Induction de sens pour enrichir des ressources lexicales

Mohammad Nasiruddin, Didier Schwab, Andon Tchechmedjiev, Gilles Sérasset, Hervé Blanchon

Induction de sens pour enrichir des ressources lexicales

Mohammad Nasiruddin, Didier Schwab, Andon Tchechmedjiev
Gilles Sérasset, Hervé Blanchon
Univ. Grenoble Alpes
{Mohammad.Nasiruddin, Didier.Schwab, Andon.Tchechmedjiev, Gilles.Serasset,
Hervé.Blanchon}@imag.fr

Résumé. En traitement automatique des langues, les ressources lexico-sémantiques ont été incluses dans un grand nombre d'applications. La création manuelle de telles ressources est consommatrice de temps humain et leur couverture limitée ne permet pas toujours de couvrir les besoins des applications. Ce problème est encore plus important pour les langues moins dotées que le français ou l'anglais. L'induction de sens présente dans ce cadre une piste intéressante. À partir d'un corpus de texte, il s'agit d'inférer les sens possibles pour chacun des mots qui le composent. Nous étudions dans cet article une approche basée sur une représentation vectorielle pour chaque occurrence d'un mot correspondant à ses voisins. À partir de cette représentation, construite sur un corpus en bengali, nous comparons plusieurs approches de clustering (k-moyennes, clustering hiérarchique et espérance-maximisation) des occurrences d'un mot pour déterminer les différents sens qu'il peut prendre. Nous comparons nos résultats au *Bangla WordNet* ainsi qu'à une référence établie pour l'occasion. Nous montrons que cette méthode permet de trouver des sens qui ne se trouvent pas dans le *Bangla WordNet*.

Abstract. In natural language processing, lexico-semantic resources are used in many applications. The manual creation of such resources is very time consuming and their limited coverage does not always satisfy the needs of applications. This problem is further exacerbated with lesser resourced languages. However, in that context, Word Sense Induction (WSI) offers an interesting avenue towards a solution. The purpose of WSI is, from a text corpus, to infer the possible senses for each word contained therein. In this paper, we study an approach based on a vectorial representation of the cooccurrence of word with their neighbours across each usage context. We first build the vectorial representation on a Bangla (also known as Bengali) corpus and then apply and compare several ing algorithms (k-Means, Hierarchical Clustering and Expectation Maximisation) that elicit clusters corresponding to the different senses of each word as used within a corpus. We wanted to use *Bangla WordNet* to evaluate the clusters, however, the coverage of *Bangla WordNet* being restrictive compared to Princeton WordNet (23.65%), we find that the clustering algorithms induce correct senses that are not present in *Bangla WordNet*. Therefore we created a gold standard that we manually extended to include the senses not covered in *Bangla WordNet*.

Mots-clés : Induction de sens, bengali, Weka, Clustering.

Keywords: Word Sense Induction, Bangla, Weka, Clustering.

1 Introduction

En traitement automatique des langues, les ressources lexico-sémantiques ont été incluses dans un grand nombre d'applications. La création manuelle de telles ressources est consommatrice de temps humain et leur couverture limitée ne permet pas toujours de couvrir les besoins des applications. Ce problème est encore plus important pour les langues moins dotées que le français ou l'anglais. L'induction de sens présente dans ce cadre une intéressante piste. À partir d'un corpus de texte, il s'agit d'inférer les sens possibles pour chacun des mots qui le composent. Nous étudions dans cet article une approche basée sur une représentation vectorielle pour chaque occurrence d'un mot correspondant à ses voisins. À partir de cette représentation, construite sur un corpus en bengali, nous comparons plusieurs approches de clustering des occurrences d'un mot pour déterminer les différents sens qu'il peut prendre. Nous montrons que cette méthode permet de trouver des sens qui ne se trouvent pas dans le *Bangla WordNet*.

Dans cet article nous commençons par présenter le bengali, le *Bangla WordNet* et le Wikipédia bengali qui constitue notre corpus. Nous présentons ensuite l'approche qui nous a permis de construire la représentation vectorielle pour chacune des occurrences de notre corpus. Enfin nous présentons les deux méthodes d'évaluation que nous utiliserons pour comparer les différents algorithmes de clustering.

2 Le bengali et *Bangla WordNet*

2.1 Le bengali

Le bengali, également appelé bangla est la septième langue la plus parlée au monde avec environ 200 millions de locuteurs et la plus orientale des langues indo-européennes. Elle est essentiellement parlée au Bangladesh (75% de la population) et la deuxième la plus parlée en Inde où elle est langue officielle de trois des vingt trois états (Garry & Rubino, 2001). Le bengali est écrit à l'aide de caractères dérivant du Brahmi. Comme le français, les mots bengalis sont séparés par des espaces et les signes de ponctuation sont les mêmes à part le dari (।) qui remplace le point pour la segmentation des phrases.

2.2 *Bangla WordNet*

Bangla WordNet (Dash, 2011; Niladri Sekhar Dash & Banerjee, 2011) a été conçu en suivant les principes du WordNet de Princeton pour l'anglais. Il fait partie du projet Indradhanush et a ainsi été développé en utilisant hindi WordNet(Somesh Jha & Bhattacharyya, 2010) comme pivot. Classiquement, un synset du *Bangla WordNet* est composée ainsi :

- un numéro d'identification du synset : un numéro unique provenant de l'hindi WordNet.
- une catégorie grammaticale : nom, verbe, adjectif, adverbe
- une glose : la description du concept par une définition et des exemples.
- un ensemble de mots, synonymes entre eux, dont les traits communs sont sensés correspondre au sens particulier décrit pas le synset

Par exemple, le synset associé à *Humâyûn* est :

ID : : 19673
CAT : : NOUN
CONCEPT : : একটি মুঘল শাসক যিনি বাবররে পুত্র ছলিনে (Dirigeant Mongol, fils de Bâbur)
EXAMPLE : : আকবর হুমায়ুনরে পুত্র ছলিনে। (Akbar était le fils de Humâyûn)
SYNSET-BENGALI : : হুমায়ুন, নাসরিুদ্দনি_হুমায়ুন (Humâyûn, Nasiruddin_Humâyûn.)

Les traductions en français ne se trouvent pas dans le *Bangla WordNet* et sont données uniquement pour faciliter la compréhension des lecteurs.

Parties du discours	Synsets	Nombre de sens	Nombre de sens (mots simples)	Nombre de sens (mots composés)	Mots monosémiques	Mots polysémiques
Nom	27 281	44 854	34 496	10 358	29 760	5 829
Verbe	2 804	4 448	318	4 130	2 260	671
Adjectif	5 815	10 264	569	9 695	6 813	1 385
Adverbe	445	906	159	747	721	79
Total	36 345	60 472	36 723	34 749	39 563	7 965

Table 1 – Statistiques sur le *Bangla WordNet*

2.3 Wikipédia bengali

Alors que le bengali est la septième langue la plus parlée au monde, le Wikipédia bengali [1] n'est que le quatre-vingt cinquième en nombre d'articles. Au 1er mai 2014 à midi-UTC il comprenait 29 756 articles contre, par exemple, 106 097 articles pour le latin ou 178 760 articles pour le basque, deux langues bien moins parlées

1. https://bn.wikipedia.org

au quotidien que le bengali [2]. Le faible nombre d'articles en bengali est évidemment explicable par le niveau d'éducation dans les régions où il est parlé. Par exemple, au Bangladesh et dans l'état du Bengale-Occidental qui représentent à eux deux trois-quarts des locuteurs du bengali, le taux de scolarisation dans le secondaire est inférieurs à 50% selon l'UNICEF [3] et l'OCDE (OCDE, 2011).

Nous allons utiliser le Wikipédia bengali comme corpus de textes dans notre expérience. Utiliser un Wikipédia est, en effet, un moyen simple d'obtenir des textes libres de droits pour toutes les langues où il en existe un. Dans cette expérience, nous utilisons la sauvegarde de la base de données du Wikipédia bengali [4] du 28 décembre 2013 [5] qui comporte 28 393 articles.

3 Construction de la matrice des voisinages

Dans cette section, nous présentons l'expérience réalisée qui a consisté à construire la matrice des voisinages puis à catégoriser chacune des instances de mots en fonction de leur contexte.

Nous appelons *occurrence* une suite de caractères délimitée par des séparateurs (espace,virgule, dari, *etc.*). Nous appelons *forme*, l'ensemble de toutes les occurrences ayant en commun leur suite de caractères. Nous appelons S l'ensemble des formes du corpus.

Si on considère le corpus constitué des deux phrases suivantes "*le chat mange la souris*" "*le chat mange le fromage*", il est composé de 10 mots et S contient 6 éléments. L'élément ‹*chat*› de S a deux instances que nous noterons ‹*chat#1*› et ‹*chat#2*›.

3.1 Préparation des données

Le texte brut des pages du Wikipédia bengali a été extrait et normalisé selon la forme normale NFC (*Normalization Form Canonical Decomposition*). Chaque phrase des pages Wikipedia est considérée comme un document indépendant. Nous obtenons donc un ensemble de 34 251 documents correspondant aux 28 393 pages originales. Enfin, chaque document est segmenté en une séquence d'instances.

Le Bangla ne distingue pas les majuscules des minuscules et a une morphologie relativement peu productive. Ainsi, aucun autre pré-traitement linguistique (lemmatisation, annotations en partie du discours, *etc.*) n'a été appliqué au corpus.

Avant pré-traitement	Après pré-traitement	Traduction indicative
বাংলা ভাষার ইতহিাসকে সাধারণত তনি ভাগে ভাগ করা হয : # প্রাচীন বাংলা (৯০০/১০০০ খ্রস্টিাব্দ – ১৪০০ খ্রস্টিাব্দ) — লখিতি নদির্শনরে মধ্যে আছে চর্যাপদ, ভক্তমিূলক গান; আমি, তুমি, ইত্যাদি সর্বনামরে আবর্িভাব; ক্রযিার্ভক্তি - ইলা, -ইবা, ইত্যাদি। ওড়য়িা ও অসমীয়া এই পর্বে বাংলা থকেে আলাদা হয়ে যায়।	বাংলা ভাষার ইতহিাসকে সাধারণত তনি ভাগে ভাগ করা হয : # প্রাচীন বাংলা (৯০০ / ১০০০ খ্রস্টিাব্দ – ১৪০০ খ্রস্টিাব্দ) — লখিতি নদির্শনরে মধ্যে আছে চর্যাপদ ,ভক্তমিূলক গান; আমি , তুমি , ইত্যাদি সর্বনামরে আবর্িভাব; ক্রযিার্ভক্তি - ইলা , - ইবা , ইত্যাদি । ওড়য়িা ও অসমীয়া এই পর্বে বাংলা থকেে আলাদা হয়ে যায় ।	L'histoire de la langue bengalie est habituellement divisée en trois étapes : Le bengali ancien (900/1000AD - 1400 AD) — Dans la langue écrite celà se manifeste par les Charyapada, chansons rituelles ; l'apparition du pronom personnel "Je", "Vous", etc. ; les inflexions verbales -ila, -iba, etc. Durant cette période, l'oriya et l'assamais sont des langues différentes du bengali.

Table 3 – Extrait du corpus avant et après le pré-traitement

2. https://meta.wikimedia.org/wiki/List_of_Wikipedias
3. http://www.unicef.org/french/infobycountry/bangladesh_bangladesh_statistics.html
4. http://dumps.wikimedia.org/backup-index-bydb.html
5. http://dumps.wikimedia.org/bnwiki/20131228/

3.2 Mise en œuvre

Après le pré-traitement du corpus, nous construisons la matrice formes-contextes et nous comparons plusieurs algorithmes de clustering pour la création des sens.

3.2.1 Construction de la matrice tf-tdf

Dans cette matrice, les lignes correspondent aux éléments de S et les colonnes correspondent à leurs occurrences. Chaque case de la matrice contient le nombre de fois où ces occurrences se trouvent dans le même document que l'élément. Par exemple, si on considère que notre corpus n'est composé que des deux documents suivants "*Le chat mange la souris.*" "*Le chat mange le fromage.*", la matrice résultat sera celle présentée dans la table 4. Notre corpus fait 34 251 documents qui contiennent 3 957 075 instances et 373 685 formes est le cardinal de

	le#1	chat#1	mange#1	la#1	souris#1	le#2	chat#2	mange#2	le#3	fromage#1
le	0	1	1	1	1	1	2	2	1	2
chat	1	0	1	1	1	1	0	1	1	1
mange	1	1	0	1	1	1	1	0	1	1
la	1	1	1	0	1	0	0	0	0	0
souris	1	1	1	1	0	0	0	0	0	0
fromage	0	0	0	0	0	1	1	1	1	0

Table 4 – Matrice obtenue avec l'exemple

l'ensemble S. Nous avons utilisé une machine de 32 cœurs Intel Xeon E5-2650 à 2.0 GHz équipée de 256 Go de ram. La génération de cette matrice a pris environ 48 heures mais n'a utilisé aucune parallélisation.

3.2.2 Construction des clusters

L'objectif de ce travail est de voir dans quelle mesure il est possible d'inférer des sens à partir de clusters pour créer une ressource lexicale. Nous voulons ainsi comparer plusieurs algorithmes de clustering pour savoir lequel ou lesquels seraient les plus performants pour cette tâche. Nous avons ainsi utilisé la suite de logiciels d'apprentissage automatique Weka (*Waikato Environment for Knowledge Analysis*) pour créer les clusters de sens. Nous expérimentons ici trois algorithmes : k-moyennes, clustering hiérarchique et espérance-maximisation qui utilisent une distance euclidienne.

k-moyenne (Hartigan & Wong, 1979) est un algorithme numérique, non-supervisé, probabiliste et itératif de partition de données. Il permet de générer un nombre de clusters k donné en paramètre. Aucune instance ne peut appartenir à deux clusters à la fois.

La clustering hiérarchique (Johnson, 1967) est un algorithme qui unit les clusters les plus proches jusqu'à ce que le nombre de clusters voulu soit atteint.

L'algorithme d'espérance-maximisation (EM) (Jin & Han, 2010) estime par maximum de vraisemblance les paramètres d'un modèle probabiliste ayant des variables latentes. Une gaussienne de paramètre inconnu modélise l'ensemble des points d'un cluster. Une distribution modélise la vraisemblance d'appartenance des points aux clusters. EM estime conjointement les paramètres des gaussiennes afin de maximiser la vraisemblance d'appartenance aux clusters. Contrairement à k-moyennes et le clustering hiérarchique, EM travaille sur les distributions des points, ce qui est une approche orthogonale qui peut amener des résultats différents et meilleurs.

4 Évaluation des catégories

4.1 Le *Bangla WordNet* et ses sens dans le corpus

Comparé au Princeton WordNet (Miller, 1995), le *Bangla WordNet* (BWN) a une couverture de seulement 23,65% (Dash, 2011). Nous avons voulu essayer d'évaluer sur un petit sous-ensemble des entrées combien de sens manquaient. Nous avons choisi 7 mots au hasard et un natif bengali a annoté les occurrences de ces mots

Mots	Nombre d'occurrences	Sens *Bangla WordNet*	Sens manquant	Total BWN + corpus
সমাগত (samāgata – répétition)	8	2	0	2
অংক (aṅka – math)	26	2	4	6
অচল (acala – obsolète)	40	5	1	6
পদবী (padabī – titre)	113	1	6	7
জনপদ (janapada – communauté)	83	1	3	4
যুক্তাক্ষর (yuk-tākṣara – conjoint)	12	1	0	1
অটল (aṭala – régulier)	43	3	6	9
Total	325	15	20	35

Table 5 – Statistiques sur sept mots pris au hasard dans le corpus

dans les 258 documents et 14 817 phrases dans lesquelles ils apparaissent. Si le sens existe dans le *Bangla WordNet*, il a annoté avec ce sens sinon il a créé de nouveaux sens. Cette annotation lui a pris environ 8 heures.

La table 5 présente les résultats obtenus avec ces 7 mots. On le voit, pour seulement sept mots, avec un corpus assez simple, on trouve 20 sens manquants à *Bangla WordNet* soit au moins 57% des sens possibles pour ces mots. L'enrichissement de cette ressource est donc un objectif assez primordial. Nous étudions dans la partie suivante dans quelle mesure nous pouvons le faire toujours en nous basant sur nos 7 mots.

4.2 Évaluation

L'évaluation se fait par rapport à deux références : le *Bangla WordNet* et le *Bangla WordNet* enrichi par les sens manquants découverts par l'annotateur. L'élément le plus important pour l'évaluation est le choix de mesures de la qualité du la clustering par rapport aux références.

4.2.1 Mesures

Nous utilisons 4 mesures standard pour l'évaluation de la qualité du clustering. Le score de chaque cluster est calculé comme la somme de la mesure entre le cluster et chacun des clusters de référence. Sur l'ensemble des clusters, on donne la moyenne des scores pour chaque cluster.

- La F1-mesure est la moyenne harmonique entre la précision (P) et le rappel (R) et s'exprime comme $F1 = \frac{2 \cdot P \cdot R}{P+R}$. P est le nombre de points correctement assignés (vrais positifs, VP) sur le nombre total de points (somme de VP + les faux positifs, FP). R est le nombre de points correctement assignés sur le nombre de points attendus (somme de VP et des faux négatifs, FN).
- L'indice de Jaccard sert à calculer la similarité entre un cluster et un cluster de référence en comptant le nombre d'éléments communs sur le nombre total d'éléments des deux clusters. $JI(C, C_{ref}) = \frac{TP}{VP+FP+FN}$
- L'indice de Rand permet de calculer le pourcentage d'éléments assignés correctement, $RI(C, C_{ref}) = (VP + VN)/(VP + FP + FN + VN)$.
- L'indice de Rand ajusté est un indice Rand qui est ajusté pour prendre en compte l'assignation correcte ou incorrecte due au hasard en prenant en compte la distribution de l'ensemble des indices de rand sur les clusters : $ARI(C, C_{ref}) = \frac{Indice - IndiceEspr}{IndiceMaximum - IndiceEspr}$

4.2.2 Résultats

La table 6 présente les résultats de trois algorithmes k-moyennes, clustering hiérarchique et EM. À droite, la référence est le *Bangla WordNet* tandis qu'à gauche, la référence est le standard que nous avons créé. Pour

Algorithmes	F1	JI	RI	ARI
k-moyennes	**54.91**	42.85	42.85	2.36
Hiérarchique	46.06	**61.34**	**55.41**	**4.78**
EM	49.19	39.73	39.98	2.30

Algorithmes	F1	JI	RI	ARI
k-moyennes	50.69	36.62	40.69	3.16
Hiérarchique	**52.05**	**58.09**	**62.54**	**5.80**
EM	44.21	34.40	43.42	2.63

Table 6 – Résultats obtenus avec la référence *Bangla WordNet* (à gauche) avec la référence créée (à droite)

la référence *Bangla WordNet*, si l'on s'en tient au F-score, k-moyennes obtient la meilleure performance par rapport à EM et au clustering hiérarchique (résp. +8,85, +4,72). En revanche, avec les trois autres mesures c'est le clustering hiérarchique qui est de loin le meilleur (par rapport au second meilleur, JI — +18,49, RI — +12,56, ARI — +2,52). k-moyennes et EM sont proches, mais k-moyennes reste toujours devant. Pour la référence que nous avons créée, nous constatons que pour toutes les mesures le clustering hiérarchique est meilleur (par rapport au second meilleur, F-score — +1.36, JI — +21,47, RI — +19,12, ARI — +2,64). k-moyennes est deuxième pour toutes les mesures sauf pour RI. En fonction des applications et de l'importance des vrais et faux négatifs, certaines mesures seront plus représentatives que d'autres. ARI est dans le cas général considéré comme la mesure la plus représentative. Pour l'objectif que nous nous sommes fixés, améliorer la base lexicale, EM semble ainsi être la voie la plus prometteuse.

5 Conclusions et perspectives

Dans cet article, nous avons présenté une première approche d'enrichissement d'une base lexico-sémantique en particulier pour des langues moins dotées que l'anglais comme l'est le bengali pourtant septième langue la plus parlée au monde. Par l'étude des occurrences de sept mots dans un corpus constitués des textes du Wikipédia du bengali, nous avons montré que 20 sens n'étaient pas répertoriés dans le *Bangla WordNet* contre 15 qui s'y trouvaient (soit un taux d'absents de 57%). La mise en œuvre d'une méthode très simple nous a permis de découvrir automatiquement des sens qui ne se trouvaient pas dans la ressource initiale. Nos travaux actuels visent à améliorer la construction des clusters en particulier en exploitant les informations issues de la ressource initiale et à les évaluer *in vivo*, c'est-à-dire dans une application comme la traduction automatique.

Références

Dash N. S. (2011). Problems in defining language specific synsets (lss) in bengali for the indradhanush indo-wordnet : Some theoretical and practical issues. In *Proceedings of the 2nd National Workshop of Indradhanush WordNet Consortium*, p. 4–18.

Garry J. & Rubino C. (2001). Facts about the world's languages. *HW Wilson*.

Hartigan J. A. & Wong M. A. (1979). Algorithm as 136 : A k-means clustering algorithm. *Applied statistics*, p. 100–108.

Jin X. & Han J. (2010). Expectation maximization clustering. In *Encyclopedia of Machine Learning*, p. 382–383. Springer.

Johnson S. C. (1967). Hierarchical clustering schemes. *Psychometrika*, **32**(3), 241–254.

Miller G. A. (1995). Wordnet : a lexical database for english. *Communications of the ACM*, **38**(11), 39–41.

Niladri Sekhar Dash, Abhisek Sarkar D. B. & Banerjee S. (2011). Problems and challenges in translation of hindi synsets into bengali in indradhanush wordnet. In *Proceedings of the 2nd National Workshop of Indradhanush WordNet Consortium*, p. 19–38.

OCDE (2011). *Études économiques de l'OCDE : Inde 2011*. Rapport interne, OCDE.

Somesh Jha, Darren Narayan P. P. & Bhattacharyya P. (2010). A wordnet for hindi. In *Proceedings of the International Workshop on Lexical Resources in Natural Language Processing*.