Validation sur le Web de reformulations locales: application à la Wikipédia

Houda Bouamor Aurélien Max Gabriel Illouz Anne Vilnat
LIMSI-CNRS
Univ. Paris Sud 11
Orsay, France
prenom.nom@limsi.fr

RÉSUMÉ
Ce travail présente des expériences initiales en validation de paraphrases en contexte. Les révisions de Wikipédia nous servent de domaine d'évaluation : pour un énoncé ayant connu une courte révision dans l'encyclopédie, nous disposons d'un ensemble de réécritures possibles, parmi lesquelles nous cherchons à identifier celles qui correspondent à des paraphrases valides. Nous abordons ce problème comme une tâche de classification fondée sur des informations issues du Web, et parvenons à améliorer la performance de plusieurs techniques simples de référence.

ABSTRACT
Assisted rephrasing for Wikipedia contributors through Web-based validation

This works describes initial experiments on the validation of paraphrases in context. Wikipedia's revisions are used : we assume that a set of possible rewritings are available for a given phrase that has been rewritten in the encyclopedia's revision history, and we attempt to find the subset of those rewritings that can be considered as valid paraphrases. We tackle this problem as a classication task which we provide with features obtained from Web data. Our experiments show that our system improves performance over a set of simple baselines.

MOTS-CLÉS : paraphrase, Wikipédia, aide à la rédaction.

KEYWORDS: paraphrasing, Wikipedia, authoring aids.

1 Introduction

Il existe plusieurs scénarios dans lesquels il est souhaitable de pouvoir faire produire du texte par la machine. Ce problème a traditionnellement été abordé comme une tâche de génération de texte à partir de concepts. Toutefois, ces besoins s'appliquent parfois à des cas où un nouveau texte devrait être dérivé de certains textes existants, par exemple lorsqu'il s'agit de transformer un texte afin qu'ils aient certaines propriétés souhaitables pour un usage particulier (Zhao *et al.*, 2009). Par exemple, on peut souhaiter qu'un texte soit condensé (Cohn et Lapata, 2008), adapté à certains profils de lecteur (Zhu *et al.*, 2010), conforme à certaines normes spécifiques (Max, 2004), voire même simplement plus adapté pour des tâches de traitement automatique ultérieures.

Le mécanisme de réécriture de texte doit donc produire un texte dont le sens est compatible avec la définition de la tâche à accomplir, tout en garantissant que celui-ci demeure grammatical.

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN, pages 197–210,
Grenoble, 4 au 8 juin 2012.

La complexité de la *génération texte-à-texte*, par opposition à la génération *concepts-à-texte*, provient essentiellement du fait que la correspondance sémantique entre deux textes est difficile à contrôler, car les réécritures mises en jeu sont très dépendantes du contexte. En effet, la grande diversité des techniques d'acquisition de paraphrases *sous-phrastiques* (Madnani et Dorr, 2010), la polysémie de ces unités linguistiques ainsi que les contraintes pragmatiques associées à leur substitution font qu'il est impossible de garantir que des paires de paraphrases candidates seront substituables quel que soit le contexte de réécriture. Ce problème a été déjà décrit au niveau lexical (Zhao *et al.*, 2007; McCarthy et Navigli, 2009) ; la validation automatique en contexte de reformulations de segments demeure une question fondamentale pour la réécriture de texte.

Dans ce travail, nous abordons le problème sous l'angle d'un *paraphrasage ciblé* [1], défini comme la réécriture d'un segment d'un énoncé. Bien que ce problème soit plus simple que la réécriture d'une phrase complète, son étude se justifie par la nécessite de bien comprendre ce niveau moins complexe avant d'aborder la réécriture d'unités plus étendues, ce qui en outre facilite la tâche complexe de l'évaluation.

Nous présentons ici un scénario en *révision interactive de textes* dans lequel des paraphrases sous-phrastiques doivent être proposées en tenant compte du contexte. Les paraphrases candidates considérées sont obtenues à partir d'un répertoire existant, et sont validées en contexte à l'aide d'informations obtenues sur le Web. Les expériences que nous avons menées ciblent plus particulièrement les contributeurs de l'encyclopédie Wikipédia dans leurs tâches de révision des articles. Nous avons pour cela utilisé un ensemble de segments ayant fait l'objet de réécritures dans l'historique des articles de Wikipédia, que nous substituons par des paraphrases connues à l'avance. Étant donné la grande variété de segments possibles et de leurs paraphrases, nous ne nous appuyons pas sur des modèles de substituabilité préétablis, mais nous les construisons *à la volée* à partir du Web.

Dans cet article, nous allons tout d'abord décrire la tâche de révision de texte sous forme de paraphrasage ciblé (section 2). Nous passerons ensuite en revue les principaux travaux précédents portant sur l'acquisition de paraphrases sous-phrastiques et décrirons les sources de connaissances que nous avons utilisées dans ce travail (section 3). Nous détaillerons ensuite notre méthode de calcul des modèles de substitution de segments en contexte exploitant des informations issues du Web (section 4). Les expériences menées pour valider les paraphrases contenues dans le répertoire existant et leurs résultats seront finalement présentés (section 5). Notre article se conclura par une discussion de ces résultats et une présentation des principales voies de recherche (6).

2 Paraphrasage ciblé pour la révision de texte

La reformulation d'un énoncé, ou d'un segment plus précis, est une activité importante en révision de texte. Certaines modifications locales ont ainsi vocation à améliorer sa qualité générale, en le rendant par exemple plus facile d'accès (Zhu *et al.*, 2010) ou en l'adaptant au niveau d'expertise de ses lecteurs (Deléger et Zweigenbaum, 2009). Les modifications de ce type, qui n'altèrent pas le sens des textes, incluent non seulement la synonymie lexicale mais également des transformations lexico-syntaxiques plus complexes.

1. Ce terme est utilisé par Resnik *et al.* (2010) pour décrire l'obtention (manuelle) de paraphrases pour des segments jugés difficiles à traduire.

On trouve notamment de telles reformulations dans les historiques de révision de textes, qui sont désormais disponibles en grandes quantités avec l'émergence de ressources collaboratives sur le Web telles que l'encyclopédie Wikipédia. L'historique des révisions des articles de cette ressource constitue en effet une source importante de phénomènes de *réécriture naturelle*. L'étude de Dutrey *et al.* (2011) a notamment montré que cet historique contient une variété importante de phénomènes de reformulation, dont de nombreuses paraphrases. Cette étude a également montré, au travers d'une tentative d'identification automatique à base de règles, les difficultés pour parvenir à une bonne couverture de l'ensemble des phénomènes paraphrastiques présents.

Peu de travaux, ont, à notre connaissance, porté sur l'utilisation du paraphrasage contextuel dans le cadre de l'aide à la rédaction. Max et Zock (2008) présentent une méthode proposant aux rédacteurs des paraphrases sous-phrastiques candidates pour les segments qu'ils souhaitent reformuler. L'approche utilisée pour la génération des paraphrases est fondée sur l'équivalence de traduction (Bannard et Callison-Burch, 2005). Les travaux de Bernstein *et al.* (2010) portent eux sur l'externalisation de diverses tâches d'édition de texte, dont la révision, *via* le *crowdsourcing*.

Par ailleurs, la réécriture d'un texte peut être destinée plus spécifiquement à une application automatique. Dans (Resnik *et al.*, 2010), des reformulations pour des segments jugés difficiles à traduire sont acquises *via le crowdsourcing* : des contributeurs monolingues de la langue source proposent ainsi des reformulations en contexte pour ces unités [2]. Les reformulations collectées sont ensuite utilisées en entrée dans un système de traduction automatique, qui peut ainsi bénéficier de la variété d'expressions pour produire de meilleures traductions (Schroeder *et al.*, 2009). Par exemple, le segment *une optique festive* dans *L'usage intervient alors dans* **une optique festive** peut être réécrit en : 1) *un cadre festif*, 2) *une perspective de fête*. Ces réécritures sont grammaticalement correctes et ont des significations raisonnablement proches de la formulation d'origine.

Outre la reformulation des segments de texte, la réécriture d'énoncés a aussi été à l'origine de plusieurs travaux (Barzilay et Lee, 2003; Quirk *et al.*, 2004; Zhao *et al.*, 2010; Ganitkevitch *et al.*, 2011). Cependant, celle-ci pose de nombreux autres défis, notamment au niveau de l'évaluation des reformulations produites. Le jugement par des humains devient alors encore plus complexe et n'autorise pas des distinctions fines ni des accords inter-annotateurs satisfaisants. La génération de paraphrases d'énoncés peut toutefois être évaluée indirectement dans le cadre de leur utilisation dans une application plus complexe. Par exemple, Madnani *et al.* (2008) parviennent à améliorer les performances d'un système de traduction automatique statistique en fournissant des paraphrases automatiques des traductions de référence lors de l'apprentissage des paramètres du système. Cependant, les améliorations observées n'indiquent pas clairement les liens avec la qualité des paraphrases utilisées.

Nous abordons dans ce travail la tâche plus modeste de paraphrasage sous-phrastique appliqué à la révision de texte. Afin d'éviter tout biais, nous utilisons des réécritures *écologiques* (que nous entendons ici comme : produites naturellement) extraites d'une mémoire de rédaction des articles de Wikipédia. Nous utilisons pour cela le corpus WiCoPaCo (Max et Wisniewski, 2010), qui contient de nombreux phénomènes de réécriture, dont de nombreuses instances de reformulations lexicales, syntaxiques et sémantiques (Dutrey *et al.*, 2011). Ce dernier type de reformulation est illustré dans l'exemple suivant, où le remplacement du segment *un mode d'expression* par sa paraphrase possible *une figure de rhétorique* permet de préciser et d'affiner le

2. Nous notons toutefois que les contributeurs ne reçoivent aucune indication directe de l'utilité des reformulations qu'ils proposent.

sens voulu par le contributeur initial :

L'antiphrase est ***[un mode d'expression → une figure de rhétorique]*** *consistant à dire le contraire de ce que l'on pense.*

Ce corpus est pertinent à plusieurs titres pour la tâche que nous visons. Tout d'abord, le fait qu'il contienne des réécritures obtenues hors du cadre d'expériences offre une source riche et intéressante d'unités textuelles réécrites en contexte. De plus, les instances de réécriture où le sens n'a pas été modifié offrent directement une paraphrase candidate qui peut être considérée comme *correcte*, donnée pouvant s'avérer utile pour l'apprentissage automatique du processus de validation en contexte.

3 Acquisition de paraphrases sous-phrastiques

La disponibilité grandissante de masses de données textuelles a rendu possible un grand nombre de travaux en acquisition et en génération de paraphrases (Madnani et Dorr, 2010). Les techniques proposées apparaissent néanmoins assez fortement liées aux types de ressources auxquelles elles s'appliquent. Les types de corpus utilisés pour sont principalement :

- des paires de paraphrases d'énoncés (**corpus monolingues parallèles**), qui permettent d'obtenir des paraphrases précises, mais en faible quantité (Barzilay et McKeown, 2001; Pang *et al.*, 2003; Cohn *et al.*, 2008; Bouamor *et al.*, 2011) ;
- des paires d'énoncés en relation de traduction (**corpus multilingues parallèles**), qui permettent de générer de nombreuses paraphrases candidates (Bannard et Callison-Burch, 2005; Kok et Brockett, 2010) ;
- des paires d'énoncés en relation partielle (**corpus monolingues parallèles**), qui permettent sur le principe d'acquérir de nombreuses paraphrases (Barzilay et Lee, 2003; Pasça et Dienes, 2005; Bhagat et Ravichandran, 2008; Deléger et Zweigenbaum, 2009).

Bien que la précision de ces techniques d'acquisition peut se mesurer sur la base d'une référence attendue portant sur une collection de paires d'énoncés (Cohn *et al.*, 2008), il est plus utile de pouvoir la mesurer au travers de la question de *substituabilité en contexte*, laquelle a déjà été abordée au niveau lexical (Connor et Roth, 2007; Zhao *et al.*, 2007) où elle a fait l'objet de campagnes d'évaluation (McCarthy et Navigli, 2009). Celle-ci pose des défis supplémentaires, dûs au fait que les segments sont plus rares que les mots en corpus.

4 Validation contextuelle sur le Web

4.1 Cadre d'évaluation

Le présent travail porte sur la tâche de validation automatique de paraphrases sous-phrastiques en contexte. Pour cela, nous avons eu recours à un répertoire existant de paires de paraphrases. Comme décrit plus haut, nous avons utilisé le corpus WiCoPaCo comme corpus de reformulations sous-phrastiques naturelles. La réécriture contenue dans cette ressource peut être utilisée comme paraphrase potentielle. Afin d'obtenir d'autres paraphrases candidates de différentes qualités, nous avons utilisé deux autres méthodes d'acquisition, qui fourniront des paraphrases aux

instances extraites de WiCoPaCo qui ne seront pas nécessairement substituables en contexte : *a*) une traduction automatique par pivot, et *b*) une acquisition manuelle de paraphrases.

La génération de paraphrases par traduction s'effectue simplement en traduisant automatiquement un segment dans une langue pivot, puis en le rétrotraduisant dans la langue d'origine, et en retenant la première hypothèse différente du segment d'origine. Si cette technique n'offre aucune garantie sur la qualité des résultats, elle est aisée à mettre en œuvre et produit des résultats variés. En outre, l'utilisation d'une langue pivot proche de la langue d'origine augmente la probabilité d'obtenir de bonnes paraphrases (ceci sera étudié lors de nos expériences, décrites dans la section 5).

Nous avons défini l'acquisition manuelle de paraphrases de la façon suivante : un corpus d'extraits de documents du Web contenant les segments à réécrire est tout d'abord constitué, en s'assurant que ce corpus ne contient pas de données provenant de Wikipédia. Pour chaque segment à réécrire dans ce corpus, des locuteurs natifs du français proposent *via* une interface web une réécriture possible. Ainsi, les contextes utilisés pour faire l'acquisition de paraphrases des segments sont possiblement différents de ceux, extraits de WiCoPaCo, sur lesquels portera l'évaluation : notre système de validation en contexte aura donc à considérer des paraphrases *potentiellement* valides [3] mais qui ne le sont pas dans le contexte d'une réécriture particulière.

Ces deux méthodes, dont la mise en œuvre est aisée, nous permettent de simuler la disponibilité d'un répertoire existant de paraphrases sous-phrastiques, qui nous servira pour l'évaluation de la performance de notre technique de validation en contexte.

4.2 Classification automatique de réécritures en contexte

Nous décrivons maintenant l'approche que nous proposons pour réaliser une validation de réécritures en contexte, fondée sur une classification binaire exploitant des modèles calculés à partir d'informations du Web. Le recours au Web semble indispensable : seule une telle échelle nous permet d'accéder à des exemple en nombre suffisants pour certains segments. En outre, il a été montré qu'un certain nombre d'applications de Traitement Automatique des Langues peuvent être améliorées grâce à l'exploitation de fréquences de *n*-grammes sur le Web (Lapata et Keller, 2005).

Considérant un ensemble de contextes de réécritures pour des segments ainsi qu'un répertoire existant contenant des paraphrases pour ces segments, notre tâche consiste à classer (i.e. *paraphrase* vs. *pas paraphrase*) chaque paraphrase possible pour chaque contexte original. Une instanciation concrète possible de cette tâche est la proposition de Max et Zock (2008), où de telles reformulations candidates sont présentées dans un ordre décroissant de pertinence à un utilisateur d'un éditeur de texte, et donc éventuellement lors de la révision d'un article de Wikipédia.

La tâche d'identification automatique de paraphrases a été déjà abordée par classification automatique dans des travaux précédents, en utilisant des modèles calculés sur des corpus collectés (Brockett et Dolan, 2005) et sur des documents issus du Web (Zhao *et al.*, 2007). Cependant, ces travaux se sont limités à l'identification de paraphrases lexicales (McCarthy et Navigli, 2009). Une difficulté importante est que certains mots sont absents ou très peu fréquents

3. On les suppose ici valides parce que obtenues par réécriture manuelle d'un segment dans un texte. Ceci repose cependant fortement sur la capacité de nos contributeurs natifs à bien réaliser la tâche demandée.

dans les index des moteurs de recherche, et *a fortiori* dans des corpus spécialisés, difficulté qui s'amplifie lorsque l'on considère des segments. [4]

De façon analogue aux travaux de Brockett et Dolan (2005), nous considérons l'identification de paraphrases comme une tâche de classification : étant donné un segment d'origine s dans le contexte d'une phrase p, nous cherchons à déterminer si une paraphrase candidate s' serait une paraphrase *grammaticale* de s dans le contexte de p. Nous avons abordé ce problème avec un classifieur de type séparateur à vaste marge (SVM) exploitant les traits décrits ci-dessous.

Distance d'édition Les approches les plus répandues en calcul de pertinence d'un document relativement à une requête exploitent des mesures de similarité de surface, qui peuvent dans certains cas être de bons indicateurs de proximité sémantique. Un coût de transformation entre chaînes de caractères peut par exemple être celui donné par la mesure TER (Snover *et al.*, 2010), initialement développée pour mesurer la similarité entre une hypothèse de traduction et une traduction de référence. Cette mesure se base sur des opérations d'édition (substitution, déplacement, insertion, suppression) plus informatives que les méthodes basées sur des intersections lexicales [5]. Nous effectuons en outre ce calcul sur les lemmes plutôt que sur les formes de surface, que nous avons obtenus à l'aide du TreeTagger (Schmid, 1994) [6]. Nous retenons donc le score suivant, calculé entre un segment d'origine seg_{orig} et une paraphrase seg_{para}, où la fonction Lem produit une forme lemmatisée de son argument :

$$h_{edit} = \text{TER}(Lem(seg_{orig}), Lem(seg_{para})) \quad (1)$$

Il convient de noter que, contrairement aux autres modèles, celui-ci ne dépend pas d'informations provenant du Web.

Score de modèle de langue La vraisemblance d'une phrase peut être un relativement bon indicateur de sa grammaticalité locale (Mutton, 2006). Les probabilités données par un modèle de langue peuvent désormais être obtenues à l'aide de comptes provenant du Web. Nous avons pour cela utilisé le Service Web N-gram de Microsoft (Wang *et al.*, 2010) dans sa déclinaison à des fins de recherche [7]. Afin de pouvoir utiliser correctement ce service sur des textes français, nous avons dû supprimer tous les diacritiques : un examen précis des paraphrases candidates classées a montré que cette transformation, bien qu'abérante, nous a permis d'obtenir des résultats cohérents. [8]

4. Nous faisons cependant l'hypothèse que des segments absents ou très peu fréquents sur le Web présentent un intérêt moindre pour la réécriture, et n'accordons donc pas pour cette étape de nos travaux d'attention particulière à ce problème. Il est toutefois possible d'argumenter que ces segments pourraient être *mal écrits* (par exemple, par un locuteur non natif, un apprenant, voire une machine) et donc possiblement non connus des moteurs de recherche, pour lesquels une assistance à la réécriture serait tout à fait pertinente. Cela représente néanmoins une problématique en soi.

5. Il faut noter que les opérations de racinisation et de correspondance sémantique utilisant WordNet n'ont pas été prises en compte car nos expériences portent sur le français.

6. Ce calcul de lemmatisation se fait, pour le segment original et sa paraphrase, dans le contexte de la substitution testée : il est toutefois possible que la lemmatisation produise des erreurs.

7. `http://research.microsoft.com/en-us/collaboration/focus/cs/web-ngram.aspx`

8. La description du service n'était pas très explicite lorsque nous l'avons utilisé : il semblerait que l'intention de son fournisseur était avant tout de proposer un service pour l'anglais.

Un simple score de modèle de langue pour un énoncé après réécriture n'est toutefois pas suffisant, car il ne tient pas compte de l'énoncé d'origine. Nous avons donc utilisé le rapport entre le score de modèle de langue de l'énoncé paraphrasé phr_{para} et le score de modèle de langue de l'énoncé d'origine phr_{orig}, normalisé par la longueur des énoncés (Onishi *et al.*, 2010) :

$$h_{ML} = \frac{ML(phr_{para})^{1/longueur(phr_{para})}}{ML(phr_{orig})^{1/longueur(phr_{orig})}} \tag{2}$$

Scores de similarité thématique hors contexte Les techniques mises en œuvre pour calculer une notion de *similarité* entre unités textuelles sont fréquemment fondées sur le calcul de représentations des contextes d'occurrences de ces unités sur lesquelles sont calculées des mesures de similarité. Nous avons suivi ce type d'approche pour mesurer une similarité thématique entre paraphrases entre profils de mots cooccurrents. Nous construisons tout d'abord des profils *hors contexte* de la manière suivante : un moteur de recherche est interrogé afin de récupérer les N premiers extraits de documents (*snippets*) pertinents pour le segment seg_{orig}. La fréquence des mots pleins présents dans ces extraits est alors calculée et est utilisée pour obtenir les valeurs de chaque dimension d'un vecteur de profil lexical T, dont la valeur pour un mot m est définie ainsi :

$$T_{orig}[m] = \frac{freq(seg_{orig}, m)}{freq(seg_{orig})} \tag{3}$$

Pour le calcul des fréquences, $freq(u)$ correspond au nombre d'extraits de documents retournés contenant l'unité u, et $freq(u, v)$ au nombre d'extraits de documents rapportés contenant les deux simultanément. Nous construisons de façon analogue un profil thématique pour chaque paraphrase possible seg_{para}, en se limitant aux dimensions du vecteur pour le segment d'origine :

$$T_{para}[m] = \frac{freq(seg_{para}, m)}{freq(seg_{para})} \tag{4}$$

Enfin, nous mesurons la similarité entre le profil du segment d'origine et de chacune de ses paraphrases possibles à l'aide du cosinus entre les vecteurs de leur profil thématique :

$$h_{them} = \frac{T_{orig} \cdot T_{para}}{||T_{orig}|| * ||T_{para}||} \tag{5}$$

Pour l'ensemble de nos expériences, nous avons utilisé le service Web Yahoo ! Search BOSS [9] pour obtenir le nombre de documents du Web indexés contenant une expression littérale (typiquement, un segment d'intérêt) ainsi que les extraits de documents à partir desquels nous construisons les vecteurs de profils thématiques. En supposant que la distribution des mots pleins cooccurrents n'est pas biaisée par l'ordre des résultats du moteur de recherche, notre modèle mesure donc un certain type de similarité thématique entre seg_{orig} et seg_{para}.

9. `http://developer.yahoo.com/search/boss/`

Scores d'un modèle thématique contextuel Nous définissons également un modèle thématique contextuel de la façon suivante : considérant $cont_{orig}$, constituée des deux sous-chaînes de phr_{orig} privée de seg_{orig}, nous construisons un vecteur de profil T^{cont} ayant pour dimension uniquement pour les mots pleins du contexte de la phrase où a lieu la réécriture. Les valeurs associées à chaque dimension correspondent à des rapports de fréquence obtenus comme précédemment par interrogation du moteur de recherche. La similarité thématique contextuelle utilisée est finalement définie par :

$$h^{cont}_{them} = \frac{T^{cont}_{orig} \cdot T^{cont}_{para}}{||T^{cont}_{orig}|| * ||T^{cont}_{para}||} \quad (6)$$

5 Expériences et résultats

Dans cette section, nous détaillons les expériences que nous avons menées afin d'évaluer les performances de l'approche de validation automatique de paraphrases en contexte.

5.1 Description des données utilisées

Nous avons extrait aléatoirement 150 énoncés en français du corpus WiCoPaCo et leur réécriture pour des exemples annotés comme "paraphrases" lors d'une annotation manuelle réalisée par une étudiante en linguistique francophone. Un sous-ensemble de 100 énoncés a été utilisé comme corpus d'apprentissage, les 50 énoncés restants ayant servi pour l'évaluation. Les segments originaux ainsi que leur paraphrase dans le corpus d'évaluation sont décrits dans la figure 1.

taille segment	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
# segments originaux	0	3	29	8	6	2	2	0
# paraphrases	39	64	74	36	21	10	5	1

FIGURE 1 – Répartition du nombre de segments par taille (nombre de tokens) dans le corpus d'évaluation

Nous disposons finalement de 5 paraphrases par segment d'origine :

- **WICOPACO** : la paraphrase associée au segment dans le corpus WiCoPaCo ;
- **HUMAIN** : deux paraphrases candidates proposées par des contributeurs humains pour d'autres contextes issus du Web ;
- **PIVOT$_{ES}$** and **PIVOT$_{ZH}$** : deux paraphrases candidates obtenues par traduction par pivot. Nous avons utilisé le système de traduction automatique sur le Web GOOGLE TRANSLATE [10], avec une langue proche du français comme pivot (l'espagnol), et une autre plus distante (chinois).

La partie évaluation de nos expériences a impliqué 4 évaluateurs humains [11], tous francophones. Ceux-ci ont participé à la collecte manuelle des paraphrases (HUMAIN) pour la moitié du corpus d'apprentissage et d'évaluation. Afin d'évaluer le caractère approprié de l'utilisation des

10. http://translate.google.com
11. La personne ayant réalisé l'annotation originelle de WiCoPaCo n'a pas pris part à ce nouveau travail.

paraphrases issues des paraphrases collectées dans les contextes de réécriture sélectionnés, les phrases d'origine et leurs différentes paraphrases ont été présentées dans un ordre aléatoire aux deux évaluateurs ayant initialement travaillé sur l'autre moitié des corpus. Une interface sur le Web, illustrée sur la figure 2, permet alors aux évaluateurs d'indiquer quelles substitutions sont acceptables, à la fois au niveau de la conservation du sens et de la grammaticalité du nouvel énoncé.

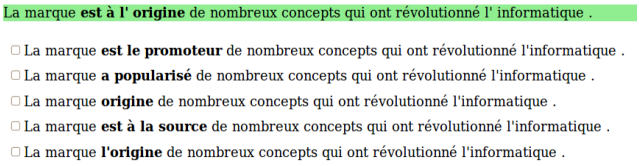

FIGURE 2 – Exemple d'une phrase d'origine (sur fond vert) et de ses 5 paraphrases candidates (présentées dans un ordre aléatoire). Le segment en gras dans la phrase d'origine, *est à l'origine*, est ici paraphrasé par *est le promoteur* , *a popularisé* , *origine* , *est à la source* et *l'origine*.

La valeur d'accord inter-annotateur [12] sur l'ensemble des énoncés annotés est de $\kappa = 0,65$, ce qui correspond à un accord *fort* selon les grilles de Landis et Koch (1977). Nous pensons que le fait d'aborder tout d'abord des tâches relativement certaines du point de vue de l'accord entre humains comme celle-ci est nécessaire avant de s'attaquer à des problèmes plus complexes, tels que l'identification de paraphrases d'énoncés ou encore l'identification d'implications textuelles.

Notre technique de validation étant très dépendante de la fréquence des segments considérés sur le Web, nous avons décidé dans ces premières expériences de ne conserver que les segments ayant une fréquence minimale de 10 occurrences pour le moteur de recherche utilisé. Le nombre d'exemples du corpus d'apprentissage a ainsi été réduit de 750(=150*5) à 434, et celui du corpus d'évaluation de 250(=50*5) à 215. L'atténuation de cette limitation devra bien évidemment faire partie de la suite de nos travaux.

Nous détaillerons nos résultats pour les 3 conditions suivantes :

- **Possibles** : les exemples annotés comme "paraphrases" par au moins l'un des juges sont utilisés : l'ensemble d'évaluation correspondant comprend 116 cas positifs et 99 cas négatifs.
- **Sûres** : les exemples que les deux juges n'ont pas annotés comme "paraphrases" ou "non paraphrases" ne sont pas retenus : l'ensemble d'évaluation correspondant comprend 76 cas positifs et 139 cas négatifs.
- **Sûres++** : seuls les exemples pour lesquels les deux juges proposent la même annotation sont retenus. Ceci réduit nos ensembles d'apprentissage et d'évaluation à respectivement 287 et 175 exemples, ce qui ne permet pas une comparaison directe avec les deux autres conditions. L'ensemble d'évaluation correspondant comprend 76 cas positifs et 99 cas négatifs.

12. Nous avons utilisé le logiciel R (`http://www.r-project.org`) pour calculer la valeur de κ de Cohen. Cette valeur est calculée sur l'ensemble des données, chaque moitié étant annotée par les deux mêmes annotateurs.

5.2 Techniques de référence

Nous présentons ici brièvement les techniques de référence auxquelles nous comparerons notre système.

Fréquence sur le Web Les deux premières techniques sont fondées sur des calculs de fréquences sur le Web. La première, ML_WEB considère un énoncé comme paraphrase d'un énoncé d'origine si son score de modèle de langue issu du Web est plus élevé que celui de l'énoncé d'origine. La deuxième technique, ML_FRONTIÈRES, considère qu'un énoncé est paraphrase d'un énoncé d'origine si la fréquence sur le Web des bigrammes traversant les frontières gauche et droite après substitution est supérieure à 10.

Conservation de dépendances syntaxiques Lors de la réécriture d'une partie d'un énoncé, la conservation des dépendances syntaxiques entre un segment d'origine et son contexte d'une part, et sa paraphrase avec le même contexte d'autre part, peut renseigner sur la substituabilité grammaticale des deux segments (Zhao *et al.*, 2007; Max et Zock, 2008). Nous avons calculé les dépendances syntaxiques pour les deux segments à l'aide de la version française (Candito *et al.*, 2010) de l'analyseur probabiliste de Berkeley (Petrov et Klein, 2007). Nous considérons donc le sous-ensemble des dépendances qui existent entre les mots du segment d'origine et son contexte (Dep_{orig}) et entre les mots de la paraphrase et ce contexte (Dep_{para}). Cette technique, DEPCONT, retient la paraphrase candidate si et seulement si $Dep_{para} = Dep_{orig}$.

5.3 Résultats et analyse

Nous avons utilisé un séparateur à vastes marges (SVM) avec les traits décrits dans la section 4 [13] Les performances des différentes techniques sur les 3 conditions décrites précédemment sont données dans la figure 3.

	ML_WEB	LM_FRONTIÈRES	DEPCONT	CLASSIFIEUR
POSSIBLES	**62,79**	54,88	48,53	57,67
SÛRES	68,37	36,27	51,90	**70,69**
SÛRES++	56,79	51,41	42,69	**62,85**

FIGURE 3 – Résultats de la performance de la classification (*exactitude*) pour les 3 techniques de référence et notre classifieur sur le corpus d'évaluation et les 3 conditions. Il convient de noter que la condition SÛRES++ n'est pas directement comparable aux autres conditions puisque les tailles des corpus d'apprentissage et d'évaluation sont différentes à celles des deux autres conditions.

La première observation que nous pouvons faire est que la tâche de classification de paraphrases est une tâche difficile : la meilleure performance (*exactitude*) obtenue par l'un des systèmes est de 70,69 pour la condition SÛRES. En outre, il existe une variation importante entre les

13. Nous avons utilisé l'implémentation LIBSVM (Chang et Lin, 2001).

différentes conditions testées avec un résultat faible pour notre classifieur de 57,67 dans la condition POSSIBLES (cas de désaccord entre annotateurs, où un seul reconnaît le statut de paraphrase).

D'une manière plus générale, la technique ML_WEB et notre classifieur sont plus performants que les autres techniques de références. ML_FRONTIÈRES et DEPCONT ne modélisent que des contraintes grammaticales locales, ce qui fait qu'il n'est pas surprenant que ces informations ne permettent pas la reconnaissance de variations sémantiques licites entre paraphrases candidates. WEBLM, qui se limite à la comparaison de scores de modèles de langue dérivé du Web, apparaît donc comme une technique relativement compétitive [14], mais sa performance est peu élevée (56,79) pour la condition SURE++. Puisque cette condition ne prend en compte que les annotations consensuelles pour l'apprentissage et l'évaluation, nous considérons cette condition comme la plus utile pour l'interprétation des résultats de ces travaux préliminaires. Ici, notre système obtient la meilleur performance, avec un avantage de 6,06 points par rapport à WEBLM. Ceci montre que la seule utilisation d'un modèle de langue, aussi bien estimé soit-il, est trop limitée pour rendre compte correctement de l'ensemble des phénomènes de paraphrases présents dans notre corpus d'évaluation, ce qui confirme des résultats précédents où les modèles de langue n'étaient pas issus de comptes du Web (Bannard et Callison-Burch, 2005).

Finalement, la figure 4 détaille les performances atteintes par chacune des méthodes d'acquisition de paraphrases pour chacune des 3 conditions. Il n'est tout d'abord pas surprenant que les reformulations extraites de WICOPACO soient largement identifiées comme de bonnes paraphrases en contexte, en particulier dans les conditions POSSIBLES et SÛRES++. Ces paraphrases sont le résultat de reformulations par des contributeurs de Wikipédia dans le contexte d'évaluation, et avaient déjà été reconnues comme telles par une première annotatrice.

	WICOPACO	HUMAIN	PIVOT_{ES}	PIVOT_{ZH}
POSSIBLES	**89,33**	67,00	47,33	20,66
SÛRES	**64,00**	44,50	31,33	10,66
SÛRES++	**86,03**	57,34	37,71	12,60

FIGURE 4 – Performance (valeurs d'*exactitude*) de nos différentes méthodes d'acquisition pour nos trois conditions d'évaluation.

Les paraphrases obtenues par collecte manuelle sur des contextes issus du Web, donc d'un contexte possiblement différent de celui de l'évaluation, obtiennent une performance relativement acceptable. Les résultats confirment cependant le fait attendu que la substituabilité des paraphrases dépend fortement du contexte. Par exemple, la substitution du segment *de l'éditeur* par *publiée par les éditions* dans le contexte de l'énoncé "*Neopolis est une collection de bandes dessinées* ***de l'éditeur*** *Delcourt.* [15]" permet de conserver le sens d'origine ainsi que la grammaticalité de l'énoncé. *A contrario*, la substitution par le segment *du logiciel* n'est pas adaptée à ce contexte.

Finalement, les paraphrases obtenues automatiquement par traduction par pivot ne sont pas de bonne qualité. Nous notons cependant que la proximité de la langue pivot avec la langue

14. Une explication peut résider dans le fait que nos méthodes d'acquisition de paraphrases utilisant Google Translate commme un traducteur automatique par pivot ont tendance à produire des segments ayant une forte valeur de probabilité dans le modèle de langue utilisé, qui est certainement assez comparable à celui utilisé dans nos expériences.

15. Une réécriture est extraite de l'historique de révision de l'article "Neopolis" sur Wikipédia accessible sur : `http://fr.wikipedia.org/w/index.php?title=Neopolis&diff=45811975&oldid=2017149`.

de réécriture joue un rôle important : l'utilisation de l'espagnol mène ainsi à de bien meilleurs résultats que l'utilisation du chinois [16].

6 Conclusions et perspectives

Nous avons présenté dans cet article une approche de paraphrasage en contexte appliqué à la révision de texte, un scénario soutenu par les données extraites des réécritures contenues dans la Wikipédia francophone. La méthode d'identification que nous avons proposée prend en entrée un répertoire existant de paraphrases sous-phrastiques, et détermine par classification automatique exploitant des données issues du Web si les paraphrases connues peuvent se substituer à un segment dans un contexte particulier. Nous avons simulé différents niveaux de qualité pour les paraphrases existantes, en exploitant des paraphrases provenant de Wikipédia, des contributions humaines acquises dans d'autres contextes, et des paraphrases obtenues par traduction automatique par pivot.

Nos expériences ont montré que la version actuelle de notre classifieur est plus performante que les différentes techniques de référence utilisées lorsque l'on ne considère que les paraphrases obtenant des jugements consensuels dans la référence utilisée. Bien que ces premières expériences soient positives, nous sommes conscients que leurs résultats peuvent être améliorés sur différents aspects. Tout d'abord, il est possible d'élargir l'exploration des différentes caractéristiques que nous mettons en jeu dans le classifieur. Nous comptons intégrer d'autres traits, dont des modèles mettant en jeu des dépendances syntaxiques calculées sur des données du Web. Nous allons également analyser plus finement nos résultats afin d'identifier les cas problématiques, dont certains ne peuvent pas être modélisés sans avoir recours à des connaissances du monde, ce qui suggérera notamment l'intégration de connaissances du domaine, éventuellement dérivées de méta-informations provenant des articles Wikipédia concernés. L'ensemble de ces expériences pourra être conduit en plusieurs langues, les données utilisées et les méthodes employées pouvant facilement être transposées. Finalement, nous sommes également intéressés par le fait d'utiliser l'approche décrite ici comme un cadre pour l'évaluation des systèmes d'acquisition de paraphrases.

Références

BANNARD, C. et CALLISON-BURCH, C. (2005). Paraphrasing with bilingual parallel corpora. *In Actes de ACL*, Ann Arbor, USA.

BARZILAY, R. et LEE, L. (2003). Learning to paraphrase : an unsupervised approach using multiple-sequence alignment. *In Actes de NAACL-HLT*, Edmonton, Canada.

BARZILAY, R. et MCKEOWN, K. (2001). Extracting paraphrases from a parallel corpus. *In Actes de ACL*, Toulouse, France.

BERNSTEIN, M. S., LITTLE, G., MILLER, R. C., HARTMANN, B., ACKERMAN, M. S., KARGER, D. R., CROWELL, D. et PANOVICH, K. (2010). Soylent : a word processor with a crowd inside. *In Proceedings of the ACM symposium on User interface software and technology*.

16. Bannard et Callison-Burch (2005) ont montré que l'utilisation simultanée de plusieurs langues pivots permettait de diminuer de façon importante les phénomènes de bruit.

BHAGAT, R. et RAVICHANDRAN, D. (2008). Large scale acquisition of paraphrases for learning surface patterns. *In Actes de ACL-HLT*, Columbus, États-Unis.

BOUAMOR, H., MAX, A. et VILNAT, A. (2011). Monolingual alignment by edit rate computation on sentential paraphrase pairs. *In Proceedings of ACL, Short Papers session*, Portland, USA.

BROCKETT, C. et DOLAN, W. B. (2005). Support vector machines for paraphrase identification and corpus construction. *In Proceedings of The 3rd International Workshop on Paraphrasing IWP*, Jeju Island, South Korea.

CANDITO, M., CRABBÉ, B. et DENIS, P. (2010). Statistical French dependency parsing : treebank conversion and first results. *In Proceedings of LREC*, Valletta, Malta.

CHANG, C.-C. et LIN, C.-J. (2001). *LIBSVM : a library for support vector machines*. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

COHN, T., CALLISON-BURCH, C. et LAPATA, M. (2008). Constructing corpora for the development and evaluation of paraphrase systems. *Comput. Linguist.*, 34(4).

COHN, T. et LAPATA, M. (2008). Sentence compression beyond word deletion. *In Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, Manchester, UK.

CONNOR, M. et ROTH, D. (2007). Context sensitive paraphrasing with a single unsupervised classifier. *In Proceedings of ECML*, Warsaw, Poland.

DELÉGER, L. et ZWEIGENBAUM, P. (2009). Extracting lay paraphrases of specialized expressions from monolingual comparable medical corpora. *In Proceedings of the 2nd Workshop on Building and Using Comparable Corpora : from Parallel to Non-parallel Corpora*, Singapore.

DUTREY, C., BOUAMOR, H., BERNHARD, D. et MAX, A. (2011). Paraphrases et modifications locales dans l'historique des révisions de wikipédia. *In Actes de TALN 2011*, Montpellier, France.

GANITKEVITCH, J., CALLISON-BURCH, C., NAPOLES, C. et VAN DURME, B. (2011). Learning sentential paraphrases from bilingual parallel corpora for text-to-text generation. *In Proceedings of EMNLP*, Edinburgh, UK.

KOK, S. et BROCKETT, C. (2010). Hitting the Right Paraphrases in Good Time. *In Proceedings of NAACL*, Los Angeles, USA.

LANDIS, J. et KOCH, G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, pages 159–174.

LAPATA, M. et KELLER, F. (2005). Web-based Models for Natural Language Processing. *ACM Transactions on Speech and Language Processing*, 2(1):1–31.

MADNANI, N. et DORR, B. J. (2010). Generating Phrasal and Sentential Paraphrases : A Survey of Data-Driven Methods . *Computational Linguistics*, 36(3).

MADNANI, N., RESNIK, P., DORR, B. et SCHWARTZ, R. (2008). Are multiple reference translations necessary ? investigating the value of paraphrased reference translations in parameter optimization. *In Proceedings of AMTA*, Waikiki, Hawai'i.

MAX, A. (2004). From controlled document authoring to interactive document normalization. *In Proceedings of COLING*, Geneva, Switzerland.

MAX, A. et WISNIEWSKI, G. (2010). Mining Naturally-occurring Corrections and Paraphrases from Wikipedia's Revision History. *In Proceedings of LREC*, Valletta, Malta.

MAX, A. et ZOCK, M. (2008). Looking up phrase rephrasings via a pivot language. *In Proceedings of the COLING Workshop on Cognitive Aspects of the Lexicon*, Manchester, United Kingdom.

McCarthy, D. et Navigli, R. (2009). The english lexical substitution task. *Language Resources and Evaluation*, 43(2).

Mutton, A. (2006). *Evaluation of sentence grammaticality using Parsers and a Support Vector Machine*. Thèse de doctorat, Macquarie University.

Onishi, T., Utiyama, M. et Sumita, E. (2010). Paraphrase lattice for statistical machine translation. *In Proceedings of the ACL 2010 Conference, Short Paper session*, Uppsala, Sweden.

Pang, B., Knight, K. et Marcu, D. (2003). Syntax-based alignement of multiple translations : Extracting paraphrases and generating new sentences. *In Actes de NAACL-HLT*, Edmonton, Canada.

Pasça, M. et Dienes, P. (2005). Aligning Needles in a Haystack : Paraphrase Acquisition Across the Web. *In Proceedings of IJCNLP*, Jeju Island, South Korea.

Petrov, S. et Klein, D. (2007). Improved inference for unlexicalized parsing. *In Proceedings of NAACL-HLT*, Rochester, USA.

Quirk, C., Brockett, C. et Dolan, W. B. (2004). Monolingual machine translation for paraphrase generation. *In Proceedings of EMNLP*, volume 149, Barcelona, Spain.

Resnik, P., Buzek, O., Hu, C., Kronrod, Y., Quinn, A. et Bederson, B. B. (2010). Improving translation via targeted paraphrasing. *In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, Cambridge, MA.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. *In Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK.

Schroeder, J., Cohn, T. et Koehn, P. (2009). Word Lattices for Multi-Source Translation. *In Proceedings of EACL*, Athens, Greece.

Snover, M., Madnani, N., Dorr, B. J. et Schwartz, R. (2010). TER-Plus : paraphrase, semantic, and alignment enhancements to Translation Edit Rate. *Machine Translation*, 23(2-3).

Wang, K., Thrasher, C., Viegas, E., Li, X. et Hsu, B.-j. P. (2010). An Overview of Microsoft Web N-gram Corpus and Applications. *In Proceedings of the NAACL-HLT Demonstration Session*, Los Angeles, USA.

Zhao, S., Lan, X., Liu, T. et Li, S. (2009). Application-driven Statistical Paraphrase Generation. *In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, Suntec, Singapore.

Zhao, S., Liu, T., Yuan, X., Li, S. et Zhang, Y. (2007). Automatic acquisition of context-specific lexical paraphrases. *In Proceedings of IJCAI*, Hyderabad, India.

Zhao, S., Wang, H., Liu, T., et Li, S. (2010). Leveraging multiple mt engines for paraphrase generation. *In Proceedings of COLING*, Beijing, China.

Zhu, Z., Bernhard, D. et Gurevych, I. (2010). A monolingual tree-based translation model for sentence simplification. *In Proceedings of COLING*, Beijing, China.