Intérêt des ressources morphologiques pour la recherche d'information précise

Anne-Laure Ligozat, Delphine Tribout, Brigitte Grau

Intérêt des ressources morphologiques pour la recherche d'information précise

Anne-Laure Ligozat* — **Delphine Tribout**** — **Brigitte Grau***

** LIMSI-CNRS, 91403 Orsay cedex, ENSIIE, 91000 Evry*
*** LLF, 75013 Paris*

RÉSUMÉ. *Cet article présente la construction automatique, le filtrage et la validation d'une ressource morphologique concernant les noms d'agents déverbaux. Cette validation utilise différentes ressources et corpus pour tester l'appartenance des verbes et noms à la même famille morphologique, ainsi que leur lien, méthode qui peut se généraliser à d'autres ressources du même type. Hormis une méthode de construction et d'aide à la validation d'une ressource, nous montrerons l'intérêt de disposer de ressources morphologiques pour la recherche de courts passages en questions-réponses.*

ABSTRACT. *This paper presents how to build a reliable morphological knowledge base, by mean of an automatic generation, the filtering of the obtained words and their validation. The generated base concerns verbs and their agents. The validation method relies on different resources and corpora in order to verify that a verb and a noun belong to the same family, and are related by an agent link. The method can be generalized to other kinds of resources. We will also show in what extend using such morphological resources is useful for selecting passages in a question answering system.*

MOTS-CLÉS : *ressources morphologiques, validation de ressource, système de question-réponse*

KEYWORDS: *morphological resources, resources validation, question-answering system*

1. Introduction

Des connaissances morphologiques sont utilisées dans de nombreuses applications de traitement automatique des langues et de recherche d'information. L'apport de ces connaissances a été évalué dans plusieurs types d'applications, comme la reconnaissance de la parole (Creutz *et al.,* 2007), la traduction automatique (Koehn *et al.,* 2007) ou la recherche d'information (Hahn *et al.,* 2003). Les systèmes de questions-réponses visent à retourner une réponse précise à une question posée en langage naturel. Dans un tel système, des connaissances morphologiques peuvent être utilisées à différents niveaux du processus de réponse aux questions.

Dans cet article, nous présenterons les différentes types de morphologie impliqués en questions-réponses, les ressources disponibles pour le français, et en particulier, une ressource que nous avons construite contenant des noms d'agents déverbaux. Puis, nous évaluerons les performances d'un système particulier, QAVAL, en fonction des ressources morphologiques utilisées.

2. Morphologie en QA

2.1. *Types de relations morphologiques entre question et documents*

Ce travail est dans la continuité de celui présenté par (Bernhard *et al.,* 2011), qui ont étudié les relations dérivationnelles les plus fréquentes entre des questions et des documents réponses en français. Entre une question et des documents contenant sa réponse, des variations peuvent en effet exister. Ainsi, pour la question «Quand la dynamite a-t-elle été inventée ?», une réponse possible est «Alfred Nobel fait breveter cette invention le 25 novembre 1867, sous le nom de dynamite.». Pour relier question et document réponse, il faut notamment pouvoir détecter la relation (morphologique) entre «inventée» et «invention».

(Bernhard *et al.,* 2011) ont annoté un corpus de questions et passages (extraits de documents) réponses, en notant le type de relation morphologique présente entre les deux textes, et ont ainsi montré que les relations les plus présentes dans un corpus de domaine général sont des relations flexionnelles et dérivationnelles. En ce qui concerne les dérivations, les types les plus courants en domaine général sont les adjectifs dénominaux (région/régional), et les nominalisations, en particulier les noms d'action (inaugurer/inauguration) et les noms d'agents (réaliser/réalisateur).

Nous avons souhaité mettre à profit ces observations dans un système de questions-réponses, et étudier si l'ajout des connaissances morphologiques appropriées pouvait améliorer les performances du système.

2.2. *Utilisation de connaissances morphologiques dans QAVAL*

2.2.1. *Description générale du système QAVAL*

QAVAL (Grappy *et al.,* 2011) est un système de questions-réponses développé pour le français. Il utilise le moteur de recherche Lucene [1] pour sélectionner de courts passages (au lieu de documents), qui sont ensuite analysés par un analyseur terminologique de surface, Fastr (Jacquemin, 1999) ce qui permet de les pondérer en fonction des termes trouvés et de n'en retenir qu'un sous-ensemble. Des réponses candidates sont ensuite extraites de ces passages et différents critères sont ensuite appliqués par un système de validation de réponse par apprentissage afin d'ordonner ces réponses. Fastr permet d'annoter des termes complexes et leurs variations, celles-ci étant décrites par un ensemble de règles, s'appuyant sur des lexiques. La reconnaissance de variantes de termes complexes permet de restreindre les variations dans ces contextes d'utilisation. Par ailleurs, nous annotons aussi les passages par les variantes des termes simples.

2.3. *Intégration de connaissances morphologiques dans le processus de réponse aux questions*

Les variations morphologiques sont gérées à deux niveaux : lors de l'interrogation de la collection et lors de la sélection des passages, quand les passages sont annotés par Fastr. La collection est indexée en utilisant le module de stemming présent dans Lucene. L'interrogation est ensuite faite en gardant les mots pleins des questions, qui subissent eux-aussi un stemming, mais où les noms propres sont aussi gardés tels quels. De cette manière les passages renvoyés sont susceptibles de contenir des variations morphologiques des termes des questions. De manière à étudier l'effet de l'utilisation de ressources morphologiques, nous avons annoté les passages avec les mots des questions uniquement, plus précisément avec les lemmes, et avec les mots et variantes dérivationnelles. L'étude présentée dans la section résultat compare les performances du module de sélection de passage dans les deux cas de figure.

3. Ressources dérivationnelles pour le français

Verbaction [2] est une ressource lexicale regroupant tous les noms d'actions dérivés d'un verbe (Hathout *et al.,* 2002a, Hathout *et al.,* 2002b). Elle contient un total de 9 393 paires de nom-verbe.

Prolexbase [3] est un dictionnaire multilingue de noms propres (Tran *et al.,* 2006, Bouchou *et al.,* 2008). Bien qu'elle ne contienne pas explicitement de connaissances morphologiques, cette ressource fournit des informations sur les noms relationnels

1. http ://lucene.apache.org
2. *http ://w3.erss.univ-tlse2.fr :8080/index.jsp ?perso=hathout&subURL=verbaction*
3. *http ://www.cnrtl.fr/lexiques/prolex/*

et les adjectifs associés aux noms propres. Par exemple, *Français* et *français* sont explicitement associés à l'entrée *France*. Au total, Prolexbase contient 76 118 lemmes et 20 614 relations dérivationnelles.

Des ressources dérivationnelles existent donc pour le français, mais, par rapport aux types de dérivation observés en corpus de questions-réponses, une ressource concernant les noms d'agents déverbaux fait défaut. Nous avons donc construit semi-automatiquement une telle ressource.

4. Construction et validation d'une ressource de noms d'agents déverbaux

4.1. *Construction d'une ressource par heuristiques*

La méthode de détection automatique de noms d'agent déverbaux repose exclusivement sur les propriétés formelles des noms. En français, on peut en effet identifier certains suffixes qui semblent corrélés à la formation de noms d'agent déverbaux, par exemple le suffixe *-eur*, comme dans *danseur* dérivé du verbe *danser*, ou le suffixe *-ant*, comme dans *dirigeant* dérivé du verbe *diriger*. Nous avons identifié neuf suffixes liés à des règles de formation de noms d'agent déverbaux :

1) *-eur* (*danser* > *danseur*)
2) *-euse* (*chanter* > *chanteuse*)
3) *-rice* (*inspecter* > *inspectrice*)
4) *-eresse* (*défendre* > *défenderesse*)
5) *-aire* (*signer* > *signataire*)
6) *-ant* (*attaquer* > *attaquant*)
7) *-ante* (*diriger* > *dirigeante*)
8) *-ent* (*adhérer* > *adhérent*)
9) *-ente* (*présider* > *présidente*)

Pour récupérer les noms d'agent sur la base des propriétés formelles des noms nous avons utilisé le lexique Morphalou [4]. Celui-ci est un lexique librement accessible de formes fléchies du français, constitué automatiquement à partir de la nomenclature du TLF. Il contient 539 413 formes fléchies correspondant à 68 075 lemmes. La liste de noms d'agent déverbaux a été constituée en deux temps. Nous avons tout d'abord récupéré tous les noms de Morphalou se terminant par l'un des neufs suffixes présentés ci-dessus. Puis nous avons vérifié, pour chaque nom, qu'un verbe formellement proche existait dans le lexique Morphalou. La vérification a été effectuée au moyen d'heuristiques basées sur la forme des noms et des verbes. Par exemple lorsque le nom se termine par -eur la règle la plus générale permettant d'obtenir le verbe dont il dérive est la suivante :

4. http ://www.cnrtl.fr/lexiques/morphalou/

supprimer le suffixe *-eur* puis ajouter *-er*

Cette règle permet par exemple de récupérer le verbe *chanter* à partir du nom *chanteur*. D'autres règles sont nécessaires pour rendre compte de relations formellement plus complexes entre le nom et le verbe, comme pour le nom *formateur* et le verbe *former*, ou le nom *finisseur* et le verbe *finir*, gérées par les règles suivantes :

supprimer le suffixe -ateur puis ajouter -er

supprimer le suffixe -isseur puis ajouter -ir

Au total une vingtaine de règles ont été établies, grâce auxquelles 4 067 paires nom-verbe dont le nom se termine par l'un des suffixes mentionnés ci-dessus ont été récupérées. Comme cela a été mentionné plus haut, cette méthode de récupération des noms d'agent déverbaux pose quelques problèmes. En effet, une ressemblance formelle entre un nom et un verbe ne garantit pas que les deux sont morphologiquement reliés. Par exemple la paire *accentuer*/*accentueur* 'oiseau du genre passereau' est récupérée alors que le nom *accentueur* n'est pas morphologiquement lié au verbe *accentuer*, mais dérive du latin *accentor*. Dans d'autres cas le nom et le verbe appartiennent bien à la même famille dérivationnelle, mais le nom n'est pas dérivé du verbe. C'est le cas par exemple de la paire *rougir*/*rougeur* qui est récupérée par l'une des heuristiques établies. Dans ce cas, le nom et le verbe sont bien morphologiquement liés, mais ils ne le sont pas directement : ils dérivent tous deux de l'adjectif *rouge*. À l'issue de cette étape de constitution de la ressource, une validation des paires récupérées est donc nécessaire. Cette validation a été commencée manuellement, mais elle est très coûteuse en temps. Nous avons donc également essayé de mettre en œuvre des méthodes d'enrichissement de notre ressource, afin d'automatiser la validation.

4.2. *Validation manuelle*

En un premier temps nous avons vérifié manuellement que le nom était effectivement dérivé du verbe et qu'il désignait bien un agent. La vérification du lien sémantique entre le nom et le verbe a été réalisée grâce au TLFi lorsque le nom était trop rare ou inconnu de nous, par exemple pour *amodiateur* "propriétaire qui cède une terre, une exploitation rurale par amodiation", dérivé du verbe *amodier* "donner à ferme un bien foncier, une exploitation rurale".

363 paires ont ainsi été examinées. La validation manuelle de l'échantillon a révélé que 76% des paires de VerbAgent étaient correctes, c'est-à-dire qu'elles étaient bien constituées d'un verbe et d'un nom d'agent dérivé. 24% des paires étaient en revanche incorrectes.

Parmi les erreurs, il est notable que la moitié est constituée de noms en *-ant* ou en *-aire*, qui sont bien dérivés du verbe, mais qui ne dénotent pas un agent, comme *adoucissant* ou *aliénataire*, dérivés respectivement de *adoucir* et *aliéner*. Il est possible que les heuristiques de récupération des noms d'agent incluant ces deux suffixes ne soient pas assez contraignantes d'un point de vue sémantique. Nous les avions pour-

tant incluses afin de ne pas perdre des noms d'agent comme *dirigeant* ou *signataire*. Cependant il est évident que l'inclusion de ces suffixes engendre du bruit, que nous espérons toutefois éliminer grâce aux autres méthodes de validation. Quant aux autres paires erronées, il s'agit dans 19% des cas de noms en *-eur* qui sont bien déverbaux mais qui dénotent un instrument, comme *accélérateur* ou *aspirateur*. Enfin, les 31% restants sont des erreurs d'analyse comme *actionner/actionnaire* ou *aigrir/aigreur*.

Cette validation manuelle est relativement fiable mais nécessiterait le travail de plusieurs personnes et la confrontation de leurs différentes validations, afin de minimiser au maximum les erreurs de jugement personnel. Cependant une telle validation serait très coûteuse. C'est pourquoi, sur la base de la partie validée manuellement, nous avons réfléchi à la mise au point d'une méthode de validation automatique qui nous permettrait de limiter de manière automatique les erreurs engendrées par les heuristiques formelles, et de réduire ainsi la validation manuelle qui restera certainement à faire. Pour cela nous avons recherché les paires verbe-nom créées de manière automatique et validées manuellement, dans d'autres ressources. Nous avons vu que les problèmes qui se posent lors de la validation sont de deux types : 1) les mots sont de la même famille mais ne sont pas liés par une relation agent ; 2) les mots ne sont pas de la même famille. Les ressources que nous avons utilisées permettent de traiter ces problèmes à la fois, ou seulement l'un des deux.

4.3. *Extraction des définitions*

Les agents sont généralement définis par rapport à l'action qu'ils permettent de faire, d'où l'idée d'exploiter les définitions de dictionnaire. Nous avons utilisé le XMLittré [5], une version électronique du Littré présentée dans un format XML. Cette ressource contient les données du dictionnaire de la langue française d'Emile Littré, qui comprend 78 423 entrées, et, pour chacune, différentes informations comme la prononciation, la nature, et plusieurs définitions (appelées variantes).

L'extraction de noms d'agent déverbaux à partir des définitions du Littré s'est faite en deux étapes. Dans un premier temps nous nous sommes basés uniquement sur la sémantique des définitions. Puis, nous avons ajouté une contrainte morphologique au patron de définition des noms d'agent, afin d'être sûrs de ne récupérer que les noms d'agents déverbaux.

Pour extraire de façon automatique les noms d'agent d'après leurs définitions, nous avons, lors de la première étape, uniquement pris en compte la sémantique des noms d'agent. Pour cela nous avons tout d'abord dû repérer la façon dont sont généralement définis les noms d'agent dans le dictionnaire. Nous avons donc étudié les définitions de noms d'agent prototypiques, comme chanteur, danseur, président, dirigeant... ce qui nous a permis de définir deux patrons de définition des noms d'agent : "Celui, celle qui" ou "Celui qui" suivi généralement du verbe base. Ainsi, pour le nom d'agent chanteur, l'une des définitions est : "Celui, celle qui chante, qui fait métier de chanter".

5. http ://francois.gannaz.free.fr/Littre/

Grâce à ces patrons de définition nous avons extrait 2 944 noms. Cependant, comme cela a été mentionné plus haut, le patron de définition des noms d'agent ne garantit pas que le nom est réellement dérivé du verbe qui suit "Celui, celle qui" dans la définition. Par exemple cette méthode d'extraction a retourné des noms d'humains qui ne sont pas des noms déverbaux, comme académicien dont l'une des définitions commence par "Celui qui fait partie d'une société de gens de lettres", ou encore pianiste défini comme "Celui, celle qui joue du piano". C'est pourquoi, nous avons ensuite restreint les noms extraits lors de la première étape, en ajoutant une contrainte morphologique entre le verbe suivant *qui* dans la définition et le nom vedette. En réalité, cette contrainte était formelle plus que morphologique, car elle exigeait simplement que les deux premiers caractères du nom et du verbe soient identiques. Cette seconde étape nous a permis de rejeter les noms comme *académicien* et *pianiste*, dont le verbe suivant *qui* dans la définition ne commence pas par les deux mêmes caractères que le nom, respectivement *ac* et *pi*, mais par *fa* et *jo*. Cette seconde extraction nous a permis de recueillir 1 121 noms. Certes, cette liste de noms d'agents obtenue après la seconde étape est plus restreinte, et comporte nécessairement des manques. Ainsi, le nom *agresseur* défini comme "Celui qui attaque le premier" n'est pas récupéré parce que sa définition ne correspond pas à la contrainte formelle rajoutée lors de la deuxième étape, alors qu'il s'agit bien d'un nom d'agent dérivé du verbe *agresser*. Mais on peut supposer qu'elle sera plus précise, ce que nous confirmerons par comparaison avec une partie validée manuellement de notre ressource. Cette liste extraite du Littré devrait nous permettre à la fois de valider les paires verbe-nom établies avec la première méthode, et de les compléter éventuellement avec des noms d'agents qui ne correspondraient pas aux heuristiques ayant permis de récupérer les paires.

Nous avons comparé les paires créées par heuristiques et validées manuellement avec les noms d'agents extraits du Littré lors de la première étape, c'est-à-dire sans la contrainte formelle. Cela a fait ressortir 92 noms communs aux deux méthodes de construction de la ressource. Sur ces 92 noms, 87 sont des noms ayant été considérés, lors de la validation manuelle, comme des noms d'agent déverbaux. Nous avons ensuite comparé les paires créées par heuristiques avec les noms d'agents extraits du Littré lors de la seconde étape, c'est-à-dire avec la contrainte formelle entre le nom et le verbe. Nous avons alors obtenu 60 noms communs aux deux méthodes de constitution de la ressource. Mais ces 60 noms étaient tous des noms validés comme corrects lors de la validation manuelle. Les données de ces deux comparaison sont résumées dans le tableau 1.

Si l'on compare ces résultats avec la validation manuelle de l'échantillon de VerbAgent, qui comporte 275 couples verbe-nom corrects, cette validation automatique par comparaison avec les données extraites des définitions du Littré ne présente donc pas un très bon rappel. En effet, celui-ci est d'environ 22% pour le second patron. En revanche cette validation est très précise. Le faible rappel s'explique par le fait que certaines définitions de noms d'agents déverbaux ne suivent pas les patrons que nous avons spécifiés, comme *agresseur* par exemple. Mais il s'explique aussi grandement par le fait que certains noms d'agents sont absents du Littré, comme *avaliseur*.

	Patron de définition "Celui qui, celle qui" ou "Celui qui" uniquement	Ajout de la contrainte formelle entre le verbe et le nom
Nombre de mots en commun	92	60
Nombre de noms d'agents déverbaux en commun	87	60

Tableau 1. *Comparaison des résultats obtenus par l'extraction du Littré avec la méthode à base d'heuristiques*

4.4. *Comparaison de contextes sous forme de cooccurrents*

Afin dévaluer si deux mots sont sémantiquement proches, et donc appartiennent à la même famille morphologique, nous pouvons nous appuyer sur leurs contextes d'utilisation. Aussi, une autre ressource qu'il nous a semblé intéressant d'exploiter dans ce cadre, et qui était à notre disposition pour le français, est un réseau de cooccurrences lexicales construit à partir de corpus du journal Le Monde (Ferret, 1998 :281-288). Ce réseau a été construit sur un corpus de 24 mois du Monde, en utilisant une fenêtre de 20 mots, et en ne tenant pas compte de l'ordre au sein des cooccurrences. Seules les cooccurrences de fréquence supérieure à 5 ont été conservées, de sorte que le réseau contient 31 000 mots. Une mesure de cohésion entre deux mots est calculée par estimation de l'information mutuelle. Les cooccurrents d'un mot sont ensuite classés par ordre décroissant de leur valeur de cohésion. Notre hypothèse est que si une paire verbe-nom possède des cooccurrents communs, elle sera reliée sémantiquement, et sera donc plus susceptible d'être issue d'une dérivation.

Nous avons donc extrait, pour chaque paire verbe-nom, leurs cooccurrents les plus proches [6], et considéré qu'une paire était reliée si elle avait au moins un cooccurrent en commun. Le principal inconvénient de cette méthode est que la taille du corpus est limitée, et de nombreux mots sont absents du réseau. Ainsi, sur l'ensemble de la ressource, seules 869 paires sont retrouvées dans le réseau, c'est-à-dire qu'il n'y a que 869 paires pour lesquelles à la fois le verbe et le nom apparaissent dans le corpus. Afin d'évaluer la pertinence des cooccurrences, nous avons comparé les paires présentant au moins un cooccurrent commun avec la partie validée de VerbAgent. 85 paires ont été trouvées dans le réseau de cooccurrents, parmi lesquelles 56 ont un cooccurrent commun. Sur ces 56 paires, 45 ont effectivement été validées comme correctes dans VerbAgent, et 11 n'ont pas été validées, comme *accablant-accabler* ou *accélérateur-accélérer*. On peut noter que ces paires sont bien reliées sémantiquement et morphologiquement, mais que les noms ne correspondent pas à des noms d'agents. Cette

6. Les noms féminins ont été passés au masculin si nécessaire, car le TreeTagger, utilisé pour construire le réseau de cooccurrences, lemmatise ainsi. En revanche, les noms pouvant être des participes présents, comme *tranchant* n'ont pas été modifiés.

méthode semble donc donner un indice sur la relation entre le nom et le verbe, mais nécessiterait un corpus de plus grande taille pour fournir des résultats plus complets.

4.5. *Usage en contexte : N-grammes de mots*

Une autre idée de validation en contexte repose sur l'usage que l'on peut faire du verbe ou de l'agent pour exprimer une même notion dans un multi-terme, qui correspondent alors à des réécritures. Le but est d'identifier des réécritures du type *chanteur d'opéra* et *chanter un opéra*.

Nous avons de ce fait utilisé des n-grammes de mots pour déterminer s'ils pouvaient permettre de valider les paires verbe-nom constituées de manière automatique par les heuristiques. Le corpus utilisé pour cette étude est issu des Google Ngrams, qui comprend des n-grammes de mots extraits de la numérisation de livres.

Pour réaliser la validation, nous avons dans un premier temps constitué des n-grammes de tous les noms et verbes de VerbAgent, puis nous avons comparé les mots apparaissant dans les n-grammes des nom et verbe constituant une paire. Tout d'abord nous avons extrait, pour les noms, tous les trigrammes constitués d'un nom de VerbAgent, suivi du déterminant *du*, *des* ou *de*, et d'un autre mot. Le trigramme a ainsi la forme "nom+du/des/de+mot". Pour les verbes, nous avons extrait tous les trigrammes constitués d'un verbe de VerbAgent, suivi du déterminant *un/une/le/les/des/son/ses*, et d'un autre mot, de sorte que le trigramme a la forme "verbe+un/une/le/les/des/son/ses+mot". Puis, pour chaque paire verbe-nom de VerbAgent, nous avons compté le nombre de termes étant des variantes, c'est-à-dire dont le troisième mot est identique. Nous avons ainsi extrait 1 795 variantes de termes, correspondant à 231 paires. L'évaluation sur l'échantillon validé manuellement de VerbAgent montre que 19 paires sont trouvées grâce à cette méthode, dont une seule n'est pas une paire validée. La méthode semble donc précise, mais elle manque réellement de couverture, et ne semble pas, de ce fait, constituer un bon moyen de valider une ressource constituée de manière automatique.

4.6. *Recouvrement et combinaison des méthodes*

littré	*		*	*	*			
coocs	*	*		*		*		
termes	*	*	*				*	
# paires trouvées	170	191	163	79	790	161	223	2290
validées	15	9	11	7	54	14	16	179
invalidées	0	10	0	0	5	1	12	61

Tableau 2. *Nombre de paires nom-verbe trouvées par chaque méthode*

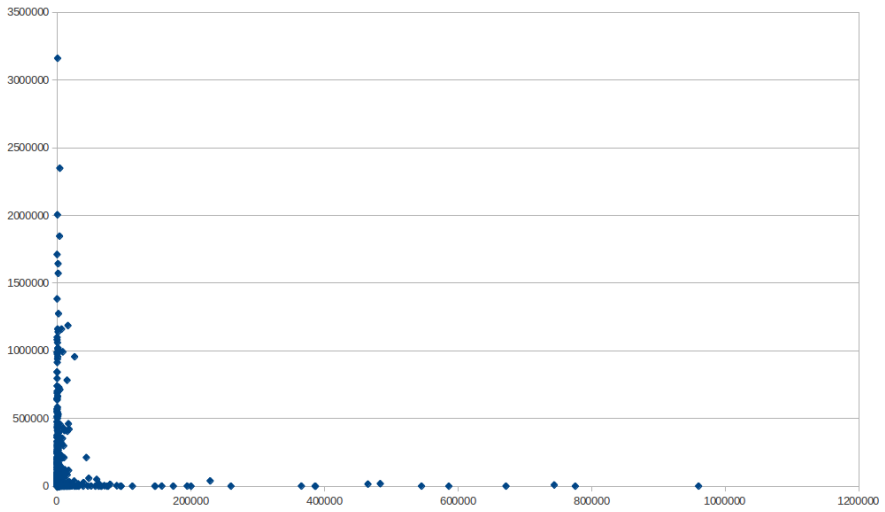

Figure 1. *Distribution des paires non trouvées*

Le tableau 2 présente le nombre de paires nom-verbe validées par chaque méthode, ainsi que par la combinaison de plusieurs méthodes. La seconde colonne indique par exemple que 170 paires ont été validées par les trois méthodes, parmi lesquelles 15 avaient été validées manuellement, et aucune n'avait été invalidée.

On peut constater que les définitions du Littré permettent de trouver de nombreuses paires, avec une bonne précision. Toutefois, 2290 paires ne sont retrouvées par aucune méthode, principalement car la fréquence du verbe ou du nom est trop faible.

La figure 1 montre les fréquences (provenant des Google NGrams) du nom et du verbe des paires non trouvées par les trois méthodes.

Ce graphique montre que pour les paires non trouvées, la fréquence du nom ou du verbe est relativement faible, et que ces méthodes ne sont donc pas adaptées pour les paires possédant un terme peu fréquent.

Une possibilité pour parvenir à valider ces termes peu fréquents serait d'utiliser de plus grands corpus, en effectuant par exemple des recherches sur le Web à partir de ces mots-clés.

5. Apport des connaissances morphologiques dérivationnelles pour le processus de réponse aux questions

Les tests ont été menés en travaillant sur deux collections, la collection constituée dans le cadre du projet Quæro, contenant des documents issus du Web, et la collection

collection	#quest.	#q OK	#q SS	#q VAR	#q Exist-VAR	MRR SS	MRR VAR
clef05	197	187	175	174	125	0,6298	0,6486
clef07	156	92	86	82	49	0,5269	0,5484
equer	126	117	105	105	96	0,6782	0,7039
quæro	147	125	106	113	76	0,3984	0,4347
total	626	521	472	474	346	0,5778	0,6027

Tableau 3. *Résultats de QAVAL pour la sélection des passages, avec et sans connaissances morphologiques*

fournie par CLEF, lors des campagnes d'évaluation des systèmes de question-réponse, contenant des articles de journaux.

Les questions testées sont les 147 questions factuelles de l'évaluation Quæro 2010 et 479 questions factuelles provenant des campagnes CLEF et EQUER. Nous avons retenu les 150 premiers passages retournés par le moteur de recherche et gardé ensuite les 50 premiers suite à leur annotation et pondération par les termes des questions (avec et sans variations).

Afin de comparer les performances du système dans les deux conditions, nous avons déterminé le nombre de questions auquel il était possible de répondre après passage de Lucene (colonne *q OK* tableau 5). Ensuite, nous avons retenu parmi celles-ci celles dont au moins un passage comportait des variantes (colonne ExistVAR, sous ensemble de la colonne précédente).

Nous avons ensuite calculé le MRR (Mean Reciprocal Rank, soit la moyenne des inverses des rangs des premières bonnes réponses) du système sur les 10 premières réponses à ces questions, sur les passages sans et avec variations (resp. colonnes MRR SS et MRR VAR). Un passage répond à une question si il contient la réponse attendue à cette question. En effet, au delà de la dixième position, il est assez difficile d'extraire une réponse correcte qui puisse être ensuite classée dans les premiers rangs.

Les résultats sont présentés dans le tableau 5 où les deux premières colonnes indiquent la collection et le nombre de questions testées. La colonne *q SS* indique le nombre de questions qui ont une réponse lorsque les passages ne sont sélectionnés que sur les lemmes identiques à ceux des questions, et *q VAR* les résultats totaux lorsqu'on annote les passages avec les variations des termes des questions.

On peut constater que la reconnaissance de variations morphologiques entraîne systématiquement une hausse du MRR, et ce quelque soit la collection utilisée. Globalement, le nombre total de documents corrects après reconnaissance des variantes de termes n'est pas meilleur que lorsqu'on n'utilise pas les variantes (colonnes SS et VAR), sauf dans la collection des documents du Web. Mais, pour toutes les collections, le classement des documents est amélioré. Ce type d'amélioration, même faible, est

importante dans un système de question-réponse. En effet il est primordial de savoir apparier question et réponses, surtout lorsqu'elles diffèrent dans leur formulation. Un bon rapprochement ne peut être réalisé qu'en s'appuyant sur des ressources fiables et ayant une bonne couverture, qu'elles soient morphologique ou sémantiques.

6. Conclusion

Nous avons décrit dans cet article la construction automatique d'une ressource morphologique précise, permettant non seulement de trouver des mots de la même famille, mais aussi de typer la relation qu'ils entretiennent. Afin d'aider à la validation des couples engendrés, nous avons proposé l'utilisation de corpus et ressources disponibles (lexique, réseau de cooccurrences et n-grammes de mots), et montré leurs apports.

De plus, afin de montrer l'intérêt de disposer de ressources morphologiques en recherche d'information, nous avons mené différentes expérimentations de recherche et sélection de passages dans le système de question-réponse QAVAL. Nous avons montré que l'utilisation de telles connaissances a un impact pour classer les passages retrouvés par un moteur de recherche.

Remerciements

Ce travail a été en partie réalisé dans le cadre du projet ANR CSOSG 2008 FILTRAR-S.

7. Bibliographie

Bernhard D., Cartoni B., Tribout D., « Evaluating Morphological Resources : a Task-Based Study for French Question Answering », *WoLeR 2011 at ESSLLI, International Workshop on Lexical Resources*, 2011.

Bouchou B., Maurel D., « Prolexbase et LMF : vers un standard pour les ressources lexicales sur les noms propres », *Traitement Automatique des Langues*, vol. 49, n° 1, p. 61-88, 2008.

Creutz M., Hirsimäki T., Kurimo M., Puurula A., Pylkkönen J., Siivola V., Varjokallio M., Arisoy E., Saraçlar M., Stolcke A., « Morph-based speech recognition and modeling of out-of-vocabulary words across languages », *ACM Transactions on Speech and Language Processing (TSLP)*, vol. 5, n° 1, p. 1-29, 2007.

Grappy A., Grau B., Falco M.-H., Ligozat A.-L., Robba I., Vilnat A., « Selecting Answers to Questions from Web Documents by a Robust Validation Process », *Web Intelligence*, 2011.

Hahn U., Honeck M., Shulz S., « Subword-Based Text Retrieval », *Proceedings of the 36th Hawaii International Conference on System Sciences (HICSS'03)*, Big Island, Hawaii, January 06 - 09, 2003.

Hathout N., Namer F., Dal G., *Many Morphologies*, Cascadilla Press, chapter An Experimental Constructional Database : The MorTAL Project, p. 178-209, 2002a.

Hathout N., Tanguy L., « Webaffix : Discovering Morphological Links on the WWW », *Proceedings of the Third International Conference on L anguage Resources and Evaluation*, Las Palmas de Gran Canaria, Espagne, p. 1799-1804, 2002b.

Jacquemin C., « Syntagmatic and paradigmatic representations of term variation », *Proceedings of the 37th annual meeting of ACL*, 1999.

Koehn P., Hoang H., « Factored Translation Models », *Proceedings of EMNLP-CoNLL 2007*, Prague, Czech Republic, p. 868-876, 2007.

Tran M., Maurel D., « Prolexbase : un dictionnaire relationnel multilingue de noms propres », *Traitement Automatique des Langues*, vol. 47, n° 1, p. 115-139, 2006.

ANNEXE POUR LE SERVICE FABRICATION

A FOURNIR PAR LES AUTEURS AVEC UN EXEMPLAIRE PAPIER
DE LEUR ARTICLE ET LE COPYRIGHT SIGNE PAR COURRIER
LE FICHIER PDF CORRESPONDANT SERA ENVOYE PAR E-MAIL

1. ARTICLE POUR LA REVUE :
   L'objet. Volume 8 – n°2/2005

2. AUTEURS :
   Anne-Laure Ligozat[*] — Delphine Tribout[**] — Brigitte Grau[*]

3. TITRE DE L'ARTICLE :
   Intérêt des ressources morphologiques pour la recherche d'information précise

4. TITRE ABRÉGÉ POUR LE HAUT DE PAGE MOINS DE 40 SIGNES :
   Ressources morphologiques en Q-R

5. DATE DE CETTE VERSION :
   22 février 2012

6. COORDONNÉES DES AUTEURS :

   – adresse postale :

   [*] LIMSI-CNRS, 91403 Orsay cedex, ENSIIE, 91000 Evry
   [**] LLF, 75013 Paris

   – téléphone : 00 00 00 00 00

   – télécopie : 00 00 00 00 00

   – e-mail : Roger.Rousseau@unice.fr

7. LOGICIEL UTILISÉ POUR LA PRÉPARATION DE CET ARTICLE :
   LaTeX, avec le fichier de style `article-hermes.cls`,
   version 1.2 du 03/03/2005.

8. FORMULAIRE DE COPYRIGHT :
   Retourner le formulaire de copyright signé par les auteurs, téléchargé sur :
   `http://www.revuesonline.com`

SERVICE ÉDITORIAL – HERMES-LAVOISIER
14 rue de Provigny, F-94236 Cachan cedex
Tél : 01-47-40-67-67
E-mail : revues@lavoisier.fr
Serveur web : http://www.revuesonline.com