Identification de cognats à partir de corpus parallèles français-roumain

Mirabela Navlea Amalia Todiraşcu

(1) Université de Strasbourg, 22 rue René Descartes, BP, 80010, 67084 Strasbourg, cedex
navlea@unistra.fr, todiras@unistra.fr

Résumé Cet article présente une méthode hybride d'identification de cognats français - roumain. Cette méthode exploite des corpus parallèles alignés au niveau propositionnel, lemmatisés et étiquetés (avec des propriétés morphosyntaxiques). Notre méthode combine des techniques statistiques et des informations linguistiques pour améliorer les résultats obtenus. Nous évaluons le module d'identification de cognats et nous faisons une comparaison avec des méthodes statistiques pures, afin d'étudier l'impact des informations linguistiques utilisées sur la qualité des résultats obtenus. Nous montrons que l'utilisation des informations linguistiques augmente significativement la performance de la méthode.

Abstract This paper describes a hybrid French - Romanian cognate identification method. This method uses lemmatized, tagged (POS tags) and sentence-aligned parallel corpora. Our method combines statistical techniques and linguistic information in order to improve the results. We evaluate the cognate identification method and we compare it to other methods using pure statistical techniques to study the impact of the used linguistic information on the quality of the results. We show that the use of linguistic information in the cognate identification method significantly improves the results.

Mots-clés : cognat, identification de cognats, corpus parallèles alignés au niveau propositionnel
Keywords: cognate, cognate identification, sentence-aligned parallel corpora

1 Introduction

Les cognats sont des indices lexicaux importants pour différentes applications multilingues, et notamment pour des systèmes d'alignement de corpus parallèles et de traduction automatique statistique. Nous définissons comme cognats les équivalents de traduction ayant une forme identique d'une langue à l'autre ou présentant des similarités aux niveaux orthographique ou phonétique (mots d'étymologie commune, emprunts). Les cognats sont nombreux entre des langues apparentées comme le français et le roumain, deux langues latines avec une morphologie flexionnelle riche. Mais, l'identification de cognats à partir des textes multilingues parallèles est une tâche difficile. Ceci est dû aux similarités orthographiques ou phonétiques importantes entre des mots ayant un sens différent.

Ainsi, de nombreux travaux se concentrent sur l'identification de cognats pour différentes paires de langues. Plusieurs approches exploitent les similarités orthographiques entre les mots d'une paire bilingue. Une approche simple et efficace est la méthode appelée 4-grammes: deux mots sont considérés comme cognats s'ils possèdent au moins 4 caractères et leurs premiers 4 caractères sont identiques (Simard et al., 1992). D'autres méthodes exploitent le coefficient de Dice (Adamson, Boreham, 1974 ; Brew, McKelvie, 1996). Ce score d'association calcule le rapport entre le nombre de caractères des bigrammes communs aux deux mots considérés et le nombre total des bigrammes des deux mots. Afin d'identifier les cognats, certaines méthodes calculent le rapport entre le nombre de caractères (ordonnés et pas nécessairement contigus) de la sous-chaine maximale commune aux deux mots et la longueur du mot le plus long (Melamed, 1999 ; Kraif, 1999). De manière similaire, d'autres méthodes calculent la distance entre deux mots, qui représente le nombre minimum de substitutions, insertions et suppressions utilisées pour transformer un mot dans un autre (Wagner, Fischer, 1974). D'autre part, certaines approches estiment la distance phonétique entre deux mots appartenant à une paire bilingue (Oakes, 2000). Kondrak (2009) propose des méthodes identifiant trois caractéristiques des cognats : les correspondances de sons récurrents, la similarité phonétique et l'affinité sémantique.

Notre méthode exploite des similarités orthographiques et phonétiques entre les mots d'une paire bilingue. Nous combinons méthodes n-grammes avec des stratégies de désambigüisation des données d'entrée (calcul de fréquences des cognats, extraction itérative des cognats considérés les plus fiables). Notre méthode vise en premier l'obtention d'une précision élevée, afin d'être intégrée dans un système d'alignement lexical.

Dans la section suivante, nous présenterons notre module d'identification de cognats. Ainsi, nous présenterons le corpus utilisé et les prétraitements effectués (sous-section 2.1), ainsi que la méthode développée (sous-section 2.2). Nous présenterons aussi l'évaluation des résultats obtenus et la comparaison avec d'autres méthodes statistiques dans la section 3. Nos conclusions et nos perspectives figurent dans la section 4.

2 Identification de cognats à partir de corpus parallèles bilingues

2.1 Le prétraitement du corpus

Nous avons mené nos expériences sur un corpus parallèle bilingue extrait automatiquement à partir de l'*Acquis Communautaire*. Ce corpus est composé d'un ensemble de normes européennes (1950-présent) étant disponible pour 231 paires de langues obtenues à partir de 22 langues officielles de l'UE. Le corpus utilisé contient 1000 phrases alignées 1:1. Le corpus français comprend 33 036 tokens et le corpus roumain 28 645 tokens. Nous avons sélectionné automatiquement des phrases bien formées (commençant par une majuscule et finissant par un signe de ponctuation). Nous avons limité la longueur des phrases à 80 mots. De plus, nous avons prétraité le corpus de travail en appliquant l'étiqueteur *TTL*[1] (Ion, 2007). Ainsi, le corpus est tokénisé, lemmatisé, étiqueté au niveau morphosyntaxique et annoté au niveau des chunks (groupes nominaux, groupes verbaux simples, etc.). *TTL* utilise l'ensemble d'étiquettes MSD[2] proposé par le projet Multext[3] pour le français (Ide, Véronis, 1994) et le roumain (Tufiş, Barbu, 1997). Nous donnons un exemple de phrases alignées 1:1 et prétraitées avec *TTL* dans la Figure 1 ci-dessous :

(fr) Les États membres communiquent à la Commission les méthodes qu'ils utilisent.
(ro) Statele membre comunică Comisiei metodele pe care le utilizează.

Français	Roumain
<seg lang="fr"><s id="ttlfr.41"> <w lemma="le" ana="Da-mp" chunk="Np#1">Les</w> <w lemma="état" ana="Ncmp" chunk="Np#1">États</w> <w lemma="membre" ana="Ncmp" chunk="Np#1">membres</w> <w lemma="communiquer" ana="Vmsp3p" chunk="Vp#1">communiquent</w> <w lemma="à" ana="Spa" chunk="Pp#1">à</w> <w lemma="le" ana="Da-fs" chunk="Pp#1,Np#2">la</w> <w lemma="commission" ana="Ncfs" chunk="Pp#1,Np#2">Commission</w> <w lemma="le" ana="Da-fp" chunk="Np#3">les</w> <w lemma="méthode" ana="Ncfp" chunk="Np#3">méthodes</w> <w lemma="que" ana="Pr">qu'</w> <w lemma="il" ana="Pp3mp" chunk="Vp#2">ils</w> <w lemma="utiliser" ana="Vmip3p" chunk="Vp#2">utilisent</w> <c>.</c></s></seg>	<seg lang="ro"><s id="ttlro.41"> <w lemma="stat" ana="Ncfpry" chunk="Np#1">Statele</w> <w lemma="membru" ana="Ncfp-n" chunk="Np#1">membre</w> <w lemma="comunica" ana="Vmis3s" chunk="Vp#1">comunică</w> <w lemma="comisie" ana="Ncfsoy" chunk="Np#2">Comisiei</w> <w lemma="metodă" ana="Ncfpry" chunk="Np#3">metodele</w> <w lemma="pe" ana="Spsa">pe</w> <w lemma="care" ana="Pw3--r">care</w> <w lemma="el" ana="Pp3fpa--------w" chunk="Vp#2">le</w> <w lemma="utiliza" ana="Vmip3" chunk="Vp#2">utilizează</w> <c>.</c></s></seg>

Figure 1 Phrases alignées français - roumain prétraitées avec TTL

Dans l'exemple présenté dans la Figure 1, l'attribut *lemma* garde les informations sur les lemmes, l'attribut *ana* contient les propriétés morphosyntaxiques et l'attribut *chunk* marque les groupes nominaux, adjectivaux, prépositionnels, verbaux et adverbiaux, non-récursifs.

[1] Tokenizing, Tagging and Lemmatizing free running texts

[2] Morpho-Syntactic Descriptors

[3] http://aune.lpl.univ-aix.fr/projects/multext/

2.2 La méthode d'identification de cognats

Notre méthode exploite le corpus parallèle lemmatisé, étiqueté et aligné au niveau propositionnel décrit antérieurement et applique les informations linguistiques associées aux unités lexicales, telles que les lemmes et les étiquettes morphosyntaxiques. Ainsi, nous considérons comme cognats les paires bilingues de mots qui remplissent les conditions linguistiques suivantes :

- les lemmes sont des équivalents de traduction dans deux phrases parallèles ;

- les lemmes sont identiques ou présentent des similarités orthographiques ou phonétiques ; Afin de détecter ces similarités (cf. Figure 2), nous nous concentrons plutôt sur le début des candidats cognats et nous ignorons leurs terminaisons ;

- les lemmes sont des mots contenu (noms, adjectifs, verbes, etc.) ayant la même partie du discours ou appartenant à la même classe d'équivalence de catégorie lexicale (par exemple, un nom peut être traduit par un nom, un verbe ou un adjectif) ; Ainsi, nous filtrons les mots courts comme les prépositions et les conjonctions afin de diminuer le bruit. Nous éliminons aussi les paires ambigües comme *lui* (pronom personnel) (fr) vs. *lui* (déterminant possessif) (ro), *ce* (déterminant démonstratif) (fr) vs. *ce* (pronom relatif) (ro). Nous détectons également des cognats courts comme *il* vs. *el* (pronom personnel), *cas* vs. *caz* (nom).

Notre méthode vise prioritairement l'obtention d'une précision élevée, afin d'être intégrée dans un système d'alignement lexical français - roumain. Ainsi, pour augmenter la précision de la méthode, nous combinons les conditions linguistiques mentionnées ci-dessus avec d'autres stratégies de désambigüisation des données d'entrée (extractions itératives des cognats considérés les plus fiables, extraction des candidats cognats les plus fréquents pour les cas ambigus). Nous présentons plus loin la configuration choisie (cf. Figure 3) dans ce sens.

Au niveau orthographique, nous classifions les cognats identifiés dans plusieurs catégories :

1. transfuges (nombres, certains sigles et acronymes, ainsi que les signes de ponctuation) ;

2. cognats identiques (*document* vs. *document*) ;

3. cognats similaires remplissant une des conditions suivantes :

- 4-grammes (Simard et al., 1992) ; Les cognats ont au moins les 4 premiers caractères du lemme identiques. La longueur des lemmes est égale ou supérieure à 4 (***prod****uit* vs. ***prod****us*) ;

- 3-grammes ; Les cognats ont les 3 premiers caractères identiques et la longueur de leurs lemmes est égale ou supérieure à 3 (***acte*** vs ***act***) ;

- 8-bigrammes ; Les cognats possèdent une sous-chaine de caractères ordonnés commune parmi les 8 premiers bigrammes au niveau du lemme. Au moins un caractère de chaque bigramme est commun aux deux lemmes. Cette condition permet les sauts d'un caractère différent (***souscrire*** vs. ***subscrie***). Dans ce cas, les lemmes ont une longueur supérieure à 7 ;

- 4-bigrammes ; Les cognats possèdent une sous-chaine de caractères ordonnés commune parmi les 4 premiers bigrammes au niveau du lemme. Au moins un caractère de chaque bigramme est commun aux deux lemmes. Dans ce cas, nous considérons aussi bien les lemmes courts (longueur égale ou inférieure à 7) (***gr****o****up****e* vs ***grup***) que les lemmes longs (longueur supérieure à 7) (***homol****ogué* vs. ***omol****ogat*).

Premièrement, nous appliquons un ensemble d'ajustements orthographiques constitué empiriquement, au niveau des lemmes, tels que : l'élimination des diacritiques, le repérage des correspondances phonétiques, etc. (cf. Figure 2). Comme le français a une écriture étymologique et le roumain possède une écriture généralement phonétique, nous identifions des correspondances phonétiques au niveau des lemmes et ensuite nous effectuons des ajustements orthographiques du français vers le roumain. Par exemple, les cognats *phase* vs. *fază* deviennent ***f****aze* vs. ***f****aza*. Dans cet exemple, nous réalisons deux ajustements : le groupe consonantique *ph* [f] du français devient [f] comme en roumain et le *s* [z] intervocalique du français devient [z] comme en roumain. Nous faisons aussi des ajustements dans les cas ambigus (*ch* ([ʃ] ou [**k**])) en considérant les deux variantes en étapes successives : *ma****ch****ine* vs. *ma****ş****ină*; ***ch****lorure* vs. ***c****lorură*.

Niveaux d'ajustements orthographiques	FR	RO	Exemples
signes diacritiques	é-e ; à-a…	î-i ; â-a, ă-a…	dépôt - depozit
lettres identiques contigües	cc-c ; dd-d…	cc-c ; nn-n…	rapport - raport
groupes consonantiques	ph	f [f]	phase - fază
	th	t [t]	méthode - metodă
	dh	d [d]	adhérent - aderent
	cch	c [k]	bacchante - bacantă
	ck	c [k]	stockage - stocare
	cq	c [k]	grecque - grec
	ch	ș [ʃ]	machine - mașină
	ch	c [k]	chlorure - clorură
q	q (final)	c [k]	cinq - cinci
	qu(+i) (médial)	c [k]	équilibre - echilibru
	qu(+e) (médial)	c [k]	marquer - marca
	qu(+a)	c(+a) [k]	qualité - calitate
	que (final)	c [k]	pratique - practică
s intervocalique	v + s + v	v + z + v	présent - prezent
W	w	v	wagon - vagon
y	y	i	yaourt - iaurt

Figure 2 Ajustements orthographiques appliqués au corpus parallèle français - roumain

Deuxièmement, nous appliquons sept étapes d'extractions itératives de cognats par catégories identifiées, dans l'ordre qui permet l'obtention d'une précision élevée de chaque étape (cf. Figure 3). Nous appliquons cette procédure aux candidats qui sont des mots contenu ayant la même catégorie lexicale (N-N, V-V etc.).

De plus, pour limiter le bruit des résultats, nous appliquons deux stratégies supplémentaires de désambiguïsation des données d'entrée.

Tout d'abord, nous filtrons les candidats ambigus (un même lemme source apparaît avec plusieurs candidats cible) en calculant la fréquence des paires candidates dans le corpus étudié. Ainsi, nous gardons la paire candidate la plus fréquente. Cette opération est très efficace pour augmenter la précision des résultats, mais dans certains cas, elle décroît le rappel. En effet, concernant les déverbaux, les lemmes français présentant une seule forme ont comme équivalents de traduction en roumain deux formes différentes : *information* vs. *informaţie* ou *informare*. Nous récupérons ces paires en utilisant des expressions régulières basées sur les terminaisons spécifiques des lemmes (*ion* (fr) vs. *ţie*|*re* (ro)).

Nous utilisons une autre stratégie de désambiguïsation des données d'entrée, et notamment la suppression du corpus des cognats considérés fiables (précision élevée) à la fin de chaque étape d'extraction. Par exemple, les cognats identiques *transport* vs. *transport* obtenus pendant l'étape d'extraction correspondante et supprimés des données d'entrée, éliminent l'occurrence du candidat *transport* vs. *tranzit* comme 4-grammes cognats dans l'étape suivante.

Etapes d'extraction par catégorie de cognats	Mots contenu / Même catégorie lexicale	Fréquence des candidats ambigus	Suppression des données d'entrée	Précision (%)
1 : transfuges			x	100
2 : cognats identiques	x		x	100
3 : 4-grammes (longueur des lemmes >= 4)	x	x	x	99,05
4 : 3-grammes (longueur des lemmes >= 3)	x	x	x	93,13
5 : 8-bigrammes (lemmes longs, longueur >7)	x		x	95,24
6 : 4-bigrammes (lemmes longs, longueur > 7)	x			75
7 : 4-bigrammes (lemmes courts, longueur =< 7)	x	x		65,63

Figure 3 Étapes d'extraction de cognats français - roumain

Après avoir extrait les candidats ayant la même catégorie lexicale, nous appliquons la même méthode d'extraction pour les cognats présentant des équivalences de catégorie lexicale (N-V, N-ADJ). Nous gardons seulement les cognats 4-grammes, puisque nous observons une diminution importante de la précision pour les autres catégories considérées (étapes 4-7).

3 Évaluation et comparaison de méthodes

Nous avons évalué notre méthode d'identification de cognats et nous avons fait aussi une comparaison des résultats avec ceux fournis par des méthodes statistiques générales (Tableau 1) :

- le calcul de la $mesure_{SCM}$ qui prend en compte la longueur de la sous-chaine maximale (SCM) de caractères communs aux deux mots d'une paire bilingue et la longueur du mot le plus long ; Les mots sont considérés comme cognats si la valeur de la $mesure_{SCM}$ est supérieure ou égale à 0.68 (seuil établi empiriquement) :

$$mesure_{SCM}(mot_1, mot_2) = \frac{longueur(sous\_chaîne\_commune(mot_1, mot_2))}{\max(longueur(mot_1), longueur(mot_2))}$$

- le calcul du coefficient de Dice ; Les mots sont retenus comme cognats si le coefficient est supérieur ou égale à 0,62 (valeur établie empiriquement pour notre corpus) :

$$Dice(mot_1, mot_2) = \frac{2 * nombre\_bigrammes\_communs}{nombre\_total\_bigrammes(mot_1) + nombre\_total\_bigrammes(mot_2)}$$

- 4-grammes ; Les mots sont considérés comme cognats s'ils comprennent au moins 4 caractères et si leurs premiers 4 caractères sont identiques.

Nous avons implémenté ces méthodes en utilisant le corpus contenant l'ensemble d'ajustements orthographiques effectués au préalable au niveau des lemmes (cf. Figure 2). Ces méthodes s'appliquent généralement pour des mots ayant une longueur égale ou supérieure à 4 pour réduire le bruit. Les paires de cognats sont recherchées dans des phrases parallèles alignées. Les sous-chaines recherchées doivent être quasiment parallèles (***rembourser*** vs. ***rambursare)***.

Méthodes	Précision	Rappel	F-mesure
SCM + Ajustements	**44,13%**	58,95%	**50,47%**
DICE + Ajustements	56.47%	60.91%	58.61%
4-grammes - Sans ajustements	90,85%	**47,84%**	62,68%
4-grammes + Ajustements	91,55%	72,42%	80,87%
Notre méthode	**94,78%**	**89,18%**	**91,89%**

Tableau 1 Évaluation du module développé et comparaison avec d'autres méthodes

Nous avons évalué notre module d'identification de cognats en termes de précision, rappel et performance, par rapport à une liste de référence constituée manuellement à partir du corpus parallèle décrit dans la sous-section 2.1. Cette liste contient 2 034 cognats français - roumain.

Notre méthode a extrait 1 814 cognats corrects sur 1 914 candidats cognats fournis, obtenant ainsi les meilleurs scores (précision=94,78% ; rappel=89,18% ; f-mesure=91,89%), par rapport aux autres méthodes implémentées. La méthode 4-grammes appliquée sur le corpus initial (sans ajustements orthographiques) a obtenu une bonne précision (90,85%), mais un rappel faible de 47,84%. L'étape d'ajustements orthographiques au niveau des lemmes améliore nettement le rappel de la méthode 4-grammes. Ceci s'explique par les spécificités du corpus juridique utilisé où les termes provenant du français vers le roumain

par des emprunts sont nombreux. Les scores les plus faibles sont obtenus par la *mesure*$_{SCM}$ (f-mesure=50,47%), suivie par la méthode basée sur le coefficient de Dice (f-mesure=58,61%). Ces approches générales produisent beaucoup de bruit dû aux ressemblances formelles importantes entre des mots qui ne présentent aucun lien au niveau sémantique. Leurs résultats pourraient être améliorés en combinant des techniques statistiques avec d'autres informations : équivalences de catégorie lexicale, transfuges.

4 Conclusions et perspectives

Nous avons présenté une méthode d'identification de cognats français - roumain. Cette méthode combine des techniques statistiques et des filtres linguistiques à partir de corpus parallèles juridiques lemmatisés, étiquetés et alignés au niveau propositionnel. Notre méthode fournit des résultats performants par rapport à des méthodes statistiques pures. Cependant, les résultats dépendent de la paire de langues étudiées, du domaine du corpus utilisé et aussi du volume des données. Des études similaires restent encore nécessaires sur d'autres types de corpus parallèles français - roumain, afin de pouvoir généraliser. Toutefois, notre méthode s'avère efficace pour l'identification de cognats à partir des corpus juridiques. À l'avenir, cette méthode sera intégrée dans un système d'alignement lexical français - roumain exploitant des corpus parallèles juridiques.

Références

ADAMSON G., BOREHMAN J. (1974). The use of an association measure based on character structure to identify semantically related pairs of words and document titles. *Information Storage and Retrieval*, 10(7-8), 253-260.

BREW C., MCKELVIE D. (1996). Word-pair extraction for lexicography. Proceedings of *International Conference on New Methods in Natural Language Processing*, Bilkent, Turkey, 45-55.

IDE N., VERONIS J. (1994). Multext (multilingual tools and corpora). Proceedings of *the 15th International Conference on Computational Linguistics, CoLing 1994*, Kyoto, August 5-9, pp. 90-96.

ION R. (2007). *Word Sense Disambiguation Methods Applied to English and Romanian*. PhD thesis (in Romanian). Romanian Academy, Bucharest, May 2007, 148 p.

KONDRAK G. (2009). Identification of Cognates and Recurrent Sound Correspondences in Word Lists. *Traitement Automatique des Langues (TAL)*, 50(2), 201-235.

KRAIF O. (1999). Identification des cognats et alignement bi-textuel : une étude empirique. Actes de *la 6ème conférence annuelle sur le Traitement Automatique des Langues Naturelles, TALN 99*, Cargèse, 12-17 juillet 1999, 205-214.

MELAMED D. (1999). Bitext Maps and Alignment via Pattern Recognition. *Computational Linguistics*, 25(1), 107-130.

OAKES M. (2000). Computer Estimation of Vocabulary in Protolanguage from Word Lists in Four Daughter Languages. *Journal of Quantitative Linguistics*, 7(3), 233-243.

SIMARD M., FOSTER G., ISABELLE P. (1992). Using cognates to align sentences. Proceedings of *the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation*, Montréal, pp. 67-81.

TUFIŞ D., BARBU A. M. (1997). *A Reversible and Reusable Morpho-Lexical Description of Romanian*. In Dan Tufiş and Poul Andersen (eds.), *Recent Advances in Romanian Language Technology*, pp. 83-93, Romanian Academy Publishing House, Bucharest, 1997. ISBN 973-27-0626-0.

WAGNER R., FISCHER M. (1974). The String-to-String Correction Problem. *Journal of the ACM*, 21(1), 168-173.