila 2023)

Exemples d'erreurs

- Asthme → maladie où les airways se ferment et se contractent, faisant du bruit lors de l'inspiration et de la respiration

- Céphalée avec signes de gravité → signes de gravité : trouble de la conscience, trouble de la parole, trouble de la mobilité, trouble de la vision, trouble de la coordination, trouble de la sensibilité, trouble de la sensibilité

Conclusion

- Un modèle IA (générative) s'appuie sur des fonctions mathématiques avec des millions/milliards/… de paramètres optimisés à partir de grands volumes de données

- Des performances assez spectaculaires au premier abord, mais
  - les productions peuvent être erronnées (toujours se méfier)
  - les modèles sont opaques (interprétabilité ?)

Mais aussi…

cf. les présentations de Karen Fort [https://members.loria.fr/KFort/] sur le sujet

- Un grand coût énergétique (impact carbone, consommation d'eau)
- Des stéréotypes sociaux dans les productions
- Des problématiques juridiques sur les données utilisées
- Apprentissage par renforcement humain
- …

Références bibliographiques

Prédiction de la complexité des mots en contexte

Abdelhak Kelious, Mathieu Constant, and Christophe Coeur. 2024a. Complex Word Identification: A Comparative Study between ChatGPT and a Dedicated Model for This Task. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3645–3653, Torino, Italia. ELRA and ICCL.

Abdelhak Kelious, Mathieu Constant, and Christophe Coeur. 2024b. Investigating strategies for lexical complexity prediction in a multilingual setting using generative language models and supervised approaches. In *Proceedings of the 13th Workshop on Natural Language Processing for Computer Assisted Language Learning*, pages 96–114, Rennes, France. LiU Electronic Press.

Matthew Shardlow, Richard Evans, Gustavo Henrique Paetzold, and Marcos Zampieri. 2021. SemEval- 2021 task 1: Lexical complexity prediction. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 1–16, Online. Association for Computational Linguistics.

Références bibliographiques

Identification d'expressions multi-mots

Carlos Ramisch, Agata Savary, Bruno Guillaume, Jakub Waszczuk, Marie Candito, Ashwini Vaidya, Verginica Barbu Mititelu, Archna Bhatia, Uxoa Iñurrieta, Voula Giouli, Tunga Güngör, Menghan Jiang, Timm Lichte, Chaya Liebeskind, Johanna Monti, Renata Ramisch, Sara Stymne, Abigail Walsh, and Hongzhi Xu. 2020. Edition 1.2 of the PARSEME Shared Task on Semi-supervised Identification of Verbal Multiword Expressions. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 107–118, online. Association for Computational Linguistics.

Shiva Taslimipoor, Sara Bahaadini, and Ekaterina Kochmar. 2020. MTLB-STRUCT @Parseme 2020: Capturing Unseen Multiword Expressions Using Multi-task Learning and Pre-trained Masked Language Models. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 142–148, online. Association for Computational Linguistics.

Références bibliographiques

Reformuler des termes médicaux

Ioana Buhnila. 2023. *Une méthode automatique de construction de corpus de reformulation*. Thèse de doctorat. Université de Strasbourg.

Ioana Buhnila, Aman Sinha, and Mathieu Constant. 2024. Retrieve, Generate, Evaluate: A Case Study for Medical Paraphrases Generation with Small Language Models. In *Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024)*, pages 189–203, Bangkok, Thailand. Association for Computational Linguistics.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. BioMistral: A Collection of Open-Source Pretrained Large Language Models for Medical Domains. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5848–5864, Bangkok, Thailand. Association for Computational Linguistics.

Merci pour votre attention !