CIFLI-SurviTra, deux facettes : démonstrateur de composants de TA fondée sur UNL, et phrasebook multilingue

Georges FAFIOTTE, Achille FALAISE, Jérôme GOULIAN

LIG-GETALP, UJF Grenoble 1 BP 53, 38041 – GRENOBLE cedex 9
{georges.fafiotte, achille.falaise, jerome.goulian}@imag.fr

Résumé CIFLI-SurviTra ("Survival Translation" assistant) est une plate-forme destinée à favoriser l'ingénierie et la mise au point de composants UNL de TA, à partir d'une mémoire de traduction formée de livres de phrases multilingues avec variables lexicales. SurviTra est aussi un *phrasebook* digital multilingue, assistant linguistique pour voyageurs monolingues (français, hindi, tamoul, anglais) en situation de "survie linguistique". Le corpus d'un domaine-pilote ("Restaurant") a été structuré et construit : sous-domaines de phrases alignées et classes lexicales de locutions quadrilingues, graphes UNL, dictionnaires UW++/français et UW++/hindi par domaines. L'approche, générique, est applicable à d'autres langues. Le prototype d'assistant linguistique (application Web, à interface textuelle) peut évoluer vers une application UNL embarquée sur SmartPhone, avec Traitement de Parole et multimodalité.

Mots-clés : TA via UNL, démonstrateur de composants UNL, assistant linguistique sur le Web, *phrasebook* digital multilingue, mémoire de traduction, collecte collaborative de corpus

1 CIFLI, objectifs, contexte

UNL (Universal Networking Language) est un formalisme lexico-sémantique interlingue, exprimant de manière non ambiguë tout énoncé textuel d'une langue naturelle. Un énoncé y est représenté par un graphe, dont les arcs sont des "relations sémantiques UNL", et les nœuds sont des "UWs" (Universal Words, acceptions appartenant à une Base de Connaissances UNL et à des dictionnaires d'UWs pour chaque langue), UWs eux-mêmes qualifiés par des "attributs UNL". Cette approche pivot favorise les méthodologies collaboratives de construction de composants de TA, notamment pour les langues "peu dotées" (*pi-languages*).

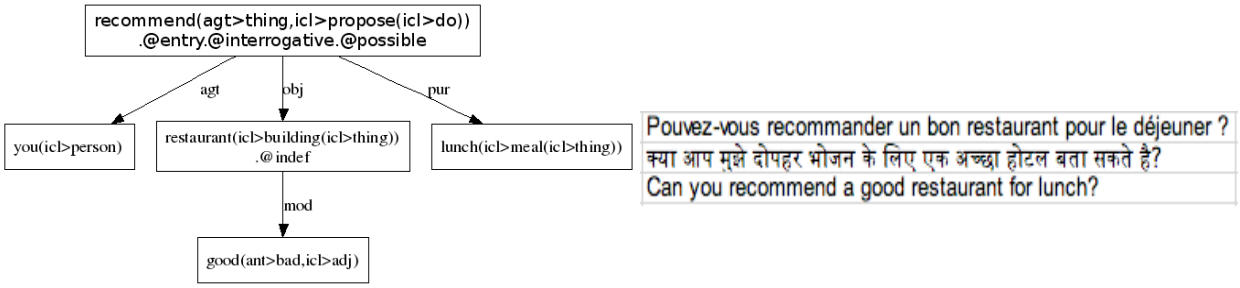

Figure 1. Graphe UNL d'un énoncé du corpus SurviTra (français-hindi-anglais)

Le protoypage de SurviTra s'inscrit dans un projet franco-indien en ingénierie de la TA fondée sur UNL, CIFLI (Communication sur Internet en Français et Langues Indiennes), entre LIG-GETALP (pilote), CFILT d'IIT-Bombay (coordinateur indien), Pondicherry University.

2 L'assistant linguistique SurviTra, phrasebook multilingue

L'application-pilote SurviTra est un *phrasebook* (livre de phrases) digital multilingue, assistant de "survie linguistique" pour voyageurs monolingues, fonctionnant à la fois comme "mémoire de traduction", et comme démonstrateur de composants UNL de TA. Il est fondé sur un corpus multilingue structuré (domaines, sous-domaines de phrases, classes lexicales de lexies), aligné (français, hindi, tamoul, anglais), avec extension dynamique modérée, en ligne. Avec le prototype Survitra-1, pour illustrer la facette "application-hôte" en **mode "assistant de base"** (hors UNL)**, sont présentés**, dans le cadre d'une interaction entre un touriste (*survivor*) et un interlocuteur (*helper*), français ou indiens : la **recherche de phrase** dans le *phrasebook* (par configuration de mots-clés, par sous-domaine, ou combinaison des deux), le traitement de **"phrases à trou(s)"** (phrases avec variants lexicaux, que l'utilisateur instancie, comme dans un *phrasebook*), l'appel de la **mémoire de traduction** (quadrilingue) pour une phrase trouvée du **domaine-pilote** ciblé (Restaurant), le traitement d'une **phrase non trouvée** (ajout dynamique pour enrichissement ultérieur de la base de phrases multilingue), et les possibilités actuelles **d'enrichissement de corpus** en mode "bac à sable", facilitant la gestion contributive des ressources lexicales utiles pour la facette "TA via UNL" du projet.

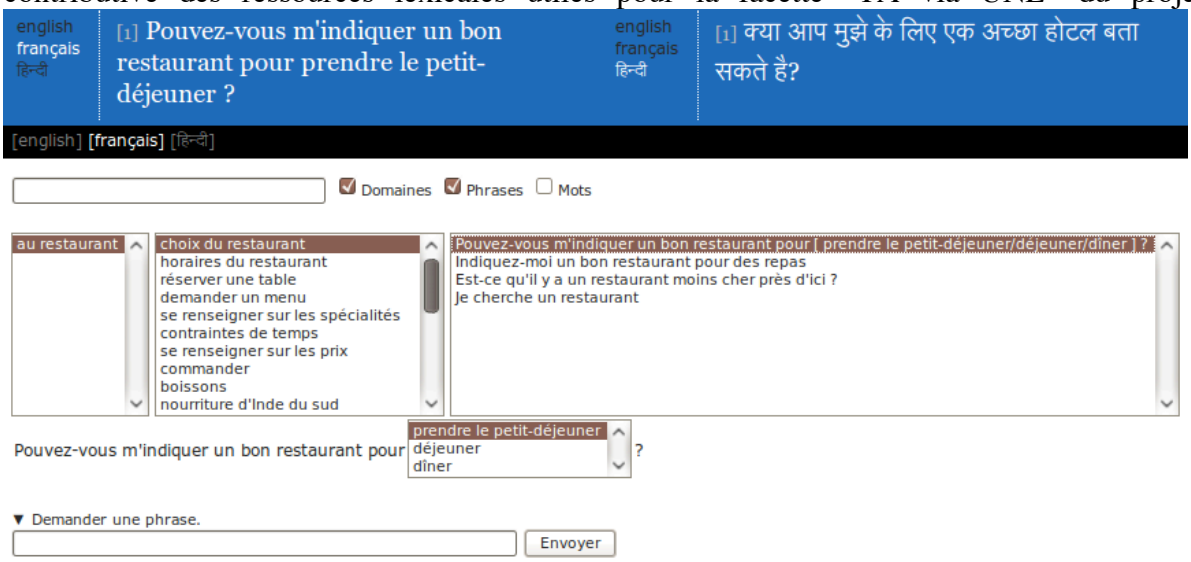

Figure 2. Question "à variable lexicale" du *survivor* français, traduite en hindi pour le *helper*

3 SurviTra, démonstrateur de composants UNL

En mode **"démonstrateur UNL" bilingue** (français-hindi), l'environnement SurviTra permet d'activer/d'inhiber les enconversions (Langue→UNL) et déconversions (UNL→Langue) disponibles, en *plug-in*. Il fonctionne en banc d'essai-test-réglage de ces composants, tout ce qui est produit étant enregistré pour analyse et validation. Ce choix fonctionnel répond à la complexité de réalisation et de mise au point des déconversions et des enconversions : dans la réalité des réalisations de TA via UNL, celles-ci ne sont pas disponibles en même temps ni avec une même couverture syntaxique pour chaque langue. La plate-forme fonctionne soit avec la mémoire de traduction (utilisant la base de phrases multilingue, dotée de graphes UNL

produits manuellement), soit par traitement effectif UNL, selon les composants disponibles. SurviTra-1 privilégie l'approche support Web monoposte. Il est opérationnel en mémoire de traduction sur un premier domaine ciblé, et en cours de finalisation pour le *plug-in* des composants UNL disponibles (avec une couverture variable en français et hindi). **Le développement incrémentiel permet de présenter** ici, pour des phrases du domaine-pilote "Restaurant, en Inde et en France", les composants d'**enconversion Français→UNL**, **déconversion UNL→Hindi**, **déconversion UNL→Français** (rétro-traduction de validation).

4 Bilan, évolution

Work in progress : Le corpus SurviTra comporte à ce jour 25 sous-domaines et 40 classes lexicales, 200 phrases et 400 lexies, en français, hindi, tamoul et anglais. L'extension est en cours, en couverture du premier domaine-pilote et en construction d'autres domaines. L'unification des dictionnaires (UW-français 45 000 entrées, UW-hindi 116 000 entrées fondé sur WordNet indien), au standard UW++, donne lieu à une recherche spécifique au CFILT (IIT-Bombay). Le dictionnaire UW++ de référence, basé sur WordNet 2.1 (200 000 entrées hors-langue), est développé par les *UNL Language Centers* espagnol et russe. Une approche pragmatique s'attache parallèlement (lors de la création manuelle des graphes UNL de la base de phrases SurviTra) à produire semi-automatiquement des sous-dictionnaires bilingues "par domaines ciblés" UW++/français/hindi. Le travail sur le corpus tamoul a commencé. Pour la TA via UNL, la déconversion du français est disponible à 75% de couverture syntaxique, l'enconversion à 20%. La déconversion de l'hindi progresse au CFILT. SurviTra-1 peut accueillir des logiciels de TA substitutive hors UNL (pour les langues où ils sont disponibles), en alternative aux modules UNL et pour favoriser les situations d'extension de corpus.

Evolution : L'approche générique SurviTra (structuration de corpus spécifiée, extensibilité multilingue, modularité) est tout à fait applicable à d'autres langues, d'abord en mémoire de traduction, puis pour autant que soit réalisé l'important travail d'enconversion-déconversion et sur les dictionnaires UNL. Un projet reprend la logique de SurviTra-1, en la ciblant (en plus du support Web) sur une application UNL embarquée pour SmartPhone ou PDA, avec multi-modalité, parole multi-accents, Traitement de Parole, écran tactile, photos illustrant des lexies.

Remerciements Recherche financée par ARCUS (Région Rhône-Alpes, Ministère des Affaires Etrangères français). Merci à tous les contributeurs (Grenoble, Bombay, Pondichéry).

Références

BEKIOS J., BOGUSLAVSKY I., CARDEÑOSA J., GALLARDO C. (2007). Automatic Construction of an Interlingual Dictionary for Multilingual Lexicography. Proceedings, *International Conf. on Machine Learning: Models, Technologies & Applications (MLMTA'07)*, 215-220.

BOITET C., BHATTACHARYYA P., BLANC E., MEENA S., BOUDHH S., FAFIOTTE G., FALAISE A., VACHANI V. (2007). Building Hindi-French-English-UNL resources for SurviTra-CIFLI, linguistic survival system under construction. *SNLP '07*, Thaïlande.

SINGH S., DALAL M., VACHANI V., BHATTACHARYYA P., DAMANI O. (2007). Hindi Generation from Interlingua. Actes de *Machine Translation Summit*. Copenhague, 421-428.