Réutilisation de traducteurs gratuits pour développer des systèmes multilingues

VO TRUNG Hung

Institut National Polytechnique de Grenoble
GETA, CLIPS, IMAG - campus
385, rue de la Bibliothèque, BP 53-38041 Grenoble Cedex 9, France
Tél : +33 4 76 51 43 55, Fax : +33 4 76 51 44 05
Mail : Hung.Vo-Trung@imag.fr

Résumé – Abstract

Nous présentons ici une méthode de réutilisation de systèmes de traduction automatique gratuits en ligne pour développer des applications multilingues et évaluer ces mêmes systèmes. Nous avons développé un outil de traitement et de traduction de documents hétérogènes (multilingues et multicodage). Cet outil permet d'identifier la langue et le codage du texte, de segmenter un texte hétérogène en zones homogènes, d'appeler un traducteur correspondant avec une paire de langue source et cible, et de récupérer les résultats traduits dans la langue souhaitée. Cet outil est utilisable dans plusieurs applications différentes comme la recherche multilingue, la traduction des courriers électroniques, la construction de sites web multilingues, etc.

We present here a method of the reuse of the free on line automatic translation systems to develop multilingual applications and to evaluate translators. We developed a tool for treatment and translation of the heterogeneous documents (multilingual and multicoding). This tool makes it possible to identify the language and the coding of the text, to segment a heterogeneous text in homogeneous zones, to call a translator corresponding with a fair of languages and to recover the results translated into desired language. We can use this tool in several different applications as multilingual research, the translation of the e-mail, the construction of the multilingual Web, etc.

Keywords – Mots Clés

Traduction Automatique, Traducteur Multilingue, Multilinguisme, Document Multilingue
Machine Translation, Multilingual Translator, Multilingualism, Multilingual Document

1 Introduction

Actuellement, il existe plusieurs versions de traducteurs automatiques non commerciaux en ligne, comme Systran, WorldLingo, Reverso, … avec la limitation générale de la longueur du texte à moins de 50-150 mots. En outre, ces traducteurs ne permettent que de traduire des textes ou des pages web monolingues et monocodage avec une paire de langues déterminées à l'avance.

Avec l'utilisation répandue d'Internet, nous recevons régulièrement des informations écrites en plusieurs langues (par exemple, courriers électroniques, catalogues, notices techniques, sites web, etc) et le besoin de la traduction de ces textes en langue d'utilisateur se pose naturellement. De plus, nous recevons des informations dont nous ne pouvons pas savoir quelles langues contenues.

La réutilisation de traducteurs gratuits pour développer des applications multilingues est donc une solution utile. L'idée est d'appeler des traducteurs existants pour traduire des textes dans la langue souhaitée. L'intégration des traducteurs dans les systèmes multilingues permet de traduire automatiquement des textes ou des messages à l'exécution et d'éviter par ailleurs le stockage d'une même donnée en plusieurs langues.

Nous présentons ici une méthode de réutilisation des systèmes de traduction automatique gratuits en ligne pour développer les applications multilingues et évaluer des traducteurs. Nous combinons plusieurs traducteurs et utilisons l'anglais comme pivot de langue pour obtenir un maximum de paires de langue. Nous avons développé un outil de traitement et de traduction des documents hétérogènes (multilingues et multicodage). Cet outil permet d'identifier la langue et le codage du texte, de segmenter un texte hétérogène en zones homogènes, d'appeler un traducteur correspondant avec une paire de langue et récupérer les résultats traduits en langue souhaitée. Ce système peut s'adapter pour plusieurs systèmes de codage différents comme BIG-5, GB-2312 (chinois), Shift-JIS (japonais), EUC-kr (coréen), KOI-8, CP 1251 (russe), etc. Nous pouvons utiliser cet outil dans plusieurs applications comme la recherche multilingue, la traduction des courriers électroniques, la construction des sites web multilingues, etc.

2 Objectifs

Notre premier objectif est la construction d'un outil d'appeler des traducteurs gratuits, disponibles sur le Web et de récupérer leurs résultats. Nous pouvons utiliser les paramètres tels que les langues (source et cible), le codage, le nom de traducteur d'obtenir des résultats traduits différents pour une même entrée.

Le deuxième objectif est l'utilisation de ces traducteurs pour développer des applications multilingues. Nous pouvons intégrer ces traducteurs dans des systèmes multilingues pour traduire des textes, des messages en exécutant du programme.

Nous avons développé un site web pour traduire des textes hétérogènes (multilingues et multilicodage) dans la langue souhaitée, même si nous ne savons pas la langue du texte d'entrée. Cet outil permet de traduire un texte quelconque dans la langue souhaitée. Si ce texte est hétérogène, notre système segmentera et identifiera quel segment est écrit en quelle langue

et quel codage pour déterminer un traducteur correspondant. Par exemple, nous pouvons copier/coller des textes en plusieurs langues (anglais, chinois, japonais, etc) et choisir le français comme langue cible. Ou, nous recherchons des sites contenant un mot (en particulier des mots techniques) de "*segment*" par google, le résultat comprend des sites en français, anglais, allemand, etc, et nous pouvons utiliser cet outil pour lire les résultats dans une même langue.

Nous pourrions aussi intégrer cet outil dans plusieurs applications différentes comme la traduction des courriers électroniques, la génération d'un message en plusieurs langues, l'évaluation de qualité des traducteurs, etc.

3 État de l'art

Cette session aborde quelques systèmes récents de traduction automatique et de traitement des documents hétérogènes en ligne. Nous pouvons appeler ces systèmes pour traiter et traduire des textes dans les systèmes multilingues.

3.1 Systran

Actuellement, SYSTRAN est un traducteur très connu et sa technologie alimentent des solutions de traduction pour Internet, PC et infrastructures de réseau avec 36 paires de langue et 20 domaines spécialisés différents. La version en ligne permet de traduire pour 34 paires de langue sur l'adresse http://www.systranbox.com/.

3.2 Reverso online

C'est un outil de chez Softissimo pour la traduction automatique en ligne de textes ou pages web. Cet outil peut fonctionner en PC, Internet, Intranet soit sous la forme d'une application autonome. L'adresse de traducteur est http://www.reverso.net/textonly/default.asp.

3.3 Sandoh

SANDOH (Système d'ANalyse des DOcuments Hétérogène) est un outil permettant analyse d'un document hétérogène en zones homogènes. Nous avons développé cet outil pour analyser un texte hétérogène selon la méthode du diagnostic de la langue et du codage n-grammes et la méthode de la segmentation progressive pour segmenter un texte hétérogène en zones homogènes. Le résultat de l'analyse est le nom d'un couple <langue, codage> si ce document est homogène, sinon, les zones et le couple <langue, codage> utilisé dans chaque zone {zone-1, langue-1, codage-1}, {zone-2, langue-2, codage-2}, …, {zone-n, langue-n, codage-n}.

Cet outil se trouve sur le site web http://www-clips.imag.fr/geta/User/hung.vo-trung/id_langue/web_fr/Index.htm.

4 Construction d'outil de traduction automatique multilingue

4.1 Architecture du système

Nous avons développé un outil pour traduire automatiquement des textes (multilingues ou monolingues) à partir de traducteurs existants en ligne. L'architecture du système est la suivante :

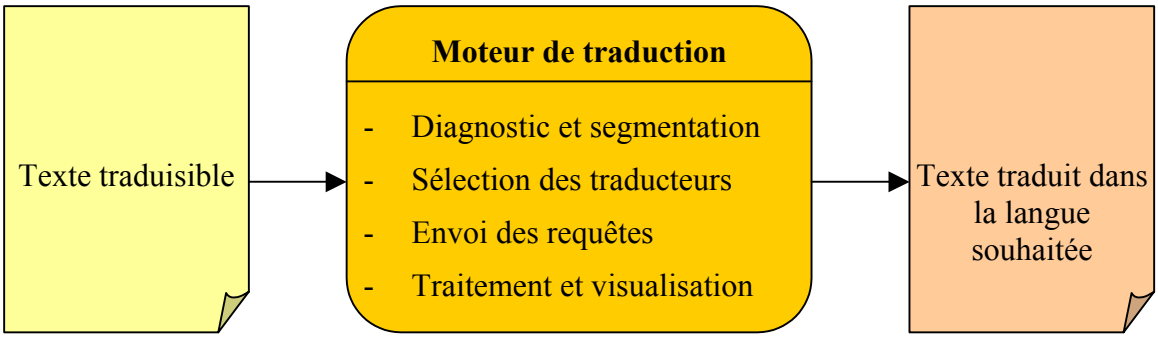

Figure 1. Système de la traduction automatique multilingue

4.2 Diagnostic et segmentation

Le texte d'entrée peut-être un texte multilingue ou multicodage (on peut copier/coller à partir de plusieurs sites web différents, ou bien les textes reçus peuvent être hétérogènes). Nous extrayons chaque paragraphe pour diagnostiquer la langue et le codage et envoyer aux serveurs de traduction parce que ces serveurs n'acceptent souvent que des textes de longueur inférieure à 150 mots. Si le paragraphe est monolingue, nous l'envoyons immédiatement au serveur, sinon il faut segmenter en zones monolingues dont chacune correspond à une ou plusieurs phrases. Après le diagnostic de la langue et du codage, il faut éventuellement convertir le codage du texte vers le système de codage accepté par le serveur de traduction.

4.3 Sélection des traducteurs

Nous réutilisons les traducteurs gratuits en ligne et composons deux traducteurs si le couple source-cible direct n'existe pas en prenant l'anglais comme pivot. En effet, si un système comme Systran, Gist-in-Time, Reverso, … a un traducteur pour la langue L, il en a toujours un entre L et l'anglais. Certainement les traductions "double" sont de qualité nettement inférieure aux traductions simples. Mais nous sommes en contexte d'accès à l'information et les résultats restent le plus souvent utilisables. Par exemple, pour traduire un texte français en japonais, nous pouvons composer deux traducteurs français → anglais et anglais → japonais (du serveur de traduction Systran). De même, pour traduire un texte arabe en français, nous pouvons composer le moteur bilingue arabe → anglais (du serveur de FreeLanguageTranslation) et anglais → français (du serveur de Systran).

Avec l'outil Sandoh, nous pouvons traduire un texte dans la langue souhaitée même si nous ne savons pas quelle est la langue du texte source. Sandoh diagnostiquera la langue du texte, puis

le système déterminera un traducteur pour la paire de langues correspondante. Cette fonction est très confortable pour traduire des courriers, des sites web, des documents qu'on n'a pas marqués la langue utilisée.

4.4 Envoi des requêtes

Après avoir déterminé le serveur de TA adéquat, nous lui envoyons une requête au serveur de traduction. Le résultat obtenu est un fichier sous la forme HTML.

Cela est réalisé par la fonction ***get_doc(),*** que nous avons écrite en Perl. Pour traduire un texte, le programme appelle cette fonction sur chaque unité de traduction déterminée par la segmentation avec en paramètres l'URL du site web de traduction, le contenu du segment à traduire, la paire de langues source et cible. Par exemple, pour traduire un texte sur le serveur Systran, nous appelons ***get_doc()*** de la façon suivante :

```
@res = get_doc("www.systranbox.com/systran/box? systran_text=$phrase[$j]&systran_lp=$ls_lc");
```

Les serveurs de TA acceptent souvent des codages différents pour les mêmes langues (par exemple, EUC_jp ou Shift-JIS pour le japonais, ISO-8958-1 ou Unicode pour le Français). Il faut convertir le codage du texte en codage accepté du serveur avant l'envoi d'une requête de la traduction.

4.5 Expérimentation

Nous avons développé un site web qui permet de traduire automatiquement 11 langues avec 55 paires de langues différentes en traduction simple. Cet outil permet de traduire des textes entrés directement dans une boîte de texte, ou contenus dans des fichiers sur le disque local. La longueur du texte à traduire n'est pas limitée. Ce texte peut être monolingue ou multilingue.

5 Applications possibles

L'application des systèmes de traduction automatique à la multilinguïsation des logiciels. Nous pensons à appliquer ces systèmes à deux niveaux. Au premier niveau, il s'agira de traduire des messages, des fichiers de messages, des menus, des fichiers d'aide, des documents en localisation des logiciels (par exemple, transcrire les fichiers de messages en Ariane G5). Au deuxième niveau, on l'utilisera comme un module interne dans des systèmes multilingues, pour traduire directement à l'exécution des messages d'une langue dans une autre. L'application des traducteurs au deuxième niveau permettra de gérer facilement l'interface d'utilisateur. Nous pourrons alors construire un logiciel multilingue qui comprend un seul code de programme, les catalogues de messages, et un module de traducteur.

L'intégration dans les systèmes de communication sur Internet. Nous prévoyons aussi d'utiliser cet outil dans les systèmes de courrier électronique (pour traduire automatiquement les courriers reçus dans la langue d'utilisateur), les systèmes de dialogue multilingue en ligne (pour dialoguer en plusieurs langues), les systèmes de commerce électronique, etc.

Le traitement de corpus multilingues. Cet outil servira enfin à produire des corpus en nouvelle langue (ou au moins à produire un premier jet, comme cela commence dans le projet TraCorpEx), à évaluer des corpus (grâce aux corpus testés), etc.

6 Perspectives

Dans le projet TraCorpEx, qui vise à évaluer la qualité de traducteurs à partir des corpus multilingues existants (par exemple, le corpus BTEC avec les langues comme anglais, français, japonais, …) et ajouter des langues à des corpus multilingues d'exemples. Nous avons commencé à traduire la partie anglaise de ce corpus en français par des traducteurs différents (Systran, WorlLingo, …). Nous évaluerons ensuite ces traductions par diverses méthodes.

À court terme, nous allons aussi étendre notre système par une interface permettant de lui "déclarer" des traducteurs disponibles sur le web, avec toutes les informations nécessaires (couples de langue, codages, formats, paramétrabilité). Nous étendrons aussi la fonction getdoc() pour permettre d'appeler des traducteurs "Pro" installés sur notre serveur, car ils sont bien plus rapides et paramétrables que les versions web.

Références

BLANCHON H., BOITET C. & CAELEN J. (1999), Participation francophone au consortium C-STAR II, *La Tribune des Industries de la Langue et du Multimédia* , Vol. 31-32 (August-December 1999) : pp. 15-23.

BOITET C., BLANCHON H. (1995), Multilingual Dialogue-Based MT for monolingual authors: the LIDIA project and a first mockup, *Machine Translation*, Vol. 9, pp 99-132.

HUTCHINS J. (2001), Machine Translation Over Fifty Years, Revue *HEL (Histoire Epistemologie Langage)*, publié par la Société d'histoire et d'épistémologie des sciences du langage (SHESL), Université de Paris VII, Vol. 23.

HUTCHINS J. (1999), The development and use of machine translation systems and computer-based translation tools, *International Symposium on Machine Translation and Computer Language Information Processing*, Beijing, Chine.

LAROSSE L. (1998), Méthodologie de l'évaluation des traductions, *Meta,* Vol. 43, N° 2, Presses de l'Université de Montréal.

VO TRUNG H. (2003), Évaluation des méthodes et outils pour identifier automatiquement la langue et le codage d'un texte homogène, Actes de la conférence *MAJECSTIC'03*, octobre 2003, Marseille, France.

VO TRUNG H. (2004), SANDOH - un système d'analyse de documents hétérogènes, Actes de la conférence *JADT 2004 (Journées internationales d'Analyse statistique des Données Textuelles)*, mars 2004, Louvain-la-Neuve, Belgique.