Etiquetage morpho-syntaxique du français à base d'apprentissage supervisé

Julien Bourdaillet, Jean-Gabriel Ganascia
LIP6 - Univesité Paris VI
8 rue du capitaine Scott - 75015 Paris
{Julien.Bourdaillet, Jean-Gabriel.Ganascia}@lip6.fr

Mots-clefs : étiquetage morpho-syntaxique, apprentissage supervisé, modèle de Markov caché, évaluation, homographes

Keywords: part-of-speech tagging, supervised learning, hidden Markov model, evaluation, homographs

Résumé Nous présentons un étiqueteur morpho-syntaxique du français. Celui-ci utilise l'apprentissage supervisé à travers un modèle de Markov caché. Le modèle de langage est appris à partir d'un corpus étiqueté. Nous décrivons son fonctionnement et la méthode d'apprentissage. L'étiqueteur atteint un score de précision de 89 % avec un jeu d'étiquettes très riche. Nous présentons ensuite des résultats détaillés pour chaque classe grammaticale et étudions en particulier la reconnaissance des homographes.

Abstract A french part-of-speech tagger is described. It is based on supervised learning: hidden Markov model and trained using a corpus of tagged text. We describe the way the model is learnt. A 89 % precision rate is achieved with a rich tagset. Detailed results are presented for each grammatical class. We specially pay attention to homographs recognition.

1 Introduction

L'étiquetage morpho-syntaxique consiste à assigner la bonne classe grammaticale, définie suivant un certain niveau de granularité, à chaque mot d'un texte en entrée. Nous présentons ici un étiqueteur (ou tagger) du français basé sur un modèle d'apprentissage supervisé. Celui-ci est une adaptation du tagger de l'analyseur syntaxique RASP de la langue anglaise. Cet étiqueteur apprend un modèle de langage à partir d'un corpus préalablement étiqueté.

Plusieurs approches ont été présentées pour l'étiquetage morpho-syntaxique du français. Le Brill Tagger (Brill, 92) apprend des règles à partir d'un corpus étiqueté et a été adapté pour le français avec WinBrill. (Giguet, 97) et (Chanod, 95) se basent sur des propriétés de la langue comme les mots noyaux ou des méthodes à base de contraintes. (Chanod, 95) présente un autre étiqueteur à base d'apprentissage non-supervisé, qui apprend un modèle de langage à partir d'un corpus non-étiqueté via une variante de l'algorithme Estimation-Maximisation (EM).

(Chanod, 95) présente les limites de ce modèle qui est fortement dépendant des conditions initiales et peut rester bloqué sur un optimum local. (Stein, 95) présente une adapatation au français du TreeTagger. Celui-ci est basé sur un modèle de Markov caché (Hidden Markov Model ou HMM) modélisé par un arbre de décisions. Nous nous situons dans la lignée de ces deux derniers travaux et utilisons également un HMM. Toutefois puisqu'il a été prouvé dans (Elworthy, 94) que l'apprentissage d'un modèle à partir d'un corpus manuellement étiqueté produit de meilleurs résultats que la procédure d'apprentissage d'EM et que nous disposions d'un tel corpus, nous avons choisi cette alternative.

2 Présentation de l'étiqueteur

Nous utilisons l'étiqueteur morpho-syntaxique de l'anglais du projet RASP (Briscoe, 02). Cet étiqueteur, détaillé dans (Elworthy, 94), est basé sur HMM du premier ordre (bigramme). Celui-ci est représenté par un lexique des formes fléchies qui associe à chaque mot ses tags potentiels et par une matrice de transition entre états. Nous l'avons adapté au traitement de la langue française.

Nous avons utilisé le corpus GRACE qui est annoté morpho-syntaxiquement. Il comporte environ 800.000 mots et est constitué pour moitié d'articles du Monde et pour moitié d'oeuvres et d'essais littéraires. Ce corpus est étiqueté avec un jeu de 312 étiquettes qui correspondent à un étiquetage très fin. Les mots sont tout d'abord regroupés en 12 grandes classes très générales : adjectif, conjonction, déterminant, mot-phrase, nom, pronom, adverbe, préposition, verbe, résidu, ponctuation et extra-lexical. Ces classes sont ensuite affinées, comme par exemple conjonction en conjonction de coordination et conjonction de subordination, verbe en verbe auxiliaire et verbe principal. On aboutit ainsi à un jeu de 36 étiquettes. Enfin, ont été adjoints à ces classes grammaticales des traits proprement morphologiques, tels que le genre, le nombre, la personne, le mode ou encore le temps qui donnent le jeu de 312 étiquettes. Nous avons ajouté, pour des facilités d'implémentation et sans dénaturer la cohérence de l'ensemble, huit étiquettes composées (qui existaient dans le corpus sous la forme d'une composition de deux étiquettes) et sommes ainsi arrivés à un jeu de 320 étiquettes.

Nous avons choisi d'effectuer l'apprentissage sur le corpus GRACE avec le jeu de 320 étiquettes. En effet, l'intérêt de ce jeu est que les classes très fines apportent beaucoup d'informations sur les mots et permettent de se passer d'analyseur morphologique en vue d'une étape ultérieure d'analyse en constituants.

Ainsi la procédure d'apprentissage permet d'apprendre un HMM basé sur 320 états. Nous avons gardé la majeure partie du corpus comme données d'apprentissage et utilisé le reste comme données de test, soit environ 26.000 mots. L'entraînement du modèle a permis d'obtenir un dictionnaire d'environ 44.000 formes fléchies.

Nous avons modifié l'algorithme d'étiquetage proprement dit pour que celui-ci prenne en compte les locutions. Pour cela, nous avons appliqué l'heuristique du motif le plus long. Au moment de lire le texte en entrée, l'étiqueteur va chercher, grâce au lexique, si la combinaison du mot suivant au mot courant forme une locution. Si tel est le cas, on applique ce principe itérativement avec le mot suivant jusqu'à ce que l'application ne soit plus possible, auquel cas, on garde la dernière locution trouvée. L'étiquetage de la phrase par Viterbi est ensuite effectué avec celle-ci.

3 Evaluation

Pour évaluer notre travail, nous utilisons la précision (proportion d'étiquetages corrects parmi les étiquetages stricts) et la décision (proportion d'étiquetages stricts parmi l'ensemble de tous les étiquetages). Dans un premier temps, nous avons développé un script Perl chargé de cette évaluation, qui sera appelé par la suite EVAL. Dans un second temps nous avons réutilisé la boîte à outils d'évaluation des analyseurs du projet ELSE[1]. Notons que EVAL comporte 150 lignes de code alors que l'évaluateur ELSE en comporte plusieurs milliers, même si ce dernier se veut plus ambitieux et traite, par exemple, divers formats de fichiers en entrée.

3.1 Résultats

Lors de tous nos tests, notre analyseur atteint un score de décision de 100% et ceci pour deux raisons. Tout d'abord, il ne renvoie qu'une seule étiquette par mot, ce qui ne génère pas d'étiquetage ambigu. Et ensuite, notre jeu d'étiquettes est identique à celui de GRACE. Il n'y a donc pas de problèmes de projection d'un jeu dans l'autre (phénomène qui entraîne des regroupements de plusieurs étiquettes en une seule ou inversement) lors de l'utilisation de l'évaluateur ELSE.

Nous avons effectué des tests avec trois niveaux de granularité de l'étiquetage. Le premier niveau correspond au jeu complet des 320 étiquettes. Le second est ce même jeu mais sans les traits morphologiques, ce qui donne 36 étiquettes. Le dernier niveau est l'étiquetage le plus général en 12 grandes classes grammaticales. Dans le tableau 1 nous présentons les scores de précision dans différents cas. La première colonne correspond à l'étiquetage du corpus de test produit par notre analyseur morpho-syntaxique avec EVAL. La deuxième fait référence au même étiquetage mais avec l'évaluateur ELSE. Et la dernière présente l'évaluation par ELSE de ce même corpus de test mais étiqueté par l'analyseur Cordial. En effet, ce dernier cas nous a semblé intéressant comme élément de comparaison puisque Cordial semble actuellement être le meilleur analyseur du français.

	EVAL	ELSE	ELSE / Cordial
320 tags	87.65 %	89.07 %	94.44 %
36 tags	92.24 %	93.52 %	94.63 %
12 tags	94.39 %	95.69 %	97.52 %

Table 1: Scores de précision

A partir de ces chiffres nous pouvons tirer plusieurs constations. Tout d'abord remarquons qu'entre l'évaluation de notre analyseur avec EVAL et celle avec ELSE, on a pour les trois niveaux environ 1.5 point d'écart. Ceci vient de la différence des méthodes de segmentation utilisées. Avec ELSE, l'alignement entre la sortie de l'étiqueteur et le corpus correctement étiqueté n'est pas très bon car ELSE utilise un segmenteur générique et non un segmenteur adapté à l'analyseur comme EVAL. Ainsi plus de tokens ne sont pas correctement réalignés avec ELSE, comme certains mots composés ou contenant une apostrophe. Ceux-ci ne sont donc pas soumis à évaluation, ce qui tend à améliorer mécaniquement la précision. D'autre part, cela souligne l'importance de la question de la segmentation et la difficulté d'y apporter une réponse qui soit valide à travers plusieurs formalismes. Avec EVAL, les tokens incorrectement réalignés sont des mots contenant des tirets, des mots composés et des expressions non présentes dans le dictionnaire. On en compte environ 0.8 % du nombre de tokens contenus dans le corpus de test.

[1] http://www.limsi.fr/TLP/ELSE

Ensuite, avec le jeu d'étiquettes complet (320 tags) et l'évaluateur ELSE, notre analyseur est moins performant que Cordial. Cependant si l'on compare ce score avec ceux présentés dans (Adda, 99), nous nous situons dans la moyenne des analyseurs évalués qui obtiennent de 82 à 96 % de précision. De plus, notons que lors de cette campagne les meilleurs résultats ont été obtenus non par des étiqueteurs mais par des analyseurs syntaxiques complets. Ceux-ci ayant un avantage certain sur les premiers pour désambiguïser les cas les plus difficiles. En effet, leur résolution peut être reportée au niveau syntaxique, ce qui améliore de quelques (mais précieux) points la précision. Au vu de nos résultats et du fait que nous nous limitons au niveau de l'étiquetage, notre approche est intéressante vis-à-vis des autres.

Enfin, nous avons effectué l'évaluation avec des jeux d'étiquettes plus concis. En effet, ceux-ci se rapprochant plus des jeux utilisés par les analyseurs de l'anglais, la comparaison est rendue possible. Notre score se situe également dans la moyenne de ceux présentés dans la littérature. Notons que dans (Briscoe, 02), donc pour RASP sur l'anglais, les auteurs atteignent une précision de 97 %. Nous nous situons en dessous, ce qui laisse entrevoir des possibilités d'amélioration. Toutefois un tel score semble difficilement atteignable par un étiqueteur seul pour le français. En effet, du fait de sa plus grande richesse morphologique, le français est plus dur à étiqueter que l'anglais. Ceci est confirmé, au vu de la littérature, par le fait que globalement les scores des étiqueteurs de l'anglais sont plus élevés, de quelques (et toujours précieux) points, que ceux du français.

3.2 Evaluation par classes

Nous cherchons dans cette partie à évaluer les points forts et les points faibles de notre étiqueteur de façon plus précise. Dans un premier temps, nous présentons dans cette section les résultats pour chaque classe grammaticale. Dans un second temps, nous présentons dans la section suivante le taux de reconnaissance des homographes.

Dans le tableau 2, nous présentons pour chaque classe grammaticale, le nombre d'occurrences dans le corpus de test, le pourcentage d'étiquetage correct et les trois types d'erreurs les plus fréquentes par ordre décroissant.

	Occurrences	% Correct	Erreur 1	Erreur 2	Erreur 3
Nom	6506	95.2 %	Adj / 2 %	V / 0.8 %	D / 0.7 %
Verbe	3184	96 %	Adj / 2.4 %	N / 1.1 %	Prep / 0.3 %
Verbe auxiliaire	669	78.5 %	Vp / 21 %	N / 0.1 %	Adv / 0.1 %
Verbe principal	2515	92.8 %	Adj / 3 %	Va / 2.3 %	N / 1.3 %
Adjectif	1773	85.4 %	V / 6.6 %	N / 5.1 %	Adv / 1.2 %
Pronom	1456	89.3 %	C / 4.5 %	D / 2 %	Adv / 1.7 %
Déterminant	3162	94.0 %	Prep / 1.8 %	Pron / 1.7 %	N / 1.6 %
Adverbe	1170	94.9 %	C / 2.3 %	Prep / 1 %	N / 0.8 %
Conjonction	856	97.9 %	Adv / 1%	Prep / 1 %	Pron / 0.6 %
Préposition	3863	81.8 %	D / 16.4 %	Pron / 0.6 %	N / 0.3 %
Ponctuation	3424	98.8 %	Adv / 0.8 %	Pron / 0.3 %	N / 0.1 %

Table 2: Scores et types d'erreurs par classes

Nous considérons que les classes présentant un taux supérieur à 94 % sont bien reconnues et que les efforts d'amélioration de l'étiqueteur devraient plutôt se porter sur la reconnaissance

des autre classes. En étudiant en particulier les deux composantes de la classe Verbe (auxiliaire et principal), on constate que les auxiliaires sont les plus mal reconnus. Néanmoins le tagger hésite essentiellement entre Verbe auxiliaire et Verbe principal, ce qui est satisfaisant pour une étape ultérieure d'analyse syntaxique. La ponctuation présente un score élevé mais toutefois insuffisant, ce qui est dû aux points de suspension mal étiquetés.

Les prépositions semblent particulièrement mal reconnues. En analysant les erreurs, on constate que ce sont les mots du type "de la, de l', des" qui posent la plupart des problèmes, c'est-à-dire les mots homographes qui sont soit des articles partitifs, soit des prépositions contractées. Cette difficulté est due à la forte ambiguïté entre les deux cas, élément caractéristique de la grammaire française. De plus, les adjectifs et les pronoms sont un point faible de notre étiqueteur.

3.3 Evaluation des homographes

En nous inspirant de (Vergne, 99) qui présente différents types d'homographes du français, nous avons cherché à affiner ces résultats. Le tableau 3 présente ci-dessous les taux de reconnaissance de différents types d'homographes. Dans la première colonne (H1) sont présentés les homographes déterminant/pronom (le, l', la, les); en H2, les homographes préposition contractée/article partitif (de, d', du, des); en H3, les homographes conjonction/pronom relatif (que, qu'); en H4, les homographes auxiliaire/nom (est, être, avoir); en H5, les homographes adverbe/nom (bien, mal, moins, plus, pas, point...); en H6, les homographes nom/adjectif et en H7, les homographes nom/verbe principal. La première ligne présente le nombre d'occurrences dans le corpus de test de la classe majoritaire (en effet, pour les cas d'homographes, une classe est largement majoritaire par rapport à l'autre, de l'ordre de 70 à 95 % des cas); la seconde, le nombre d'occurrences de la classe minoritaire; la troisième, le taux de reconnaissance de la classe majoritaire, et la dernière le taux de reconnaissance de la classe minoritaire.

	H1	H2	H3	H4	H5	H6	H7
Occ. Maj	D: 1767	Prep: 1601	C: 200	Aux: 267	Adv: 223	Adj: 290	N: 141
Occ. min	Pro: 59	Part: 132	Pro: 47	N: 6	N: 27	N: 171	VP: 96
% OK Maj	97.7 %	82.5 %	98 %	99.2 %	97.3 %	91.7 %	92.9 %
% OK min	69.5 %	62.1 %	17 %	85.7 %	85.2 %	87.7 %	83.3 %

Table 3: Reconnaissance des homographes

Au vu de ces résultats, nous constatons que les classes majoritaires sont bien reconnues (sauf en H2, où on retrouve la difficulté précédente). Pour les classes minoritaires, les résultats sont réellement encourageants pour H4, H5, H6 et H7, or ce sont ces cas qui sont les plus intéressants pour l'étape ultérieure d'analyse syntaxique. En effet, nous pensons que discriminer un nom d'un verbe (H7) apporte plus d'information qu'une préposition d'un partitif (H2), même si cela semble difficilement quantifiable. Nous avons déjà discuté de la difficulté du cas H2. Pour ce qui est de H1 et H3, dans les deux cas, la classe minoritaire est le pronom. Nous pensons atteindre ici certaines limites de l'étiquetage pour ce qui est de la reconnaissance des pronoms. En effet, celle-ci serait probablement plus pertinente au niveau de l'analyse en constituants .

4 Conclusion

A travers ce travail, nous avons cherché à développer un étiqueteur morpho-syntaxique du français se fondant sur une méthode d'apprentissage supervisé. Pour ce faire, en nous basant sur un tagger de l'anglais, nous avons adapté la procédure d'apprentissage aux exigences du traitement automatique du français. Nous obtenons un score de précision de 89 %, score inférieur à ceux des meilleurs étiqueteurs mais comme, d'une part ceux-ci sont des analyseurs syntaxiques complets et d'autre part nous n'avons effectué aucune optimisation sur notre étiqueteur, ce résultat est intéressant. Après avoir analysé en détail les erreurs d'étiquetage, il ressort que les points faibles se situent au niveau des adjectifs, pronoms et prépositions et en particulier sur ceux qui sont homographes avec une autre classe. Toutefois nous avons identifié certains cas d'homographes comme étant les plus intéressants et ceux-ci s'avèrent bien reconnus.

Dans une perspective de poursuite de ce travail, nous avons effectué des tests prospectifs avec le dictionnaire appris à partir de tout le corpus (mais avec les mêmes transitions), la précision s'est améliorée de 1.5 à 2 points pour les trois niveaux. Cela laisse à penser que l'augmentation de la taille du dictionnaire présenterait des gains significatifs de précision et, plus généralement, l'acquisition d'un modèle de langage à partir d'un corpus plus conséquent. D'autre part, nous avons réutilisé tel quel le guesser qui est optimisé pour l'anglais. Puisqu'il repose sur des règles statistiques, il fonctionne également ici mais mériterait d'être étudié plus en détail et adapté.

Remerciements

Nous tenons à remercier John Carroll pour son concours et Emmanuel Giguet pour ses commentaires sur ce travail.

Références

Adda G., Mariani J., Paroubek P., Rajman M., Lecomte J. (1999), Métrique et premiers résultats de l'évaluation GRACE des étiqueteurs morphosyntaxiques pour le français, Actes de *la Sixième Conférence sur le Traitement Automatique des Langues Naturelles*.

Brill E. (1992), A simple rule-based part of speech tagger, Actes de *Third Conference of Applied Natural Language Processing*.

Briscoe T., Carroll J.(2002), Robust accurate statistical annotation of general text. Actes de *Third International Conference on Language Resources and Evaluation*. p. 1499-1504.

Chanod J-P., Tapanainen P. (1995), Tagging French - comparing a statistical and a constraint-based method, Actes de *Seventh Conference of the European Chapter of the ACL*.

Elworthy D. (1994), Does Baum-Welch re-estimation help taggers ?, Actes de *Fourth ACL Conference on Applied NLP*.

Giguet E., Vergne J. (1997), From part of speech tagging to memory-based deep syntactic analysis, Actes de *Fifth Iternational Workshop on Parsing Technologies.*

Stein A., Schmid H. (1995), Etiquetage morphologique de textes français avec un arbre de décisions, *Revue T.A.L*, Vol. 36.

Vergne J. (1999), *Habilitation à diriger des recherches*, p. 36.