Analyse et classification d'exemples illustratifs dans le dictionnaire "Cesselin" en utilisant Google Traduction et un dictionnaire UNL-UWs

Mutsuko Tomokiyo, Mathieu Mangeot & Christian Boitet
Laboratoire LIG, équipe GETALP
Bâtiment IMAG CS 40700
38058 GRENOBLE CEDEX 9, FRANCE
{mutsuko.tomokiyo,mathieu.mangeot,christian.boitet}@imag.fr

1 Introduction

Nous présentons des travaux en cours sur un dictionnaire bilingue japonais-français, le "Cesselin" (Cesselin, 1940), qui a été publié en 1939 et 1957 au Japon. Notre objectif est de le rendre disponible en tant que dictionnaire en ligne pour les apprenants de japonais ou de français, les chercheurs en études japonaises et les développeurs de systèmes de traduction automatique japonais-français (Mangeot, 2016).

Le dictionnaire Cesselin inclut des connaissances sur la langue japonaise parlée et écrite et de riches exemples illustratifs des mots-vedette, tels que des descriptions phonétiques aujourd'hui obsolètes, les anciens caractères chinois et l'ancienne façon d'écrire les Okurigana[1]. Il n'est cependant pas satisfaisant que les exemples soient donnés sans indication concernant leur nature: sont-ils des proverbes, des exemples d'utilisation normale, des exemples de quantificateurs, ou d'autres types de collocations? Leur introduction permettrait aux développeurs de systèmes de TA de traiter méthodiquement les ambiguïtés lexicales (Tomokiyo, Boitet & Mangeot, 2017) . Cela aiderait aussi les apprenants ou les chercheurs en langues à pratiquer le japonais ou le français et à approfondir leur compréhension de la culture japonaise.

Après avoir analysé un ensemble de 500 exemples (sur 280 000 exemples estimés), nous avons défini une classification des exemples et un processus de classification. La prochaine étape envisagée est de l'automatiser pour économiser le plus de temps d'expert humain possible.

Nous rapportons une expérience faite afin d'étudier quels types d'exemples sont inclus dans le Cesselin. Puis, nous proposons une classification, et une procédure de classification d'un exemple donné. Après avoir expliqué le dictionnaire d'UWs, nous proposons une procédure de classification automatique et examinons quelques aspects techniques de son automatisation.

2 Études de cas utilisant le système Google Translation pour la classification

Pour classer tous les exemples dans le Cesselin en fonction de leurs usages linguistiques, nous avons traduit 500 exemples japonais et leurs traductions françaises en anglais en utilisant le système de traduction de Google ("GT" ci-dessous), et avons comparé les deux traductions.

Notre approche est basée sur l'hypothèse que, si un exemple est un proverbe, ou si un mot dans un exemple est utilisé dans un usage collocatif, les lexèmes dans l'exemple traduit JE devraient différer de ceux de l'exemple traduit FE, parce que, dans les proverbes et les expressions collocatives, y compris les classificateurs, les mots semblent avoir tendance à être utilisés au sens figuré (Tomokiyo, Boitet & Mangeot, 2017), et GT ne fournit pas de traductions adéquates pour eux (Tomokiyo & Boitet, 2016),

1 Un mot japonais est écrit en Kanji (漢字, caractère chinois) et Hiragana (平仮名, caractère japonais), en kanji seulement, en hiragana seulement ou en katakana (片仮名). Les okuriganas (送り仮名) sont des suffixes qui suivent suivant les racines écrites en kanji. Les règles pour les okuriganas ont changé plusieurs fois, et ont été standardisées par le gouvernement japonais en 1973.

mais plutôt traductions mot-à-mot. Ainsi, les deux sorties contiennent des lexèmes différents, et on peut distinguer les proverbes et les collocations des exemples d'utilisation ordinaires.

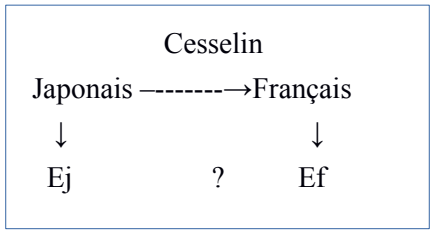

Figure 1: Hypothèse: r(Ej) ⊆ _ r(Ef) ∿ /Ø = usage collocatif, proverbial, idiomatique

3 Classer les exemples du Cesselin

Dans les exemples suivants pour le verbe ″飲む (nomu, boire)", (a) est un proverbe, (b) est une phrase incluant "飲む" dans un sens figuratif, et (c) est une phrase qui n'est pas un proverbe ni une collocation, mais un usage ordinaire. Quand on compare les lexèmes dans les traductions JE et FE par GT, alors que (a) et (b) ne contiennent aucun mot commun, en (c), les deux traductions ont deux mots ("boire" et "eau") en commun.

Cela confirme que les exemples montrant des usages ordinaires ont tendance à se traduire par la composition de mots communs. Nous avons pensé que ce phénomène pourrait être un atout pour classer tous les exemples du Cesselin, car les traductions du japonais-anglais par GT sont presque des traductions mot-à-mot et, quand un mot dans les exemples japonais est utilisé dans le sens figuratif dans les contextes, les résultats de traduction ne correspondent jamais aux résultats de la traduction français-anglais.

Parmi les exemples d'usage courant, les proverbes, les expressions collocatives, les quantificateurs et les exemples spécifiques au domaine, nous pouvons distinguer les proverbes d'autres types d'exemples en utilisant un dictionnaire de proverbes et en utilisant des labels tels que {fig} et {botanic, zoology...}, qui sont attachés à certains exemples dans le Cesselin, nous pouvons distinguer les expressions collocatives et les exemples spécifiques au domaine d'autres exemples, respectivement.

5. Vers une classification automatique

Pour effectuer une classification automatique des exemples, nous avons besoin d'un dispositif qui associe des phrases, des mots ou des lexèmes et analyse les exemples. Il passe par les étapes suivantes, suivant l'analyse manuelle précédente.

1. Si des exemples japonais apparaissent dans un dictionnaire de proverbes japonais, ils sont annotés comme {prov}, pour "proverbe".
2. Si la traduction française dans le Cesselin est étiquetée par {fig}, l'exemple est annoté comme {colexp}, pour "collocation", et si la traduction française dans Cesselin est étiquetée par {botanic, zoology,... etc.}, l'exemple est annoté comme {domexp}, pour "utilisation spécifique au domaine" indiqué.
3. Si les lexèmes de deux sorties de traduction correspondent à 100%, l'exemple est annoté comme {ordusg}, pour "exemple d'utilisation ordinaire".
4. Si les deux sorties de traduction ne sont pas dans les cas 1, 2, 3 ci-dessus, elles sont examinées avec une liste KWIC, en utilisant un logiciel, appelé Sketch Engine. Si un exemple apparaît avec certaines occurrence, il est considéré comme une collocation ou un quantificateur (Alda, 2011).
5. Contrairement à la liste KWIC, si un exemple est un syntagme nominal composé de "chiffre+ nom", alors l'exemple est annoté comme {quant} pour "quantificateur" (Tomokiyo, Boitet & Mangeot, 2017), (Miyagawa 1989).

6. Si les exemples dans les deux résultats de traduction n'apparaissent pas dans la liste KWIC avec certaines occurrences, la synonymie entre les mots dans les deux résultats de traduction est examinée en utilisant un dictionnaire UNL-UW.
7. Si le verbe prédicatif et ses actants dans les deux traductions de l'exemple correspondent, et si les mots du reste des exemples traduits correspondent au niveau de la synonymie, alors l'exemple est annoté comme {ordusg} pour "exemple d'usage ordinaire".
8. S'il n'y a pas de correspondance entre les lexèmes contenus dans les deux traductions de l'exemple, il est annoté {colexp}, pour "utilisation en collocation", sauf s'ils sont déjà classés comme proverbes, quantificateurs, ou usage spécifique au domaine.
9. Si un exemple n'appartient pas à un des cas 1, 2, 3, 4, 5, 6 7, 8, mentionnés ci-dessus, l'exemple est annoté comme {ordusg} pour "exemple d'utilisation ordinaire".

Remarque

Les exemples dans les dictionnaires ne sont en général pas très longs (environ 10 mots en moyenne), et comprennent en certaines proportions des proverbes, des groupes nominaux, des groupes adjectivaux, ou des phrases incomplètes. Donc, si nous cherchons à appliquer notre méthode de classification/annotation à d'autres documents, nous ne sommes pas sûrs de pouvoir obtenir les mêmes résultats que dans le cas d'un dictionnaire.

Références

Gustave Cesselin (1940) Dictionnaire japonais-français. Maruzen, Tokyo, juillet 1940, 2340 p.

Alda Mari. 2011. *Quantificateurs polysémiques*, Université Paris-Sorbonne, Vol.23, France.

Mathieu Mangeot-Nagata. 2016. *Collaborative construction of a good quality, broad coverage and copyright free Japanese-French dictionary*. HAL-01294566.

Mutsuko Tomokiyo, Christian Boitet et Mathieu Mangeot. 2017. *Development of a classifiers/quantifiers dictionary towards French-Japanese MT*, MTSummit 2017, Japan

Mutsuko Tomokiyo et Christian Boitet. 2016. *Corpus and dictionary development for classifiers/quantifiers towards French-Japanese machine translation,* COLING, Cog-Alex 2016, Japan

Shigeru Miyagawa. 1989. *Structure and case marking in Japanese, Syntax and Semantic*s, Syntax and Semantics, Vol.22, New York.

Uchida, H., Zhu, M., & Della Senta, T.G. 2006. *Universal Networking Language*. UNDL Foundation, Japan.