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

Utilisation d'un score de qualité de traduction pour le résumé multi-document cross-lingue

Stéphane Huet[1] Florian Boudin[1] Juan-Manuel Torres-Moreno[1,2,3]
(1) LIA, Université d'Avignon, France
(2) École Polytechnique de Montréal, Canada
(3) GIL-IINGEN, Universidad Nacional Autónoma de México, Mexique
{stephane.huet,florian.boudin,juan-manuel.torres}@univ-avignon.fr

Résumé. Le résumé automatique cross-lingue consiste à générer un résumé rédigé dans une langue différente de celle utilisée dans les documents sources. Dans cet article, nous proposons une approche de résumé automatique multi-document, basée sur une représentation par graphe, qui prend en compte des scores de qualité de traduction lors du processus de sélection des phrases. Nous évaluons notre méthode sur un sous-ensemble manuellement traduit des données utilisées lors de la campagne d'évaluation internationale DUC 2004. Les résultats expérimentaux indiquent que notre approche permet d'améliorer la lisibilité des résumés générés, sans pour autant dégrader leur informativité.

Abstract. Cross-language summarization is the task of generating a summary in a language different from the language of the source documents. In this paper, we propose a graph-based approach to multi-document summarization that integrates machine translation quality scores in the sentence selection process. We evaluate our method on a manually translated subset of the DUC 2004 evaluation campaign. Results indicate that our approach improves the readability of the generated summaries without degrading their informativity.

Mots-clés : Résumé cross-lingue, qualité de traduction, graphe.

Keywords: Cross-lingual summary, translation quality, graph.

1 Introduction

La multiplication des documents dans de nombreuses langues, en particulier sur le Web, a rendu nécessaire la mise au point de méthodes de recherche et d'extraction d'information cross-lingue. Le résumé automatique cross-lingue vise à donner à l'utilisateur un accès rapide à des contenus exprimés dans une ou plusieurs langues qu'il maîtrise mal ou ne connaît pas. Plus précisément, cette tâche consiste à générer un résumé dans une langue cible différente de celle utilisée dans les documents sources. Dans cette étude, nous nous intéressons au résumé automatique multi-document de l'anglais vers le français, la motivation première étant de permettre aux utilisateurs francophones d'accéder à la masse toujours croissante d'actualités disponibles à travers des sources majoritairement anglophones.

Plusieurs études récentes se sont intéressées aux modèles de graphes pour représenter l'information dans des applications de Traitement Automatique des Langues Naturelles (TALN) (Banea *et al.*, 2010). Dans ces modèles, les entités — qui peuvent être par exemple les mots, les phrases ou même les documents — sont représentées sous la forme de nœuds et les relations entre elles par des arêtes. Ce type d'approche a déjà été utilisé dans des applications TALN diverses tel que l'étiquetage en parties du discours, l'extraction d'information, l'analyse de sentiments ou le résumé automatique auquel nous nous intéressons ici.

Une méthodologie simple pour aborder le résumé automatique cross-lingue serait d'appliquer un système de traduction automatique (TA) directement sur les sorties d'un système de résumé automatique classique. Toutefois, cette approche n'est pas sans inconvénients puisqu'elle devient dépendante de la qualité du système de TA. Dans cet article, nous proposons de prendre en compte la qualité de traduction des phrases en français lors de la sélection des phrases retenues pour assembler le résumé, l'idée étant de minimiser l'impact des erreurs commises par le système de TA. Les phrases ainsi sélectionnées pour construire le résumé seront celles jugées à la fois informatives

par le système de résumé automatique et faciles à traduire par le système de TA. Pour ce faire, nous recourons à une méthode d'apprentissage supervisé pour prédire les scores de qualité de la traduction et intégrons ces scores durant la construction du graphe utilisé pour sélectionner les phrases informatives.

Dans la suite de cet article, nous commençons par présenter les travaux connexes aux nôtres. La section 3 est consacrée à la description de la méthode que nous proposons. Nous décrivons ensuite en section 4 nos résultats expérimentaux avant de conclure et de montrer quelques perspectives.

2 Travaux connexes

Dans cette section, nous présentons dans un premier temps les travaux existants sur la prédiction de la qualité de traduction automatique. Nous décrivons ensuite les approches de résumé automatique basées sur les modèles de graphes ainsi que les études sur le résumé automatique cross-lingue.

2.1 Prédiction de la qualité de traduction automatique

La traduction automatique est un composant naturel d'un système automatique de résumé cross-lingue de documents. Malheureusement, bien que des progrès importants aient été réalisés depuis une décennie, les systèmes de TA restent sujets à des erreurs qui peuvent dégrader fortement la qualité des résumés produits, en introduisant en particulier des informations erronées ou en rendant les phrases générées difficiles à lire. Afin de réduire ces effets, il est intéressant de prendre en compte un score jugeant de la qualité de la traduction pour filtrer les traductions incorrectes lors du résumé.

La prédiction de la qualité de la traduction a tout d'abord été vue comme un problème de classification binaire pour distinguer les bonnes traductions des mauvaises (Blatz *et al.*, 2003). Des études plus récentes ont estimé une valeur continue de score soit au niveau du mot (Raybaud *et al.*, 2009), soit au niveau de la phrase (Raybaud *et al.*, 2009; Specia *et al.*, 2009). Dans cet article, nous employons des scores calculés au niveau de la phrase, ceux-ci étant plus faciles à intégrer dans le processus de sélection de phrases pour le résumé.

Plusieurs classificateurs ont été construits pour estimer la qualité de traduction. Ces modèles statistiques ont été utilisés sur des traductions étiquetées manuellement comme correctes ou non (Quirk, 2004; Specia *et al.*, 2009), ou étiquetées par des métriques automatiques comme le taux d'erreur de mots (Blatz *et al.*, 2003), le score NIST (Blatz *et al.*, 2003; Specia *et al.*, 2009) ou BLEU (Raybaud *et al.*, 2009). Parmi les différentes caractéristiques utilisées pour le calcul des valeurs de qualité, on retrouve des traits linguistiques — dépendant ou non de ressources telles que des analyseurs syntaxiques ou Wordnet —, des mesures de similarité entre la phrase source et la phrase cible, et des caractéristiques internes au système de traduction utilisées — comme le nombre de traductions proposées par mots sources ou les scores de segments (phrases) des meilleures hypothèses de traduction.

2.2 Résumé automatique fondé sur les modèles de graphes

Ces dernières années, de nombreuses évaluations ont été conduites sur la tâche du résumé automatique multi-document, en particulier dans le cadre des campagnes internationales *Document Understanding Conference* [1] (DUC) et *Text Analysis Conference* [2] (TAC) organisées par le *National Institute of Standards and Technology* [3] (NIST). La quasi-totalité des approches proposées recourent à des méthodes d'extraction où il s'agit d'identifier les unités textuelles — le plus souvent des phrases — les plus importantes des documents. Les phrases contenant les concepts les plus importants sont sélectionnées puis assemblées selon leur degré de pertinence afin de générer les résumés. Ce type d'approche donne de bons résultats et permet de contourner les problématiques difficiles de compréhension sémantique du texte ou de génération de texte en langue naturelle.

Les travaux menés jusqu'à présent sur la tâche du résumé automatique multi-document sont basés, entre autres, sur l'utilisation du centroïde pour la sélection de phrases (Radev *et al.*, 2004), sur l'apprentissage supervisé des critères d'informativité (Wong *et al.*, 2008) ou sur la fusion d'information (Barzilay *et al.*, 1999). Dans cet article,

1. http://duc.nist.gov
2. http://www.nist.gov/tac/
3. http://www.nist.gov

nous employons une approche basée sur les modèles de graphes introduite dans (Mihalcea, 2004; Erkan & Radev, 2004). Les algorithmes de classement basés sur les graphes tels que PAGERANK (Page *et al.*, 1998) ont été utilisés avec succès dans les réseaux sociaux, l'analyse du nombre de citations ou l'étude de la structure du Web. Appliqué au résumé automatique, ce type d'approche suggère de représenter les documents par un graphe d'unités textuelles (phrases) inter-connectées par des relations issues de calculs de similarité. Les phrases sont ensuite sélectionnées selon des critères de centralité ou de prestige dans le graphe puis assemblées pour produire des extraits. Cette approche a deux principaux avantages. Premièrement, contrairement à la plupart des autres méthodes, elle ne nécessite pas de données d'apprentissage. Deuxièmement, du fait qu'elle se base sur des traitements linguistiques minimaux (segmentation en phrases et similarité inter-phrases), cette approche est facilement adaptable à d'autres langues (Mihalcea & Tarau, 2005).

2.3 Résumé automatique cross-lingue

Quelques études se sont récemment intéressées à la problématique du résumé automatique cross-lingue. Deux solutions simples à ce problème consistent soit à traduire les documents avant la phase d'extraction, soit à traduire les résumés générés. Cette seconde approche est généralement préférée à la première car la traduction au préalable des documents rend le processus de sélection de phrases plus risqué de part les erreurs potentiellement introduites par le système de TA. Orăsan & Chiorean (2008) ont ainsi proposé d'utiliser la méthode *Maximal Marginal Relevance* (MMR) (Carbonell & Goldstein, 1998) pour produire des résumés d'actualités exprimés en roumain et ensuite de les traduire automatiquement en anglais.

Plus récemment, Wan *et al.* (2010) se sont intéressés au résumé automatique mono-document, depuis l'anglais vers le chinois, en employant des méthodes supervisées pour estimer la qualité de traduction automatique. Leur étude a montré que la prise en compte de scores de qualité de traduction permet d'améliorer à la fois le contenu et la lisibilité des résumés générés. Dans notre article, nous utilisons une approche similaire en nous intéressant cette fois au résumé automatique multi-document. Contrairement aux travaux de Wan *et al.*, notre approche utilise un algorithme non supervisé et indépendant de la langue pour sélectionner des phrases (Mihalcea & Tarau, 2005). De plus, nous n'utilisons pas de corpus annotés manuellement selon leur qualité de traduction mais un indicateur calculé automatiquement à partir de traductions de références produites par des humains, ce type de corpus étant long et parfois délicat à construire.

3 Notre méthode pour le résumé cross-lingue

Notre approche pour résumer un ensemble de documents depuis l'anglais vers le français se fait en trois étapes. Chaque phrase est tout d'abord traduite automatiquement et la qualité de la traduction est estimée (section 3.1). Chaque phrase se trouve ensuite évaluée en fonction de son contenu informatif (section 3.2) et de son score de qualité de traduction (section 3.3). Puis, les phrases de plus haut score sont sélectionnées pour les inclure dans le résumé (section 3.4). La figure 1 présente un aperçu de l'architecture de notre méthode.

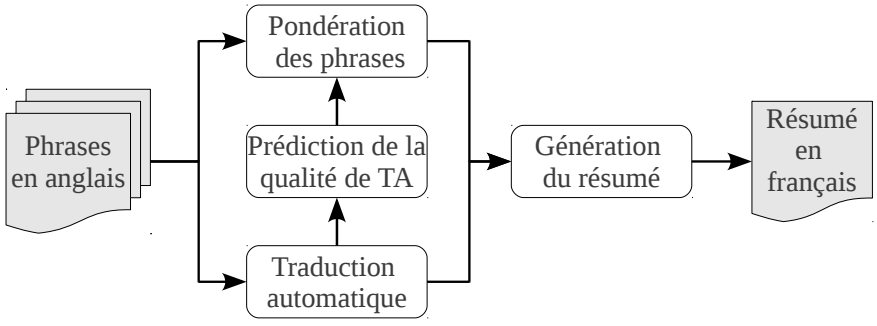

FIGURE 1 – Architecture de notre système de résumé automatique cross-lingue.

3.1 Prétraitement de documents et prédiction de la qualité de traduction

Chaque document de l'ensemble à résumer est segmenté en phrases en utilisant la méthode PUNKT de détection de changement de phrases (Kiss & Strunk, 2006) mise en œuvre dans la boîte à outils NLTK (Bird & Loper, 2004). Toutes les phrases en anglais ont été automatiquement traduites en français en utilisant le système de traduction de Google [4].

Un score de TA est calculé pour chaque phrase pour estimer la justesse et la fluidité des phrases générées en français. Pour ce faire, nous calculons pour chaque phrase 8 caractéristiques, qui donnent des informations sur la difficulté de traduction et sur la lisibilité des traductions générées :

– la longueur de la phrase source en terme de mots ;
– le ratio des longueurs des phrases source et cible ;
– le nombre de signes de ponctuation dans la phrase source ;
– la proportion des nombres et des signes de ponctuation présentes dans la phrase source qui sont retrouvées dans la phrase cible ;
– les perplexités des phrases source et cible calculées à l'aide de modèle de langue (ML) 5-grammes en avant ;
– les perplexités des phrases source et cible calculées par des ML bigrammes en arrière, *i. e.* en inversant l'ordre des mots des phrases.

Ces quatre premières caractéristiques sont parmi les traits les plus pertinents mis en exergue dans (Specia *et al.*, 2009), parmi 84 caractéristiques étudiées ; les quatre dernières ont déjà montré leur efficacité dans le calcul de mesure de confiance au niveau des mots (Raybaud *et al.*, 2009). Des ML sont construits à partir des corpus monolingues du domaine des actualités, rendus disponible pour l'atelier WMT 2010 (Callison-Burch *et al.*, 2010) et constitués respectivement de 991 et 325 millions de mots en anglais et en français. Les scores de perplexité visent à estimer la fluidité. Contrairement à d'autres études, nous nous sommes concentrés sur des caractéristiques simples ne requérant pas de ressources linguistiques comme des analyseurs syntaxiques ou des dictionnaires. En outre, nous nous sommes restreints à des scores ne dépendant par du système de traduction utilisé.

Pour prédire la qualité de la traduction, nous avons employé la méthode ϵ-SVR, qui est une extension des séparateurs à vaste marge pour faire de la régression et qui a déjà été utilisée dans le même cadre applicatif (Wan *et al.*, 2010; Raybaud *et al.*, 2009). Nous avons employé la librairie LIBSVM (Chang & Lin, 2001), en nous restreignant aux noyaux gaussiens comme recommandé par les auteurs. Le modèle de régression a deux paramètres : une erreur de coût c et le coefficient γ de la fonction noyau ; leur valeur a été optimisée par recherche par quadrillage et par validation croisée.

Le modèle ϵ-SVR devrait idéalement être appris sur un corpus étiqueté manuellement du point de vue de la qualité de traduction. Malheureusement, nous ne connaissons pas de corpus de ce genre ayant une taille suffisante pour la paire anglais-français et la production de jugements de la TA reste un processus très lent. Nous nous sommes par conséquent tournés vers un indicateur calculé automatiquement à partir de traductions de référence produites par des humains : le score NIST (Doddington, 2002). Cette métrique a en effet déjà été utilisée dans le passé dans le même objectif (Blatz *et al.*, 2003; Specia *et al.*, 2009) et s'est révélée plus corrélée avec des jugements humains au niveau de la phrase que BLEU (Blatz *et al.*, 2003). Notre corpus d'apprentissage a été obtenu à partir des traductions de référence fournies dans le domaine des actualités pour les ateliers WMT (Callison-Burch *et al.*, 2010) de 2008 à 2010, ce qui représente un ensemble de 7 112 phrases. Pour contrôler la qualité du modèle ainsi obtenu, nous avons calculé la métrique MSE (*Mean Squared Error*) : $\frac{1}{N}\sum_{j=1}^{N}(y_j - \hat{y}_j)^2$, N étant le nombre de phrases, $\hat{y}$ la prédiction estimée par le modèle et y la valeur réelle. Sur les 2 007 phrases de WMT 2007 gardées à cette fin, le MSE mesuré a été de 0,456.

3.2 Pondération des phrases

Notre système de résumé multi-document est fondé sur un graphe dirigé $G = (V, E)$ construit pour chaque ensemble de textes, V étant l'ensemble de nœuds et E les arcs (arêtes) dirigés. Un nœud est ajouté au graphe pour chaque phrase de l'ensemble de documents ; les arêtes sont définies entre ces nœuds en fonction de la mesure de similarité définie dans (Mihalcea, 2004). Cette mesure détermine le nombre de mots communs entre les représentations lexicales des deux phrases, les mots outils ayant été au préalable supprimés et les autres mots ayant été

4. http://translate.google.com

stemmés avec le *stemmeur* de Porter. Pour éviter de favoriser les phrases longues, cette valeur est normalisée par les longueurs des phrases. Si $\mathrm{freq}(w, S)$ représente la fréquence du mot w dans la phrase S, la similarité entre les phrases S_i et S_j est définie par :

$$\textit{Sim}(S_i, S_j) = \frac{\sum_{w \in S_i, S_j} \mathrm{freq}(w, S_i) + \mathrm{freq}(w, S_j)}{\log(|S_i|) + \log(|S_j|)} \quad (1)$$

Les algorithmes de classement basés sur les modèles de graphes mettent en œuvre le concept de recommandation. Les phrases sont évaluées selon des scores calculés récursivement sur l'intégralité du graphe. Dans notre étude, nous utilisons une adaptation de l'algorithme PAGERANK de Google (Page *et al.*, 1998) qui inclut les poids des arêtes :

$$p(V_i) = (1 - d) + d \times \sum_{V_j \in pred(V_i)} \frac{Sim(S_i, S_j)}{\sum_{V_k \in succ(V_i)} Sim(S_k, S_i)} \, p(V_i) \quad (2)$$

où d est un « facteur d'amortissement » (typiquement dans l'intervalle $[0.8, 0.9]$), $pred(V_i)$ représente l'ensemble des nœuds qui ont une arête en direction de V_i et $succ(V_i)$ l'ensemble des nœuds connectés à V_i par une arête sortante. La méthode employée ici, décrite dans (Mihalcea, 2004), est très similaire au PAGERANK lexical, appelé LEXRANK (Erkan & Radev, 2004).

3.3 Inclusion des scores de qualité de traduction

Pour prendre en compte l'aspect cross-lingue, la mesure de similarité inter-phrases, définie dans l'équation 1 pour les phrases d'origine en anglais, est modifiée pour inclure les scores de qualité de traduction :

$$\textit{Sim}_2(S_i, S_j) = \textit{Sim}(S_i, S_j) \times \textit{Prediction}(S_i) \quad (3)$$

où $\textit{Prediction}(S_i)$ est le score de qualité de TA de la phrase S_i calculée comme décrit en section 3.1. Cette métrique est asymétrique, contrairement à celle définie par l'équation 1. Une phrase traduite correctement et fluide voit ainsi les poids de ses arêtes sortantes renforcés et jouera par conséquent un rôle plus central dans le graphe.

Nous avons modifié l'algorithme de classement afin de tirer profit des spécificités des documents. Comme la position d'une phrase au sein d'un document est un indicateur fort sur l'importance de son contenu — les articles de journaux présentant généralement au début une description concise du sujet — le poids des arcs sortant du nœud correspondant à la première phrase a été doublé. En outre, les phrases dupliquées ainsi que les phrases contenant moins de 5 mots ont été mises de côté.

3.4 Génération de résumé

Bien souvent, les documents regroupés sous une thématique contiennent des phrases très similaires, voire même identiques. Il est donc possible que deux phrases très redondantes se retrouvent dans un résumé, dégradant à la fois sa lisibilité et son contenu informatif. Pour pallier ce problème, Carbonell & Goldstein (1998) ont proposé la méthode d'assemblage itératif *Maximal Marginal Relevance* (MMR). Cette technique, probablement la plus utilisée, consiste à réordonner les phrases en fonction de deux critères qui sont l'importance de la phrase et la redondance par rapport aux phrases déjà sélectionnées. Le résumé est ensuite construit itérativement par l'ajout des phrases maximisant l'informativité tout en minimisant la redondance.

Dans cette étude, nous avons utilisé une approche différente. Suivant la méthode proposée dans (Mihalcea & Tarau, 2005) pour la construction de graphes, aucun arc n'est ajouté entre deux nœuds dont la similarité excède un seuil maximal. De façon à réduire la redondance, une étape supplémentaire est ajoutée lors de la génération des résumés (Genest *et al.*, 2009). Nous générons pour ce faire tous les résumés candidats à partir des combinaisons des N phrases ayant les meilleurs scores, en veillant à ce que le nombre total de caractères soit optimal (*i. e.* en dessous d'un seuil donné et qu'il soit impossible d'ajouter une autre phrase sans dépasser ce seuil). Le résumé

retenu au final est celui possédant le score global le plus élevé, ce score étant calculé comme le produit du score de la diversité du résumé — estimé par le nombre de n-grammes différents — et de la somme des scores des phrases.

Afin d'améliorer la lisibilité du résumé produit, les phrases sont triées dans l'ordre chronologique de publication des documents où ils apparaissent, ce qui maximise la cohérence temporelle ; si deux phrases sont extraites à partir d'un même document, l'ordre original du document est conservé.

4 Résultats

Cette section décrit les données utilisées, les métriques d'évaluation et les résultats de notre système.

4.1 Cadre expérimental

Nous avons employé dans notre étude les ensembles de documents mis à disposition pour l'évaluation DUC 2004, ce qui représente 50 ensembles de documents en anglais. Chaque ensemble traite d'une même thématique et comporte en moyenne 10 articles de journaux produits par *Associated Press* ou le *New York Times*. La tâche consiste à générer des résumés d'au plus 665 caractères — incluant les caractères alphanumériques, les espaces et les ponctuations — contenant l'essentiel du contenu de l'ensemble de documents correspondants. Nous avons effectué sur ces données une évaluation automatique du contenu. Une évaluation manuelle de la lisibilité a également été menée sur un échantillon constitué de 16 ensembles de documents tirés aléatoirement.

4.1.1 Évaluation automatique

La plupart des méthodes d'évaluation automatique opèrent en comparant les résumés générés avec un ou plusieurs résumés de référence. La métrique que nous avons employée ici est ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*), dont on sait qu'elle est bien corrélée avec les jugements humains (Lin, 2004). ROUGE correspond en fait à plusieurs mesures, calculées à partir du nombre de n-grammes commun entre le résumé candidat et le(s) résumé(s) de référence. Nous avons calculé trois métriques au cours de nos expériences : ROUGE-1 (basée sur les unigrammes), ROUGE-2 (bigrammes) et ROUGE-SU4 (bigrammes à trou, *i. e.* des couples de deux mots contenant au plus quatre mots entre eux) [5].

Quatre résumés de référence en anglais étaient fournis pour chacun des ensembles de documents de DUC 2004. Pour évaluer notre méthode, nous avons demandé à trois annotateurs de traduire les résumés disponibles pour le sous-ensemble de 16 groupes de documents, en veillant à ce que chaque phrase du résumé soit traduite phrase par phrase sans introduire d'information supplémentaire comme la génération d'anaphores, la désambiguisation des noms propres ou la réduction des informations redondantes. 64 résumés de référence ont été ainsi produits, chaque annotateur traduisant en moyenne un résumé en 15 minutes.

L'évaluation par ROUGE étant ici réalisée dans un cadre différent des tâches habituelles — produisant des résumés en anglais à partir de documents exprimés dans la même langue —, nous avons effectué quelques modifications. Aucune contrainte stricte n'a été imposée sur la taille des résumés traduits produits en français. En revanche, nous avons fait en sorte que notre algorithme de génération construise des résumés pour lesquelles la longueur totale des phrases correspondantes en anglais respecte la contrainte imposée à DUC 2004 sur le nombre de caractères. La longueur des résumés de référence en français se trouve ainsi accrue de 25 % en moyenne par rapport aux résumés correspondants en anglais. Notons enfin que le *stemmer* de Porter utilisé dans l'évaluation ROUGE a été adapté au français.

4.1.2 Évaluation manuelle

L'évaluation de la qualité linguistique des résumés a été effectuée selon un protocole similaire à celui utilisé lors des campagnes DUC. Nous avons évalué la lisibilité des résumés sur une échelle de 1 à 5, où 5 est attribué aux résumés « faciles à lire » et 1 aux résumés « difficiles à lire ». Cinq annotateurs ont participé à cette expérience.

5. Nous avons utilisé la version 1.5.5 de ROUGE avec les paramètres par défaut indiqués pour DUC 2004.

Afin de comparer notre approche, nous avons généré deux résumés pour chaque ensemble de documents, *i. e.* pour chacune des thématiques. Le premier résumé est produit par la méthode que nous proposons tandis que le second (*baseline*) est obtenu en traduisant directement un résumé en français (obtenu par la fonction de pondération décrite en section 3.2). La tâche qui leur a été confiée était d'attribuer une note aux deux résumés d'une même thématique, l'ordre d'apparition des résumés étant aléatoire afin d'éviter tout biais.

4.2 Expériences monolingues

Les performances de notre méthode ont tout d'abord été évaluées sur une tâche de résumé monolingue. Le tableau 1 indique les scores d'évaluation automatique obtenus sur l'ensemble des données de DUC 2004 pour différentes méthodes : le plus haut score atteint lors de la campagne en 2004 (ligne 1), le score obtenu avec la méthode GRAPH-SUM décrite en section 3.2 basée sur les graphes (ligne 2) et un score calculé pour une méthode naïve prenant la première phrase des documents les plus récents de chaque ensemble à traduire. L'approche basée sur les graphes obtient de bons résultats, la différence avec le meilleur système n'étant pas statistiquement significative [6].

Système	ROUGE-1	**Rang**	ROUGE-2	**Rang**	ROUGE-SU4	**Rang**
1^{er} système	0,38244†	1	0,09218†	1	0,13323†	1
GRAPH-SUM	0,38052†	2	0,08566†	4	0,13114†	3
Méthode naïve	0,32381	26	0,06406	25	0,10291	29

TABLE 1 – Scores ROUGE moyens mesurés sur les données de DUC 2004 et rangs obtenus par rapport aux 35 participants de la campagne. Les scores indiqués par † sont statistiquement significatifs par rapport au modèle de base ($\rho < 0.001$ avec un t-test de Student).

4.3 Expériences cross-lingues

Dans cette seconde série d'expériences, nous évaluons notre approche pour le résumé automatique multi-document cross-lingue. La première partie de cette évaluation est réalisée automatiquement à l'aide des mesures ROUGE et concerne le contenu des résumés. Il s'agit d'évaluer si les résumés produits contiennent les informations les plus importantes des documents sources. Les résultats de référence sont obtenus en traduisant le résumé en anglais produit par l'approche basée sur les modèles de graphes (méthode GRAPH-SUM). Les scores ROUGE calculés avec cette méthode sont présentés à la ligne 1 du tableau 2. En utilisant notre méthode faisant intervenir la qualité des traductions automatiques (ligne 2), nous observons une légère amélioration en terme de ROUGE-2 et ROUGE-SU4. Cependant, cette évolution des scores n'est pas statistiquement significative. Ceci peut s'expliquer par le fait que notre méthode favorise les phrases dont la qualité de TA est bonne. Ainsi des phrases ayant un contenu informationnel plus faible peuvent être introduites dans le résumé, ce qui limite l'amélioration des résultats.

Système	ROUGE-1	ROUGE-2	ROUGE-SU4
Méthode de réference	0,39704	0,10249	0,13711
Notre méthode	0,39624	0,10687	0,13877

TABLE 2 – Scores ROUGE moyens calculés sur le sous-ensemble de DUC 2004 traduit en français.

La seconde partie de cette évaluation concerne la qualité linguistique des résumés générés. Il s'agit d'évaluer manuellement si les résumés produits sont lisibles mais également compréhensibles. Le tableau 3 montre les résultats de l'évaluation manuelle obtenus sur le sous-ensemble de 16 groupes de documents. Le score moyen donné par chaque annotateur est également indiqué. Tous les annotateurs ont jugé que notre méthode conduisait à une amélioration de la lisibilité des résumés produits, ce qui montre ainsi l'intérêt d'utiliser des scores de qualité de TA pour améliorer la qualité linguistique des résumés.

6. Le t-test de Student est de $\rho = 0.77$ pour ROUGE-1, $\rho = 0.17$ pour ROUGE-2 et $\rho = 0.57$ pour ROUGE-SU4.

Néanmoins, il faut noter que les scores moyens sont relativement bas. Ceci indique que les résumés générés par notre méthode, bien qu'étant meilleurs du point de vue de la lisibilité par rapport à l'approche de référence, ne sont pas encore satisfaisants. Un exemple des sorties de notre système de résumé automatique est donné en annexe. Plusieurs types d'erreurs ont été identifiées comme récurrentes. La qualité de la TA est dépendante de la difficulté de la phrase à traduire. Ainsi, par des traitements simples comme la suppression des références temporelles, la résolution des acronymes ou la normalisation des noms propres, nous espérons pouvoir réduire la difficulté des phrases sources et par conséquent réduire le nombre d'erreurs de traduction.

Annotateur	**Lisibilité**	
	Méthode de référence	Notre méthode
Annotateur 1	2,44	2,50
Annotateur 2	1,56	1,63
Annotateur 3	1,75	2,31
Annotateur 4	3,06	3,31
Annotateur 5	1,50	1,63
Moyenne	2,06	2,28

TABLE 3 – Scores moyens de lisibilité de notre méthode comparés avec une approche standard basée sur les graphes. Les scores varient selon une échelle de 1 à 5, 5 étant le plus haut score possible.

5 Conclusions et perspectives

Dans cet article, nous avons présenté une approche basée sur les modèles de graphes pour le résumé automatique multi-document cross-lingue. Nous avons proposé d'introduire des scores de qualité de traduction automatique au moment de l'étape de construction du graphe représentant les unités textuelles, un algorithme de classement par popularité étant ensuite chargé de sélectionner les phrases traduites qui sont à la fois les plus informatives mais également les plus lisibles. Cette approche a été évaluée sur un corpus de 16 ensembles de documents traduits manuellement parmi les documents mis à disposition dans le cadre de la campagne d'évaluation internationale DUC 2004. Les résultats expérimentaux montrent que notre méthode améliore sensiblement la lisibilité (*i. e.* la qualité linguistique) des résumés générés tout en maintenant un contenu informatif au niveau de l'état de l'art.

En perspectives de nos travaux, nous souhaitons dans un premier temps mener une évaluation plus complète en produisant des résumés de référence sur l'ensemble des données de la compétition DUC 2004 et en étendant l'évaluation à d'autres langues. Ceci permettra de renforcer l'importance des résultats que nous avons obtenus mais aussi d'envisager un apprentissage supervisé de la répartition entre l'informativité et à la lisibilité des phrases. Dans un deuxième temps, nous souhaitons travailler sur la réécriture des phrases sources et en étudier l'impact sur la qualité de traduction, l'idée étant de simplifier au maximum les phrases sources à l'aide de traitement linguistiques comme la résolution d'anaphores afin de faciliter le travail du système de TA. Nous souhaitons également suivre la piste de la fusion non supervisée de phrases (Filippova, 2010) afin de générer des phrases courtes et linguistiquement simples. Une dernière perspective que nous voulons étudier concerne l'utilisation de notre propre modèle de traduction automatique, ce qui permettra à la fois d'adapter ce système pour le type de documents à résumer et de prendre en compte de nouveaux indices pour prédire la qualité de la traduction.

Références

C. BANEA, A. MOSCHITTI, S. SOMASUNDARAN & F. M. ZANZOTTO, Eds. (2010). *TextGraphs-5 Workshop*. Uppsala, Suède.

BARZILAY R., MCKEOWN K. R. & ELHADAD M. (1999). Information fusion in the context of multi-document summarization. In *ACL*, College Park, MD, USA.

BIRD S. & LOPER E. (2004). NLTK : The natural language toolkit. In *ACL*, Barcelone, Espagne.

BLATZ J., FITZGERALD E., FOSTER G., GANDRABUR S., GOUTTE C., KULESZA A., SANCHIS A. & UEFFING N. (2003). *Confidence Estimation for Machine Translation*. Rapport interne, Johns Hopkins University, Batimore, MD, USA.

CALLISON-BURCH C., KOEHN P., MONZ C., PETERSON K., PRZYBOCKI M. & ZAIDAN O. (2010). Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation. In *Workshop on Statistical Machine Translation and Metrics (WMT)*, Uppsala, Suède.

CARBONELL J. & GOLDSTEIN J. (1998). The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *SIGIR*, Melbourne, Australie.

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

DODDINGTON G. (2002). Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *HLT*, San Diego, CA, USA.

ERKAN G. & RADEV D. (2004). LexRank : Graph-based lexical centrality as salience in text summarization. *JAIR*, **22**(1), 457–479.

FILIPPOVA K. (2010). Multi-sentence compression : Finding shortest paths in word graphs. In *Coling*, Pékin, Chine.

GENEST P., LAPALME G., NERIMA L. & WEHRLI E. (2009). A symbolic summarizer with 2 steps of sentence selection for TAC 2009. In *TAC Workshop*, Gaithersburg, MD, USA.

KISS T. & STRUNK J. (2006). Unsupervised multilingual sentence boundary detection. *Computational Linguistics*, **32**(4), 485–525.

LIN C.-Y. (2004). Rouge : A package for automatic evaluation of summaries. In *ACL Workshop on Text Summarization Branches Out*, Barcelone, Espagne.

MIHALCEA R. (2004). Graph-based ranking algorithms for sentence extraction, applied to text summarization. In *ACL*, Barcelone, Espagne.

MIHALCEA R. & TARAU P. (2005). A language independent algorithm for single and multiple document summarization. In *IJCNLP*, Jeju Island, Corée du Sud.

ORĂSAN C. & CHIOREAN O. A. (2008). Evaluation of a cross-lingual romanian-english multi-document summariser. In *LREC*.

PAGE L., BRIN S., MOTWANI R. & WINOGRAD T. (1998). *The pagerank citation ranking : Bringing order to the web*. Rapport interne, Stanford Digital Library Technologies Project.

QUIRK C. B. (2004). Training a sentence-level machine translation confidence measure. In *LREC*, Lisbonne, Portugal.

RADEV D., JING H., STY M. & TAM D. (2004). Centroid-based summarization of multiple documents. *Information Processing & Management*, **40**(6), 919–938.

RAYBAUD S., LANGLOIS D. & SMAÏLI K. (2009). Efficient combination of confidence measures for machine translation. In *Interspeech*, Brighton, UK.

SPECIA L., CANCEDDA N., DYMETMAN M., TURCHI M. & CRISTIANINI N. (2009). Estimating the sentence-level quality of machine translation systems. In *EAMT*, Barcelone, Espagne.

WAN X., LI H. & XIAO J. (2010). Cross-language document summarization based on machine translation quality prediction. In *ACL*, Uppsala, Suèede.

WONG K.-F., WU M. & LI W. (2008). Extractive summarization using supervised and semi-supervised learning. In *Coling*, Manchester, UK.

Annexe

Méthode de référence (Score de lisibilité moyenne de 2,6)
Leaders de l'opposition du prince Norodom Ranariddh et Sam Rainsy, invoquant des menaces de Hun Sen à l'arrestation de l'opposition, après deux tentatives présumées sur sa vie, a dit qu'ils ne pouvaient pas négocier librement au Cambodge et a appelé à des pourparlers à la résidence de Sihanouk à Pékin. (*Opposition leaders Prince Norodom Ranariddh and Sam Rainsy, citing Hun Sen's threats to arrest opposition figures after two alleged attempts on his life, said they could not negotiate freely in Cambodia and called for talks at Sihanouk's residence in Beijing.*) Le parti de Hun Sen a récemment demandé à Ranariddh pour retourner à la table des négociations et a déclaré qu'il était disposé à faire une "concession appropriées" pour sortir de l'impasse de former un gouvernement. (*Hun Sen's party recently called on Ranariddh to return to the negotiation table and said it was willing to make an "appropriate concession" to break the deadlock over forming a government.*) La semaine dernière, Hun Sen Parti du peuple cambodgien et le parti Ranariddh FUNCINPEC ont convenu de former une coalition qui laisserait Hun Sen comme Premier ministre seul et faire le prince président de l'Assemblée nationale. (*Last week, Hun Sen's Cambodian People's Party and Ranariddh's FUNCINPEC party agreed to form a coalition that would leave Hun Sen as sole prime minister and make the prince president of the National Assembly.*)
Notre méthode (Score de lisibilité moyenne de 3,2)
Le parti au pouvoir a soutenu l'action de la police dans sa déclaration, en notant que les biens publics ont été endommagés par des manifestants et que des grenades ont été lancées sur la maison de Hun Sen après Sam Rainsy a suggéré dans un discours que le gouvernement américain devrait tirer des missiles de croisière à Hun Sen. (*The ruling party supported the police action in its statement, noting that public property was damaged by protesters and that grenades were thrown at Hun Sen's home after Sam Rainsy suggested in a speech that the U.S. government should fire cruise missiles at Hun Sen.*) Politiciens cambodgiens a exprimé l'espoir lundi qu'un nouveau partenariat entre les parties de l'homme fort Hun Sen et son rival, le prince Norodom Ranariddh, dans un gouvernement de coalition ne mettrait pas fin à plus de violence. (*Cambodian politicians expressed hope Monday that a new partnership between the parties of strongman Hun Sen and his rival, Prince Norodom Ranariddh, in a coalition government would not end in more violence.*) Le roi Norodom Sihanouk a salué mardi les accords sur le Cambodge les deux principaux partis politiques précédemment amère rivaux pour former un gouvernement de coalition dirigé par l'homme fort Hun Sen. (*King Norodom Sihanouk on Tuesday praised agreements by Cambodia's top two political parties previously bitter rivals to form a coalition government led by strongman Hun Sen.*)

TABLE 4 – Example de résumés en français généré pour l'ensemble D30001T de DUC 2004 par la méthode de référence et notre approche.