TALN 2010, Montréal, 19–23 juillet 2010

Apprentissage non supervisé de la morphologie d'une langue par généralisation de relations analogiques

Jean-François Lavallée et Philippe Langlais
Université de Montréal
C.P. 6128, succ. Centre-ville
Montréal, Qc, Canada, H3C 3J7
{lavalljf,felipe}@iro.umontreal.ca

Résumé. Bien que les approches fondées sur la théorie de l'information sont prédominantes dans le domaine de l'analyse morphologique non supervisée, depuis quelques années, d'autres approches ont gagné en popularité, dont celles basées sur *l'analogie formelle*. Cette dernière reste tout de même marginale due notamment à son coût de calcul élevé. Dans cet article, nous proposons un algorithme basé sur l'analogie formelle capable de traiter les lexiques volumineux. Nous introduisons pour cela le concept de *règle de cofacteur* qui permet de généraliser l'information capturée par une analogie tout en contrôlant les temps de traitement. Nous comparons notre système à 2 systèmes : *Morfessor* (Creutz & Lagus, 2005), un système de référence dans de nombreux travaux sur l'analyse morphologique et le système analogique décrit par Langlais (2009). Nous en montrons la supériorité pour 3 des 5 langues étudiées ici : le finnois, le turc, et l'allemand.

Abstract. Although approaches based on information theory are prominent in the field of unsupervised morphological analysis, in recent years, other approaches have gained in popularity. Those based on *formal analogy* remain marginal partly because of their high computational cost. In this paper we propose an algorithm based on formal analogy able to handle large lexicons. We introduce the concept of *cofactor rule* which allows the generalization of the information captured by analogy, while controlling the processing time. We compare our system to 2 others : *Morfessor* (Creutz & Lagus, 2005), a reference in many studies on morphological analysis and the analogical system described by Langlais (2009). We show the superiority of our approach for 3 out of the 5 languages studied here : Finnish, Turkish, and German.

Mots-clés : Analyse morphologique non supervisée, Analogie formelle, Approche à base de graphe.

Keywords: Unsupervised Learning of Morphology, Formal Analogy, Graph-Based Approach.

1 Introduction

Dans le domaine de l'analyse morphologique, il existe 4 grandes familles d'approches solidement établies. La plus répandue consiste à utiliser des statistiques fondées sur l'entropie, un concept très connu en théorie de l'information, pour identifier la segmentation la plus probable. L'idée de base exploitée par ce type d'approche est qu'une lettre difficile à prédire dans un mot marque vraisemblablement une frontière de morphème. Harris (1955) a décrit un système basé sur cette idée il y a plus de 50 ans. La deuxième

approche courante consiste à regrouper les mots en paradigme et à enlever les affixes communs à ces groupes. *Paramor* (Monson *et al.*, 2007) qui utilise une variante de cette approche obtient régulièrement de bons résultats aux ateliers Morpho Challenge [1]. D'autres chercheurs abordent l'analyse morphologique comme un problème de compression. Ils décomposent les mots de façon à permettre l'encodage optimal du lexique. Les systèmes les plus connus du domaine, *Linguistica* (Goldsmith, 2001) et *Morfessor* (Creutz & Lagus, 2005) utilisent cette technique. Finalement, la quatrième approche établie consiste à utiliser le contexte des mots à analyser dans un corpus. L'idée étant que les mots ayant des contextes similaires ont souvent des morphèmes en commun. Yarowsky & Wicentowski (2000) utilisent cette technique pour faire le lien entre un verbe irrégulier et sont lexème. L'apprentissage analogique a récemment fait l'objet de plusieurs études en traitement automatique des langues (Lepage & Denoual, 2005; Denoual, 2007; Langlais & Patry, 2007). En particulier, plusieurs auteurs ont appliqué l'analogie formelle (ou analogie pour simplifier) à l'analyse morphologique. Stroppa & Yvon (2005) ont démontré que l'apprentissage analogique permet d'apprendre de manière supervisée à identifier la racine d'un mot et ses caractéristiques morphologiques. Hathout montre qu'il est possible de regrouper automatiquement les mots d'un lexique en familles morphologiques en combinant l'analogie à des données sémantiques (Hathout, 2002, 2008). Langlais (2009) a utilisé l'analogie pour analyser morphologiquement les mots de la base de données anglaise et allemande de Celex.

L'approche décrite dans (Langlais, 2009) ne permet pas de traiter des lexiques de la taille de ceux que nous manipulons dans cette étude. Nous introduisons le concept de *règle de cofacteur* (RC) qui permet à notre système de généraliser l'information capturée par l'analogie tout en réduisant significativement les temps de calculs. Nous décrivons un algorithme de regroupement de mots basé sur les graphes qui fait usage des RCs apprises en corpus. Les mots regroupés sont alors décomposés automatiquement en morphèmes à l'aide d'un algorithme dédié. Nous appliquons notre système sur différentes langues, dont des langues à morphologie riche et montrons sa supériorité sur le système décrit par Langlais (2009). Nous montrons également que notre système produit des résultats compétitifs avec ceux de *Morfessor*.

La suite de cet article est structurée comme suit. Nous décrivons en section 2 le modèle qui décompose de manière non supervisée les mots d'un lexique en leurs morphèmes. Nous présentons en section 3 notre protocole expérimental et les résultats obtenus. Nous concluons ce travail en section 4 et proposons des pistes de développements futurs.

2 Modèle

Le système que nous présentons ici est basé sur le moteur analogique développé par Langlais (2009). En particulier, nous continuons à nous appuyer sur l'analogie formelle pour extraire de manière non supervisée l'information morphologique d'un lexique. Bien qu'il existe plusieurs définitions de l'analogie formelle (voir par exemple (Pirrelli & Yvon, 1999)), nous reprenons celle utilisée par l'auteur, initialement proposée par Yvon *et al.* (2004) qui lient la définition d'une analogie entre formes à celle de leur *factorisation*.

Le lecteur intéressé trouvera dans ces études les détails de cette définition. Nous l'illustrons simplement à l'aide d'un exemple. La relation entre les 4 mots `cordially`, `cordial`, `lovely` et `love` est une analogie, ce que l'on note $[\texttt{cordially} : \texttt{cordial} = \texttt{lovely} : \texttt{love}]$, car il existe une factorisation

1. http://www.cis.hut.fi/morphochallenge2009/

de ces 4 chaînes qui met en jeu des *alternances*. La figure 1 illustre deux factorisations de ces chaînes. Celle de gauche met en œuvre 4 facteurs par factorisation, tandis que celle de droite n'en utilise que 2, illustrant les alternances `love`/`cordial` et `ly`/ϵ capturées par cette analogie. Le nombre de facteurs de la plus petite factorisation est appelé le degré de l'analogie ; soit 2 dans l'exemple.

$f_{cordially}$	$\equiv$	cordia	l	l	y	$f_{cordially}$	$\equiv$	cordial	ly
$f_{cordial}$	$\equiv$	cordia	ϵ	l	ϵ	$f_{cordial}$	$\equiv$	cordial	ϵ
f_{lovely}	$\equiv$	love	l	ϵ	y	f_{lovely}	$\equiv$	love	ly
f_{love}	$\equiv$	love	ϵ	ϵ	ϵ	f_{love}	$\equiv$	love	ϵ

FIGURE 1 – Deux factorisations de l'analogie [`cordially` : `cordial` = `lovely` : `love`].

L'hypothèse fondatrice de notre système est que les analogies identifient implicitement deux paires de formes qui sont morphologiquement liées. Par exemple, l'analogie de la figure 1 relie `love` à `lovely` et `cordial` à `cordially`. Le système décrit par Langlais (2009) repose entièrement sur cette hypothèse. Chaque fois qu'un mot est impliqué dans une analogie, sa factorisation est calculée et une distribution des factorisations observées est maintenue. L'analyse morphologique résultante pour un mot du lexique est la *segmentation* la plus fréquemment observée. Ce système requiert donc de calculer toutes les analogies impliquant les mots d'un lexique $\mathcal{L}$, ce qui requiert de vérifier $O(|\mathcal{L}|^4)$ analogies. Même en utilisant plusieurs optimisations (Stroppa, 2005; Langlais & Yvon, 2008), les temps de calculs sont trop importants lorsque le lexique dépasse quelques dizaines de milliers de mots.

Le problème est que l'information capturée par ce système est fortement lexicale. Ainsi, l'analogie [`love` : `lovely` = `cordial` : `cordially`] ne nous donne aucune information sur la relation entre `love` et `loved` ou entre `live` et `lively`. Le cœur de notre système consiste à introduire un procédé capable de généraliser l'information capturée par une analogie.

2.1 CoFacteur et RC

Les *cofacteurs* d'une analogie de degrée d [`x` : `y` = `z` : `t`] sont les d alternances intervenant dans les factorisations des termes de l'analogie (Langlais & Patry, 2007). Dans notre exemple, nous avons 2 cofacteurs : `love`/`cordial` et `ly`/ϵ. Il est important de noter que ces paires de *facteurs* ne sont pas dirigées, c'est-à-dire que la paire `ly`/ϵ est égale à la paire ϵ/`ly`. Le cofacteur `ly`/ϵ capture en anglais le phénomène productif où un adverbe est dérivé d'un substantif ou d'un adjectif en lui ajoutant le suffixe `ly`. Il permettrait par exemple de rendre compte de la forme `flabbergastedly` à partir de la forme `flabbergasted` (ou l'inverse) même si cette dernière n'est pas présente dans le lexique. Il permet malheureusement de lier à tort les formes `flyable` et `fable`.

La généralisation offerte par les cofacteurs peut donc introduire du bruit si on l'applique aveuglément. Ceci nous a amené à introduire le concept de *règle de cofacteur (RC)*, une règle de réécriture contextualisée. Plus précisément, une RC est une règle de réécriture $\langle \alpha \rightarrow \beta \rangle$ où α et β sont les deux facteurs d'un cofacteur et tel que $|\alpha| \geq |\beta|$ [2], de sorte que l'application d'une RC à un mot produise toujours un mot de taille inférieure ou égale. Afin de réduire le bruit inhérent à l'application d'une telle règle, nous ajoutons le symbole $\star$ à gauche et/ou à droite de ces facteurs pour indiquer l'existence d'un facteur non vide à

2. Si les deux facteurs sont de la même taille, l'ordre alphabétique est utilisé.

gauche et/ou à droite de l'alternance considérée. Ceci est fait dans le but de distinguer les opérations de préfixation, de suffixation et d'infixation communes à de nombreuses langues. Dans notre exemple, les RCs $\langle \star ly \rightarrow \star\epsilon \rangle$ et $\langle love\star \rightarrow cordial\star \rangle$ sont produites. La présence du symbole $\star$ à gauche de ly dans la première RC interdit par exemple de lier les formes $flyable$ et $fable$.

Nous notons $\mathcal{R}(x)$, l'application de la RC $\mathcal{R}$ sur une forme x. Par exemple, si $\mathcal{R}$ est $\langle \star ly \rightarrow \star\epsilon \rangle$, alors $\mathcal{R}(elderly) = elder$. Par extension directe, $[\mathcal{R}_1, \ldots, \mathcal{R}_n](x)$ dénote la forme [3] résultante de l'application de n RCs : $\mathcal{R}_n(\ldots \mathcal{R}_2(\mathcal{R}_1(x)) \ldots)$. Dans le cas où la règle ne s'applique pas, on a $\mathcal{R}(x) = x$.

2.2 Extraction des RCs

Obtenir les RCs requiert d'identifier un ensemble d'analogies mettant en œuvre les mots du lexique. Dans la mesure où les RCs opèrent une généralisation de l'information capturée par analogie, il n'est pas nécessaire d'identifier l'ensemble de ces analogies. En pratique cependant, nous en avons calculé un grand nombre par lexique (voir la section 3).

Nous avons considéré différentes façon d'associer un score à une RC. Se baser seulement sur la fréquence crée un biais envers les RCs composées de facteurs courts. Par exemple, en anglais, la RC $\langle anti\text{-}\star \rightarrow \epsilon\star \rangle$ est observée 2 472 fois. Cependant la RC $\langle ka\star \rightarrow \epsilon\star \rangle$, qui est fort probablement fortuite, est vue 13 839 fois. Pour résoudre ce problème, nous utilisons une deuxième métrique : la *productivité* d'une règle $\mathcal{R}$, notée $prod(\mathcal{R})$, est définie par le ratio du nombre de fois où l'application de cette règle mène à une forme valide du lexique $\mathcal{L}$ au nombre de formes de $\mathcal{L}$ auxquelles elle peut-être appliquée. Formellement :

$$prod(\mathcal{R}) = \frac{|\{x \in \mathcal{L} : \mathcal{R}(x) \in \mathcal{L}\}|}{|\{x \in \mathcal{L} : \mathcal{R}(x) \neq x\}|}$$

Les RCs $\langle anti\text{-}\star \rightarrow \epsilon\star \rangle$ et $\langle ka\star \rightarrow \epsilon\star \rangle$ ont des productivités respectives de 0,9490 et 0,2472. La productivité et la fréquence de quelques RCs extraites d'un lexique anglais utilisé dans nos expériences sont rapportées en table 1. Comme le nombre de RCs rencontrées est très élevé, nous appliquons un filtre qui élimine celles ayant une productivité inférieure à ρ (voir la section 3) en espérant qu'elles ne jouent pas un rôle important dans la morphologie de la langue étudiée.

TABLE 1 – Fréquence et productivité de RCs extraites d'un lexique anglais.

RCs	Fréquence	Productivité
$\langle \star's \rightarrow \star\epsilon \rangle$	2 225 258	0.93
$\langle \star a\star \rightarrow \star\epsilon\star \rangle$	288 743	0.04
$\langle \star ing \rightarrow \star ed \rangle$	4 669	0.54
$\langle \star\text{-}based \rightarrow \star\epsilon \rangle$	3 226	0.98
$\langle \star co \rightarrow \star as \rangle$	68	0.10
$\langle \star able's \rightarrow \star able \rangle$	65	1.00

3. Afin de simplifier l'exposé, nous omettons le cas où l'application d'une RC peut générer plusieurs formes.

2.3 Construction des arbres de dérivation de mots (ADM)

L'ensemble des RCs collectées à l'étape précédente constitue la matière première de notre système. À l'aide de ces RCs, nous formons des regroupements de mots partageant la même racine. Nos regroupements sont structurés hiérarchiquement sous la forme d'arbres (ADM) comme celui de la figure 2. Les noeuds de ces arbres sont les mots du lexique. Un arc entre deux noeuds n_a et n_b, noté $n_a \rightarrow n_b$, est étiqueté d'une séquence de RCs qui lorsque appliquées au mot n_a résultent en le mot n_b. Dans notre exemple, les séquences de RCs sont toutes des singletons.

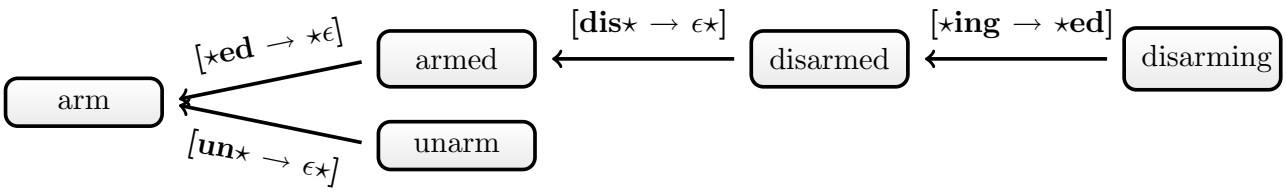

FIGURE 2 – ADM des mots anglais dérivés de `arm`.

La construction des ADMs est un processus glouton qui applique les trois étapes suivantes jusqu'à ce que tous les mots du lexique soient traités. Le résultat de ce processus est une forêt d'ADMs où chaque arbre regroupe, du moins nous l'espérons, les mots du lexique reliés morphologiquement.

1. Choisir un mot n du lexique $\mathcal{L}$ qui n'a pas encore été traité. [4]
2. Calculer $\mathcal{S}(n)$: l'ensemble des mots de $\mathcal{L}$ qui peuvent être atteints en appliquant un nombre strictement positif de RCs au mot n.
3. Ajouter un arc $n \rightarrow b$, où $b \equiv \text{argmax}_{w \in \mathcal{S}(n)}\, score(n, w)$ est le mot de $\mathcal{S}(n)$ maximisant un score (décrit ci-après) , si ce score est supérieur à un seuil donné τ (voir section 3).

La figure 3 illustre la construction de $\mathcal{S}(n \equiv \texttt{disarmed})$ à l'étape 2. Comme on peut l'observer, il est fréquent que plusieurs chemins (séquences de RCs) entre deux mots existent. Le score d'un lien entre deux mots n et w est donc calculé en sommant le score de chacun des chemins entre ces deux mots :

$$score(n, w) = \sum_{[\mathcal{R}_1,\ldots,\mathcal{R}_m](n) \equiv w} score([\mathcal{R}_1, \ldots, \mathcal{R}_m])$$

où le score d'une séquence de RCs $[\mathcal{R}_1, \ldots, \mathcal{R}_m]$ est calculé par le produit des productivités de chaque RC :

$$score([\mathcal{R}_1, \ldots, \mathcal{R}_m]) = \prod_{i=1}^{m} prod(\mathcal{R}_i)$$

Le mot b retenu à l'étape 3 est ajouté à l'ADM et l'arc entre n et b est étiqueté par la séquence de RCs $[\mathcal{R}_1, \ldots, \mathcal{R}_m]$ du chemin entre n et b ayant le plus haut score. Dans notre exemple, la forme `disarm` totalisant un score de $0,96$ est celle qui est sélectionnée à l'étape 3, et la séquence $[\langle \star\texttt{med} \rightarrow \star\texttt{m} \rangle]$ étiquette dans l'ADM l'arc `disarmed` $\rightarrow$ `disarm`.

4. L'ordre dans lequel les mots de $\mathcal{L}$ sont choisis n'a pas d'importance.

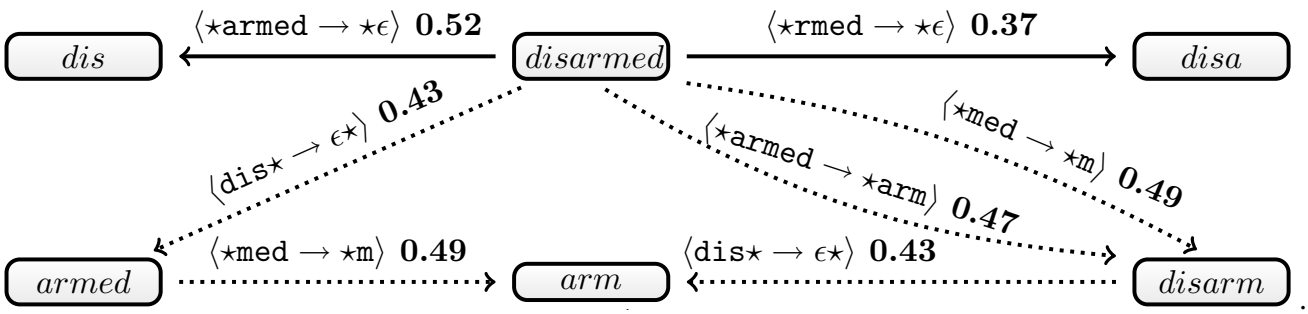

FIGURE 3 – Graphe des mots atteignables depuis le mot anglais *disarmed*. Les arcs en pointillés sont ceux considérés lors du calcul du score entre *disarmed* et *arm*.

2.3.1 Segmentation en morphèmes

Chaque noeud de l'ADM contient la segmentation en morphèmes (potentiels) du mot qui lui est associé. Dans le cas de la racine, le seul morphème est le mot lui-même. Pour tout autre noeud n, l'ensemble des morphèmes est constitué en regroupant les morphèmes du noeud parent p et de ceux des RCs étiquetant l'arc $p \rightarrow n$. Par exemple, dans le cas d'un ADM qui contient l'arc *disarmed* $\rightarrow$ *arm* étiqueté par [$\langle$*dis*$\star \rightarrow \epsilon\star\rangle$,$\langle\star$*ed* $\rightarrow \star\epsilon\rangle$], les morphèmes de *disarmed* sont [*arm*,*dis*,*ed*] ; *dis* et *ed* provenant de leur RC respective. Ce procédé simple en apparence implique un algorithme assez compliqué que nous ne détaillons pas, faute de place. Un point intéressant de notre approche est qu'elle permet de gérer des cas plus complexes que la simple segmentation. Par exemple, l'analyse fournie pour le mot *abilities* est [*ability*,*ies*] car l'arc entre *abilities* et *ability* est étiqueté par [$\langle\star$*ies* $\rightarrow \star y\rangle$].

3 Expériences

3.1 Tâche

Les expériences décrites ici ont été réalisées en partie dans le cadre de notre participation à Morpho Challenge 2009. La tâche à laquelle nous nous intéressons (tâche 1 de Morpho Challenge) consiste à identifier dans un lexique les paires de mots reliés morphologiquement, c'est-à-dire les mots partageant au moins un morphème. Un lexique de référence connu des seuls organisateurs liste l'ensemble des paires de mots reliés et le nombre de morphèmes qu'ils partagent. La qualité des systèmes est évaluée en terme de précision, rappel et f-mesure selon une procédure d'échantillonnage décrite dans (Kurimo & Varjokallio, 2008). Un crédit supérieur est donné au système lorsqu'il identifie le même nombre de morphèmes partagés par une paire de mots que dans la référence. Cette procédure d'évaluation ne fait aucunement usage de l'identité des morphèmes puisque les approches évaluées sont non supervisées, ce qui exclue à priori un jeu de morphèmes commun aux différentes approches.

Cette tâche diffère d'une tâche de segmentation en morphèmes. Par exemple, les segmentations *chien+s* et *hibou+x* seraient valides dans le cas d'une évaluation par segmentation, tandis qu'ici, il faut reconnaître que *s* et *x* sont 2 représentations du même morphème "pluriel d'un nom". De plus, il faut différencier les morphèmes partageant la même représentation comme *s* qui marque aussi bien le pluriel d'un nom en français que la deuxième personne du singulier d'un verbe.

3.2 Données

Nous avons étudié notre système sur les lexiques de 5 langues (anglais, allemand, turc, finnois et arabe) mis à disposition aux participants de Morpho Challenge 2009. Comme le montre la table 2, la taille de ces lexiques est très variable. Le lexique finnois comporte plus de 2 millions de mots alors que le lexique arabe en contient un peu moins de 15 000. Nous signalons que ces lexiques, extraits de textes réels, contiennent de nombreux noms propres ainsi que des erreurs typographiques qui compliquent la tâche d'analyse.

TABLE 2 – Taille des lexiques de Morpho Challenge 2009.

	ANG.	ALL.	TUR.	FIN.	ARB.
Nb. de mots	384 903	1 266 159	617 298	2 206 719	14 957

L'évaluation de notre système pour ces 5 langues a été effectuée par les organisateurs de Morpho Challenge 2009. Après la campagne, nous avons évalué l'impact des méta-paramètres contrôlant notre modèle à l'aide de références extraites de la base de données de Celex (Baayen *et al.*, 1995) pour l'allemand et l'anglais. La table 3 fait état des principales caractéristiques des deux références que nous avons mises au point.

TABLE 3 – Principales caractéristiques de la référence Celex. De gauche à droite : la taille en nombre de mots du lexique, le nombre de morphèmes distincts, la moyenne du nombre d'analyses possibles pour un mot, le nombre de morphèmes moyens pour une analyse, le nombre moyen de mots avec lequel un mot partage au moins 1 morphème.

	Nb. Mots	Nb. de type de morphème	analyses/mot	morphèmes/analyse	paires/mot
ANG.	72 628	16 388	1.07	2.15	21.93
ALL.	311 000	13 102	1.23	3.35	327.75

3.3 Résultats

Pour chaque langue, nous avons exécuté notre moteur analogique pendant une semaine sur les lexiques fournis par les organisateurs de Morpho Challenge 2009. Nous avons ainsi recueilli un grand nombre d'analogies : de 11 (arabe) à 52 (turc) millions d'analogies ont ainsi été extraites par langue. Bien que ces quantités puissent paraître élevées, il est important de noter que les analogies que nous avons calculées ne constituent qu'une faible partie de celles existant au sein d'un lexique. De ces ensembles d'analogies ont ensuite été extraites les RCs utilisées par notre système.

La table 4 montre les résultats obtenus par notre système pour la référence Celex (anglais et allemand) selon les différentes valeurs de méta-paramètres qui le contrôlent : ρ (la productivité minimale d'une RC) et τ (le score minimum qu'un chemin entre deux mots doit avoir pour que ces mots soient regroupés dans un même ADM). Comme on pouvait s'y attendre, augmenter le seuil τ ou la productivité minimale ρ améliore la précision au détriment du rappel. La valeur optimale du seuil τ pour l'anglais est de 0,3

contrairement à 0,25 pour l'allemand. Cette différence s'explique par le fait qu'en moyenne, le nombre de mots en relation est plus élevé pour l'allemand que pour l'anglais (voir la table 3). Donc en réduisant le seuil, plus de relations sont créés ce qui avantage les langues morphologiquement plus complexes.

TABLE 4 – Impact des méta-paramètres ρ et τ sur la Précision (Pr.), le Rappel (Rp.) et la F-Mesure (F1) de notre système. Les analyses soumises à Morpho Challenge 2009 sont soulignés et les meilleurs résultats sont en gras.

		$\rho = 0.20$			$\rho = 0.25$			$\rho = 0.30$		
		Pr.	Rc.	F1	Pr.	Rc.	F1	Pr.	Rc.	F1
$\tau = 0.20$	ANG.	51.98	52.13	51.81	59.55	47.94	52.92	66.43	44.8	53.32
	ALL.	60.01	42.59	49.39	66.99	37.71	47.81	70.72	32.11	43.71
$\tau = 0.25$	ANG.	61.02	48.85	54.04	60.59	47.9	53.31	65.67	44.76	53.05
	ALL.	**68.89**	**41.66**	**52.24**	67.07	40.15	49.61	70.56	32.11	43.70
$\tau = 0.30$	**ANG.**	**65.97**	**46.75**	**54.52**	66.20	45.29	53.59	66.97	44.72	53.46
	ALL.	68.89	41.66	51.38	69.87	36.24	47.02	71.03	32.11	43.82
$\tau = 0.35$	ANG.	71.46	43.09	53.54	71.27	40.81	51.65	72.45	39.62	50.94
	ALL.	70.66	33.42	44.75	72.14	30.01	41.60	74.05	25.99	37.89
$\tau = 0.40$	ANG.	74.22	41.15	52.76	76.10	38.88	51.26	76.72	37.61	50.25
	ALL.	71.48	32.43	44.00	74.99	28.98	41.09	75.99	24.38	36.30

La table 5 donne les résultats officiels obtenus à Morpho Challenge 2009 par les deux systèmes *Rali-Cof* et *Rali-Ana* que nous avons soumis aux organisateurs pour les 5 langues, ainsi que le système *Morfessor* [5], un système jouissant d'une large distribution et qui est fréquemment employé comme référence dans le domaine. Le premier système est celui que nous avons décrit dans cette étude avec un seuil de productivité ρ de 0.25 et un seuil de score τ de 0,3 pour toutes les langues. Le deuxième système est celui proposé par Langlais (2009). Nous soulignons que ces deux systèmes font usage du même ensemble d'analogies décrit plus haut.

On observe que notre système *Rali-Cof* surpasse *Rali-Ana* et *Morfessor* sur trois des cinq langues : l'allemand, le finnois et le turc. Le faible rappel obtenu par *Rali-Ana* s'explique par le fait qu'une petite partie seulement des analogies a été calculée (voir la section 3.2).

Nous observons aussi que les résultats de *Rali-Cof* sont similaires pour toutes les langues à l'exception de l'arabe qui a un rappel très faible. Il se pourrait que ce soit causé par la très petite taille du lexique fourni. Comme l'apprentissage analogique se base sur les schémas observés pour identifier les morphèmes, plusieurs morphèmes valides ont pu être mis de coté dû à leur faible fréquence dans l'ensemble de test.

Il est difficile d'attribuer un rang global à un système dans la compétition Morpho Challenge, mais nous pouvons tout de même faire les observations suivantes. Premièrement, aucun compétiteur n'a réussi à obtenir des résultats satisfaisants sur l'arabe. Deuxièmement, notre approche a terminé parmi les meilleures pour les autres langues ; exception faite de l'anglais où plusieurs systèmes (dont *Morfessor*) sont très bons. *Paramor* est le seul système nous ayant surpassé sur ces 4 langues [6]. Troisièmement, *Rali-Cof* a obtenu des résultats supérieurs à ceux du système à base de graphe *MorphoNet* (Bernhard, 2010) sur toutes les

5. Nous avons utilisé la version 1.0 disponible sur http ://www.cis.hut.fi/projects/morpho/.
6. Nos résultats ont été produits sans ajustement des méta-paramètres.

langues autre que l'arabe. Comme ces deux systèmes se rejoignent sur plusieurs points, ceci laisse présager que l'analogie prodigue un gain réel. Finalement, malgré le fait que *Morfessor* a de meilleurs résultats pour l'anglais, notre approche le surpasse sur les langues morphologiquement plus complexes. Ceci est particulièrement important car les bénéfices potentiels de l'analyse morphologique sont plus importants pour ces langues.

TABLE 5 – Précision (Pr.), Rappel (Rp.) et f-mesure (F1) des systèmes analogiques et du système de référence *Morfessor* à l'atelier Morpho Challenge 2009.

	Rali-Cof			*Rali-Ana*			*Morfessor*		
	Pr.	Rp.	F1	Pr.	Rp.	F1	Pr.	Rp.	F1
ANG.	68.32	46.45	55.30	64.61	33.48	44.10	**74.93**	**49.81**	**59.84**
FIN.	74.76	**26.20**	**38.81**	60.06	10.33	17.63	**89.41**	15.73	26.75
TUR.	48.43	**44.54**	**46.40**	69.52	12.85	21.69	**89.68**	17.78	29.67
ALL.	67.53	**34.38**	**45.57**	61.39	15.34	24.55	**81.70**	22.98	35.87
ARB.	**94.56**	2.13	4.18	92.40	4.40	8.41	91.77	**6.44**	**12.03**

4 Conclusion

Nous avons présenté le système *Rali-Cof* que nous avons développé dans le cadre de Morpho Challenge 2009. Bien que *Rali-Cof* et *Rali-Ana* soient tous les deux basés sur le principe d'analogie formelle, *Rali-Cof* se différencie par sa capacité à généraliser l'information capturée passivement par une analogie. Un des avantages de cette abstraction opérée par les RCs est que *Rali-Cof* n'utilise qu'un (petit) sous-ensemble des analogies impliquant les mots d'un lexique et est donc applicable à des lexiques beaucoup plus volumineux que ceux analysables par le système de segmentation suggéré par Langlais (2009).

Notre système surpasse le système *Morfessor* pour le turc, le finnois et l'allemand, trois langues à la morphologie complexe. Ces résultats appuient donc notre hypothèse que l'analogie formelle peut être utilisée efficacement pour réaliser de manière non supervisée l'analyse morphologique d'une langue.

Nous avons également montré qu'il est possible d'influencer le niveau de précision et de rappel de notre système en modifiant les valeurs de ses méta-paramètres. Ceci permet d'adapter nos analyses à la tâche à accomplir.

Plusieurs pistes de recherche restent à explorer à l'issue de ce travail. Premièrement, même si les RCs capturent plus d'information contextuelle que les cofacteurs, d'autres alternatives pourraient être considérées, comme les expressions régulières utilisées par Bernhard (2010). Il reste cependant à vérifier que ce type de règle n'introduit pas trop de bruit. Deuxièmement, nous avons observé une tendance à la sursegmentation des ADM que nous produisons. Nous sommes confiants que nous pouvons améliorer la qualité de ces regroupements. Enfin, nous souhaitons étudier l'impact de la quantité d'analogies extraites d'un lexique sur les performances de notre système.

Remerciements

Cette étude a été partiellement financée par le programme MITACS.

Références

BAAYEN R. H., PIEPENBROCK R. & GULIKERS L. (1995). The CELEX lexical database (release 2). CD-ROM, Linguistic Data Consortium, Univ. of Pennsylvania, USA.

BERNHARD D. (2010). Morphonet : Exploring the use of community structure for unsupervised morpheme analysis. In *10th CLEF Workshop*, Corfu,Greece.

CREUTZ M. & LAGUS K. (2005). Unsupervised morpheme segmentation and morphology induction from text corpora using morfessor 1.0. In *Publications in Computer and Information Science, Report A81*, Helsinki University of Technology.

DENOUAL E. (2007). Analogical translation of unknown words in a statistical machine translation framework. In *MT Summit, XI*, Copenhagen.

GOLDSMITH J. (2001). Unsupervised learning of the morphology of a natural language. *Computational Linguistics*, **27**, 153–198.

HARRIS Z. S. (1955). From phoneme to morpheme. *Language*, **31**(2), 190–222.

HATHOUT N. (2002). From wordnet to celex : acquiring morphological links from dictionaries of synonyms. In *3rd LREC*, p. 1478–1484, Las Palmas de Gran Canaria.

HATHOUT N. (2008). Acquistion of the morphological structure of the lexicon based on lexical similarity and formal analogy. In *3rd Textgraphs workshop*, p. 1–8, Manchester, United Kingdom.

KURIMO M. & VARJOKALLIO M. (2008). Unsupervised morpheme analysis evaluation by a comparison to a linguistic gold standard – Morpho Challenge 2008. In *CLEF 2008 Workshop*.

LANGLAIS P. (2009). Étude quantitative de liens entre l'analogie formelle et la morphologie constructionnelle. In *16è Conférence sur le Traitement Automatique des Langues Naturelles (TALN'09)*, Senlis, France.

LANGLAIS P. & PATRY A. (2007). Translating unknown words by analogical learning. In *EMNLP-CoNLL*, p. 877–886, Prague, Czech Republic.

LANGLAIS P. & YVON F. (2008). *Scaling Up Analogical Learning*. Rapport interne, Paritech, INFRES, IC2, Paris, France.

LEPAGE Y. & DENOUAL E. (2005). Purest ever example-based machine translation : Detailed presentation and assessment. *Machine Translation*, **29**, 251–282.

MONSON C., CARBONELL J., LAVIE A. & LEVIN L. (2007). Paramor : Minimally supervised induction of paradigm structure and morphological analysis. In *Proceedings of 9th SIGMORPHON Workshop*, p. 117–125, Prague, Czech Republic : ACL.

PIRRELLI V. & YVON F. (1999). The hidden dimension : a paradigmatic view of data-driven NLP. *Journal of Experimental & Theroretical Artifical Intelligence*, **11**, 391–408.

STROPPA N. (2005). *Définitions et caractérisations de modèles à base d'analogies pour l'apprentissage automatique des langues naturelles*. PhD thesis, ENST, ParisTech, Télécom, Paris, France.

STROPPA N. & YVON F. (2005). An analogical learner for morphological analysis. In *CoNLL*, p. 120–127, Ann Arbor, MI.

YAROWSKY D. & WICENTOWSKI R. (2000). Minimally supervised morphological analysis by multimodal alignment. In *ACL '00 : Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, p. 207–216, Morristown, NJ, USA : Association for Computational Linguistics.

YVON F., STROPPA N., DELHAY A. & MICLET L. (2004). *Solving analogical equations on words*. Rapport interne D005, ENST, Paris, France.