Étiquetage morpho-syntaxique pour des mots nouveaux

Ingrid Falk Delphine Bernhard Christophe Gérard Romain Potier-Ferry
LiLPa, Université de Strasbourg
ifalk, dbernhard, christophegerard@unistra.fr ; romainpotierferry@gmail.com

Résumé. Les outils d'étiquetage automatique sont plus ou moins robustes en ce qui concerne l'étiquetage de mots inconnus, non rencontrés dans le corpus d'apprentissage. Il est important de connaître de manière précise la performance de ces outils lorsqu'on cible plus particulièrement l'étiquetage de néologismes formels. En effet, la catégorie grammaticale constitue un critère important à la fois pour leur identification et leur documentation. Nous présentons une évaluation et une comparaison de 7 étiqueteurs morphosyntaxiques du français, à partir d'un corpus issu du Wiktionnaire. Les résultats montrent que l'utilisation de traits de forme ou morphologiques est favorable à l'étiquetage correct des mots nouveaux.

Abstract. Part-of-speech (POS) taggers are more or less robust with respect to the labeling of unknown words not found in the training corpus. It is important to know precisely how these tools perfom when we target part-of-speech tagging for formal neologisms. Indeed, grammatical category is an important criterion for both their identification and documentation. We present an evaluation and comparison of 7 POS taggers for French, based on a corpus built from Wiktionary. The results show that the use of form-related or morphological features supports the accurate tagging of new words.

Mots-clés : étiquetage morphosyntaxique, évaluation, néologie formelle.

Keywords: part-of-speech tagging, evaluation, formal neologisms.

1 Introduction

Les outils d'étiquetage morphosyntaxique procèdent par apprentissage automatique à partir d'un corpus annoté manuellement. De fait, ils sont plus ou moins robustes en ce qui concerne l'étiquetage de mots inconnus, non rencontrés dans le corpus d'apprentissage. Cette problématique nous intéresse tout particulièrement dans le cadre de l'identification de néologismes formels, c'est-à-dire les nouvelles formes qui apparaissent quotidiennement. En effet, la catégorie grammaticale constitue un critère important à la fois pour l'identification et la documentation de ce type de néologismes.

L'objectif de ce travail est de présenter une analyse détaillée de 7 étiqueteurs pour le français, en comparant leurs performances pour l'étiquetage de néologismes issus du Wiktionnaire. L'objectif final est d'identifier l'outil ayant les meilleurs résultats afin de l'incorporer à un système d'identification automatique de néologismes formels. Nous souhaitons avant tout disposer d'un outil qui convient à nos besoins : (i) prêt à l'emploi (ii) sans entraînement supplémentaire (iii) simple d'utilisation et (iv) librement disponible. Cet outil doit par ailleurs avoir les meilleures performances possibles pour l'étiquetage des néologismes dans ces conditions, et sa performance pour les mots connus n'est pas décisive dans le cadre de notre projet.

L'article est organisé comme suit : dans un premier temps, nous présentons le corpus utilisé dans le cadre de nos expériences, puis les outils comparés. La Section 4 détaille le protocole utilisé pour nos expériences et la Section 5 propose une discussion des résultats obtenus.

2 Corpus de référence

Nous avons constitué un corpus de néologismes contextualisés (phrases ou paragraphe), extrait de la liste fournie par le Wiktionnaire [1]. Ce corpus se constitue de 152 entrées [2] toutes conformes au trois critères suivants :

i) Le néologisme et son contexte sont attestés, et non pas forgés.

ii) Il existe au moins une source en format numérique aisément accessible.

iii) La date d'apparition du néologisme est postérieure à 2000.

Techniquement ce corpus se présente sur la forme d'un fichier XML dont un extrait est montré en Figure 1. Les informations utilisés dans nos expériences sont : le lemme du néologisme (balise `word`), sa catégorie grammaticale (attribut `pos`) ainsi que les phrases où il apparaît (balise `texte`). Dans ces phrases le néologisme est marqué par la balise `<neologisme>`.

```
<?xml version="1.0" encoding="UTF-8"?>
<corpus>
  <entry id="L00001">
    <word pos="adv">abracadabrantesquement</word>
    <examples>
      <texte url="http://www.bruxelles-francophone.be/topic2035.html"
             type="forum" date="2009" id="L00001-1">
        Son appétit institutionnel est <neologisme>abracadabrantesquement</neologisme> vorace.
      </texte>
    </examples>
  </entry>
  ....
  <entry id="L00003">
    <word pos="verbe">accessibiliser</word>
    <examples>
      <texte url="http://fplanque.net/Blog/itTrends/2003/01/22/accessibiliser_l_internet_oui_mais_pas_a"
             date="2003" type="blog" id="L00003-1">
        <neologisme>Accessibiliser</neologisme> l'Internet... oui, mais pas à moitié!
      </texte>
      <texte url="http://blog.temesis.com/post/2009/10/25/Marcel-fais-peter-ton-tableau"
             date="2009" type="blog" id="L00003-2">
        Ne pas passer une journée à tenter d'<neologisme>accessibiliser</neologisme>
        deux-trois tableaux hallucinants, pour un gain trop faible.
      </texte>
    </examples>
  </entry>
  ...
```

FIGURE 1: Extrait du corpus de néologismes utilisé dans nos expériences.

La Table 1 résume quelques propriétés de notre corpus.

#	total	noms	verbes	adj	adverbes	locutions	mots à tiret
lemmes	158 [3]	84	36	29	3	6	11
occurences	459	293	68	81	4	13	28
phrases	220	115	55	38	4	8	19

TABLE 1: Propriétés du corpus de néologismes.

1. http://fr.wiktionary.org/wiki/Cat\%E9gorie:N\%E9ologismes_en_fran\%C3\%A7ais

2. Signalons toutefois qu'il est possible que certaines de ces entrées se trouvent déjà dans des ressources lexicales plus récentes. Ainsi le mot *assistanat* est présent dans le Lefff (Sagot, 2010) et *animaliste* peut être trouvé dans le lexique Wortschatz (Quasthoff *et al.*, 2006). Nous estimons néanmoins que cette collection présente un échantillon de formes assez récentes et par conséquence assez représentatif pour permettre une évaluation de l'étiquetage automatique appliquée à de nouveaux mots.

3. Le nombre de formes distinctes est bien de 152, mais nous comptons plusieurs fois celles qui ont plusieurs catégories grammaticales (par exemple animaliste, nom et animaliste, adjective.)

3 Présentation des étiqueteurs comparés

Pour notre projet de collecte de néologismes formels, nous devrons utiliser les étiqueteurs quotidiennement, sur des textes journalistiques.

Les logiciels que nous avons pu identifier et qui répondent aux conditions énoncées dans l'introduction sont présentés brièvement dans cette section. Il s'agit de LGtagger et SEM (Constant & Sigogne, 2011; Constant *et al.*, 2011), LIA_tagg (Nasr *et al.*, 2004), l'étiqueteur de Stanford (Toutanova *et al.*, 2003), MElt (Denis & Sagot, 2010), Talismane (Urieli & Tanguy, 2013) et le plus ancien mais encore largement utilisé, le TreeTagger (Schmid, 1994). Ces outils se distinguent non seulement par les algorithmes appliqués mais leurs performances dépendent aussi d'autres facteurs, comme par exemple :

– le corpus d'apprentissage (et le jeu d'étiquettes utilisées) ;
– des ressources lexicales externes utilisées ou non ;
– du pré- et/ou post traitement.

A leur tour, les algorithmes d'apprentissage et d'étiquetage s'appuient largement sur un ensemble de traits extraits du corpus d'apprentissage, qui ont aussi un effet important sur leur performance.

LGtagger et SEM (Constant & Sigogne, 2011; Constant *et al.*, 2011) LGTagger et SEM ont un fonctionnement similaire : ils procèdent par apprentissage automatique avec des CRF (*Conditional Random Fields* / champs markoviens conditionnels) et peuvent exploiter des lexiques externes. Ces outils visent également à combiner segmentation en unités et étiquetage, en reconnaissant les unités polylexicales. Les deux outils diffèrent toutefois par les attributs et ressources externes utilisées. L'exactitude (*accuracy*) de SEM varie de 81,6% (corpus oral) à 95.6% (corpus blog) [4]. Celle de LGTagger est de 97.7% (sans segmentation incorporée). Nous avons utilisé la version 1.1 de LGTagger, qui intègre plusieurs ressources lexicales (cf. Tableau 2) et où la segmentation est incorporée. SEM a été utilisé avec un modèle appris sur un corpus ne contenant pas d'unités multi-mots et en prenant en compte des ressources linguistiques externes (le Lefff). Comme pour LGTagger, la segmentation était incorporée.

LIA_tagg (Nasr *et al.*, 2004) LIA_tagg est un étiqueteur morpho-syntaxique pour le français prêt à l'emploi développé au Laboratoire Informatique d'Avignon. Il utilise un jeu de 103 étiquettes et implémente une approche probabiliste basée sur les Chaînes de Markov Cachées (HMM). Nous n'avons pas plus d'information sur le corpus d'apprentissage et le taux de précision obtenu.

Stanford (Toutanova *et al.*, 2003) Le *Stanford tagger* est un étiqueteur développé en 2003 pour l'anglais. Ils a été étendu à d'autre langues (français, arabe, chinois, allemand ...), constamment amélioré et est distribué librement à `http://nlp.stanford.edu/software/tagger.shtml#Download`. Sa particularité est d'appliquer un modèle markovien *bidirectionel* à maximisation d'entropie et d'utiliser des propriétés morphologiques : *n*-grammes de suffixes et préfixes, présence de tirets, majuscules, etc. Nous utilisons le modèle français entrainé sur la French Treebank (FTB, (Abeillé *et al.*, 2003)) avec un jeu de 15 étiquettes. Nous n'avons pas d'informations sur la performance de ce modèle.

MElt (Denis & Sagot, 2010) MElt est un étiqueteur basé sur un modèle markovien à maximisation d'entropie qui intègre également un lexique exogène à large couverture (le Lefff (Sagot, 2010)). La méthode d'étiquetage et les traits utilisés sont similaires à ceux de l'étiqueteur de Stanford (cf. Section 3). Dans nos expériences nous avons utilisé la version librement disponible à `http://gforge.inria.fr/projects/lingwb/` qui a été entrainé sur la French Treebank (FTB, (Abeillé *et al.*, 2003)) avec un jeu de 29 étiquettes. Evalué sur le FTB, MElt atteint un taux de précision de 97.75% (91.36% sur les mots inconnus).

Talismane (Urieli & Tanguy, 2013) Talismane est un analyseur syntaxique qui intègre les différentes étapes d'analyse : segmentation en mots et en phrases, étiquetage morphosyntaxique et finalement repérage des dépendances syntaxiques. Pour chaque module, un modèle est appris à partir d'un corpus d'entraînement (FTB). Au moment de l'analyse, des règles peuvent être appliquées pour contraindre les résultats. L'exactitude de Talismane pour le FTB est de 97,8%.

4. `http://www.lattice.cnrs.fr/sites/itellier/SEM.html`

Outil	Méthode	Corpus d'apprentissage (étiquettes)	Ressources lexicales	Utilisation de la forme	Particularité
LGTagger	CRF (i)	FTB (ii)	DELA, Lefff, Prolex, Organisations, Prénoms	Oui (iv)	Segmentation incorporée
SEM	CRF (i)	FTB (ii)	Lefff	Oui (iv)	
LIA_tagg	HMM (iii)	103	lexique 10 000 mots	Non	
Stanford	CMM à maximisation d'entropie.	FTB (ii, 14 étiquettes [6])	–	Oui (iv)	bidirectionnel
MElt	comme *stanford*	FTB (ii, 29 étiquettes)	Lefff	Oui (iv)	
Talismane	classifieur par entropie maximale	FTB	Lefff	Oui	
TreeTagger	Arbres de décisions	43 834 mots, 33 étiquettes	–	Oui (v)	

TABLE 2: Les étiqueteurs morpho-syntaxique utilisés.

(i) CRF : Conditional Random Fields (Champs Aléatoires Conditionnels)

(ii) FTB : French Treebank (Abeillé *et al.*, 2003)

(iii) HMM : Hidden Markov Model (Modèle de Markov caché)

(iv) Utilisation de traits « mots inconnus » : *n*-gram suffixes et préfixes (*n* de 1 à 4), tirets, majuscules, chiffres ...

(v) Utilise un lexique associant à des suffixes la probabilité des étiquettes.

TreeTagger (Schmid, 1994) Le TreeTagger est un des plus anciens outils d'étiquetage automatique. Il est basé sur un modèle probabiliste d'arbre de décision. La version (française) que nous avons utilisé est un modèle issu d'un apprentissage sur un corpus d'environ 44 000 mots et un jeu de 33 étiquettes réalisé par Achim Stein [5]. Dans nos expériences nous n'avons pas utilisé de ressource lexicale supplémentaire.

Le Tableau 2 présente un résumé comparatif des outils employés dans nos expériences.

4 Expériences

Afin d'évaluer l'étiquetage automatique pour la détection et la documentation de néologismes formels, nous appliquons les étiqueteurs décrits en Section 3 au corpus de néologismes que nous avons constitué (présenté en Section 2). Nous comparons ensuite l'étiquetage des outils à celui du corpus de référence.

Étiquettes. Les néologismes du corpus n'ayant que les catégories nom, verbe, adjectif et adverbe, nous projetons les sorties des étiqueteurs, qui sont souvent plus détaillés, sur ces catégories principales. Le nombre d'étiquettes ayant un effet important sur la qualité de l'étiquetage, ce choix d'étiquettes risque de pénaliser les outils produisant un étiquetage plus détaillé. Nous considérons l'étiquetage d'une locution comme correct si les étiquettes des composantes correspondent à celles de la référence.

Prétraitement. Le prétraitement, plus spécifiquement la segmentation en mots, est une étape décisive dans l'étiquetage automatique et chaque étiqueteur a sa propre stratégie pour l'aborder. En général, les outils distribués sont assortis de scripts pour d'abord découper le texte en phrases et ensuite en unités de formes (des mots), qui convient plus ou moins à notre format d'entrée. Pour d'autres outils, comme LGtagger et SEM, la segmentation est intégrée à l'étiquetage.

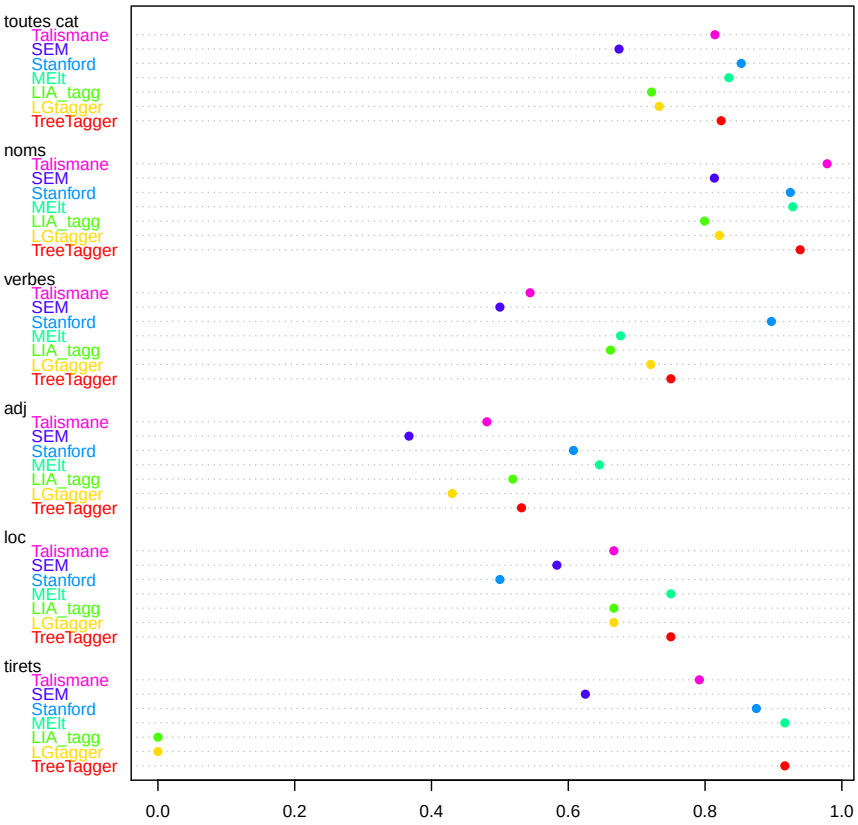
5. http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

6. Si la FTB utilise un ensemble de 15 étiquettes morpho-syntaxiques, nous avons constaté que le Stanford Tagger ne distingue pas les noms propres (NP) des noms communs (NC).

5 Résultats et discussion

Étiqueteur	toutes	noms (293)	verbes (68)	adjectifs (81)	locutions (13)	mots à tiret (28)
LGtagger	73.30	82.08	72.06	43.04	66.67	0.00
LIA_tagg	72.17	79.93	66.18	51.90	66.67	0.00
MElt	83.26	92.83	67.65	64.56	75.00	91.67
SEM	67.42	81.36	50.00	36.71	58.33	62.50
Stanford	85.29	92.47	89.71	60.76	50.00	87.50
Talismane	81.45	97.85	54.41	48.10	66.67	79.17
TreeTagger	82.35	93.91	75.00	53.16	75.00	91.67
majorité	86.43					

(a) Pourcentages d'étiquettes correctes.



(b) Représentation graphique.

FIGURE 2: Résultats par catégorie grammaticales.

Les résultats sont présentés en Figure 2. Nous montrons le pourcentage d'étiquettes correctes pour toutes les occurrences de mots nouveaux dans notre corpus et également pour les catégories principales (nom, verbe, adjectif). À titre indicatif est affiché également le pourcentage d'étiquetages correctes pour les locutions mais nous considérons que la taille de cet échantillon ne permet pas de conclusions finales. Les néologismes formels sont souvent formés par composition avec des tirets, pour cette raison nous considérons également les performances des étiqueteurs sur ce type de phénomène [7].

Les performances observées sont, d'une manière générale, plus basses que celles rapportés dans des publications. Ceci

7. A titre indicatif, l'échantillon étant trop limité.

est un effet attendu, vu que les performances publiées sont calculées sur un corpus test issu du même corpus global. Les meilleurs étiquetages sont ceux du Stanford tagger (85.29%) suivis de ceux du MElt (83.26%) et du TreeTagger (82.35%). En mutualisant ces résultats par un vote majoritaire le taux a pu être augmenté à 86.43%.

Par rapport à l'algorithme d'étiquetage, les CMM utilisés pour Stanford et MElt semblent plus favorables dans le cadre de notre application que les CRF (employés par LGTagger et SEM) ou les classifieurs à entropie maximale.

Pour ce qui est de ressources lexicales, leur emploi n'a pas d'effet positif dans notre cas. En effet, alors que Stanford et MElt s'appuient sur des algorithmes et traits similaires, l'utilisation d'une ressource lexicale externe par MElt ne lui permet pas d'améliorer ses performances. Par contre, l'utilisation de traits de forme ou morphologiques (préfixe, suffixe, etc.) semble favorable : Stanford, MElt et TreeTagger, qui les intégrent obtiennent les meilleurs taux d'étiquetage correctes. Il a par ailleurs été montré que ces traits jouent un rôle important pour l'identification de néologismes formels (Sagot *et al.*, 2013; Falk *et al.*, 2014).

A son tour LIA_tagg semble défavorisé par le grand nombre d'étiquettes utilisées.

6 Conclusion

Au vu de ces résultats, nous concluons que le *Stanford Tagger* est l'outil le plus plus adapté à notre application et nous allons donc l'utiliser dans notre projet. Comme c'était à prévoir l'utilisation de ressources lexicales externes n'a pas d'impact. Par ailleurs, la combinaison de la segmentation et de l'étiquetage est plutôt défavorable. Par contre les traits de forme et morphologiques ont un impact positif.

Remerciements. Ces travaux ont été financés par l'Université de Strasbourg dans le cadre de l'Initiative d'Excellence (IdEx) 2012 (projet Logoscope).

Références

ABEILLÉ A., CLÉMENT L. & TOUSSENEL F. (2003). Building a Treebank for French. In A. ABEILLÉ, Ed., *Treebanks*, number 20 in Text, Speech and Language Technology.

CONSTANT M. & SIGOGNE A. (2011). MWU-aware part-of-speech tagging with a CRF model and lexical resources. In *Proceedings of the Workshop on Multiword Expressions : from Parsing and Generation to the Real World, ACL 2011*.

CONSTANT M., TELLIER I., DUCHIER D., DUPONT Y., SIGOGNE A. & BILLOT S. (2011). Intégrer des connaissances linguistiques dans un CRF : application à l'apprentissage d'un segmenteur-étiqueteur du français. In *Actes de TALN 2011*.

DENIS P. & SAGOT B. (2010). Exploitation d'une ressource lexicale pour la construction d'un étiqueteur morphosyntaxique état-de-l'art du français. In *Actes de TALN 2010*.

FALK I., BERNHARD D. & GÉRARD C. (2014). From Non Word to New Word : Automatically Identifying Neologisms in French Newspapers. In *Proceedings of LREC 2014*, Reykjavik, Islande.

NASR A., BÉCHET F. & VOLANSCHI A. (2004). Tagging with Hidden Markov Models Using Ambiguous Tags. In *Proceedings of COLING 2004*, COLING '04, Stroudsburg, PA, USA.

QUASTHOFF U., RICHTER M. & BIEMANN C. (2006). Corpus Portal for Search in Monolingual Corpora. In *Proceedings of LREC 2006*, Genoa.

SAGOT B. (2010). The Lefff, a Freely Available and Large-coverage Morphological and Syntactic Lexicon for French. In *Proceedings of LREC'10*, Valletta, Malta.

SAGOT B., NOUVEL D., MOUILLERON V. & BARANES M. (2013). Extension dynamique de lexiques morphologiques pour le français à partir d'un flux textuel. In *Actes de TALN 2013*.

SCHMID H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, UK.

TOUTANOVA K., KLEIN D., MANNING C. D. & SINGER Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of ACL 2003*.

URIELI A. & TANGUY L. (2013). L'apport du faisceau dans l'analyse syntaxique en dépendances par transitions : études de cas avec l'analyseur talismane. In *Actes de TALN 2013*.