TALN 2010, Montréal, 19–23 juillet 2010

L'évaluation des paraphrases : pour une prise en compte de la tâche

Jonathan Chevelu[1, 2] Yves Lepage[1] Thierry Moudenc[2] Ghislain Putois[2]
(1) Université de Caen, France
(2) Orange Labs, France
{jonathan.chevelu, thierry.moudenc, ghislain.putois}@orange-ftgroup.com,
yves.lepage@info.unicaen.fr

Résumé. Les définitions des paraphrases privilégient généralement la conservation du sens. Cet article démontre par l'absurde qu'une évaluation uniquement basée sur la conservation du sens permet à un système inutile de production de paraphrase d'être jugé meilleur qu'un système au niveau de l'état de l'art. La conservation du sens n'est donc pas l'unique critère des paraphrases. Nous exhibons les trois objectifs des paraphrases : la conservation du sens, la naturalité et l'adaptation à la tâche. La production de paraphrase est alors un compromis dépendant de la tâche entre ces trois critères et ceux-ci doivent être pris en compte lors des évaluations.

Abstract. Meaning preservation is generally rooted in the paraphrase definitions. This article proves by *reductio ad absurdum* that an evaluation based only on the meaning preservation can rank a dummy and useless system better than a state-of-the-art system. Meaning preservation is therefore not the one and only criterion for a paraphrase system. We exhibit the three objectives of paraphrase : meaning preservation, sentence naturalness and adequacy for the task. Paraphrase generation consists actually in reaching a task-dependent compromise between these three criteria, and they have to be taken into account during each evaluation process.

Mots-clés : Générateur de paraphrase, évaluation des paraphrases.

Keywords: Paraphrase generator, paraphrase evaluation.

1 Introduction

Des définitions variées de la paraphrase existent dans le domaine de la génération automatique de paraphrases. (Barzilay & McKeown, 2001) définissent « *Les paraphrases [comme] des moyens alternatifs de transmettre la même information* » [1] ou « *le même sens* » [2] pour (Zhao *et al.*, 2009). (Sekine, 2005) la définit ainsi : « *Une « paraphrase », à savoir un ensemble de phrases qui expriment la même chose ou le même évènement* »[3]. Ces définitions insistent sur une relation de conservation – qu'elle soit du sens, de l'information, du message ou de l'intention communicative – entre la phrase d'origine et la paraphrase produite.

Cette focalisation sur la conservation du sens fait que les systèmes de génération et leurs évaluations oublient de prendre en compte la globalité des objectifs de la production automatique de paraphrases : la conservation du sens, la naturalité, mais aussi l'adaptation à la tâche.

Dans la section 2, nous démontrons par l'absurde que l'évaluation d'un système qui ne prend pas en compte l'ensemble des objectifs n'est pas pertinente. Dans la section 3, nous montrons que notre segmentation des objectifs de la génération en trois catégories couvre les principaux usages visés par le domaine.

2 Un (trop) bon générateur de paraphrases

L'évaluation de générateurs de paraphrase est un problème difficile que beaucoup de travaux abordent par des méthodes différentes et incompatibles entre elles (Callison-Burch *et al.*, 2008).

Un générateur de paraphrases est souvent un composant d'un système plus complexe. Une façon de l'évaluer est de mesurer les performances du système auquel il appartient au moyen d'un critère global. Le problème est que pour les applications où la paraphrase est présentée à l'utilisateur final, il n'est pas possible de se contenter d'une évaluation orientée tâche, comme dans (Cahill *et al.*, 2009). Par exemple, un générateur de paraphrases peut être utilisé conjointement à un système de synthèse vocale. Si celui-ci introduit une négation, il se peut que la synthèse vocale soit améliorée, mais le générateur dénature le message d'origine. Un tel message ne peut être considéré comme paraphrase. De plus, une évaluation comportant uniquement un critère lié à une tâche ne permet pas de généraliser les résultats à d'autres problèmes. Ceci fait qu'il y a peu de protocoles d'évaluation orientés tâche (Callison-Burch *et al.*, 2008).

Lorsqu'ils ne sont pas évalués en fonction des performances globales d'un système, les générateurs sont systématiquement évalués selon le critère de conservation du sens (Barzilay & McKeown, 2001; Bannard & Callison-Burch, 2005; Max, 2008; Chevelu *et al.*, 2009). La naturalité est, elle aussi, souvent évaluée pour s'assurer que les paraphrases sont syntaxiquement correctes (Bannard & Callison-Burch, 2005; Max, 2008; Chevelu *et al.*, 2009) .

L'objectif de l'expérience suivante est de démontrer qu'une évaluation focalisée sur la conservation du sens n'est pas pertinente. De plus, l'ajout d'un critère d'évaluation lié à la naturalité est certes appréciable mais pas suffisant. Nous proposons de réaliser cette démonstration par l'absurde, en comparant deux systèmes uniquement sur des critères de conservation de sens et de naturalité. Nous montrons qu'un système inutile

[1] Paraphrases are alternative ways to convey the same information.
[2] the same meaning.
[3] "paraphrase", i.e. a set of phrases which express the same thing or event.

Système	*Référence*	*Test*
Sens préservé	$48/100$	**99 / 100**
Syntaxe correcte	$51/100$	$51/100$
Syntaxe correcte et sens préservé	$35/100$	**50 / 100**

TAB. 1 – Le système *Test* est meilleur en termes de conservation du sens et aussi lorsque les deux critères sont combinés. Le coefficient Kappa d'accord entre les juges est de $0,58$.

peut, dans une telle évaluation, être jugé meilleur qu'un système au niveau de l'état de l'art.

Le premier système considéré est le générateur de paraphrase statistique de référence décrit dans (Chevelu *et al.*, 2009). Ce système *Référence* est censé avoir des performances au niveau de l'état de l'art (Callison-Burch *et al.*, 2008; Zhao *et al.*, 2009). Le générateur de paraphrase repose sur un décodeur statistique et sur une table de paraphrase construite à l'aide d'une langue pivot. Nous utilisons l'anglais comme langue pivot pour produire une table de paraphrases du français.

Le second système est conçu spécialement pour cette expérience. Puisque les systèmes ne seront comparés qu'en fonction du critère de conservation du sens et de la correction syntaxique des paraphrases produites, l'objectif est de concevoir un système prenant le moins de « risques » possibles. Nous imposons tout de même aux systèmes de proposer uniquement des paraphrases différentes de la phrase d'origine. Le générateur de test consiste à supprimer les virgules de la phrase origine s'il y en a. Si la phrase ne comporte pas de virgule, nous utilisons un second système composé lui aussi du décodeur statistique `Moses`. La table de paraphrases utilisée est celle du système de référence où sont conservées uniquement les entrées qui n'entraînent pas de modification du segment ou qui se contentent d'ajouter une virgule. Ce système *Test* est donc capable de supprimer des virgules ou d'en ajouter à la phrase d'origine.

Le corpus d'entraînement est constitué de $1\,576\,897$ phrases extraites des débats au Parlement Européen et regroupées dans le corpus Europarl (Koehn, 2005). 100 phrases sélectionnées aléatoirement sont retirées du corpus d'entrainement pour constituer un corpus d'évaluation.

L'évaluation est réalisée sur une plateforme en ligne dédiée à l'évaluation des paraphrases comme présentée dans (Chevelu *et al.*, 2009). Les évaluateurs peuvent répondre par « Oui », « Non » ou « Ne sais pas ». Une paraphrase est jugée comme correcte pour une question uniquement si tous les évaluateurs la jugent correcte. Pour cette expérience, chaque paraphrase est évaluée par deux francophones.

Les résultats sont présentés dans le tableau 1. Le coefficient Kappa d'accord entre les juges est de $0,58$ (p-valeur $< 10^{-3}$), ce qui est traditionnellement interprété comme « modéré ». Le corpus de test extrait aléatoirement semble plus difficile pour le générateur de paraphrase que ceux d'expériences précédentes (Chevelu *et al.*, 2009). Les résultats montrent que le système *Test* est significativement meilleur que le générateur de référence sur ce corpus de test en termes de conservation du sens. L'ajout du critère de naturalité réduit les performances des deux systèmes mais le générateur *Test* reste meilleur de 142%. On peut tout de même se demander si ces résultats reflètent bien l'intérêt de chaque système. Évidemment, en minimisant les risques pris, le système qui manipule uniquement des virgules arrive à produire de meilleurs résultats mais est probablement inutile pour la majorité des applications réelles de la génération de paraphrase.

Cette expérience démontre qu'une évaluation basée uniquement sur la conservation du sens n'est pas pertinente, même en ajoutant un critère sur la naturalité.

3 Les buts de la génération de paraphrase

Si la conservation du sens et la naturalité seules ne permettent pas d'évaluer un générateur de paraphrases c'est que cette activité doit avoir d'autres buts. Nous posons les objectifs suivants comme les buts de la production de paraphrase :
– la conservation du sens : c'est le but premier de la paraphrase. Ma phrase d'origine sert de référence en termes de sens à produire ;
– la naturalité : il est nécessaire que la paraphrase soit syntaxiquement correcte, afin qu'elle ait un sens ;
– l'adéquation à la tâche : la génération automatique de paraphrase n'est pas une activité en soi (contrairement à la traduction par exemple). La génération de paraphrase est toujours associée à une tâche et intégrée dans un processus plus vaste. Les paraphrases produites doivent être adaptées à l'usage qu'il en sera fait.

Nous définissons la génération de paraphrase comme un compromis entre ces trois objectifs, le réglage de ce compromis étant dépendant de la tâche. À ce compromis s'ajoute une séparation en deux classes selon que le lieu des modifications est imposé ou non.

Ces trois objectifs se retrouvent dans les différentes classes d'applications qui utilisent ou pourraient utiliser un générateur de paraphrases. Nous segmentons ces applications en fonction des traitements réalisés, après génération, sur les paraphrases proposées.

La première catégorie correspond aux cas où les paraphrases ne sont plus modifiées par le système après génération. La paraphrase est une sortie du système global. Les applications fréquemment citées sont :
– la compression de phrase (Zhao *et al.*, 2009) : le but est de produire une phrase plus courte en nombre de caractères que la phrase d'origine. Cette application est une version simplifiée du résumé de texte ;
– la synthèse de la parole (Boidin *et al.*, 2009; Cahill *et al.*, 2009) : le but est de modifier la phrase d'origine afin que la paraphrase produite, une fois vocalisée, ait une plus grande pertinence. Les travaux actuels se concentrent sur l'amélioration de l'acoustique. Cette application appartient à cette catégorie car le synthétiseur vocal ne modifie pas la phrase ni, en particulier, son sens.

La seconde catégorie comprend les systèmes où la paraphrase générée peut être modifiée. Mais, la phrase en sortie du système global doit conserver le sens de la phrase d'entrée du générateur de paraphrase. Pour les applications de ce type, c'est souvent un opérateur humain qui réalise les modifications :
– l'aide à l'écriture (Max, 2008) : le but est de fournir des alternatives à un rédacteur lorsque celui-ci n'est pas satisfait d'une partie de la phrase (répétition, terminologie, ...) ;
– l'aide à la conception des messages dans un système de dialogue (Boidin *et al.*, 2009) : le but est de proposer des alternatives à un concepteur de système de dialogue humain-machine pour les messages du système. Par exemple, ces alternatives peuvent prendre en compte l'ensemble des messages déjà construits pour conserver une certaine homogénéité.

Dans la troisième catégorie, la paraphrase n'est pas directement reliée à la sortie du système global. Elle est une étape de calcul intermédiaire. En général, la paraphrase sert d'entrée à un système de traitement automatique de la langue :
– un système de recherche d'information (Sekine, 2005) : le but est d'améliorer la couverture des schémas que le système sait reconnaître dans un texte en « normalisant » les phrases d'entrée grâce à la relation de paraphrase ;
– un système de traduction automatique (Callison-Burch *et al.*, 2006) : le but est de remplacer les séquences difficiles à traduire pour le système en séquence de mots plus facile à traduire. La majorité des travaux actuels se concentrent sur le remplacement des mots hors vocabulaires.

Chacune de ces applications cherche à conserver le sens de la phrase d'entrée. Les applications de la première catégorie contraignent fortement la qualité syntaxique des paraphrases à produire puisqu'elles sont présentées directement à un utilisateur final. La contrainte sur la conservation du sens et sur la naturalité des paraphrases produites est en revanche moins forte pour les applications de la seconde catégorie puisque les paraphrases peuvent être corrigées par la suite. Ceci est encore plus vrai pour la troisième catégorie où les paraphrases ne sont pas des sorties du système. Le réglage du compromis servant à définir ce qu'est une paraphrase correcte pour une application donnée est donc dépendant de la tâche.

De plus, nous distinguons deux types de problèmes de génération. En fonction de la tâche, les mots modifiables peuvent être imposés (Sekine, 2005; Callison-Burch *et al.*, 2006; Max, 2008). C'est le cas, par exemple, pour le problème de traduction automatique où ce sont les mots ou expressions hors vocabulaire qui sont à modifier. D'autres tâches n'imposent pas le lieu de la phrase d'origine à modifier (Barzilay & McKeown, 2001; Chevelu *et al.*, 2009; Zhao *et al.*, 2009). C'est le cas de la compression de texte par exemple. Ces deux types de problèmes engendrent deux types de générateurs potentiellement différents et difficilement comparables entre eux. C'est pourquoi nous ne caractérisons pas cette séparation comme un objectif de la génération de paraphrase mais comme deux classes de problèmes distincts.

4 Conclusion et discussion

Trop peu de travaux utilisent un critère lié à la tâche lors de l'évaluation et se contentent généralement d'une évaluation de la conservation du sens et parfois de la naturalité. En comparant un système inutile, manipulant uniquement des virgules, à un système au niveau de l'état de l'art, nous avons montré que les évaluations basées principalement sur la conservation du sens ne sont pas pertinentes. Contrairement à d'autres problèmes de traitement automatique des langues, l'absence de critère de tâches peut valoriser des systèmes aberrants. Cette spécificité est due à la phrase d'origine qui est optimale en termes de conservation du sens mais n'est pas valide uniquement car les paraphrases doivent lui être différentes.

Nous avons mis en évidence trois objectifs communs aux problèmes de production de paraphrases qui se retrouvent dans les applications. Nous avons pu définir la génération de paraphrase comme la production d'une phrase réalisant un compromis entre trois critères : la conservation du sens, la naturalité et l'adaptation à une tâche. Le lieu des modifications peut être imposé ou libre.

Afin de pouvoir comparer différents systèmes, nous pensons qu'il est nécessaire d'utiliser un critère générique de tâche lors de la description des performances d'un système. De préférence, celui-ci doit être le plus simple possible et ne pas nécessiter de ressources externes. Nous envisageons d'utiliser la maximisation du *taux d'erreur en caractères* (CER) :

$$\text{CER}(\text{Paraphrase}|\text{Origine}) = \frac{\text{distance d'édition}(\text{Paraphrase}, \text{Origine})}{|\text{Origine}|}$$

En effet, nous pensons que plus une paraphrase est différente de la phrase d'origine en termes de forme, et plus il est difficile de s'assurer qu'elle soit correcte syntaxiquement et sémantiquement. Ainsi, un générateur capable d'avoir de bonnes performances sur ces trois critères montrerait ses facultés à produire des paraphrases variées mais serait aussi probablement capable de produire des paraphrases plus simples si une tâche applicative l'imposait. Notons qu'il ne semble pas pertinent de définir un niveau minimal de transformation en dehors de l'interdiction de l'identité. En effet, pour certaines tâches, comme la synthèse vocale ou la normalisation de texte, l'impact d'une virgule peut être très important. Nous préfèrerons donc une mesure continue comme le CER qu'un seuil arbitraire difficilement réglable. Pour les systèmes présentés dans cet article, le CER moyen du système *Référence* est de $0{,}178 \pm 0{,}094$ contre $0{,}040 \pm 0{,}036$ pour

le système *Test*. Comme attendu, le système *Test* affiche des performances plus faibles pour le critère de tâche générique, ce qui le rend objectivement moins intéressant pour beaucoup d'applications.

Nous avons montré que la prise en compte simultanée des trois critères est nécessaire à l'évaluation correcte des paraphrases. Nous pensons qu'il en est de même avec la production de paraphrase et qu'il faut adapter les modèles et les générateurs pour remplir conjointement les trois objectifs.

Références

BANNARD C. & CALLISON-BURCH C. (2005). Paraphrasing with bilingual parallel corpora. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, p. 597–604, Ann Arbor, Michigan, USA : Association for Computational Linguistics.

BARZILAY R. & MCKEOWN K. R. (2001). Extracting paraphrases from a parallel corpus. In *Proceedings of 39th Annual Meeting of the Association for Computational Linguistics*, p. 50–57, Toulouse, France : Association for Computational Linguistics.

BOIDIN C., RIESER V., VAN DER PLAS L., OLIVER L. & CHEVELU J. (2009). Predicting how it sounds : Re-ranking dialogue prompts based on TTS quality for adaptive Spoken Dialogue Systems. In *Proceedings of the Interspeech Special Session : Machine Learning for Adaptivity in Spoken Dialogue*, p. 2487–2490, Brighton, UK : ISCA.

CAHILL P., DU J., WAY A. & JULIE C.-B. (2009). Using Same-Language Machine Translation to Create Alternative Target Sequences for Text-To-Speech Synthesis. In *Proceedings of the 10th Annual Conference of the International Speech Communication Association (Interspeech 2009)*, Brighton, UK : ISCA.

CALLISON-BURCH C., COHN T. & LAPATA M. (2008). Parametric : an automatic evaluation metric for paraphrasing. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, p. 97–104, Manchester, UK : Coling 2008 Organizing Committee.

CALLISON-BURCH C., KOEHN P. & OSBORNE M. (2006). Improved statistical machine translation using paraphrases. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, p. 17–24, Morristown, NJ, USA : Association for Computational Linguistics.

CHEVELU J., LAVERGNE T., LEPAGE Y. & MOUDENC T. (2009). Introduction of a new paraphrase generation tool based on Monte-Carlo sampling. In *Proceedings of the ACL-IJCNLP Conference Short Papers*, p. 249–252, Singapore : Association for Computational Linguistics.

KOEHN P. (2005). Europarl : a parallel corpus for statistical machine translation. In *Procedings of MT summit X, the tenth machine translation summit*, p. 79–86, Phuket, Thailand : Asia-Pacific Association for Machine Translation.

MAX A. (2008). Génération de reformulations locales par pivot pour l'aide à la révision. In *Actes de TALN 2008 (Traitement automatique des langues naturelles)*, Avignon : ATALA.

SEKINE S. (2005). Automatic paraphrase discovery based on context and keywords between ne pairs. In *Proceedings of International Workshop on Paraphrase (IWP2005)*.

ZHAO S., LAN X., LIU T. & 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*, p. 834–842, Suntec, Singapore : Association for Computational Linguistics.