Extraction des séquences fermées fréquentes à partir de corpus parallèles : Application à la traduction automatique

Chiraz Latiri*, Cyrine Nasri**, Kamel Smaili***, Yahya Slimani**

*Unité de Recherche URPAH, Faculté des Sciences de Tunis, Tunisie
chiraz.latiri@gnet.tn
**Unité de Recherche MOSIC, Faculté des Sciences de Tunis, Tunisie
cyrine.nasri@gmail.com, yahya.slimani@fst.rnu.tn
***LORIA, Groupe PAROLE, Vandoeuvre, France
kamel.smaili@loria.fr

Résumé. Dans cet article, nous abordons la problématique d'extraction de séquences fréquentes à partir de corpus de textes parallèles en prenant en compte l'ordre d'apparition des mots dans une phrase. Notre finalité est d'exploiter ces séquences dans la traduction automatique (TA). Nous introduisons ainsi la notion de règles associatives inter-langues (RAIL) et nous définissons notre modèle de traduction à base de ces associations. Nous décrivons également les différentes expérimentations conduites sur le corpus EUROPARL afin de construire à partir des RAIL une table de traduction bilingue qui est intégrée par la suite dans un processus complet de TA.

1 Introduction

Initialement introduit dans (Srikant et Agrawal, 1996), l'extraction de motifs séquentiels reste intuitivement applicable à tout domaine dans lequel il existe une relation d'ordre entre les éléments. Dans cet article, nous proposons d'aborder la problématique d'extraction de séquences fréquentes, en se plaçant dans le domaine de la fouille de données textuelles (Berry, 2008) et en prenant en compte l'ordre d'apparition des mots dans une phrase, et ce à partir d'un corpus parallèle aligné au niveau de la phrase. Nous nous intéressons particulièrement aux approches basées sur un parcours en largeur et dédiées à l'extraction *des motifs séquentiels fermés fréquents* (Yan et al., 2003; Wang et Han, 2004; Chang, 2004). Ces dernières évoquent le problème de la redondance des sous-séquences extraites et ayant le même support que d'autres super-séquences fréquentes.

Notre finalité est de déployer les séquences fréquentes de mots dans la traduction automatique (TA) (Brown et al., 1993). Notre choix pour la TA est justifié par le fait que des travaux récents en traduction automatique statistique confirment que les modèles fondés sur des séquences de mots (Och et al., 1999; Koehn, 2004) obtiennent des performances significativement meilleures que ceux fondés sur des mots simples (Brown et al., 1993).

La suite de l'article est structuré comme suit : la section 2 présente un bref aperçu du processus de recherche des séquences fermées fréquentes adapté aux corpus textuels, suivie de la section 3 qui décrit l'application à la TA en introduisant la notion de règles associatives

inter-langues (RAIL) et le processus de construction de la table de traduction à partir des RAIL. L'évaluation expérimentale de notre approche est illustrée dans la section 4. Nous terminons l'article par une conclusion ainsi que les travaux en cours.

2 Extraction des séquences fermées fréquentes à partir d'un corpus de textes

L'algorithme d'extraction des séquences fermées fréquentes à partir d'un corpus de textes que nous avons utilisé est une adaptation de l'algorithme BFSM décrit dans Chang (2004).

Nous considérons qu'un *contexte d'extraction textuel* est un triplet $\mathfrak{K} = (\mathcal{P}, \mathcal{T}, \mathcal{R})$ où $\mathcal{P}$ représente un ensemble fini de phrases, $\mathcal{T}$ est un ensemble fini de termes et $\mathcal{R}$ une relation binaire (*i.e.*, $\mathcal{R} \subseteq \mathcal{P} \times \mathcal{T}$). Chaque couple $(p, t) \in \mathcal{R}$ signifie que la phrase $p \in \mathcal{P}$ contient le terme $t \in \mathcal{T}$. Un *termset*[1] est un ensemble non vide de termes noté par $(t_1, t_2 ... t_k)$.

Définition 1 *Une séquence $\mathcal{S}\langle t_i, \ldots, t_j, \ldots, t_n\rangle$ tel que t_k un terme appartenant à $\mathcal{T}$, est un termset* ***ordonné****,* i.e.*, l'ordre d'apparition de ces termes dans une phrase est respecté.*
S est dite ***fréquente*** *si son support,* i.e.*, $Supp(S)_{S \subseteq p} = \|p \in \mathcal{P}\|$, est supérieur ou égal à un seuil de support minimal noté par $minsupp$. Elle est dite* ***fermée*** *s'il n'existe pas une super-séquence fréquente du contexte d'extraction textuel $\mathfrak{K}$ ayant le même support que S.*

En faisant référence à l'algorithme BFSM (Chang, 2004), le processus d'extraction des séquences fermées fréquentes de termes ($\mathcal{SFFT}$) à partir d'un contexte d'extraction $\mathfrak{K}$, se base sur deux idées clés, à savoir :

1. *L'extension de séquence :* Pour étendre une k-séquence S_k à une $(k+1)$-séquence S_{k+1}, on procède à une **extension de séquence** qui consiste à ajouter un terme comme étant un nouvel élément de la séquence. La séquence S_{k+1} est le résultat de la jointure de la séquence S_k avec une séquence S_α appartenant à la liste des 2-séquences (Srikant et Agrawal, 1996).
2. *L'élagage de l'ensemble des séquences fréquentes pour ne garder que les fermées :* Soient deux séquences fréquentes S_k et $S_{(k+n)}$. Si $S_{(k+n)}$ est une super séquence de S_k qui est générée avant $S_{(k+n)}$ et elles ont les mêmes k premiers termes et le même support, alors $S_{(k+n)}$ est dite **une super séquence arrière** de S_k. Dans ce cas, $S_{(k+n)}$ est élaguée de l'ensemble des $\mathcal{SFFT}$.

Le processus d'extraction des séquences fermées fréquentes est décrit en détail dans (Chang, 2004).

3 Application à la traduction automatique statistique : les règles associatives inter-langues

Dans ce qui suit, nous proposons de déployer les séquences de termes fermées fréquentes dans la traduction automatique en introduisant la notion de *règles associatives inter-langues* (RAIL). Nous décrivons ensuite, comment les $\mathcal{SFFT}$ ainsi que les RAIL sont intégrées dans un processus complet de traduction automatique.

[1] Terminologie proposée par analogie à celle utilisée en fouille de données, à savoir *itemset*.

3.1 Définition de la traduction automatique statistique

L'objectif d'un système de traduction automatique est de proposer pour une phrase $f = f_1, \ldots, f_i$ en une langue source sa traduction en une phrase $e = e_1, \ldots, e_j$ dans une langue cible. L'approche statistique consiste à choisir la phrase la plus probable, parmi les phrases possibles. Le problème est formalisé comme suit :

$$\hat{e} = \arg\max_{e} P(e)P(f|e) \quad (1)$$

Dans l'équation 1, $P(e)$ est une probabilité estimée par un *modèle de langage* (Jelinek, 2001). Son rôle est de proposer la phrase supposée correcte dans la langue cible. Notons que la probabilité $P(f|e)$ est calculée à partir d'un *modèle de traduction* dont la finalité est de refléter l'exactitude de la traduction. Un décodeur (Koehn, 2004) produit ensuite la meilleure hypothèse en cherchant un compromis entre, au moins, ces distributions de probabilités.

Afin d'évaluer un système de TA, la mesure BLEU (BiLingual Evaluation Understudy) (Papineni et al., 2001) est la plus utilisée par la communauté de TA. Pour estimer la valeur du BLEU, il suffit de fournir à un décodeur[2] un modèle de langage et un modèle de traduction, qui procédera ensuite à la traduction de la phrase d'une langue source vers une langue cible.

3.2 Dérivation des règles associatives inter-langues

L'idée de dérivation des *règles associatives inter-langues* est inspirée des travaux de (Lavecchia et al., 2007) qui ont introduit le concept des *triggers inter-langues* en traduction automatique statistique.

Nous considérons un corpus parallèle français-anglais, à partir duquel les séquences fermées fréquentes $\mathcal{SFFT}$ sont générées.

Définition 2 *Une règle associative inter-langues (RAIL) est une implication de la forme : $R : S_{fr} \longrightarrow S_{en}$ tel que S_{fr} et S_{en} sont deux séquences fermées fréquentes de termes, de tailles respectives n et m mots. Le support de R, noté par $Supp(R)$, exprime la fréquence avec laquelle deux séquences S_{fr} et S_{en} co-occurrent ensemble dans le corpus parallèle. La confiance de R, notée par $Conf(R)$, exprime la probabilité conditionnelle pour qu'une phrase contienne la séquence S_{en}, sachant qu'elle contient la séquence S_{fr}.*

En TA, cette règle signifie que la séquence en anglais S_{en} est une traduction candidate de la séquence en français S_{fr}.

Nous avons mis en place un algorithme itératif qui considère en entrée les deux ensembles de séquences fermées fréquentes, groupées par taille en français $\mathcal{S}_k^{fr}$ et en anglais $\mathcal{S}_k^{en}$, ainsi que le seuil minimal de confiance *minconf*. Pour chaque *k*-séquence fermée fréquente en français $S \in \mathcal{S}_k^{fr}$, il dérive toutes les RAIL de la forme : $S \longrightarrow S^{'}$, telle que $S^{'} \in \mathcal{S}_k^{en}$ est une *l*-séquence fermée fréquente en anglais. Durant ce parcours itératif, l'algorithme vérifie si $\frac{Supp(S \cup S^{'})}{Supp(S)} \geq minconf$. Si cette condition est vérifiée, la séquence $S^{'}$ est insérée dans la liste des conclusions valides représentant les traductions potentielles de S. L'algorithme s'arrête lorsque l'ensemble $\mathcal{S}_k^{fr}$ est vide.

[2] PHARAOH et MOSES sont deux décodeurs populaires pour la TA à base de séquences de mots. Ils sont distribués gratuitement respectivement sur les sites http ://www.isi.edu/licensed-sw/pharaoh/ et http ://www.statmt.org/moses/.

Nous proposons dans ce qui suit, d'utiliser l'ensemble des règles d'association inter-langues pour mettre en place notre modèle de traduction à base de séquences de mots.

3.3 Construction de la table de traduction

L'idée de construction d'une table de traduction à partir des RAIL, notée dans la suite par RAIL-DIC-n-to-m, consiste à considérer que les traductions potentielles d'une séquence en français S_f qui apparaît dans la prémisse d'une ou plusieurs RAIL sont obtenues en sélectionnant les séquences en anglais $S_{e_1}, S_{e_2}, \ldots, S_{e_n}$ qui représentent les conclusions de ces mêmes RAIL.

Formellement, une entrée dans la table de traduction bilingue RAIL-DIC-n-to-m est définie comme suit :

$$\begin{array}{c} S_f \longrightarrow S_{e_1}, S_{e_2}, S_{e_j}, \ldots, S_{e_n} \in \text{RAIL-DIC-}n\text{-to-}m; \forall\, j \in [1 \ldots n] \\ r_j : S_f \longrightarrow S_{e_j} \wedge Conf(r_j) \geq minconf \end{array} \quad (2)$$

La conversion de la valeur de la confiance de l'association inter-langues en probabilité se fait comme suit :

$$\forall S_f, S_{e_i} \in PT(S_f), P(S_{e_i}|S_f) = \frac{Conf(S_f \longrightarrow S_{e_i})}{\sum_{S_{e_i} \in PT(S_f)} Conf(S_f \longrightarrow S_{e_i})} \quad (3)$$

sachant que $PT(S_f)$ représente les traductions potentielles en anglais de la séquence en français S_f.

4 Evaluation expérimentale de la traduction automatique à base des associations inter-langues

Nous avons mené nos expérimentations sur EUROPARL, un corpus parallèle bilingue et aligné au niveau de la phrase (Koehn, 2005)[3]. Ce corpus provient des actes du Parlement Européen entre mars 1996 et septembre 2003. Les statistiques du corpus utilisé sont résumées dans le tableau 1.

		français	**anglais**
Apprentissage	Phrases	$596K$	
	Mots	$17,3M$	$15,8M$
Développement	Phrases	1444	
	Mots	$15,0K$	$14,0K$
Test	Phrases	500	
	Mots	$5,2K$	$4,9K$

TAB. 1 – *Description quantitative du corpus* EUROPARL

Le tableau 2 illustre des exemples de règles associatives inetr-langues, *i.e.*, du français vers l'anglais, générées à partir du corpus EUROPARL.

[3]Disponible gratuitement sur le site http ://www.statmt.org/europarl/.

S_f	$PT(S_f)$	$Conf(S_f \longrightarrow S_{e_i})$	$P_{RAIL}(S_{e_i}\|S_f)$
toujours possible	possible to	0.22	0.36
	always possible	0.28	0.46
	it is not always	0.11	0.18

TAB. 2 – *Exemples de règles associatives inter-langues*

Afin d'évaluer la pertinence de notre approche, nous avons utilisé le décodeur PHARAOH (Koehn, 2004) et nous avons considéré comme modèle de langage, un modèle trigrammes, révélé très performant pour modéliser les différentes langues européennes (Jelinek, 2001), pour traduire automatiquement un corpus de test de 500 phrases en anglais. L'optimisation des paramètres du décodeur a été conduite sur le corpus d'apprentissage. Les traductions produites sont ensuite comparées à l'aide de la mesure BLEU (Papineni et al., 2001).

Lors de la génération des RAIL, nous avons fixé en amont des seuils de $minsupp$ très faibles. De ce fait, certains termes du vocabulaire sont élagués lors du processus de fouille et ne figurent pas dans la table de traduction générée. Pour pallier cette limite, nous avons fixé des seuils de $minconf$ très faibles afin de favoriser la dérivation d'un nombre élevé de traductions potentielles pour chaque terme et chaque séquence fermée fréquente. Le tableau 3 montre que plus le seuil de *minsupp* est élevé et plus le score BLEU diminue. Ceci est justifié par le nombre des traductions candidates qui sont éliminées en faisant augmenter le dit seuil.

Notons que dans le cadre de nos expérimentations, nous n'avons imposé aucune limite de taille pour les séquences générées. Nous avons remarqué qu'au delà d'une certaine taille (séquences de 14 mots), le score BLEU cesse d'augmenter. Ce constat est expliqué par le fait que ces séquences sont très peu fréquentes dans le corpus de développement et que leur support dans le corpus de test est encore plus faible. Par conséquent, elles n'interviennent pas dans le calcul du score BLEU.

minsupp	*minconf*	Nbr($\mathcal{SFFT}$_fr)	Nbr($\mathcal{SFFT}$_en)	Score BLEU
20	0,1	219090	187207	**34,18**
30	0,1	140082	120317	33,79
100	0,1	39024	33765	32,10

TAB. 3 – *Evaluation des traductions automatiques avec les RAIL-DIC-n-to-m en faisant varier le seuil de $minsupp$.*

5 Conclusion et travaux en cours

Dans cet article, nous avons présenté une approche qui permet de générer à partir d'un corpus parallèle aligné au niveau de la phrase, les séquences fermées fréquentes de termes. Ces séquences ont été déployées ensuite dans le domaine de la traduction automatique pour définir le concept de règles associatives inter-langues (RAIL). Afin d'évaluer la pertinence de ces associations en tant que traductions potentielles, nous avons mis en place un modèle de traduction à base de RAIL. Les tests menés sur le corpus EUROPARL ont donné un score BLEU égal à 34,18 et ont confirmé la faisabilité de l'utilisation des règles associatives inter-langues en traduction automatique. Nous proposons à court terme de coupler notre table de traduction

fondée sur les RAIL avec la table de traduction à base des triggers inter-langues (Lavecchia et al., 2007). Notre but est d'étudier la pertinence système en terme d'amélioration de score BLEU.

Références

Berry, M. W. (2008). *Survey of Text Mining* II *: Clustering, Classification, and Retrieval*. Springer-Verlag.

Brown, P. F., V. J. D. Pietra, S. A. D. Pietra, et R. L. Mercer (1993). The mathematics of statistical machine translation : parameter estimation. *Computational Linguistics 19*(2), 263–311.

Chang, K. Y. (2004). Efficient sequential pattern mining by breadth-first approach. Master degree, National Taiwan University.

Jelinek, F. (2001). Aspects of the statistical approach to speech recognition. In *Proceedings of the IEEE International Symposium on Information Theory, Washington D.C., USA*.

Koehn, P. (2004). PHARAOH : A beam search decoder for phrase-based statistical machine translation models. In *Proceedings of* 6^th *Conference of The AMTA, Washington, DC, USA*, Volume 3265 of *LNCS*, pp. 115–124. Springer.

Koehn, P. (2005). EUROPARL : A multilingual corpus for evaluation of machine translation. In *MT Summit, Thailand*.

Lavecchia, C., K. Smaili, D. Langlois, et J. P. Haton (2007). Using inter-lingual triggers for machine translation. In *Proceedings of the Tenth Interspeech, Antwerp, Belgium*.

Och, F. J., C. Tillmann, et H. Ney (1999). Improved alignment models for statistical machine translation. In *Joint conference of Empirical Methods in Natural Language Processing and Very Large Corpora, University of Maryland, College Park, MD*, pp. 20–28.

Papineni, K., S. Roukos, T. Ward, et W.-J. Zhu (2001). BLEU : a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pp. 311–318.

Srikant, R. et R. Agrawal (1996). Mining sequential patterns : Generalizations and performance improvements. In *Proceedings of the* 5^{th} *International Conference on Extending Database Technology (EDBT'96)*, pp. 3–17.

Wang, J. et J. Han (2004). Bide : Efficient mining of frequent closed sequences. In *Proceedings of the International Conference on Data Engineering (ICDE 04), Boston, M.A, USA*.

Yan, X., J. Han, et R. Afshar (2003). Clospan : Mining closed sequential patterns in large databases. In *Proceedings of the SDM 03 Conference, San Francisco*, pp. 166–177.

Summary

This paper studies the problem of mining closed frequent sequences from corpora where the word order in the sentences is considered. Our goal is to use them in machine translation (MT). In this respect, we introduce the concept of inter-lingual association rules and we present the way to built a translation table from these associations. The carried out experiments on EUROPARL corpus highlight the practical feasibility of our approach in the context of MT.