Apport des traitements morpho-syntaxiques pour l'alignement des définitions par une classification SVM

Laura Dioşan[*,**], Alexandrina Rogozan[*], Jean-Pierre Pécuchet[*]

[*]LITIS (EA 4108) - INSA Rouen, France
[**]Babeş Bolyai University, Computer Science Department, Cluj Napoca, Romania
lauras@cs.ubbcluj.ro, arogozan@insa-rouen.fr, pecuchet@insa-rouen.fr

Résumé. Cet article propose une méthode d'alignement automatique de définitions destinée à améliorer la fusion entre des terminologies spécialisées et un vocabulaire médical généraliste par un classifieur de type SVM (Support Vecteur Machine) et une représentation compacte et pertinente d'un couple de définitions par concaténation d'un ensemble de mesures de similarité, afin de tenir compte de leur complémentarité, auquelle nous ajoutons les longueurs de chacune des définitions. Trois niveaux syntaxiques ont été investigués. Le modèle fondé sur un apprentissage à partir des groupes nominaux de type *Noms-Adjectifs* aboutit aux meilleures performances.

Les systèmes de recherche d'informations reposent sur une terminologie spécifique d'un domaine d'application que seuls les experts possèdent. En effet, les utilisateurs naïfs utilisent un langage généraliste pour formuler leurs requêtes. Pour qu'un système de recherche puisse répondre efficacement aux requêtes de ces derniers, il devrait pouvoir tirer parti des liens sémantiques entre des concepts véhiculés dans le langage généraliste et dans le langage spécialisé. Une des tâches du projet *VODEL* est de réaliser un alignement automatique de définitions, c'est-à-dire de mettre en correspondance des définitions associées à un même concept, mais ayant des vedettes différentes. Le cadre choisi étant celui du domaine médical, les ressources terminologiques de spécialité sont tirées du thésaurus *MeSH* et du dictionnaire *VIDAL*, alors que le vocabulaire généraliste est représenté par des définitions appartenant à l'encyclopédie *Wikipédia* et au réseau sémantique LDI de *Memodata*[1].

Aligner deux définitions revient à résoudre efficacement un problème de classification binaire supervisée. Notre modèle d'alignement passe par deux étapes : premièrement, une représentation compacte des définitions et deuxièmement, une classification supervisée de couples de définitions. Chaque définition a été représentée par un sac des mots, après un traitement linguistique (segmentation, lemmatisation et étiquetage morpho-syntaxique) permettant de filtrer les mots vides et de ne garder que les noms (*N*), les noms et les adjectifs (*NA*), et respectivement les noms, les adjectifs et les verbes (*NAV*). Nous proposons une représentation compacte et pertinente d'un couple de définitions par concaténation d'un ensemble de mesures de similarité classiques (Matching, Dice, Jaccard, Overlap, Cosine), afin de tenir compte de leur complémentarité, auquel nous ajoutons les longueurs de chacune des définitions. Nous proposons un alignement des terminologies par un classifieur de type SVM (Séparateur à Vaste

[1]Le corpus de définitions a été réalisé dans le cadre du projet **VODEL** par G. Lortal, I. Bou Salem et M. Wang.

Marge) Vapnik (1995) à noyau RBF (Radial Basis Fonction) qui est à la base des classifieurs les plus performants. Le hyper-paramètre C représentant la pénalité de l'erreur de classification du *SVM* et le hyper-paramètre représentant la bande du noyau RBF ont été optimisés par une technique de validation croisée. Nous adaptons ainsi automatiquement le classifieur *SVM* au problème donné, à savoir l'alignement de définitions.

Nous avons mis en oeuvre deux expérimentations à partir de six séries de couples de définitions. Pour la $1^{ère}$ expérimentation, l'apprentissage et le test sont réalisés sur des ensembles disjoints tirés d'une même série des couples de définitions. Les meilleurs résultats (cf. Tableau 1) sont obtenus pour l'alignement du thésaurus de spécialité *VIDAL* avec le dictionnaire généraliste *Wikipédia* grâce aux modèles *N* et *NA*, ainsi que pour l'alignement des dictionnaires généralistes *Memo vs. Wiki* avec le modèle *NAV*. Le but de la $2^{ème}$ expérience est d'étudier ce qui se produit lorsque nous apprenons l'alignement sur trois séries (*Memo-MeSH*, *Memo-Vidal*, *MeSH-Vidal*) et nous testons l'alignement sur des couples des définitions issues des autres trois séries : *Memo-Wiki*, *MeSH-Wiki*, *Vidal-Wiki* (voir Tableau 1). Les résultats obtenus dans la $2^{ème}$ expérimentation sont statistiquement meilleurs que les résultats obtenus dans la $1^{ère}$ expérience puisque les intervalles de confiance ne se chevauchent pas. Cette amélioration provient du fait que la taille de l'ensemble d'entraînement est 3 fois plus grande que celle utilisée dans la $1^{ère}$ expérience et peut-être parce que nous avons entraîné le SVM sur un vocabulaire généraliste *Memo* et le *VIDAL*) contre une seule terminologie de spécialité (*MeSH*).

	Memo vs. MeSH	*Memo vs. Vidal*	*Memo vs. Wiki*	*MeSH vs. Vidal*	*MeSH vs. Wiki*	*Vidal vs. Wiki*	*Memo, MeSH, Vidal vs. Wiki*
N	81±3.15	82±3.09	84±2.95	81±3.15	80±3.21	87±2.70	98±0.69
NA	77±3.38	74±3.52	86±2.79	80±3.21	85±2.87	88±2.61	98±0.69
NAV	77±3.38	79±3.27	85±2.87	80±3.21	80±3.21	84±2.95	98±0.69

TAB. 1 – *Les F-measures et leur intervalles de confiance pour le système d'alignement.*

En conclusion, l'alignement automatique des définitions est tout à fait réalisable avec un modèle fondé sur un apprentissage SVM avec optimisation des hyper paramètres. L'information véhiculée par les Noms et les Adjectifs semblerait plus pertinente que celle provenant des Verbes. Cependant, ces conclusions doivent aussi être vérifiées sur des corpus de taille plus importante et d'autres méthodes seront testées.

Références

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. Springer.

Summary

A method of automatically alignment of definitions is proposed in order to improve the fusion between specialized medical terminologies and a general one. An SVM classifier and a compact representation are used. The model trained on the nominal group of Noun-Adjectives reaches the best performances.