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

Prise en compte de la sous-catégorisation verbale dans un lexique bilingue anglais-japonais

Alexis Kauffmann
LATL, Université de Genève
2, Rue de Candolle, 1211 Genève, Suisse
alexis.kauffmann@unige.ch

Résumé. Dans cet article, nous présentons une méthode de détection des correspondances bilingues de sous-catégorisation verbale à partir de données lexicales monolingues. Nous évoquons également la structure de ces lexiques et leur utilisation en traduction automatique (TA) à base linguistique anglais-japonais. Les lexiques sont utilisés par un programme de TA fonctionnant selon une architecture classique dite "à transfert", et leur structure permet une classification précise des sous-catégorisations verbales. Nos travaux ont permis une amélioration des données de sous-catégorisation des lexiques pour les verbes japonais et leurs équivalents anglais, en utilisant des données linguistiques compilées à partir d'un corpus de textes extrait du web. De plus, le fonctionnement du programme de TA a pu être amélioré en utilisant ces données.

Abstract. In this paper, we present a method for the detection of bilingual correspondences of verb subcategorization from monolingual lexical data. We also mention the structure of the lexicons and examples making use of such data in linguistics-based English-Japanese machine translation (MT). The lexicons are used by a MT system with a classical transfer-based architecture, and their structure allow an accurate classification of verb subcategorization. Our work has improved the lexical data about subcategorization of Japanese verbs and their English equivalents, using linguistic data compiled from a corpus of web extracted texts. Furthermore, the MT system could also be improved by the use of this data.

Mots-clés : bases de données lexicales, sous-catégorisation verbale, traduction automatique à base linguistique, japonais.

Keywords: lexical databases, verb subcategorisation, linguistics-based machine translation, Japanese.

1 Introduction

Connaître la sous-catégorisation des verbes [1] peut être utile en traduction automatique afin d'améliorer la traduction des verbes et de leurs différents arguments. Ce sujet a souvent donné lieu a la création de fichiers monolingues (Raza, 2010), (Kawahara & Kurohashi, 2010) et aussi de fichiers multilingues (Mangeot & Kuroda, 2003).

Nous allons aborder ici le problème de la détection automatique et de l'enregistrement de telles données dans des bases de données lexicales, à un niveau monolingue et surtout au niveau bilingue. Nous verrons aussi comment de telles données peuvent être utilisées en TA, avec le programme de TA à base linguistique Its-2, en traduction anglais-japonais.

L'article est organisé ainsi : dans la deuxème partie, nous décrirons brièvement l'architecture du programme de TA Its-2 et la structure de ses bases de données lexicales ; dans la troisième partie, nous présenterons le fichier lexical "Case Frames" qui décrit des structures argumentales de verbes japonais ; ensuite, dans la quatrième partie, nous présenterons notre méthode de détection des correspondances bilingues de sous-catégorisations verbales et la mise à jour de nos lexiques ; enfin, en cinquième partie, nous évoquerons deux améliorations apportées au programme de TA grâce aux données sur la sous-catégorisation.

1. La sous-catégorisation verbale (ou "cadres de valence syntaxique") décrit le comportement des verbes à un niveau syntaxique, en connaissant l'ensemble de leurs compléments. Cela permet de savoir si un verbe est transitif ou intransitif, s'il peut prendre une phrase comme complément, s'il peut prendre un complément d'objet indirect, etc.

2 Architecture du programme Its-2

Its-2 (Wehrli *et al*., 2009) est un programme de traduction multilingue à base linguistique (Kinoshita *et al*., 1992). Il a tout d'abord été utilisé pour traduire les principales langues d'Europe de l'Ouest (allemand, anglais, espagnol, français, italien). Depuis 2008, il a aussi été utilisé pour la traduction de l'anglais vers le japonais, et depuis 2009, pour la traduction du français vers le japonais.

2.1 Principe de fonctionnement

L'architecture d'Its-2 ayant été déjà présentée en détails dans d'autres travaux comme (Wehrli *et al*., 2009) et (Wehrli & Nerima, 2008), nous n'en ferons ici qu'une très brève présentation. Cette architecture suit le modèle classique dit "à transfert". Le texte source, dans le cas présent en français ou en anglais, est d'abord analysé par l'analyseur syntaxique Fips (Wehrli, 2007).

Ensuite, le processus de transfert intervient. Il est en en partie spécifique à la paire de langues et au sens de la traduction. Il permet de créer une représentation abstraite où les lexèmes (ou collocations) de la phrase source sont remplacés par leur équivalent dans la langue cible, ici le japonais. Dans cette structure cible, l'ordre des constituants, ou d'autres paramètres syntaxiques, sont modifiés grâce à un certain nombre de procédures spécifiques, qui implémentent des règles ou des algorithmes de traduction définis préalablement.

La dernière étape est l'étape de génération de la phrase cible. Elle est spécifique à la langue cible, ici le japonais. Elle permet de générer la phrase japonaise à partir de sa structure arborescente abstraite. De plus, elle effectue trois opérations essentielles aux niveaux morphologiques et syntaxiques : le choix des formes verbales ou adjectivales correspondant au temps déterminé lors de la phase de transfert, l'ajout des particules de cas aux groupes nominaux, et l'ajout de particules de liaison.

Its-2 a été implémenté dans le langage Component Pascal selon une approche orientée objet. Toutes les opérations se rapportant à une langue ou une paire de langues sont écrites dans des modules spécifiques, interagissant avec les modules génériques et avec les fichiers lexicaux issus des bases de données lexicales.

2.2 Bases de données lexicales du programme

Trois bases de données lexicales sont utilisées par le programme pour la traduction anglais-japonais : un lexique monolingue japonais, un lexique monolingue anglais, et un lexique bilingue anglais-japonais.

Le lexique monolingue japonais se compose de trois tables : la table des lexèmes, la table des formes conjuguées et la table des collocations [2]. Il a été créé à partir du fichier ENAMDICT (Breen, 2009) pour les noms propres, et de la base de données lexicale japonaise du CJKI (Halpern, 2008) pour les autres lexèmes. Il contient 207517 lexèmes, sans compter les noms propres (soit 913157 lexèmes au total).

Le lexique bilingue est formé d'une table qui contient les correspondances bilingues entre lexèmes [3] ou collocations. Le lexique anglais-japonais contient, à chaque entrée, un lexème ou une collocation en anglais et un élément équivalent possible en japonais. Un score entre 1 et 6 indique la pertinence de l'élément japonais pour la traduction de l'anglais vers le japonais. Lors du processus de transfert lexical, lorsque plusieurs traductions sont possibles, Its-2 choisit l'élément lexical ayant le score le plus élevé. Le lexique anglais-japonais contient 117365 correspondances bilingues.

3 Le fichier "Case Frames"

Le fichier "Case Frames" (Kawahara & Kurohashi, 2006) est un lexique électronique stockant des verbes (ou des adjectifs) japonais avec leurs différentes sous-catégorisations [4]. Il contient actuellement 90000 entrées et a été créé

2. Par collocation nous entendons une expression de deux ou plusieurs termes formant un constituant syntaxique.

3. Un lexème (ou lemme) est la forme "canonique" d'un mot, celle que l'on peut trouver dans le dictionnaire, voir par exemple (van der Plas, 2008). Le contraire d'une forme canonique est une forme "fléchie" ou "conjuguée".

4. Le terme principal utilisé par l'auteur est "case frames", soit en français "structures de cas". Ceci est dû en partie au fait que, dans la phrase japonaise, la plupart des arguments d'un verbes sont suffixés par une particule pouvant indiquer leur cas (voir tableau1).

à partir d'un corpus issu d'une extraction automatique de textes de sites web. La version complète de ce lexique électronique contient, en plus des structures argumentales des verbes et de leur nombre d'occurrences respectif dans le corpus, des listes de noms trouvés en position d'argument des verbes, permettant de connaître les cooccurrences noms-verbe typiques de chaque sous-catégorisation.

Nous n'avons toutefois utilisé ici qu'une version allégée de ce lexique, sans listes de noms, qui contenait environ 31000 verbes ou adjectifs, leurs différentes sous-catégorisations et le nombre d'occurences de celles-ci. Le tableau1 montre la sous-catégorisation de verbes anglais et celles de leurs équivalents japonais, telles qu'elles sont stockées respectivement dans le lexique anglais d'Its2 et dans le fichier "Case Frame".

	Verb	Argument 1 (Sujet)	Arg. 2	Arg. 3
Sous-catégorisation des verbes anglais	go	NP		
	go	NP	PP(to)	
	go	NP	PP(from)	PP (to)
	write	NP	NP	
Structure de cas des verbes japonais	行く("iku")	が(ga)		
	行く(iku)	が(ga)	に(ni)	
	行く(iku)	が(ga)	から(kara)	まで(made)
	書く(kaku)	が(ga)	を(o)	

Figure 1 – Sous-catégorisation verbale en anglais et en japonais

4 Détection des correspondance bilingues de sous-catégorisations verbales

Nous avons tout d'abord dû mettre à jour le lexique monolingue japonais avant de détecter quelles étaient les correspondances bilingues et de les insérer dans le lexique bilingue.

4.1 Amélioration du lexique monolingue Japonais d'Its2

Dans nos bases de données lexicales, chaque lexème verbal correspond à une seule sous-catégorisation (Wehrli *et al*., 2009). Ainsi, si un verbe a, par exemple, cinq sous-catégorisations différentes possibles, cinq lexèmes sont enregistrés dans le lexique monolingue.

Notre lexique anglais contenait déjà une classification détaillée des sous-catégorisations verbales (voir tableau 1), avec environ 16000 entrées enregistrées. En revanche, ce n'était pas le cas pour le lexique monolingue japonais, qui contenait seulement des indications insuffisantes pour la plupart des verbes enregistrés.

Pour compléter ce lexique, nous avons donc extrait les données du fichier "Case Frames". Par une série d'heuristiques implémentées par des requêtes SQL, nous avons mis en correspondance les verbes inclus dans le fichier "Case Frames" et ceux présents dans le notre lexique ; nous avons formaté les données qui étaient définies dans "Case Frames" à notre format de données ; et nous avons finalement inséré les sous-catégorisations manquantes dans notre lexique.

Ainsi, nous avons pu ajouter 5000 sous-catégorisations de verbes japonais aux 2500 que comptait déjà notre lexique, arrivant à un total de 7500 verbes sous-catégorisés.

4.2 Détection et enregistrement des correspondances bilingues

Notre lexique bilingue nous indiquait déjà des correspondances bilingues qui associaient les bons verbes, mais qui contenaient des erreurs au niveaux des correspondances entre sous-catégorisations. Or, les correspondances entre verbes anglais et japonais doivent toujours associer celles qui contiennent des arguments équivalents. Ces arguments peuvent parfois avoir une valeur syntaxique différente dans les deux langues : par exemple, un verbe peut avoir un complément d'objet indirect en anglais alors que son équivalent japonais peut prendre un complément d'objet direct.

Arguments du verbe anglais		Arguments du verbe japonais		Conditions supplémentaires
Arg. 2	Arg. 3	Arg. 2	Arg. 3	
NP		NP (を" o" ou parfois が"ga")		
S		S complétive (と"to")		
S		S nominalisée (のを"no o" ou のが"no ga")		pas de S (と"to") complétive en position arg2 japonais
PP		PP (に"ni",で"de",へ"he"...)		à vérifier
NP		PP (に"ni",で"de",へ"he"...)		pas de NP objet direct en arg 2 japonais
PP		NP (を"o" ou parfois が"ga")		pas de NP objet direct en arg 2 anglais
NP	PP	NP (を"o" ou parfois が"ga")	PP (に"ni",で"de",へ"he"...)	à vérifier
PP	PP	PP (に"ni",で"de",へ"he"...)	PP (に"ni",で"de",へ"he"...)	à vérifier
PP	S	PP	S complétive (と"to")	
PP	S	PP (のを"no o" ou のが"no ga")	S nominalisée	pas de S (と"to") complétive en position arg3 japonais

Figure 2 – Correspondances de sous-catégorisation automatiquement validées

Grâce à une série d'heuristiques, également implémentées par des requêtes SQL, nous avons essayé de détecter les correspondances bilingues exactes entre sous-catégorisations anglaises et japonaises, à partir des données des deux lexiques monolingues. Ainsi, pour les paires de verbes indiquées dans le lexique, nous avons validé les correspondances entre sous-catégorisations dont les arguments remplissaient les conditions affichées [5] dans le tableau2 : les correspondances entre 2 verbes transitifs directs ont été validées ; celles entre un verbe transitif direct "a" et un verbe transitif indirect "b" ont été validées dans le cas où il n'existait pas de sous-catégorisation transitive directe du verbe "b", etc. Certaines correspondances, ambiguës d'un point e vue sémantique, ont été automatiquement validées en sachant qu'une validation ou correction manuelle serait nécessaire à posteriori.

Une fois la liste des correspondances bilingues de sous-catégorisations détectées établie, nous l'avons comparée avec celles déjà présentes dans le lexique, afin de trouver parmi les correspondances de la liste lesquelles figuraient déjà dans le lexique et lesquelles devaient y être ajoutées. De plus, il a fallu déterminer quelles correspondances du lexique étaient erronées et devaient être effacées.

Il a fallu également ajuster les scores des correspondances. Lorsque deux sous-catégorisations japonaises étaient possibles pour traduire une même sous-catégorisation anglaise, nous avons donné le score maximum à celle qui avait le plus grand nombre d'occurrences dans le fichier "Case Frames".

Ensuite, les données à ajouter dans le lexique ont dû être adaptées au format nécessaire. Nous avons alors pu insérer les 8000 nouvelles correspondances bilingues dans le lexique et supprimer les correspondances erronées qui s'y trouvaient.

La dernière étape a consisté à faire corriger manuellement les correspondances sémantiquement ambiguës, c'est-à-dire celles contenant au moins un argument prépositionnel pour le verbe anglais et un argument suivi d'une particule postpositionnelle pour le verbe japonais. Ainsi, un locuteur natif japonais connaissant l'anglais a validé ou corrigé

5. Dans le tableau 2, les arguments 1 (sujets) n'ont pas été affichés. Seulement les arguments 2 et 3 (compléments) l'ont été. NP signifie "Noun Phrase" (groupe nominal), et représente les compléments d'objet direct dans ce tableau. PP signifie "Prepositional Phrase" (groupe prépositionnel) et représente les compléments d'objet indirects ou circonstanciels. S signifie "Sentence" (phrase) et représente les subordonnées complétives ou les verbes à l'infinitif ou au gérondif.

2400 des 8000 correspondances bilingues. Ces corrections ont été enregistrées dans la base de données.

5 Applications

5.1 Traduction des verbes et de leurs prépositions

La première application de l'amélioration du lexique au niveau du programme conduit à une traduction plus correcte des verbes et de leurs prépositions, comme dans l'exemple suivant, ou l'on voit l'ancienne traduction, erronée, puis la nouvelle, correcte, de la phrase anglaise "I run out of wind".

(1) I run out of wind.

* 私 は	風の	尽きる。
watashi ha	kaze no	tsukiru
Je	de vent	m'épuise

*Je m'épuise du vent.

私 は	風を	尽かす。
watashi ha	kaze o	tsukasu
Je	vent [objet direct]	épuise

Je suis à court de vent.

5.2 Conjugaison des objets verbaux

La deuxième application a concerné la génération des compléments d'objet de type verbal et de propositions subordonnées complétives dans les phrases japonaises. (Kauffmann *et al*., 2011).

En japonais, selon le verbe gouverneur, et selon le contexte, différents types de formes conjuguées existent pour les subordonnées complétives ou les objets verbaux. Les données extraites du fichier "Case Frames" nous ont montré que les verbes dont un argument est suivi par la particule conjonctive と("to") introduisent généralement une proposition subordonnée complétive,

(2)

彼は	帰った	と	思います 。
kare ha	kaetta	to	omoimasu
il	rentré à la maison	que	pense

(Je) pense qu'il est rentré à la maison.

et que les verbes transitifs directs peuvent prendre pour complément d'objet direct une phrase (ou un verbe) nominalisée.

(3)

梅酒を	飲んだのを	覚えっている 。
umeshu wo	nomda no wo	oboetteiru
liqueur de prune	avoir bu (le fait de)	se rappeler

Je me rappelle avoir bu de la liqueur de prune.

D'autres données nous ont permis d'établir une liste de semi-auxiliaires japonais, qui peuvent gouverner des verbes dont le choix de la forme fléchie sera alors déterminé : soit un gérondif en-て("Te"), soit une forme de base verbale.

(4)

私 は	やって	みます 。
watashi ha	yatte	mimasu
Je	faire [gérondif]	essaye

Je vais essayer de le faire.

Nous avons annoté notre lexique pour préciser quels verbes peuvent être semi-auxiliaires, avec quelles formes fléchies associées. Nous avons implémenté les procédures de génération correspondantes, pour assurer une bonne conjugaison des objets verbaux et des subordonnées complétives, dans tous les cas possibles.

Les tests manuels ont montré de bons résultats. Toutefois, des améliorations seront encore possibles. Dans certains cas, l'utilisation de modèles de langues pourraient permettre de déterminer le temps de l'objet verbal d'après le contexte sémantique, ce que ne permettent pas les règles lexico-syntaxiques.

Une évaluation automatique effectuée sur un corpus de 500 phrases extraites de résumés d'articles scientifiques a montré que les règles de conjugaison des objets verbaux, associées à des règles de traduction de la modalité, ont permis une amélioration de +0,03 points du score BLEU, faisant passer celui-ci de 2.49% à 2.52%.

6 Conclusion

Nous avons présenté ici une méthode qui permet d'établir automatiquement des correspondances entre sous-catégorisations verbales entre 2 langues éloignées, à partir de données strictement monolingues, essentiellement syntaxiques et statistiques. Une vérification humaine a toutefois été nécessaire, mais pour seulement 30% des sous-catégorisations détectées.

Cette méthode nous a permis de valider près de 8000 nouvelles correspondances bilingues ainsi que de corriger et améliorer notre lexique électronique. De plus, les nouvelles données ont pu être utilisées pour améliorer les phrases japonaises générées par le programme de traduction à base linguistique Its-2.

Dans de prochaines recherches, nous envisageons d'utiliser les données existantes liées aux cooccurrences ou collocations pouvant être formées par les verbes et leurs arguments, afin de réduire les erreurs de sélections lexicales dans les traductions produites.

Remerciements

J'aimerais remercier Daisuke Kawahara, Sadao Kurohashi, Eric Werhli, Luka Nerima et Christopher Laenzlinger pour leur aide. Ce travail a été financé grâce à une bourse du Fonds National Suisse de la recherche scientifique (FNS).

Références

Breen J. (2009). ENAMDICT. page web. [$csse.monash.edu.au/\ jwb/enamdict_doc.html$].

Halpern J. (2008). Japanese part of speech codes. page web. [$cjk.org/cjk/samples/jappos.htm$].

Kauffmann A., Kawahara D. & Kurohashi S. (2011). Treatment of complex sentences, modality and verbal structures in linguistics-based MT. In *Actes de NLP 2011, Toyohashi, Japon*.

Kawahara D. & Kurohashi S. (2006). Case frame compilation from the web using high-performance computing. In *Proceedings of The 5th International Conference on Language Resources and Evaluation (LREC-06)*.

Kawahara D. & Kurohashi S. (2010). Acquiring reliable predicate-argument structures from raw corpora for case frame compilation. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10), pp. 1389-1393, Valletta, Malta*.

Kinoshita S., Phillips J. & Tsujii J.-I. (1992). Interaction between structural changes in machine translation. In *Actes de COLING-92, Nantes, France*.

Mangeot M. & Kuroda K. (2003). Interlinguistic divergences in papillon multilingual dictionary. In *Actes de ASIALEX 2003, Meikai University, Urayasu, Chiba, Japan, 27-29 août 2003, pp 156-162*.

Raza G. (2010). Inferring subcat frames of verbs in urdu. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10), pp. 3689-3696, Valletta, Malta*.

van der Plas L. (2008). *Automatic lexico-semantic acquisition for question answering*. PhD thesis, University of Groningen.

Wehrli E. (2007). Fips, a "deep" linguistic multilingual parser. In *Actes de ACL 2007, Prague, Czech Republic*.

Wehrli E. & Nerima L. (2008). Traduction multilingue : le projet MulTra. In *Actes de TALN 2008, Avignon, France*.

Wehrli E., Nerima L. & Scherrer Y. (2009). Deep linguistic multilingual translation and bilingual dictionaries. In *Proceedings of the Fourth Workshop on Statistical Machine Translation, pp. 90-94*.