TALN 2010 – Session Posters, Montréal, 19–23 juillet 2010

Acquisition de paraphrases sous-phrastiques depuis des paraphrases d'énoncés

Houda Bouamor Aurélien Max Anne Vilnat
LIMSI-CNRS, Univ. Paris-Sud
Orsay, F-91403, France
{prénom.nom}@limsi.fr

Résumé. Dans cet article, nous présentons la tâche d'acquisition de paraphrases sous-phrastiques (impliquant des paires de mots ou de groupes de mots), et décrivons plusieurs techniques opérant à différents niveaux. Nous décrivons une évaluation visant à comparer ces techniques et leurs combinaisons sur deux corpus de paraphrases d'énoncés obtenus par traduction multiple. Les conclusions que nous tirons peuvent servir de guide pour améliorer des techniques existantes.

Abstract. In this article, the task of acquiring sub-sentential paraphrases (word or phrase pairs) is discussed and several automatic techniques operating at different levels are presented. We describe an evaluation methodology to compare these techniques and their combination that is applied on two corpora of sentential paraphrases obtained by multiple translation. The conclusions that are drawn can be used to guide future work for improving existing techniques.

Mots-clés : Paraphrase, Patrons de correspondances de segments monolingues.

Keywords: Paraphrase, Monolingual bi-phrase patterns.

1 Introduction

Le problème de l'équivalence de sens entre segments textuels est au cœur des besoins du Traitement Automatique des Langues. La capacité à déterminer si deux mots, ou deux groupes de mots, ont la même signification dans leur contexte respectif permet de résoudre, au moins localement, les difficultés posées par la variation en langue. L'acquisition de groupes d'équivalence permet de constituer des ressources pouvant être utiles, par exemple, en génération pour aider des auteurs à trouver des formulations plus adaptées (Max, 2008). De nombreuses techniques ont été proposées pour l'acquisition de segments en relation de *paraphrase*. La plupart de ces techniques exploitent des corpus monolingues disponibles en grande quantité, et se fondent sur l'hypothèse que des unités linguistiques apparaissant de nombreuses fois dans des contextes similaires peuvent avoir la même signification.

En comparaison, peu de travaux ont porté sur l'exploitation de corpus monolingues *parallèles*, constitués de phrases en relation de paraphrase. Ce fait peut s'expliquer par la faible disponibilité de telles ressources et par leur coût de construction. Mais elles présentent des caractéristiques qui en font les candidates les plus naturelles pour l'étude de la paraphrase sous-phrastique. C'est donc dans ce cadre que se situe le présent travail qui vise en particulier à extraire des paraphrases de qualité dans le contexte défini par des

paires de paraphrases d'énoncés. Nous passerons tout d'abord en revue les principaux travaux portant sur l'acquisition de paraphrases sous-phrastiques (section 2), puis nous décrirons plus en détails trois techniques opérant au niveau des mots, des termes et de la syntaxe de la phrase sur des corpus monolingues parallèles (section 3). Nous présenterons ensuite un cadre expérimental (section 4) visant à comparer et à combiner de façon simple ces approches et nous terminerons par une discussion des résultats obtenus et une description de nos prochains travaux (section 5).

2 Travaux précédents en acquisition de paraphrases locales

Une hypothèse largement suivie pour extraire des segments textuels équivalents est que deux mots, et par extension deux segments, qui partagent des contextes similaires peuvent être interchangeables. Cette hypothèse de *distributionnalité* a notamment été appliquée au cas de chemins de dépendances syntaxiques par Lin & Pantel (2001). Des corpus comparables monolingues, dans lesquels un même contenu est probablement décrit sous plusieurs formes, permettent de guider la mise en correspondance d'équivalences locales. Ainsi, par exemple, Barzilay & Lee (2003) introduisent une technique d'alignement multi-séquence factorisant des phrases structurellement similaires de ces corpus sous forme de graphes, qui contiennent par nature des équivalences locales. Une autre hypothèse utilisée est que des segments partageant des traductions dans une autre langue peuvent être des paraphrases dans certains contextes. Bannard & Callison-Burch (2005) ont ainsi décrit une approche exploitant plusieurs corpus parallèles, et les travaux de Callison-Burch (2008) et Max (2008) prennent en compte le contexte (monolingue) pour évaluer la qualité des équivalences apprises.

Toutes les approches précédentes sont limitées par le fait que de nombreux contextes ne correspondent pas à des emplois équivalents. Si des énoncés en relation de paraphrase sont disponibles, la tâche d'extraction de segments équivalents peut être attaquée comme une tâche d'alignement fine où la mise en correspondance est limitée aux mots et segments de la phrase. Des traductions multiples d'un même énoncé ont notamment été utilisées dans ce cadre (Pang *et al.*, 2003; Callison-Burch *et al.*, 2008).

3 Acquisition depuis des paraphrases d'énoncés

3.1 Approche fondée sur l'apprentissage statistique d'alignements entre mots (MOT)

En traduction automatique statistique, l'approche la plus utilisée pour trouver des correspondances entre segments dans des corpus parallèles bilingues consiste à apprendre des alignements entre mots dans chaque direction de traduction (Och & Ney, 2003), puis à utiliser une heuristique de combinaison pour « symétriser » les alignements obtenus. Les modèles d'alignement appliqués sur deux phrases parallèles résultent de l'apprentissage sur l'ensemble des bitextes disponibles. Ainsi, plus la quantité de données est importante et plus les données sont *parallèles*, et plus les correspondances au niveau de chaque phrase sont précises. En outre, ces modèles tendent à aligner le plus de mots possibles en suivant l'hypothèse forte de parallélisme entre phrases d'une même paire. Dans ce travail, nous constituons un corpus parallèle monolingue et réalisons toutes les mises en correspondance possibles entre énoncés appartenant à un même groupe de

paraphrases, afin d'augmenter la quantité de données d'apprentissage. Nous utilisons le système MOSES[1] pour réaliser l'alignement avec le programme GIZA++ (Och & Ney, 2003) pour l'alignement dans chaque direction de traduction puis des heuristiques de symétrisation. Une fois une matrice d'alignement de mots obtenue, nous ne considérons que les bi-segments dont les mots sources (resp. cibles) sont alignés avec au moins un mot du segment cible (resp. source) et ne sont alignés qu'avec des mots de ce segment.

3.2 Approche fondée sur l'expression symbolique de la variation de termes (TERME)

Deux énoncés en relation de paraphrase utilisent des mots communs ou des mots ou groupes de mots en relation d'équivalence. Pour chaque paire de tels groupes, et sous certaines hypothèses, il est possible d'exprimer des règles régissant les variations syntagmatiques et paradigmatiques acceptables.

L'opération d'*indexation contrôlée* du système FASTR (Jacquemin, 1999) définit les variations acceptables par un système de métarègles s'appliquant à des règles de termes. Elles permettent d'exprimer les réécritures morphosyntaxiques possibles, ainsi que les relations (au niveau des lemmes) d'ordre morphologique ou sémantique contenues dans des ressources préexistantes. Dans notre travail, nous avons utilisé cette indexation contrôlée pour trouver les alignements possibles entre les deux paraphrases d'une paire donnée et qui sont trouvés dans les deux directions. Étant données deux paraphrases d'énoncé, nous recherchons dans une première phrase (notre « corpus ») des variantes pour chacun des segments possibles de l'autre (à concurrence d'une certaine taille), puis nous inversons la recherche et retenons l'intersection des résultats.

3.3 Approche fondée sur l'alignement d'énoncés par fusion syntaxique (SYNT)

L'exploitation du caractère parallèle de deux énoncés peut être poussée encore plus loin : si ces énoncés partagent une même structure syntaxique, il est possible de réaliser un alignement fin guidé par la syntaxe permettant de faire apparaître des correspondances sous-phrastiques fines. C'est cette idée qui est mise en œuvre dans l'approche de fusion syntaxique proposée par Pang *et al.* (2003). Leur algorithme consiste essentiellement à fusionner récursivement les arbres de constituants de deux énoncés là où les listes de catégories filles sont compatibles et qu'aucune évidence de non parallélisme syntaxique (via un mécanisme de *blocage lexical*) n'est détectée. La forêt d'arbre syntaxique ainsi obtenue permet de construire un automate à états finis représentant des formulations alternatives qu'il est possible d'extraire par simple parcours de l'automate en bornant la taille des segments recherchés.

Notre réimplémentation de cet algorithme améliore sa robustesse et sa correction. Nous avons tout d'abord implémenté un mode de fusion flexible, où les parties de la phrase non concernées par un blocage lexical sont tout de même fusionnées. Par ailleurs, l'algorithme étant très dépendant de la qualité des analyses syntaxiques, nous avons implémenté un mode dans lequel les *k* meilleures analyses produites par un analyseur probabiliste sont utilisées, et où la combinaison retenue entre la i-ème analyse du premier énoncé et la j-ème analyse du second parmi les k^2 combinaisons possibles est celle minimisant le nombre de nœuds de l'automate obtenu avant *réduction* (fusion d'arcs correspondant à des préfixes ou à des suffixes communs). Pour notre implémentation, nous avons utilisé l'analyseur syntaxique probabiliste de Berkeley (Klein & Manning, 2003) appris sur le français pour produire les 5 meilleures analyses pour chaque énoncé, et nous avons effectué une recherche exhaustive de la meilleure fusion pour chaque paire.

[1] `http://www.statmt.org/moses`

4 Cadre expérimental

4.1 Constitution du corpus d'évaluation

Pour construire des corpus de développement et d'évaluation, nous avons eu recours à une collecte de traductions multiples de débats parlementaires Européens du corpus Europarl[2] vers le français via une même source depuis plusieurs langues (Bouamor, 2010). Nous avons extrait deux sous-corpus constitués de groupes de paraphrases pour les 50 mêmes énoncés afin de permettre la comparaison des techniques d'acquisition étudiées en fonction du caractère plus ou moins littéral des traductions et donc des paraphrases obtenues. Un premier corpus est constitué de paraphrases obtenues par traduction depuis une même phrase en anglais, et un second corpus est constitué par des traductions depuis l'allemand, l'espagnol, l'italien et le portugais. Nous disposons donc de deux corpus constitués de 50 groupes de 4 paraphrases. Nous considérons une paraphrase de chaque groupe comme une paraphrase « de référence », qui devra être alignée avec chacune des 3 autres paraphrases de son groupe. Trois annotateurs ont aligné au niveau des mots chacune des 300 paires de paraphrases (2 corpus x 50 groupes x 3 alignements).

4.2 Résultats expérimentaux

Nous avons suivi l'approche PARAMETRIC décrite dans (Callison-Burch *et al.*, 2008) pour évaluer les différentes techniques étudiées, dans laquelle un ensemble de bi-segments de référence est comparé aux bi-segments proposés par une technique d'acquisition opérant sur une paire d'énoncés par le calcul de valeurs de précision et de rappel, définies respectivement comme la proportion des candidats proposés appartenant à la référence (précision) et la proportion des éléments de la référence proposés (rappel). Chacune des techniques présentées dans la section 3 permet d'extraire un certain nombre de bi-segments candidats pour chaque paire d'énoncés disponible. Les bi-segments constitués d'une paire d'un même segment (indépendamment de la casse) ne sont pas considérés, et une même taille maximale de 6 *tokens* est appliquée aux deux segments. Nous avons suivi deux stratégies élémentaires pour combiner les résultats de plusieurs techniques : 1) la fusion d'hypothèses par l'union (sans doublon) des listes de bi-segments proposées par plusieurs techniques, visant à augmenter la quantité d'hypothèses produites ; 2) la fusion par l'intersection des listes de bi-segments, visant à augmenter la qualité des hypothèses retenues. La Table 1 présente les résultats obtenus pour les techniques étudiées et certaines de leurs combinaisons.

Le résultat le plus marquant concerne les différences de performance des différentes techniques, tant au niveau de la précision que du rappel, sur les deux corpus : il est beaucoup plus aisé de réaliser la tâche d'extraction de bi-segments lorsque les paraphrases utilisées sont proches. La technique fondée sur la syntaxe (SYNT) y est bien sûr très sensible, tout comme celle fondée sur l'alignement entre mots (MOT), ce qui se retrouve en traduction statistique pour des paires de langues réputées difficiles à aligner. Concernant le rappel, une forte disparité existe entre l'approche statistique d'un côté (MOT) et les approches symboliques de l'autres (TERME et SYNT), ces deux dernières proposant comparativement peu de bi-segments de référence. Il est notable que MOT parvienne tout de même à obtenir une précision correcte, bien que moins importante que celle de SYNT. Les relatives bonnes performances générales de MOT étaient attenudes, et sont prometteuses dans un contexte où davantage de données viendraient à être disponibles. TERME se spécialise dans l'extraction de bi-segments de certains types, qui sont en très petit nombre en regard de

[2] http://www.statmt.org/europarl

	MOT	**TERME**	**SYNT**	**M∪T**	**M∩T**	**T∪S**	**T∩S**	**M∪S**	**M∩S**	**M∪T∪S**
Paraphrases obtenues par traduction de l'anglais										
P	41,94	41,19	50,16	41,54	55,97	46,48	80,76	40,83	71,21	40,46
$\mathbf{R}_{/3578}$	67,07	3,07	8,77	67,66	2,48	11,26	0,58	67,83	8,02	68,41
$\mathbf{F}_1$	51,61	5,87	14,93	51,47	4,76	18,13	1,16	50,98	14,41	50,58
Paraphrases obtenues par traduction depuis plusieurs langues										
P	27,05	35,98	40,46	27,08	42,98	39,46	28,57	26,91	50,15	26,90
$\mathbf{R}_{/2517}$	51,80	3,05	8,26	52,92	1,94	11,08	0,23	53,43	6,63	54,39
$\mathbf{F}_1$	35,54	6,07	13,72	35,83	3,72	17,30	0,47	35,79	11,72	36,00

TAB. 1 – Résultats obtenus pour chaque technique et certaines combinaisons

la taille de l'ensemble de référence, mais d'une précision correcte. Les différentes combinaisons essayées font apparaître des gains attendus en rappel lorsqu'une union de résultats est réalisée, et en précision lorsqu'une intersection est réalisée. Sur le premier corpus, la précision maximale est atteinte en réalisant des intersections impliquant les résultats de TERME, et le rappel maximal en réalisant des unions impliquant les résultats de MOT. Les résultats sont sensiblement les mêmes sur le second corpus, à l'exception de l'intersection des résultats de SYNT avec ceux de TERME.

Para 1 (**de**)	En ce qui concerne les relations internationales , la communauté doit s' y attaquer de manière déterminée et s' accorder avec la politique extérieure
Para 2 (**it**)	Quant aux relations internationales , la Communauté est confrontée aux décisions relatives à la politique étrangère
REF	**(En ce qui concerne↔Quant aux) (En ce qui concerne les relations↔Quant aux relations) (la politique extérieure↔la politique étrangère) (politique extérieure↔politique étrangère) (extérieure↔étrangère)**
MOT	(En↔Quant) (ce↔à) (concerne les↔aux) (concerne les relations↔aux relations) (concerne les relations internationales↔aux relations internationales) (concerne les relations internationales ,↔aux relations internationales ,) (concerne les relations internationales , la↔aux relations internationales , la) (la communauté doit s' y attaquer↔la Communauté est confrontée aux décisions) (communauté doit s' y attaquer↔Communauté est confrontée aux décisions) (doit s' y attaquer↔est confrontée aux décisions) (déterminée↔relatives) **(la politique extérieure↔la politique étrangère) (extérieure↔étrangère)**
TERME	**(politique extérieure↔politique étrangère) (extérieure↔étrangère)**
SYNT	**(En ce qui concerne les↔Quant aux)**

TAB. 2 – Exemples de bi-segments extraits par les différentes techniques à partir d'une paire de paraphrases produites depuis l'allemand et l'italien (les bi-segments en gras appartiennent à l'ensemble de référence)

La Table 2 illustre ces résultats sur un exemple d'alignement entre deux paraphrases obtenues à partir de l'allemand et de l'italien, dont l'alignement s'avère difficile comme confirmé par le faible nombre de bi-segments dans l'ensemble de référence. MOT n'a pas pu aligner de façon fiable les mots et a produit de nombreux bi-segments incorrects. TERME et SYNT n'ont au contraire proposé que peu de candidats (appartenant à l'ensemble de référence), ce qui reflète là aussi les difficultés de mise en correspondance rencontrées par ces deux techniques. Concernant plus précisément SYNT, il est par exemple à noter que le bi-segment (la communauté↔la Communauté), correspondant à des syntagmes nominaux dans des positions syntaxiquement compatibles dans les deux phrases, a été proposé mais filtré (c'est également le cas pour TERME) ; les syntagmes verbaux, de structures différentes, n'ont eux pu être alignés.

5 Discussion et travaux futurs

Dans cet article, nous avons décrit et situé la tâche d'extraction de paraphrases sous-phrastiques à partir d'énoncés en relation de paraphrase, une condition difficile à obtenir mais qui permet de se concentrer sur un cadre naturel d'étude des phénomènes de paraphrase. En outre, les correspondances extraites peuvent être facilement associées à un contexte, ce qui peut être utilisé par la suite pour initialiser la caractérisation des contextes dans lesquels ces paraphrases sont valides. Les différentes techniques que nous avons mises en œuvre, à l'origine développées pour des besoins différents, se sont révélées relativement complémentaires, et permettent sous certaines conditions d'obtenir des résultats acceptables en termes de précision et de rappel relativement à un ensemble de référence.

Nos travaux futurs s'orientent dans trois directions : tout d'abord, nous souhaitons pouvoir généraliser les correspondances obtenues et viser l'acquisition de patrons intégrant des éléments de contexte. Ensuite, nous souhaitons étendre les techniques existantes et travailler à une hybridation efficace lors de l'acquisition elle-même, et non plus simplement sous forme de fusion de résultats. Enfin, nous souhaitons mettre en œuvre des techniques de validation à plus large échelle sur des corpus monolingues comparables des paraphrases obtenues.

Références

BANNARD C. & CALLISON-BURCH C. (2005). Paraphrasing with bilingual parallel corpora. In *Actes de ACL*, p. 597–604, Ann Arbor, USA.

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

BOUAMOR H. (2010). Construction d'un corpus de paraphrases d'énoncés par traduction multilingue multisource. In *Récital-TALN*, Montréal, Canada.

CALLISON-BURCH C. (2008). Syntactic constraints on paraphrases extracted from parallel corpora. In *Actes de EMNLP*, p. 196–205, Hawai, USA.

CALLISON-BURCH C., COHN T. & LAPATA M. (2008). Parametric : An automatic evaluation metric for paraphrasing. In *Actes de COLING*, p. 97–104, Manchester, UK.

JACQUEMIN C. (1999). Syntagmatic and paradigmatic representations of term variation. In *Actes de ACL*, p. 341–348, College Park, États-Unis.

KLEIN D. & MANNING C. (2003). A* parsing : Fast exact viterbi parse selection. In *Actes de NAACL-HLT*, p. 119–126, Edmonton, Canada.

LIN D. & PANTEL P. (2001). Discovery of inference rules for question answering. *Natural Language Engineering*, **7**(4), 343–360.

MAX A. (2008). Génération de reformulations locales par pivot pour l'aide à la révision. In *Actes de TALN*, Avignon, France.

OCH F. J. & NEY H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, p. 19–51.

PANG B., KNIGHT K. & MARCU D. (2003). Syntax-based alignement of multiple translations : Extracting paraphrases and generating new sentences. In *Actes de NAACL-HLT*, Edmonton, Canada.