Étiquetage morpho-syntaxique par classification supervisée : vers une alternative aux dictionnaires ?

Frédérick HOUBEN, François RIOULT

GREYC – UMR 6072 – Université de Caen
campus II – BP 5186
F-14032 CAEN cedex
Frederick.Houben@info.unicaen.fr
Francois.Rioult@info.unicaen.fr

Abstract

We are presenting a NLP multilingual method for low cost means for minority languages creation. This method uses very general linguistic properties, accessible from the raw text, and also engineering from data mining community, notably in supervised learning. Our first object is to validate those properties relevance through a good second tagging.

Keywords: Multilingual NLP, data mining, minority languages, means creating, tagging.

Résumé

Nous présentons une méthode multilingue permettant, entre autres, de créer des ressources à moindre coût pour des langues peu dotées. Cette méthode fait appel à des propriétés très générales des langues, accessibles depuis le texte brut, ainsi qu'à des méthodes issues de la communauté de la fouille de données, notamment en apprentissage supervisé. L'objectif est essentiellement de valider la pertinence de ces propriétés à travers un ré-étiquetage réussi.

Mots-clés : Traitements multilingues, fouille de données, langues peu dotées, création de ressources, étiquetage morpho-syntaxique.

1. Introduction

Dans cet article, nous nous attacherons surtout à essayer de décrire les éléments d'une méthode de création de ressource pour des langues peu dotées à moindre coût. Cette méthode est en cours de développement et les résultats dont nous disposons ne sont qu'embryonnaires. Elle fait appel à des principes et outils issus de la communauté de la fouille de données et utilise des connaissances très générales sur les langues, connaissances qui doivent obligatoirement être accessibles directement depuis le texte brut même si, dans un premier temps, nous nous orientons vers un système d'apprentissage supervisé nécessitant l'utilisation d'un corpus étiqueté (de petite taille, 5000 mots suffisent).

Nous nous sommes fixé comme application-cadre l'étiquetage d'un corpus d'une langue quelconque (alphabétique, non agglutinante) à partir d'un petit corpus étiqueté de cette même langue. Nous pensons qu'après une phase d'apprentissage portant sur des propriétés facilement accessibles du corpus étiqueté, il est possible de généraliser l'étiquetage à l'ensemble du corpus. Ce qui nous intéresse réellement dans cette tâche n'est pas tant la

réussite de l'étiquetage que la validation de l'hypothèse selon laquelle nous pouvons catégoriser les mots à partir de propriétés accessibles depuis le corpus brut. En cas de réussite dans cette tâche, nous pourrons alors tenter de réaliser une catégorisation lexicale sans aucune ressource, et donc sans corpus étiqueté. Dans la suite, nous appellerons aussi bien ces propriétés les attributs des mots. Il s'agit simplement de nommer les éléments constitutifs du contexte du mot qui permettront, nous l'espérons, de catégoriser correctement ce mot.

Contrairement à un étiqueteur de Brill (Brill, 1992), nous n'utilisons pas de lexique, ressource indispensable dans sa démarche, obtenue à partir d'un corpus étiqueté. Notre démarche constitue un premier pas vers une méthode sans aucune autre ressource que le seul texte brut.

Pour les mêmes raisons, nous ne souhaitons pas non plus utiliser le corpus étiqueté comme le font (Debili, 1977), (Church, 1993) ou encore (Merialdo, 1994) afin d'extraire des règles de fréquence de contiguïté de tags même si nous nous intéressons nous aussi aux rapports qui peuvent exister entre des mots consécutifs.

Enfin, nous ne souhaitons pas utiliser, comme le fait (Vergne, 1998), des règles symboliques fournies manuellement au système même si nous nous reconnaissons totalement dans son envie de minimiser les ressources nécessaires aux traitements.

Nous allons donc commencer par donné un aperçu général de la méthode puis détailler les attributs actuellement utilisés avant de nous intéresser aux outils d'apprentissage dont nous allons nous servir, introduisant notamment ce qu'est la classification supervisée, la fouille de données orientée motifs puis notre méthode actuelle. Nous présenterons enfin quelques résultats en essayant de les analyser.

2. Aperçu général de la démarche

Notre démarche se scinde en trois étapes :

1. Détermination d'un certain nombre d'attributs des mots, accessibles depuis le texte brut ;

2. Classification supervisée à partir de ces attributs. On cherche des régularités, des règles d'associations de ces attributs qui nous permettrons de déterminer la catégorie lexicale du mot ;

3. Évaluation automatique des résultats par cross-validation : on cherche à vérifier que les étiquettes attribuées à l'étape précédente correspondent aux étiquettes du corpus initial étiqueté.

A l'issue de ces trois étapes, nous pourrons alors estimer, s'il y a un bon taux de réussite, que les attributs sur lesquels nous avons travaillé permettent effectivement de discriminer les mots en fonction de rôle syntaxique. Nous pourrons alors passer à la dernière étape qui consiste en un regroupement des mots d'un corpus brut en classe d'équivalence de mots de même tag. Ce processus devient un tagging sans aucune ressource.

3. Attributs utilisés

Avant de commencer, nous souhaitons préciser que nous utilisons le terme de mot en tant que « une graphie donnée à une position particulière du corpus ». Ainsi, lorsque la même graphie se retrouve à deux endroits différents du corpus, nous parlerons de deux mots mais d'une seule graphie.

Nous nous sommes donc intéressés au contexte des mots, cherchant à trouver dans ces contextes un certain nombre de propriétés qui nous permettraient de réaliser l'apprentissage dont nous avons besoin et de discriminer correctement les différents types de mots.

Ces propriétés doivent nécessairement être accessible directement depuis le texte lui-même, sans intervention supplémentaire du locuteur, afin d'automatiser la tâche autant que possible.

Nous avons établi une première liste de cinq attributs des mots que nous allons préciser ci-dessous. Il faut cependant noter que ces attributs sont locaux, propres au mot, ce qui signifie que pour une graphie donnée, il est tout à fait possible que les propriétés qui lui sont associées ne soient pas les mêmes du fait du contexte possiblement différent de chacun des mots.

Nous allons maintenant détailler la liste de ces cinq attributs :

- Nous nous sommes d'abord penchés sur un problème déjà abordé dans (Houben, 2004) : est-ce que le mot est vide ou plein. Nous utilisons ici vide et plein dans un sens proche de (Tesnière, 1969) et considérons comme mots vides l'ensemble des mots grammaticaux ainsi que les auxiliaires. Les mots pleins sont donc tous les autres. La catégorisation en mot vide ou mot plein se fait de manière automatique en appliquant le principe de Saussure (Saussure, 1922) selon lequel « dans la langue, il n'y a que des différences » ainsi que le « principe du moindre effort » de (Zipf, 1949). Notons que cet attribut peut avoir deux autres valeurs : ponctuation et non déterminé pour les mots dont nous n'avons pas réussi à calculer s'ils étaient vides ou pleins et pour lesquels nous avons préféré ne pas faire de choix, cette incertitude étant en soi une information sur le contexte.

- Nous avons aussi souhaité conserver le type (avec les mêmes valeurs possibles que ci-dessus) du mot précédent et du mot suivant, faisant l'hypothèse que les mots vides et les mots pleins ne se succèdent pas n'importe comment. Ainsi, il y a peu de chances de trouver, en français, un déterminant avant une préposition.

- Le quatrième attribut est la position du mot dans le virgulot ((Lucas, 2003) portion de la phrase entre deux ponctuations). Cette propriété est le résultat d'une observation simple : en français, on trouve régulièrement les groupes nominaux en fin de virgulot et les groupes verbaux en début de virgulot. Nous faisons l'hypothèse que cet attribut, dans son principe, n'est pas lié qu'au français et que dans les autres langues que nous étudierons les groupes nominaux et verbaux ont une position préférentielle (hypothèse vérifiée pour un certain nombre d'autres langues).

- Enfin le dernier des attributs concerne l'influence qu'un mot vide peut avoir sur les terminaisons des mots suivants. Cette propriété n'est pas purement locale. Nous calculons, sur l'ensemble du corpus, toutes les terminaisons des mots pleins suivant immédiatement une graphie donnée. Nous estimons que le mot peut influencer ses suivants à partir du moment où la terminaison la plus fréquente qui lui est associée représente au moins la moitié des terminaisons possibles (avec tout de même un minimum d'occurrences requis). A partir de là, nous examinons localement, pour chacun des mots, si la terminaison qui suit fait partie des terminaisons répétées (sans minimum) qui lui sont associées sur l'ensemble du corpus, auquel cas nous marquons que le mot influence son suivant. Dans tout autre cas, nous indiquons qu'il n'a pas cette influence.

  Par exemple, si *les* est suivi 40 fois d'un mot terminant par *-s*, 3 fois d'un *-x* et une fois d'un -e, nous constatons qu'il peut influer sur les terminaisons des mots suivants (-s

ultra majoritaire). Nous estimons alors qu'il influe effectivement sur les mots finissant par un *-s*, mais aussi sur ceux en *-x* et pas sur celui terminant avec un *-e*.

La figure suivante présente une partie du fichier une fois que nous avons extrait les cinq attributs sur l'ensemble du corpus. Cet extrait correspond au morceau de phrase « `d' ar mab henañ` », issu de notre corpus breton.

Catégorie du mot	Type du mot	Type du précédent	Type du suivant	Position dans le virgulot	Influence sur le suivant
p	v	P	v	m	o
d	v	v	n	f	n
N	n	v	P	f	n
E	P	n	s	f	n

Figure 1 : Tableau des attributs extraits du texte brut (la première colonne de notre figure contient la catégorie du mot - p pour préposition, d pour déterminant, N pour Nom , E pour Adjectif -, résultat d'une expertise humaine sur le corpus.)

La graphie du mot n'apparaît plus. Nous avons décidé de l'ignorer. Au mieux, nous estimons qu'elle ne sert à rien dans notre cadre et qu'elle n'est en aucun cas discriminante quant à la catégorie du mot en vue d'un étiquetage. Au pire, elle risque de générer un sur-apprentissage du fait d'une propriété trop « variée ». Nous n'excluons toutefois pas d'ajouter un attribut qui sera la terminaison du mot plein, celle-ci étant discriminante dans bien des langues. Par exemple, en anglais, le -ed ou le -ing sont significatifs du verbe, -ly marquant plutôt l'adverbe.

4. Quels outils pour l'apprentissage ?

Nous décrivons brièvement dans cette section la méthode utilisée pour évaluer la pertinence des attributs pour la classification supervisée.

4.1. Classification supervisée

Nous disposons d'une base de données qui décrit les propriétés des mots du corpus, parmi lesquelles figure une valeur de classe (préposition, déterminant, nom, etc.) attribuée par une expertise externe : le problème est dit *supervisé*. Dans notre cas, l'expertise fournie provient de l'analyseur de Jacques Vergne (Vergne, 1998). La *classification* est une méthode automatique qui propose une valeur de classe pour des exemples inconnus.

Pour valider la pertinence des attributs, la base de données est divisée en deux parties :

- une base d'apprentissage, qui fournit des connaissances utiles pour réaliser la classification ;
- une base de test constituée d'exemples pour lesquels nous cherchons à déterminer automatiquement une valeur de classe, à laquelle nous comparons la valeur de référence, fournie par l'expertise externe.

4.2. Fouille de données orientée motifs

a

priori

r *A R O* *A a …a*$_n$

O o …o$_m$ *R*
A O R

o *A* *motif X*
A *présent* *o* *o* *X*
fréquence *X* *fréquent*
F X γ

Z *Z* *X→Y* *X⊂Z*
Y Z X X *Y* *X→Y* *conf (X → Y)*
X *Y* *i.e.*

$$conf(X \rightarrow Y) = \frac{F(X \cup Y)}{F(X)}$$

Y *X*

χ χ

4.3. Notre méthode

Pour nos expériences, nous avons implémenté une méthode proche de CMAR, capable de prendre en charge les règles *généralisées positives* (qui, en concluant sur un attribut de classe, entérinent la possibilité que l'exemple à classer appartienne à cette classe, si elle coïncide avec la prémisse) et les règles *généralisées négatives* (qui, en concluant sur la négation d'un attribut de classe, excluent la possibilité de classe correspondante) (Antonie, Zaïane, 2004). Les règles *généralisées* étendent le principe des règles d'association en autorisant des négations d'attributs dans la prémisse, et sont obtenues à l'aide des motifs *k*-libres (Calders, Goethals, 2003). Les règles positives sont de la forme $X\overline{Y} \rightarrow c_i$, les règles négatives de la forme $X\overline{Y} \rightarrow \overline{c}_i$, où c_i est un attribut de classe. Ces règles s'appliquent pour classer tout exemple qui contient le motif *X*, mais aucun des attributs de *Y*.

Selon le modèle de CMAR, les règles sont pondérées par le χ^2 relatif. Pour un nouvel exemple, les règles positives voient leur pondération s'ajouter au score, les règles négatives soustraient leur pondération. Au final, la classe avec le meilleur score est désignée.

5. Evaluation des résultats

5.1. Des résultats encourageants …

… mais insuffisants. Nous avons actuellement un taux moyen de 34 % de précision sur le français lors de la classification automatique. C'est déjà mieux que le hasard[1], ce qui semble montrer qu'il y a effectivement moyen de se servir de ces attributs, même si ce n'est pas encore suffisant, loin s'en faut, pour valider la méthode.

Des tests ont été effectués sur un corpus breton. Nous obtenons alors 29,40 % de réussite. Ces résultats sont à relativiser du fait de notre totale méconnaissance de cette langue et d'une projection du jeu d'étiquettes original pas forcément idéale.

À titre d'information, nous signalons que nos premiers tests ont donné 10 % de réussite alors que le deuxième et le troisième attribut (type du mot précédent / suivant) étaient regroupés en un seul attribut un peu moins informatif (nous nous contentions de vérifier que le mot était ou pas entouré de mots du même type). On s'aperçoit donc que les performances augmentent très vite grâce au simple ajout d'un attribut. Or, nous ne voyons pas de raison pour limiter le nombre d'attributs tant que chacun d'eux sera calculable sur le texte brut, sans apport de la moindre ressource extérieure, si de tels raffinements permettent l'amélioration des résultats.

5.2. Exemples de règles

Nous donnons ici des exemples avec les deux types de règles (positives et négatives).

D'abord une règle positive (rappelons qu'une règle positive nous permet d'affirmer que si les prémisses sont vérifiées alors on peut conclure sur une valeur de classe) :

type:v ∧ avant:s ∧ après:n ∧ influence:o → classe = pronom personnel

Cette règle se lit de la manière suivante : si le type du mot étudié est vide, qu'il est précédé d'un signe de ponctuation, qu'il est suivi d'un mot de type indéterminé et qu'il influe sur les terminaisons alors on peut conclure que le mot est de la classe pronom personnel. Il faut noter

[1] Nous avons 14 classes différentes soit une probabilité de 7,1 % par classe.

ici que cette règle n'est pas la seule qui nous permette de retrouver les pronoms personnels. Globalement, nous disposons toujours de plusieurs règles qui nous amènent à conclure sur une classe donnée.

Ensuite, un exemple de règle négative (à l'inverse des règles positives, si les prémisses sont vérifiées, on peut conclure sur l'impossibilité de la valeur de classe correspondante) :

Type : P $\wedge$ $\neg$ après:P $\rightarrow$ $\neg$ classe = déterminant

Cette règle indique que si le mot courant est de type plein et que le mot suivant n'est pas de type plein, alors le mot courant ne peut pas être un déterminant. En fait, cette règle est naturelle : un déterminant est un mot vide par définition et il est suivi d'un adjectif ou d'un nom, c'est-à-dire de mots pleins, dans la plupart des cas. Cette règle, parmi d'autres, est donc parfaitement en accord avec les connaissances que nous avons du domaine et confirme que notre démarche est cohérente.

Notons que quelque soit le type de règle dont on dispose, elles sont en concurrence les unes avec les autres. Cela signifie qu'une règle peut avoir une incidence faible par rapport à une autre règle de plus grande confiance.

Pour finir, nous voulons préciser que les 34 % de réussite ont été obtenu grâce à la combinaison de ces deux types de règles. Si nous n'avions utilisé que des règles positives, nous aurions eu 17 % de réussite et avec les seules règles négatives, 23 %. Il s'agit donc ici d'un problème pour lequel la combinaison des deux types est impérative.

5.3. Utiliser les matrices de confusions

Ces matrices sont obtenues lors de la phase de comparaison des étiquettes calculées avec les étiquettes fournies dans le corpus étiqueté. Elles nous donnent deux types d'information :

- la quantité de mots d'un type donné qui ont été correctement rattachés à ce type ;
- quels sont les mots qui ont été confondus avec un type donné.

Voici par exemple une des matrices obtenues lors de nos tests :

	Po	Adj	Inf	ProP	Nom	ProR	PPr	Ver	am	Det	neg	prep	Adv	Ppa
Po	33	0	0	0	0	0	0	0	0	0	0	0	0	0
Adj	0	0	0	0	15	0	0	0	0	0	0	1	2	0
Inf	0	0	0	0	9	1	0	0	0	0	1	0	0	0
ProP	0	0	1	2	0	1	0	0	0	5	0	2	0	0
Nom	0	1	0	1	56	0	0	0	0	1	1	4	0	2
ProR	3	0	1	2	1	7	0	1	0	3	0	0	0	0
PPr	0	0	0	0	0	0	0	0	0	0	0	0	1	0
Ver	0	0	0	3	20	3	0	0	0	0	0	1	1	0
am	1	1	0	0	3	5	0	1	0	0	0	0	0	0
Det	4	2	1	9	7	7	0	0	0	6	0	0	0	0
neg	0	0	0	3	2	1	0	0	0	0	0	0	0	0
prep	3	0	0	16	8	15	0	0	1	2	0	2	0	2
Adv	1	0	0	1	11	0	0	2	0	0	0	1	0	0
Ppa	0	0	0	1	7	0	0	0	0	0	0	0	0	1

Figure 2 : Une matrice de confusion. En ligne, la classe réelle du mot, en colonne l'étiquette calculée, sur la diagonale, les mots correctement étiquetés.

On voit, sur cette matrice, qu'un grand nombre de mots d'autres types sont étiquetés comme des noms communs. Suite à cela, l'analyse des règles nous permet d'essayer de comprendre les raisons de cette confusion.

On voit aussi, et c'est essentiel pour comprendre une partie des problèmes rencontrés, que les classes ne sont pas du tout équilibrées, qu'il y a beaucoup plus de noms communs que de verbes, et plus de verbe que de pronoms…

5.4. Analyse de ces résultats

Il y a deux grandes raisons au manque de précision de ces résultats : un modèle linguistique insuffisant et un outil de fouille de donnée inadapté.

Dans le premier cas, il nous faudra trouver d'autres attributs qui permettront de mieux discriminer, nous pensons notamment utiliser la terminaison des mots, et regarder aussi la place du mot dans la phrase entière et non plus simplement dans le virgulot.

Quant à l'outil de fouille de données, il présente l'inconvénient de son avantage : il était immédiatement disponible et utilisable, mais il est fait pour des problèmes qui sont autres : ainsi, la disparité de volume des classes s'avère gênante, de même que l'obligation dans laquelle nous nous trouvons de travailler sur tous les attributs simultanément.

Pour parer à cette simultanéité, nous avons envisagé de ne fournir à l'outil que des données partielles, comme par exemple la liste des mots pleins uniquement. Cela nous permettait de simuler la séquentialité dans l'étude des attributs en forçant une séparation initiale en vide / plein … puis ensuite, en regardant les autres attributs simultanément. Cette pratique à effectivement un peu améliorer nos résultats mais nous nous posons alors la question de savoir si, de cette manière, nous n'utilisons pas plutôt nos propres connaissances que celles apprises par l'outil et surtout si c'est connaissances ne limites le nombre de langues auxquels la méthode sera applicable.

Notons enfin que nous regardons aussi d'autres méthodes de classification. Par exemple, un rapide test avec des arbres de décisions nous a donné un taux de réussite de 45 % avec des points réussis et des échecs très différents de notre méthode actuelle. Pour rester plus proche de notre démarche, nous avons utilisé un classifieur ensembliste, nous pourrions nous pencher sur des méthodes à base de motifs séquentiels qui ont la particularité de prendre en compte la séquentialité inhérente à l'organisation des mots dans un texte.

6. Conclusion

Nous avons posé les bases d'une méthode de validation des propriétés des mots qui présentera l'énorme qualité de ne pas nécessiter de ressources. Cette méthode semble donc particulièrement avantageuse pour des langues peu dotées. Elle permettra non seulement de travailler sans avoir besoin de ressources mais aussi, éventuellement, de créer les ressources qui manquent à moindre coût.

À partir d'un travail sur des propriétés des mots accessibles depuis le corpus brut, nous espérons pouvoir catégoriser correctement ces mots et utiliser nos résultats, soit directement pour faire un étiquetage morpho-syntaxique du corpus, soit pour la construction de ressources utilisables dans d'autres applications.

Nous n'en sommes bien sûr qu'aux prémisses de ce travail mais avons dégagé suffisamment de points encourageants et intéressants pour estimer qu'il y a là une piste en vue d'obtenir une solution originale et efficace pour les problèmes des langues sous-équipées informatiquement.

Références

Agrawal R., Mannila H., Srikant R., Toivonen H., Verkamo A. (1996). Fast discovery of association rules. In *Advances in Knowledge Discovery and Data Mining*.

Agrawal R., Srikant R. (1994). Fast algorithms for mining association rules. In *Intl. Conference on Very Large Data Bases (VLDB'94), Santiago de Chile*.

Antonie M.-L., Zaïane O .R. (2004). An associative classifier based on positive and negative rules. In *9th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD)*.

Antonie M.-L., Zaïane O .R. (2004). Mining positive and negative association rules: An approach for confined rules. In *PKDD*.

Bernard G. (2003). Détection automatique de structures syntaxiques. In *Proceedings of the 8th International Symposium on Social Communication*

Brill E. (1992), A simple rule-based part of speech tagger, In *Proceedings of the Third Conference on Applied Natural Language Processing*, Trento, Italy.

Brill E. (1994). Some advances in transformation-based part of speech tagging. *Proceedings of the twelfth national conference on Artificial intelligence*, Vol. 1 : 722-727.

Calders T., Goethals B. (2003). Minimal k-free representations of frequent sets. In *Proceedings of PKDD'03*.

Church K., Mercer R. (1993). Introduction to the special issue on using large corpora; *Computational Linguistics, vol. 19, n°1, special issue on using large corpora : I :* 1-24.

Dominic F., Meunier J. G. (2004). Classification et catégorisation automatiques : application à l'analyse thématique des données textuelles, JADT 2004.

Houben F. (2004) Mot vide, mot plein ? Comment trancher localement, Actes de *RECITAL 2004* : 61-66.

Li W., Han J., Pei J. (2001). Cmar : Accurate and efficient classification based on multiple class-association rules. In *proceedings of the IEEE International Conference on Data Mining, ICDM 01*, San Jose, California.

Liu B., Hsu W., Ma Y. (1998). Integrating classification and association rules mining. In *proceedings of Fourth International Conference on Knowledge Discovery and Data Mining, KDD 98*, AAAI Press : 80-86.

Lucas N., Turmel L., Crémilleux B. (2003). Extraction d'associations pour la caractérisation de segments de textes en anglais avec et sans faute. *Conférence internationale sur le document électronique Cide 6*.

Mariage J.-J., Bernard G. (2004). Catégorisation de patrons syntaxiques par Self Organizing Maps, Actes de *TALN 2004*.

Vergne J., Giguet E. (1998). Regards Théoriques sur le "Tagging". Actes de *TALN 1998* : 22-31.

Zipf G. K. (1949). *Human Behaviour and the Principle of Least Effort*, New York, Harper.