

Optimisation et adaptation des modèles de langage à grande échelle (LLM) pour les langues à faibles ressources

1. Contexte

Sur près de 7 000 langues parlées dans le monde, seule une minorité bénéficie des avancées en intelligence artificielle et traitement automatique du langage. Environ 6 000 langues autochtones sont menacées, dont la moitié risque de disparaître d'ici 2100 selon l'Unesco [1]. Ces langues peu dotées en ressources souffrent d'un manque de données numériques, d'outils linguistiques de base et de ressources humaines et financières capables de faciliter leur intégration dans l'espace numérique. Par ailleurs, les modèles de langage à grande échelle (LLM) ont révolutionné le traitement automatique du langage naturel (TALN) et ont démontré des performances exceptionnelles dans diverses tâches de TALN pour les langues largement parlées. Cependant, leur efficacité dans la gestion des langues à faibles ressources demeure une préoccupation. Ce projet de thèse vise à développer un pipeline complet pour enrichir les langues peu dotées en ressources (comme le Kabyle) en exploitant toutes les données disponibles - textes, vidéos, enregistrements audio - pour créer des corpus structurés et développer des outils de traduction automatique basés sur des grands modèles de langage.

2. Défis scientifiques Challenges

Ce projet de thèse vise à adresser de manière holistique les défis du traitement automatique du Kabyle comme exemple de langue peu dotée. Il intègre plusieurs défis essentiels, de la gestion et l'enrichissement des données à l'optimisation de modèles de langage avancés.

Axe 1 : Amélioration et Préparation des Données Linguistiques

Cet axe est la pierre angulaire de tout développement TALN pour une langue à faibles ressources. Il se concentre sur la création et l'enrichissement des ressources de données nécessaires. Un aspect fondamental sera de faire un état de l'art des différentes méthodes de data augmentation et d'annotation automatique de données pour une langue, puis de les expérimenter et les évaluer spécifiquement sur le Kabyle. Ces approches sont cruciales pour pallier le manque de corpus annotés, nécessaire notamment pour le fine-tuning de modèles de langage.

Axe 2 : Systèmes de Reconnaissance et de Normalisation Linguistique

Cet axe vise à transformer des données visuelles en texte kabyle utilisable, avec une couche de correction intégrée, pouvant enrichir les corpus des langues peu dotées en ressources à travers le contenu multimodal.

Axe 3 : Adaptation et Optimisation de Modèles de Langage Avancés

Cet axe est dédié à l'ingénierie des modèles de langage, en les adaptant spécifiquement aux nuances et aux défis du Kabyle, à travers un fine-tuning d'un Modèle de Langue Adapté au Kabyle pour des Tâches Spécifiques mais aussi l'optimisation par apprentissage par renforcement (RLHF) d'un modèle de langage pour le Kabyle.

3. Problématique

La question de recherche centrale est : Comment concevoir et développer un LLM pour des langues peu dotées, c'est-à-dire, lorsque les corpus sont limités, la transcription de la langue peu maîtrisée, des données numériques non exploitables directement, etc.

4. Objectifs scientifiques

Pour répondre à cette question de recherche, on identifie plusieurs objectifs :

- **Objectif 1:** élaboration d'un corpus en s'appuyant sur les données collectées par l'Inalco et d'autres sources et enrichissement des corpus avec des mécanismes de data augmentation.
- **Objectif 2:** Normalisation linguistique à travers l'intégration de règles pour la correction automatique permettant de générer des données correctes intégrables dans les corpus élaborés.
- **Objectif 3:** Concevoir un pipeline LLM adapté au Kabyle pour améliorer la traduction entre langues peu et bien dotées en ressources.
- **Objectif 4:** La mise en place d'un processus de fine-tuning adapté, et l'évaluation du modèle avec des métriques dédiées.

- **Objectif 5:** L'identification d'une méthode de RL adaptée au contexte d'étude, la préparation de l'environnement et des ressources requises, la mise en place d'un programme de RL spécifique, et l'évaluation du modèle avec des métriques dédiées.

5. Planning

Année 1: Collecte de données et construction de corpus

- **Etat de l'art:** Réaliser une étude approfondie des méthodes de **data augmentation** et d'annotation automatique, incluant les techniques génératives comme le *paraphrasing* et la *back-translation*.
- **Collecte de données:** la collecte de matériaux non numérisés et en ligne (textes, audio et vidéos) capturant du contenu linguistique authentique.
- **Construction de corpus kabyles:** création de jeux de données annotés et d'entraînement, incluant des données de reconnaissance de caractères et des corpus parallèles combinant langues bien dotées et peu dotées en ressources.

Année 2: Méthodologie pour l'ajustement fin, l'adaptation et l'optimisation des LLM

- **Développement d'un pipeline:** un LLM léger sera conçu pour supporter la traduction et l'assistance textuelle pour les langues peu dotées en ressources, illustrés à l'aide du Kabyle.
- **Ajustement fin:** La mise en place d'un processus de fine-tuning adapté, et l'évaluation du modèle avec des métriques dédiées.
- **Optimisation:** par apprentissage par renforcement (RLHF) du modèle de langue pour le Kabyle.

Année 3: Evaluation, dissémination, et rédaction de la thèse

- **Evaluation de l'approche:** Tester l'approche développée en l'évaluant grâce à des métriques bien définies.
- **Dissémination des résultats:** Contribuer via des publications dans des revues et conférences internationales.
- **Rédaction et soutenance de la thèse.**

6. Encadrement :

- Dr. **Samia Bouzefrane**, Professeure en informatique, Laboratoire CEDRIC, CNAM, Paris. Samia Bouzefrane (<https://samia.roc.cnam.fr>) est l'auteure d'un dictionnaire informatique français-anglais-kabyle. Ses recherches portent sur l'utilisation de l'IA dans différents domaines. Depuis deux ans, elle s'intéresse à l'utilisation de l'IA générative pour la langue kabyle comme exemple de langue peu dotée.
- Dr. **Kamal Nait Zerrad**, Professeur en linguistique berbère, Inalco, Paris. Kamal Nait Zerrad possède une longue expérience dans le domaine des langues berbères. Il a contribué via différents travaux de recherche dont la notation berbère, la constitution de corpus berbères, des études linguistiques et sociologiques, etc. (<https://www.centrederechercheberbere.fr/accueil.html>).

7. Collaboration internationale

Les recherches proposées dans cette thèse seront menées en collaboration avec le Dr. Youakim Badr, Professeur en analyse de données et intelligence artificielle (<https://youakim.info>) de l'Université d'État de Pennsylvanie, États-Unis. Cette collaboration favorisera un partenariat de recherche étroit sur les LLM.

9. Références bibliographiques

- [1] UNESCO. Language vitality and endangerment. Technical report, UNESCO, 2003. <https://unesdoc.unesco.org/ark:/48223/pf0000187026>.
- [2] UNESCO. Digital initiatives for indigenous languages, 2021. URL <https://unesdoc.unesco.org/ark:/48223/pf0000387186>
- [3] Kamal Nait Zerrad. Linguistique(s) de corpus : De la constitution à l'exploitation des corpus, REB, Vol. 11, 2011.
- [4] Arturo Trujillo. Translation Engines: Techniques for Machine Translation. Applied Computing, Springer, 1999.
- [5] H. Zhang, J. Wang, J. Luo, M. Zhang and G. Zhou, "Boosting LLM's Continual Sentiment Understanding for Low-Resource Languages," in IEEE Transactions on Audio, Speech and Language Processing, vol. 33, pp. 3042-3055, 2025, doi: 10.1109/TASLPRO.2025.3581018.
- [6] R. S. Kiziltepe, E. Ezin, Ö. Yentür, A. M. Basbrain and M. Karakus, "Advancing Sentiment Analysis for Low-Resource Languages Using Fine-Tuned LLMs: A Case Study of Customer Reviews in Turkish Language," in IEEE Access, vol. 13, pp. 77382-77394, 2025, doi: 10.1109/ACCESS.2025.3566000.
- [7] H. Yang, M. Zhang and J. Guo, "From Scarcity to Sufficiency: Machine Translation Techniques for Low-Resource LLMs Enhancement," 2024 2nd International Conference on Foundation and Large Language Models (FLLM), Dubai, United Arab Emirates, 2024, pp. 36-41, doi: 10.1109/FLLM63129.2024.10852466.
- [8] S. Tahery and S. Farzi, "An Adapted Few-Shot Prompting Technique Using ChatGPT to Advance Low-Resource Languages Understanding," in IEEE Access, vol. 13, pp. 93614-93628, 2025, doi: 10.1109/ACCESS.2025.3574115.