Projet Papillon : intégration de dictionnaires existants et gestion des contributions

Mathieu Mangeot

Projet Papillon□intégration de dictionnaires existants et gestion des contributions

Mathieu MANGEOT-LEREBOURS,
National Institute of Informatics (NII)
2-1-2-1314, Hitotsubashi
Chiyoda-ku Tokyo 101-8430, Japan
Tel.□□+81-3-4212-2672 - Fax.□+81-3-3556-1916
Mél.□mangeot@nii.ac.jp

Résumé

Le projet Papillon[1] a pour but de créer un environnement pour le développement et la consultation d'une base lexicale multilingue. Cet environnement coopéératif doit être libre, gratuit, et personnalisable. De plus, il doit pouvoir être accessible sur le web en permanence. La motivation initiale du projet est le manque de dictionnaires, à la fois pour humains et machines, entre le français et de nombreuses langues asiatiques. En particulier, même s'il existe de gros dictionnaires d'usage imprimés, ils ne sont utilisables que par des japonophones car ils ne contiennent jamais à la fois la prononciation transcrite en romaji et kana et lécriture en kanji (idéogrammes). Ceci est aussi vrai pour le thaï, le vietnamien, le lao, etc. Afin de réduire les coûts de création de cette connaissance lexicale, nous adoptons une méthode inspirée des projets "open-source" afin de créer une base lexicale multilingue grâce à des contributeurs bénévoles.

Le projet a déjà été présenté aux JST [2,7]. Cette année, nous présentons donc les avancées dans l'import de dictionnaires existants, ainsi que la méthodologie retenue pour créer un nouveau squelette de dictionnaire à partir de ces dictionnaires puis la gestion des contributions volontaires.

Abstract

The PAPILLON project[1] aims at creating a cooperative, free, permanent, web-oriented and personalizable environment for the development and the consultation of a multilingual lexical database. The initial motivation is the lack of dictionaries, both for humans and machines, between French and many Asian languages. In particular, although there are large F-J paper usage dictionaries, they are usable only by Japanese literates, as they never contain both original (kanji/kana) and romaji writing. This applies as well to Thai, Vietnamese, Lao, etc. In order to cut the costs of the creation of this lexical knowledge, we adopt a strategy inspired from "open-source" projects to allow volunteers to collaborate in the creation of a multilingual lexical database.

The project has already been presented during previous JST [2,7]. This year, we present the advances for importing existing dictionaries and the methodology used to create a new dictionary skeleton from these imported dictionaries and then the management of the voluntary contributions.

Introduction

1. Présentation du projet

Ce projet a pour but de créer une base lexicale multilingue comprenant entre autres l'anglais, le français, le japonais, le malais, le lao, le thaï et le vietnamien. L'accès est gratuit pourvu que l'usage ne soit pas commercial (licence de logiciel libre). Notre projet se veut utile et ouvert à la collaboration de toutes les personnes ayant un intérêt pour ces langues. La macrostructure du dictionnaire est composée d'un volume monolingue pour chaque langue et d'un volume pivot contenant des liens interlingues reliant les sens des mots composant les volumes monolingues[5,6]. La miscrostructure des articles est basée sur la lexicographie combinatoire extraite de la théorie sens-texte [3,4]. La base lexicale est accessible principalement par le Web. Toute personne peut consulter les données existantes et ensuite corriger/compléter ces données.

2. Import de dictionnaires existants

La base lexicale distingue 3 niveaux diffèrents pour la gestion des dictionnaires existants : les limbes, le purgatoire et le paradis. Les limbes sont constituées de dictionnaires stockés dans leur format original. Le purgatoire ne contient que des dictionnaires au format XML mais ayant leur structure d'origine. Enfin, le paradis contient lui, les volumes consituant le dictionnaire Papillon.

[1] http://www.papillon-dictionary.org/

Pour télécharger un dictionnaire existant au purgatoire sur le serveur Papillon, l'utilisateur doit décrire ce dictionnaire ainsi que les volumes qu'il contient et la structure des articles à l'aide d'un fichier XML de métadonnées. L'import se fait alors automatiquement à partir de ce fichier.

3. Intégration des dictionnaires importés

L'intégration des dictionnaires importés consiste à les transférer du purgatoire au paradis. Pour cela, il faut d'une part convertir les informations identifiées vers le format papillon et ensuite vérifier que l'article n'existe pas déjà. La division d'un article de dictionnaire en sens de mots est toujours subjective. Il suffit d'ouvrir 2 dictionnaires différents et de comparer quelques articles pour s'apercevoir que la division en sens de mots est rarement la même.

Il est alors très difficile de fusionner plusieurs dictionnaires existants au niveau du sens des mots. Pour résoudre ce problème, nous utilisons les vecteurs conceptuels : chaque sens de mot est représenté par un vecteur conceptuel. L'espace vectoriel est construit à partir d'un thésaurus existant. Pour le français, nous avons utilisé le thésaurus Larousse de 873 concepts. Dès lors, l'espace vectoriel aura 873 dimensions (une pour chaque concept). Pour chaque dimension d'un vecteur, on note la distance sémantique entre le sens de mot représenté par le vecteur et le concept.

Ensuite, après avoir calculé les vecteurs conceptuels pour chaque sens de mot des 2 dictionnaires que l'on veut fusionner, on utilise la distance angulaire entre 2 vecteurs pour déterminer si on fusionne ou non les sens de mots qu'ils représentent.

4. Gestion des contributions volontaires

Chaque utilisateur de la base peut à tout moment corriger un article existant, ajouter des informations ou une nouvelle traduction. Pour éviter de polluer la base, les contributions sont d'abord stockées dans l'espace privé de l'utilisateur avant d'être révisées par un spécialiste et intégrées dans la base. Cependant, en attendant la révision de ses contributions, lorsque l'auteur consultera le dictionnaire, les contributions seront affichées comme si elles étaient déjà intégrées. L'auteur de la contribution peut aussi la faire partager à un groupe de collègues qui la verront aussi s'ils consultant le dictionnaire. Par contre, un autre utilisateur ne verra cette contribution que lorsqu'elle sera révisée. Nous avons donc implémenté une ébauche d'interface d'édition des articles et une interface de gestion des contributions paramétrée pour chaque utilisateur. Il est possible de voir toutes les contributions, d'en rechercher certaires par pluseiurs critères (date, langue, contenu) ainsi que d'en supprimer certaines.

Conclusion

Le cadre théorique de la base lexicale, la macrostructure et la microstructure est bien défini. [2] Il constitue une base solide pour l'expérimentation. Des avancées importantes ont été réalisées dans la mise en œuvre. Cependant, pour éviter le découragement des contributeurs bénévoles potentiels face à un environnement partiellement défini, nous pensons qu'il est très important de bien mettre au point un tel environnement avant son lancement officiel devant le grand public

Bibliographie

[1] Mathieu Mangeot-Lerebours (2001) *Environnements centralisés et distribués pour lexicographes et lexicologues en contexte multilingue*. PhD Thesis in Computer Sciences Université Joseph Fourier Grenoble I, 27 September 2001, 280 p.

[2] Mathieu Mangeot-Lerebours ' Gilles Sérasset (2001) Projet Papillon : architecture du serveur Web et structure des articles. JST'2001 Journées Science et Technologie, National Olympic Memorial Youth Center, Tokyo, Japon, lundi 19 ' mardi 20 décembre, vol 1/1, pp 149-150.

[3] Alain Polguère (1998) *La théorie Sens-Texte* .Dialangue, Vol. 8-9, Université du Québec à Chicoutimi, pp 9-30.

[4] Alain Polguère (2000) *Towards a theoretically-motivated general public dictionary of semantic derivations and collocations for French*. Proc. EURALEX'2000, Stuttgart, pp 517-527.

[5] Gilles Sérasset (1994) *Interlingual Lexical Organisation for Multilingual Lexical Databases in NADIA*. In Proc. COLING-94, Kyoto, 5-9 August 1994, M. Nagao ed. vol. 1/2 : pp. 278-282.

[6] Gilles Sérasset ' Mathieu Mangeot-Lerebours (2001) *Papillon Lexical Database Project: Monolingual Dictionaries ' Interlingual Links*. Proc. NLPRS'2001, Hitotsubashi Memorial Hall, National Center of Sciences, Tokyo, Japan, 27-30 November 2001, vol 1/1, pp. 119-125.

[7] Mutsuko Tomokiyo et al. (2000) *Papillon : a Project of Lexical Database for English, French and Japanese, using Interlingual Links*. Journées Science et Technologie de l'ambassade de France au Japon, 13 November 2000, Tokyo, Japan, 3 p.