

Etiquetage morpho-syntaxique POS tagging pour le Bengali

Shami THIRION SEN

INALCO, Paris

Abstract

Cette étude explore l'état d'art de quelques outils existants pour l'étiquetage morpho-syntaxique et en compare trois pour l'étiquetage en POS de Bengali.

Mots-clés: POS tagging, Spacy, banglanltk,

1. Introduction

Bien que le bengali soit une langue avec un grand nombre de locuteurs, il est encore peu exploité dans le domaine du TAL (Sarker, 2021), avec très peu de données annotées. Le but principal de cette étude est une première exploration de l'état d'art et une comparaison entre Spacy, banglanltk et bnlp pour l'étiquetage morpho-syntaxique (POS), ainsi qu'une exploration des possibilités de futurs travaux.

IF needed : FOOTNOTE ¹ BOLDTEXT **bold text**
BOLDTEXT

2. Etat de l'Art

Plusieurs tentatifs de création d'outils ont été réalisés via des approches différentes. Cette étude a tenté une brève exploration de certains. Des méthodes supervisées et partiellement supervisées par exemple avec MMC (Alam et al.) ont menés à des résultats satisfaisantes - avec entre 85% et 91% d'accuracy.

Il existe plusieurs bibliothèques Python tel que banglanltk², bnlp³, bnltk⁴.

- Après une première essai, l'outil bnltk semble avoir un bug et donc ne fonctionne pas pour l'étiquetage.
- banglanltk et bnlp proposent l'étiquetage morpho-syntaxique en partie du discours.

En ce qui concerne Spacy, il n'existe pas de modèle officiellement proposé pour cette langue. Et le modèle ⁵ n'est pas encore fonctionnel, l'essai d'installation a été sans succès. L'étude menée par Sen et al. présente une analyse globale des

recherches dans le domaine, qui ne peuvent être explorés ici. Néanmoins, cette étude serait un point de départ pour toute future recherche.

1. `xelatex your_paper_name.tex`

2. `bibtex your_paper_name.aux`

3. Corpus

Le plus grand défi du projet et ainsi le processus le plus long a été de trouver un corpus suffisamment grand et adapté pour l'entraînement des modèles d'apprentissage. Malheureusement, le seul corpus normalisé en bengali se trouve sur Universal Dependencies comprend 320 tokens avec 55 phrases. Dans la perspective d'une proximité linguistique, le corpus hindi, mis à disposition par IIIT-Hyderabad, se trouvant également sur Universal Dependencies a été utilisé, partant de l'hypothèse que faisant partie des langues indo-européennes et ayant une certaine similarité linguistique, ce dernier pouvait être utile.

Un autre corpus annoté ⁶ mis à disposition sur le git (sous nom de nltr) a été nettoyé afin d'extraire de de créer des fichiers conllu pour train et dev. Le but était de servir de ces fichiers pour l'entraînement des outils. Malheureusement la conversion spacy a échoué.

4. Outils et résultats

Ce présent projet tente quelques différents outils de POS tagging en Bengali afin de comparer les résultats, certains conçus pour le bengali, alors que d'autres nécessitent un corpus annoté pour l'entraînement des modèles.

4.1. Spacy

Spacy étant la bibliothèque découverte dès le début du projet, était le premier outil essayé pour la tâche.

¹This is the example of a footnote.

²<https://pypi.org/project/banglanltk/>

³<https://pypi.org/project/bnlp-toolkit/>

⁴<https://pypi.org/project/bnlk/>

⁵bn_core_news_sm

⁶https://github.com/abhishekgupta92/bangla_pos_tagger/tree/master/d

Or, actuellement Spacy ne fournit pas de modèle dédié pour le bengali. Une première approche a été d'entraîner un modèle avec un corpus train en hindi comportant qui a donné un accuracy score de 0,35. Ceci nous menant à écarter la possibilité d'entraînement avec les formes spacy. Ulérieurement, d'autres essais peuvent être envisagé avec soit une lemmatisation ou une stemmatisation des mots en combinaison avec la vectorisation. Mais pour l'instant cette voie est peu concluante.

Des modèles multilingues "xx_sent_ud_sm", "xx_ent_wiki_sm" ne comportant pas d'analyseur morpho-syntaxique n'ont donné aucun résultat.

4.2. **bnltk**

La bibliothèque **bnltk** en Python, malgré une visible simplicité d'utilisation, contient des bugs, et donc n'aboutit à aucun résultat qui peut être comparée.

4.3. **banglanltk**

En effet, avec banglanltk nous réussissons à un étiquetage direct avec l'écriture bengali sans passer par la romanisation. Après avoir extrait le texte brut et la romanisation du corpus test de bengali « bn_bru-ud-test.conllu », nous avons obtenu les étiquettes en parties du discours de chaque token. Nous avons ensuite reconstruit le corpus sous format **Corpus** du module bn_eval_basique afin de pouvoir réutiliser les fonctions « compute_accuracy » et « print_report ». L'extraction du texte brut et du texte romanisé ont été nécessaire pour le formatage de Corpus, en vu de réutiliser les fonctions actuelles. Malheureusement, cette approche n'a pas eu beaucoup de succès.

Avec une extraction et une comparaison avec l'étiquetage originale du corpus text , nous obtenons un score de 40%. Ce qui est une amélioration en comparaison avec l'entraînement Spacy.

4.4. **Bnlp et banglanltk**

Le script textbfpos_tagging_with_banglanltk réalise une étiquetage de fichier test de POS avec bnlp et banglanltk afin d'obtenir les prédictions. Ensuite il compare les résultats avec les annotation du corpus test. Avec banglanltk nous arrivons à 42,95% alors qu'avec bnlp nous atteignons 76,25% d'accuracy.

4.5. **Tables**

The instructions for tables are the same as for figures.

Level	Tools
Morphology	Pitrat Analyser
Syntax	LFG Analyser (C-Structure)
Semantics	LFG F-Structures + Sowa's Conceptual Graphs

Table 1: The caption of the table

5. **Conclusion**

Le manque du temps, une connaissance encore insuffisante du domaine et un corpus inadapté pour la tâche entreprise, ont rendu impossible l'exploration et la compréhension globale de l'état de l'art. Les résultats obtenus sont donc loin d'être optimal. Malgré le fait qu'il s'agit d'un projet orientée vers les résultat ce présent travail reste à un stade encore thoirique.

Cependant, ce projet a été très enrichissnte et a permis de une première entrée en matière. Plusieurs aspects restent encore à explorer, tels que l'apprentissage supervisé ou semi-supervisé, plongement lexicaux (Magistry et al.) pour des langues peu dotés, création et entraînement des modèles.

Plusieurs approches sont envisageable pour améliorer les résultats mais aussi le travail globale du domaine. Création ou obtention des données manuellement annotés serait un premier pas indispensable.

- étude copréhensive des recherches déjà effectués qui ne peuvent être qu'exploré qu'avec le temps. bnltk

6. **Liens Utiles**

kjhnjb,jhbjhb

1. [Modèle Spacy](#)
2. [UD_{Bengali} – BRU](#)
3. [bnlp](#)
4. [Corpus Brut: Central Institute of Indian Languages \(CIIL\)](#)

7. **Bibliographie**

SAGOR SARKER. 2021. BNLN: Natural. Language Processing Toolkit for Bengali Language.

PIERRE MAGISTRY, ANNE-LAURE LIGOZAT, SOPHIE ROSSET. 2018. Étiquetage en parties du discours de langues peu dotées par spécialisation des plongements lexicaux.

OVISHAKE SEN 1, MOHTASIM FUAD1, MD.

NAZRUL ISLAM¹, JAKARIA RABBI ¹, MEHEDI MASUD², MD. KAMRUL HASAN ³, MD. ABDUL AWAL⁴, AWAL AHMED FIME¹, MD. TAHMID HASAN FUAD¹, DELOWAR SIKDER¹, AND MD. AKIL RAIHAN IFTEE. 2022. Bangla Natural Language Processing: A Comprehensive Analysis of Classical, Machine Learning, and Deep Learning Based Methods

FIROJ ALAM, MD. ARID HASAN, TANVIRUL ALAM, AKIB KHAN, JANNATUL TAJRIN, NAIRA KHAN, SHAMMUR ABSAR CHOWDHURY. 2021. A Review of Bangla Natural Language Processing Tasks and the Utility of Transformer Models.