TALN 2011, Montpellier, 27 juin – 1er juillet 2011

Découverte de patrons paraphrastiques en corpus comparable: une approche basée sur les n-grammes

Bruno Cartoni[1] Louise Deléger[2]
(1)Département de Linguistique, Université de Genève
(2)Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center
bruno.cartoni@unige.ch, louise.deleger@cchmc.org

Résumé. Cet article présente l'utilisation d'un corpus comparable pour l'extraction de patrons de paraphrases. Nous présentons une méthode empirique basée sur l'appariement de n-grammes, permettant d'extraire des patrons de paraphrases dans des corpus comparables d'une même langue (le français), du même domaine (la médecine) mais de registres de langues différents (spécialisé ou grand public). Cette méthode confirme les résultats précédents basés sur des méthodes à base de patrons, et permet d'identifier de nouveaux patrons, apportant également un regard nouveau sur les différences entre les discours de langue générale et spécialisée.

Abstract. This paper presents the use of a comparable corpus for extracting paraphrase patterns. We present an empirical method based on n-gram matching and ordering, to extract paraphrase pattern in comparable corpora of the same language (French) and the same domaine, but of two different registers (lay and specialised). This method confirms previous results from pattern-based methods, and identify new patterns, giving fresh look on the difference between specialised and lay discourse.

Mots-clés : Identification de paraphrases, extraction de patrons, type de discours, domaine médical, corpus comparable monolingue.

Keywords: paraphrase identification, lexico-syntactic pattern discovery, discourse type, medical domain, monolingual comparable corpora.

1 Introduction

Cet article présente une étude basée sur un corpus comparable composé de textes de deux types de discours différents (spécialisé ou grand public) mais d'un même domaine (médecine) et dans une même langue (français) permettant d'explorer empiriquement le phénomène de la paraphrase et de valider des travaux antérieurs. On peut définir les paraphrases comme des expressions linguistiques possédant une signification similaire. La compréhension des mécanismes de paraphrase est un élément-clé de nombreuses applications du TALN comme l'extraction d'information (Shinyama & Sekine, 2003), le résumé automatique (Barzilay, 2003) et la simplification de textes (Elhadad & Sutaria, 2007). Différentes approches ont été employées pour la détection de paraphrases : elles peuvent varier au niveau du type de corpus utilisés (parallèle (Barzilay & McKeown, 2001), comparable (Shinyama & Sekine, 2003)), du type de paraphrases recherchées (phrastique (Barzilay & Lee, 2003), sous-phrastique (Elhadad & Sutaria, 2007)) et du type de technique employée (alignement de graphes lexicaux (Barzilay & Lee, 2003), similarité distributionnelle (Elhadad & Sutaria, 2007), patrons lexico-syntaxiques (Jacquemin, 1999)). Parmi les approches en corpus comparable, beaucoup s'appuient sur des corpus journalistiques relatant les mêmes informations mais provenant de différentes sources (Barzilay & Lee, 2003; Shinyama & Sekine, 2003). Dans le domaine médical, (Elhadad & Sutaria, 2007) travaillent sur un corpus comparable presque parallèle, composé d'articles scientifiques et de leur version "grand public", pour extraire des paraphrases entre deux types de discours, grand public vs. spécialisé. Dans nos précédents travaux (Deléger & Zweigenbaum, 2008; Deléger & Cartoni, 2010), nous avons également étudié l'extraction de paraphrases entre ces deux types de discours, à l'aide de patrons lexicaux pré-définis basés sur des ancrages de type morphosémantique (verbe / nom déverbal, adjectif relationnel / nom). Dans la présente étude, nous prolongeons ces travaux en adoptant une approche moins "supervisée" permettant de découvrir de nouveaux patrons de paraphrases, et de confirmer la pertinence des paraphrases utilisées dans nos approches à base de patrons. Cette approche se fonde sur le repérage d'identité entre des n-grammes

lexicaux racinisés. Une fois repéré les paraphrases candidates, une étape de généralisation est effectuée, basée sur les parties du discours, pour identifier des patrons. Ces patrons sont ensuite évalués en terme de fréquence et de qualité, ainsi que dans leur spécialisation, dans le discours spécialisé (ci-après *spé*) ou grand public (*gp*).

2 Description de la méthode

Corpus comparable utilisé : Le terme "corpus comparable" définit une collection de textes partageant des caractéristiques communes. Il s'agit souvent de textes rédigés dans deux langues différentes mais sur des sujets identiques. Mais les corpus comparables peuvent également être monolingues (Barzilay & Lee, 2003; Shinyama & Sekine, 2003; Elhadad & Sutaria, 2007). Ici nous utilisons un corpus comparable monolingue en français composé de textes de deux types de discours (spécialisé ou grand public) appartenant au même domaine (la médecine). Ce corpus porte sur trois sous-domaines précis : le diabète, le cancer et le tabac (la constitution du corpus est décrite plus en détail dans (Deléger & Zweigenbaum, 2008)). Le tableau 1 présente les informations statistiques sur le corpus. Le corpus a également été étiqueté morpho-syntaxiquement et lemmatisé, en utilisant le Treetagger [1] et le lemmatiseur du français Flemm [2].

	Diabète		**Tabac**		**Cancer**		**Corpus total**	
	S	G	S	G	S	G	S	G
phrases	29,692	44,799	25,460	40,840	10,838	13,389	65,990	99,028
occurrences	581,100	581,712	604,206	604,309	228,784	228,793	**1,414,090**	**1,414,805**

TABLE 1 – Tailles des corpus (Nbr. de mots ; S=Spécialisé, G=grand public)

Extraction des n-grammes : Nous avons tout d'abord extrait des n-grammes dans chaque partie du corpus (*spé* et *gp*) en utilisant le "Ngram Statistics Package" (Banerjee & Pedersen, 2003). Nous n'avons conservé que des n-grammes de longueur 2 à 6, que nous avons ensuite filtrés en fonction de critères linguistiques pour qu'ils puissent correspondre à un groupe syntaxique (exclusion des ponctuations, des n-grammes se terminant par une préposition ou un déterminant, des n-grammes contenant moins de 2 mots pleins).

Correspondance lexicale : Nous avons racinisé tous les mots pleins en nous basant sur le package perl Lingua : :Stem [3] (phase de "stemming"). Nous avons ensuite mis en correspondance les n-grammes de deux parties du corpus, en nous basant sur leur identité de chaînes de caractères, en utilisant une approche "sac de mot" (sans prendre en compte l'ordre d'apparition des mots dans les n-grammes), et en nous restreignant à une identité exacte entre les racines. Nous obtenons ainsi un ensemble de paires de n-grammes qui sont nos paraphrases candidates.

Identification de patrons de paraphrases : L'étape suivante consiste à généraliser les paires de n-grammes pour obtenir des patrons basés sur les parties du discours, en gardant une trace du lien entre les mots pleins (par des indices numériques). Dans l'exemple ci-dessous, nous montrons une paire composée d'un n-gramme de la partie spécialisé (à gauche) et d'un n-gramme de la partie grand public (à droite), et le patron qui en est extrait :

traitement du patient $\rightarrow$ *traiter un patient*

N_1 Prep $N_2 \rightarrow V_1$ Det N_2

L'orientation des patrons est conservée (la partie gauche correspond toujours à la partie spécialisée), ce qui signifie que nous pouvons trouver des patrons qui présentent une configuration inversée d'un autre patron.

Filtrage et classification des patrons : Les patrons ont ensuite été triés en fonction de leur fréquence, c'-à-d. le nombre de paraphrases sur lequel ils sont basés. Nous avons conservé uniquement les patrons basés sur plus de 10 paraphrases, considérant les autres comme peu significatifs. Chaque patron a ensuite été soigneusement évalué par deux annotateurs indépendants (les co-auteurs), en comptant le nombre de paraphrases correctes sur lequel le

1. http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger
2. http://www.univ-nancy2.fr/pers/namer/Telecharger_Flemm.htm
3. http://search.cpan.org/~snowhare/Lingua-Stem-0.83

patron était basé. Cette évaluation fournit une mesure de précision de chaque patron, permettant ainsi d'exclure ceux dont la précision est nulle. Ensuite, nous avons classé les patrons selon 4 catégories, en fonction du type de modification mise en œuvre entre les deux parties du patrons, à savoir : une "modification morphosémantique" (comme la dérivation) , une "inversion simple", une "flexion verbale" (incluant la modification passif-actif), et une "variation zéro" (qui comprend les patrons où seules les terminaisons des mots varient, donc essentiellement des variations de type flexionnel). Enfin, nous nous sommes également intéressés à l'orientation des patrons (*spé* $\rightarrow$ *gp* et/ou *gp* $\rightarrow$ *spé*), en distinguant les patrons dont le patron "inversé" avait également été trouvé (p.ex. N Adj $\rightarrow$ N Prep N et N Prep N $\rightarrow$ N Adj ont tous deux été extraits), des patrons apparaissant dans un seul sens (p.ex. seul le sens de patron N Adj $\rightarrow$ V Adv a été trouvé dans le corpus, dans le sens *spé* $\rightarrow$ *gp*). Pour les patrons bi-directionnels, nous mesurons la différence de fréquence entre les deux directions, permettant d'évaluer la préférence d'orientation des patrons de paraphrases, et parfois une préférence significative dans l'un des deux types de discours.

3 Résultats

L'extraction des n-grammes et l'appariement lexical, tel que décrit à la section 2, a permis d'extraire 10 123 patrons, dont une grande partie provenait d'un très petit nombre de paires de n-grammes [4]. En excluant les patrons basés sur moins de 10 paraphrases, nous obtenons un total de 119 patrons, établis à partir d'un total de 4 976 paraphrases. Cette extraction a été soigneusement évaluée, comme décrit précédemment.

3.1 Identification des patrons : tri et classification

L'évaluation des patrons en terme de précision nous a permis d'exclure les patrons dont la précision était nulle (dans 10 cas). Nous obtenons ainsi 109 patrons, regroupant 4 800 paraphrases différentes. La plupart des patrons sont composés de bi-grammes, avec quelques tri-grammes. Nous avons ensuite trié les patrons en fonction du nombre de paragraphes correctes, et en fonction de la typologie présentée à la section 2. Le tableau 2 présente les résultats de la classification et le tableau 3 présente les 20 premiers patrons. [5]

Type de patrons de paraphrase	Nombre	%	Précision moyenne
Morphosémantique (M)	43	39.55	0.65
Inversion simple (I)	32	29.36	0.63
Flexion verbale (V)	25	22.95	0.76
Variation zéro (N)	9	8.26	0.55
Total	109	100%	0.66

TABLE 2 – 4 catégories de patrons de paraphrases

La catégorie la plus importante est celle des patrons "morphosémantiques" (ce qui n'est pas surprenant pour une approche basée sur des racines), suivi des "variations simples" et des "flexions verbales". De plus, cette catégorie est la plus fréquente dans les patrons les plus haut classés. Ces patrons "morphosémantiques" impliquent un changement de catégorie lexicale pour un (ou plusieurs) élément(s) du patron, souvent dû à un procédé dérivationnel, dans la plupart des cas une dérivation verbe$\rightarrow$nom et dans une proportion plus faible, une dérivation nom$\rightarrow$adj. La catégorie "flexion verbale" contient principalement des paraphrases où le verbe est simplement fléchi différemment, mais également des variations plus complexes, quand le participe passé est employé comme adjectif (provoquant une inversion, comme dans le patron de rang 10 dans le tableau 3) ou des cas de variations passif-actif. Les cas d'inversion simple ont tous une relativement bonne précision, et les erreurs sont principalement dues à la préposition employée entre les deux éléments intervertis, qui peut mener à une paraphrase erronée. Ce type de patrons pourrait donc être employé à condition d'appliquer des contraintes adéquates sur les prépositions. Les variations "zéro" ne présentent aucune variation particulière et leur précision se révèle très peu satisfaisante.

4. 7 492 patrons n'étaient établis qu'à partir d'une seule paire

5. Les patrons sont présentés selon les abréviations standard, à savoir N=nom, V=verbe, Vpper=participe passé des verbes, p=préposition, A=adjectif, ADV=adverbe. Les indices (1, 2) indiquent quels éléments ont été mis en correspondance dans les deux parties du patrons. Rappelons également que les déterminants ne sont pas pris en compte dans les patrons

Rang	Para. corr.	Préc.	Patrons (spéc./grand public)	Exemple	Cat.
1	331	0.88	N_1 P N_2 / V_1 N_2	retard de cicatrisation / retarder la cicatrisation	M
2	183	0.94	V_1 N_2 / N_1 p N_2	calculer les coûts / calcul de coût	M
3	177	0.66	N_1 p N_2 / N_1 p N_2	infirmière à domicile / infirmiers à domicile	N
4	140	0.82	N_1 p N_2 / V_1 p N_2	traitement de l' affection / traiter des affections	M
5	121	0.72	N_1 A_2 / N_1 p N_2	apports caloriques / apport en calories	M
6	116	0.92	N_1 Vpper_2 / V_2 N_1	moyens adaptés / adapter les moyens	V
7	96	0.64	N_2 p N_1 / N_1 A_2	complexité du problème / problème complexe	M
8	92	0.81	N_1 p N_2 / N_2 Vpper_1	administration de produit / produit administré	M
9	88	0.98	V ADV_2 Vpper_1/V_1 A_2	a aussi permis / permettra aussi	V
10	85	1	V_2 N_1 / N_1 Vpper_2	restreindre l' accès / accès restreints	V
11	83	0.99	V1 ADV2/V ADV2 Vpper1	fournit aussi / être aussi fournie	V
12	76	0.65	N_1 p N_2 / N_1 A_2	contrôles de la glyémie / contrôles glycémiques	M
13	67	0.61	V_1 p N_2 / N_1 p N_2	importer des produits / importation de produits	M
14	65	0.9	V_2 N_1 / N_1 V Vpper_2	calculer les doses / doses sont calculées	V
15	62	0.91	V_2 ADV_1 / ADV_1 V_2	améliore encore / encore améliorer	I
16	60	0.67	N_2 Vpper_1 / N_1 p N_2	aliments transformés / transformation des aliments	M
17	54	0.82	N_1 Vpper_2 / V_2 p N_1	expérience acquise / acquérir de l' expérience	V
18	52	0.95	N_1 A_2 / V_1 ADV_2	traitement adéquat / traiter adéquatement	M
19	51	0.81	ADV_1 V_2 / V_2 ADV_1	ainsi faciliter / facilite ainsi	I
20	51	0.74	N_1 V Vpper_2 / V_2 N_1	efficacité est renforcée / renforcer son efficacité	V

TABLE 3 – Les 20 patrons les plus fréquents - Les catégories de paraphrases sont : M = morphosémantique, N = Variation zéro, I = inversion , V= flexion verbale

Notons également que la précision moyenne de tous les patrons est de 0,66 (avec une médiane à 0,69), ce qui est satisfaisant pour des patrons extraits automatiquement. De plus, il faut souligner que 29 patrons ont une précision entre 0.9 et 1 (dont 15 qui ont une précision à 1).

3.2 Les patrons bi-directionnels

Parmis tous les patrons extraits et évalués, nous avons distingué les patrons qui apparaissaient dans les deux directions (*spé* $\rightarrow$ *gp* et *gp* $\rightarrow$ *spé*) des patrons qui n'apparaissaient que dans une direction. Environ 70% des patrons (76) apparaissent dans les deux sens, et pour les autres, il se peut que l'équivalent dans le sens inverse existe mais ait été exclu dans les étapes de filtrage précédentes. Parmi les patrons bi-directionnels, certains montrent une préférence pour une direction, comme nous pouvons le constater en regardant le pourcentage des différences entre le nombre des paraphrases acquises dans l'une et l'autre des directions. Le tableau 4 montre les patrons qui ont les différences les plus significatives (les pourcentages négatifs indiquent une préférence pour la direction *gp* $\rightarrow$ *spé*, et les positifs une préférence pour la direction *spé* $\rightarrow$ *gp*). Ces différences permettent de mettre en évidence des patrons particulièrement adéquats pour l'acquisition de paraphrases entre types de discours. Parmi eux, nous trouvons un certain nombre de variations nom-verbe (comme N_1 P N_2 / N_2 V_1, N_1 P N_2 / V_1 P N_2 etc.) pour la direction spécialisé $\rightarrow$ grand public, ce qui confirme les hypothèses sur lesquelles étaient fondées nos précédents travaux (Deléger & Zweigenbaum, 2008). Le patron adjectif-nom (N_1 A_2 / N_1 P N_2), également utilisé dans notre étude (Deléger & Cartoni, 2010) montre une différence de 37.19%, ce qui reste une préférence pour l'adjectif dans la partie spécialisée, bien que la préférence ne soit pas aussi importante qu'attendue [6].

4 Discussion

Comme nous l'avons déjà mentionné, l'extraction de paraphrases en corpus comparable (et particulièrement de deux types de discours différents) peut s'effectuer sur la base de patrons lexicaux pré-définis. La méthode basée

6. Bien qu'au delà de notre seuil, ce patron est inclus dans le tableau 4.

Patrons	différence (%)
N_1 PROrel V_2 / V_2 N_1	-180.00%
N_1 V_2 / Vpper_2 P N_1	-176.92%
ADV_1 N_2 / N_2 A_1	-142.86%
N_1 A_2 / N_2 P N_1	-140.00%
N_1 Vpper_2 / V Vpper_2 P N_1	-90.00%
N_1 P N_2 / N_2 P V_1	-66.67%
N_1 Vpper_2 / Vpper_2 P N_1	-63.64%
N_1 A_2 / V_2 P N_1	-60.00%
N_1 A_2 / N_1 P N_2	37.19 %
N_1 P N_2 / N_2 V_1	50.00%
N_1 P N_2 / V_1 P N_2	52.14%
N_1 P N_2 / V Vpper_1 N_2	62.50%
A_1 N_2 / V_2 P A_1	66.67%
N_1 P N_2 / N_1 P V_2	70.00%
N_1 A_2 / V_1 ADV_2	73.08 %

TABLE 4 – Patrons bidirectionnels avec différence significative

Rank	Préc.	Patrons	Exemples
1	0.88	N_1 P N_2 / V_1 N_2	application de principes / appliquer le principe
2	0.94	V_1 N_2 / N_1 P N_2	aggraver une pathologie / aggravation de pathologie
4	0.82	N_1 P N_2 / V_1 P N_2	ajout d'insuline / ajouter de l'insuline
8	0.81	N_1 P N_2 / N_2 Vpper_1	administration de produit / produit administré
13	0.61	V_1 P N_2 / N_1 P N_2	adopter des stratégies / adoption d' une stratégie

TABLE 5 – Exemples de patrons impliquant une variation nom-verbe

sur les n-grammes présentée ici permet de valider empiriquement les intuitions qui étaient à l'origine des patrons lexicaux pré-définis, et de découvrir de nouveaux patrons de paraphrases.

Validation de patrons pré-définis : Cette méthode permet de confirmer les patrons pré-définis, non seulement en terme de fréquences mais également en terme de préférences. Pour les variations de type nom-verbe, différents types de patrons ont été trouvés, surtout parmi les patrons les mieux classés, comme montré dans le tableau 3. Le tableau 5 reprend les patrons les plus fréquents impliquant une variation nom-verbe et montre clairement que la variation nom-verbe apparaît dans une grande variété de cas, soit avec le verbe/nom dans la position de tête, ou avec un changement d'ordre quand le verbe est un participe passé qui prend la place de l'adjectif (p.ex dans l'exemple au rang 8).

Concernant les patrons impliquant des variations adjectif-nom, les adjectifs relationnels semblent favorisés dans la partie spécialisée du corpus (la direction N_1 A_2 $\rightarrow$ N_1 P N_2 est 37.19% plus fréquent que le patron inverse), ce qui confirme nos travaux antérieurs (Deléger & Cartoni, 2010).

Découverte de nouveaux patrons : Comparé à une approche basée sur des patrons prédéfinis, la méthode décrite ici permet de mettre en évidence des nouveaux patrons morphosémantiques fréquents. Le tableau 6 présente quelques uns de ces nouveaux patrons, qui impliquent un adjectif qualificatif (deux premiers exemples) ou deux variations morphosémantiques, à savoir une variation nom-verbe impliquant également un changement au niveau du "modifieur" : le nom étant modifié par un adjectif, le verbe doit être modifié par l'adverbe correspondant (*total* $\rightarrow$ *totalement*, rang 18). Le patron impliquant un adjectif qualificatif (au rang 23 du tableau 6) montre une préférence d'orientation : l'adjectif qualificatif est clairement préféré dans la partie grand public, alors que la partie spécialisée préfère la nominalisation de l'adjectif.

Rang	Préc.	Patrons	Exemples
7	0.64	N_2 P N_1 / N_1 A_2	efficacité d' un soutien / soutien efficace
23	0.33	N_1 A_2 / N_2 P N_1	accès facile / facilité d' accès
18	0.95	N_1 A_2 / V_1 ADV_2	traitement adéquat / traiter adéquatement
53	0.82	V_1 ADV_2 / N_1 A_2	arrêter totalement / arrêt total

TABLE 6 – Exemples de nouveaux patrons

5 Conclusion

Nous avons présenté une méthode d'extraction de patrons de paraphrases en corpus comparable monolingue, qui permet d'une part de valider d'autres méthodes d'extraction basées sur des patrons (construits à partir d'intuitions linguistiques) et d'autres part de découvrir d'autres patrons. Cette méthode confirme la fréquence des patrons investigués précédemment et permet d'étudier de manière très large les préférences de direction des patrons (specialisé→ grand public et/ou grand public→ spécialisé), apportant un nouveau regard sur la distinction entre ces différents types de discours. Nous avons ainsi pu confirmer la pertinence des patrons impliquant une variation nom-verbe, et la préférence pour la nominalisation dans la partie spécialisée, là où la partie grand public préfère les verbes. Dans une moindre mesure, la préférence pour l'adjectif relationnel dans la partie spécialisée a également été confirmée. D'un point de vue plus large, cette méthode confirme également l'importance de la morphologie dérivationnelle dans le phénomène de la paraphrase. Le nombre important de patrons mettant en jeu un adverbe doit également être souligné. A notre connaissance, aucune étude ne s'est penchée sur l'adverbialisation dans un tel contexte.

L'approche présentée ici est basée sur l'appariement de n-grammes ayant le même nombre de mots pleins (le même "n"). A l'avenir, nous envisageons d'étendre cette méthode à des appariements non-univoques (comme pour des paires 2-grammes=3-grammes, et vice versa). Évidemment, cette méthode pourrait également être appliquée à d'autres types de corpus comparable monolingue, construits autour d'autres distinctions que celle spécialisé/grand public, permettant ainsi de découvrir d'autres patrons de paraphrases spécifiques à d'autres types de discours.

Références

BANERJEE S. & PEDERSEN T. (2003). The design, implementation, and use of the ngram statistics package. In *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics*, p. 370–381, Mexico City.

BARZILAY R. (2003). *Information Fusion for Multidocument Summarization : Paraphrasing and Generation*. PhD thesis, Columbia University.

BARZILAY R. & LEE L. (2003). Learning to paraphrase : An unsupervised approach using multiple-sequence alignment. In *HLT-NAACL*, p. 16–23, Edmonton, Canada.

BARZILAY R. & MCKEOWN K. (2001). Extracting paraphrases from a parallel corpus. In *ACL/EACL*, p. 50–57.

DELÉGER L. & CARTONI B. (2010). Adjectifs relationnels et langue de spécialité : vérification d'une hypothèse linguistique en corpus comparable médical. In *Proceedings of TALN 2010*.

DELÉGER L. & ZWEIGENBAUM P. (2008). Paraphrase acquisition from comparable medical corpora of specialized and lay texts. In *Proceedings of the AMIA Annual Fall Symposium*, p. 146–150, Washington, DC.

ELHADAD N. & SUTARIA K. (2007). Mining a lexicon of technical terms and lay equivalents. In *ACL BioNLP Workshop*, p. 49–56, Prague, Czech Republic.

IBRAHIM A., KATZ B. & LIN J. (2003). Extracting structural paraphrases from aligned monolingual corpora. In *Proceedings of the second international workshop on Paraphrasing*, p. 57–64, Sapporo, Japan.

JACQUEMIN C. (1999). Syntagmatic and paradigmatic representations of term variation. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, p. 341–348, College Park, Maryland.

SHINYAMA Y. & SEKINE S. (2003). Paraphrase acquisition for information extraction. In *Proceedings of the second international workshop on Paraphrasing (IWP)*, p. 65–71, Sapporo, Japan.