

De l'interrogation interlingue à la traduction automatique

**FLUHR Christian
CEA/LIST**

Résumé:

Cet article présente comment un système d'interrogation interlingue qui était à l'origine utilisé pour interroger en une langue un ensemble de documents rédigés dans des langues différentes va servir de base à une nouvelle approche de la traduction automatique.

On présente tout d'abord la technologie de l'interrogation interlingue. On montre ensuite comment on peut traduire d'un ensemble de langues vers une langue cible dans la mesure où celle-ci est largement représentée sur le web.

1. Introduction :

Le domaine de la traduction automatique a subi ces dernières années une évolution rapide. Le domaine de la traduction est né avec l'informatique et a subi de nombreux soubresauts. Toutefois, les différents systèmes proposés étaient basés sur une analyse linguistique du langage source, un transfert qui dépend du couple de langues et une génération en langage cible.

De l'interrogation interlingue à la traduction automatique

Une première évolution s'est produite par la création des mémoires de traductions. Cette technologie permet d'utiliser la connaissance de documents déjà traduits pour aider à traduire de nouveaux documents dans le même domaine. On aligne phrase à phrase les textes traduits. Pour traduire une nouvelle phrase, on compare celle-ci à l'ensemble des phrases dont on a stocké la traduction dans la mémoire de traduction. S'il y a une phrase identique, on a la traduction. Sinon, le traducteur peut s'inspirer des parties communes pour construire sa nouvelle traduction. Ces technologies sont particulièrement utilisées pour traduire des versions successives de documentations techniques. Cela permet de diminuer drastiquement les coûts de traduction et assure une cohérence de vocabulaire et de style entre versions et à l'intérieur d'une même version.

Les derniers développements portent sur des systèmes de traduction automatique statistiques qui se basent eux aussi sur la connaissance de textes traduits. En effet depuis l'avènement des mémoires de traduction la quantité de documents bilingues sur internet s'est accrue de manière très importantes. Cette masse devient suffisante pour certains couples de langues et pour certains domaines pour permettre une traduction automatique. Ces systèmes se basent sur les probabilités de corrélation entre des successions de mots dans des phrases de la langue source et des successions de mots dans les phrases correspondantes de la langue cible.

L'approche que nous proposons se base sur la disponibilité d'un dictionnaire bilingue qui permet de

De l'interrogation interlingue à la traduction automatique

traduire tout mot ou sous-structure de mots qui ne se traduit(uisent) pas mot à mot. Une technologie d'interrogation interlingue permet de trouver des phrases qui représentent le même sens ou au moins dont certaines parties représentent le sens de certaines parties de la phrase qui sert de critère d'interrogation.

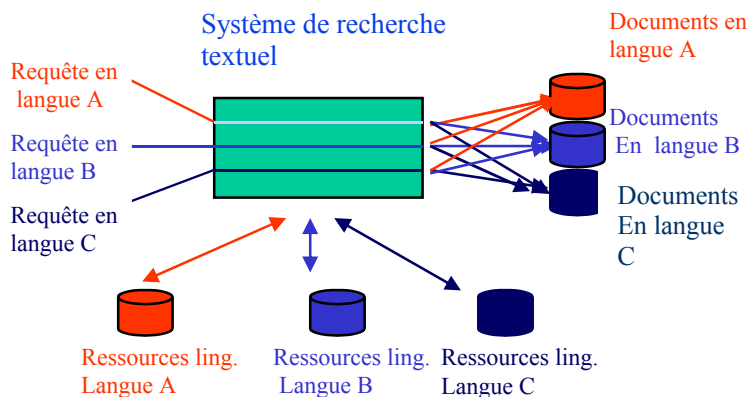
Ayant trouvé des exemples de traductions partielles de la phrase à traduire, il suffit à un traducteur humains de raccorder les morceaux de phrases traduites. Toutefois nous pensons qu'il sera possible de faire ce raccordement par des moyens automatiques.

Cette approche a l'avantage de ne demander que la connaissance de la langue cible. Les corpus en langue cible sont beaucoup plus nombreux que les corpus bilingues. La technologie employé par nous s'appuie sur une analyse syntaxique et est donc beaucoup plus inférente que la prise en compte des seules successions de mots tel qu'ont peut le trouver dans des traductions statistiques utilisant des n-grams. On va donc s'appuyer sur une indexation de la totalité du web de la langue cible (le français par exemple).

2. Interrogation interlingue :

Un système permettant une interrogation interlingue est un système multilingue qui peut traiter des documents dans plusieurs langues dans le même corpus et qui permet une interrogation de ces documents dans différentes langues.

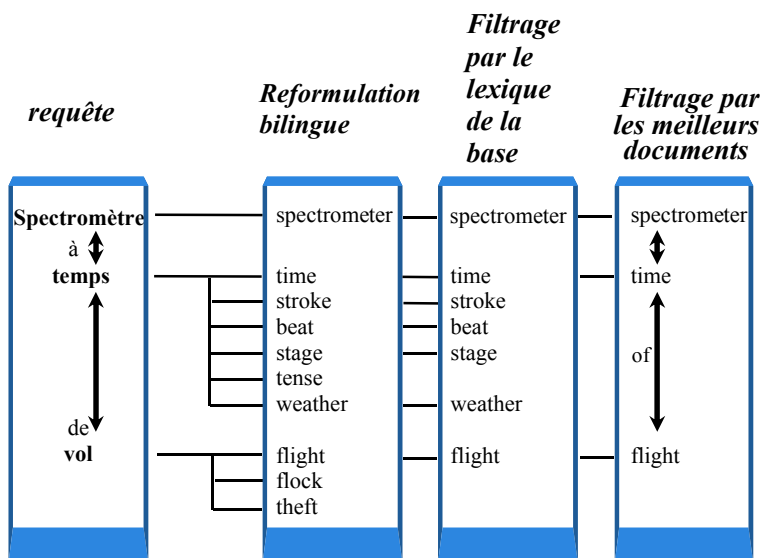
De l'interrogation interlingue à la traduction automatique



Après une analyse syntaxique de la langue de la requête chaque concept contenu dans celle-ci est traduit mot à mot ou globalement si la traduction mot à mot n'est pas possible.

Grace au dictionnaire bilingue toutes les traductions sont essayées. Certaines traductions sont éliminées parce qu'elles ne sont pas présentes dans la base. Ensuite, le document qui a la meilleure intersection sémantique (si celle-ci est suffisante) donne la bonne traduction. Bien entendu tout ceci suppose que la requête soit suffisamment longue pour que chaque mot ait un contexte qui détermine son sens.

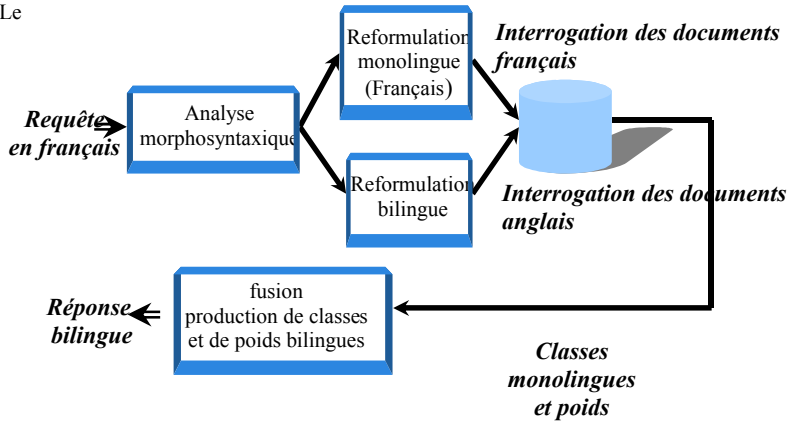
De l'interrogation interlingue à la traduction automatique



Le schéma ci-dessous montre comment se passe une interrogation en français sur une base contenant des documents en français et en anglais. Les traductions possibles de chaque concept sont d'abord filtrées par la base. Celle-ci est le catalogue des bibliothèques du CEA qui ne contient ni flock ni theft. Enfin dans les documents qui contiennent tous les concepts avec les bonnes relations de dépendance on trouve la bonne traduction.

De l'interrogation interlingue à la traduction automatique

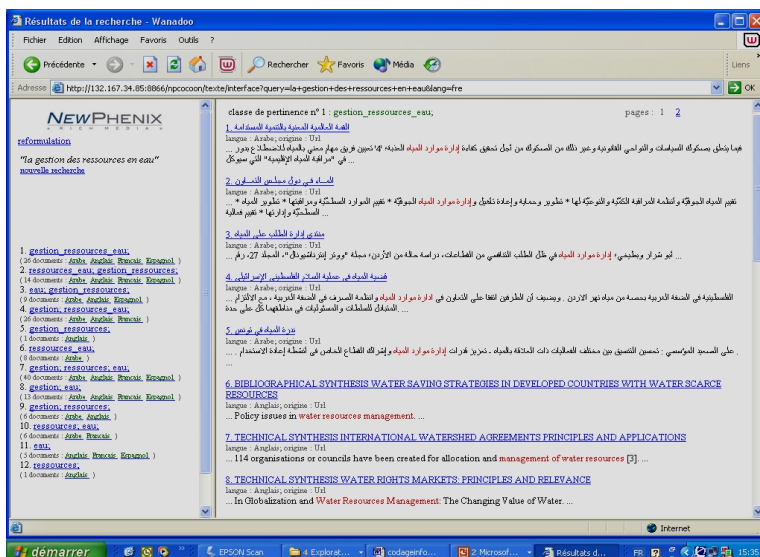
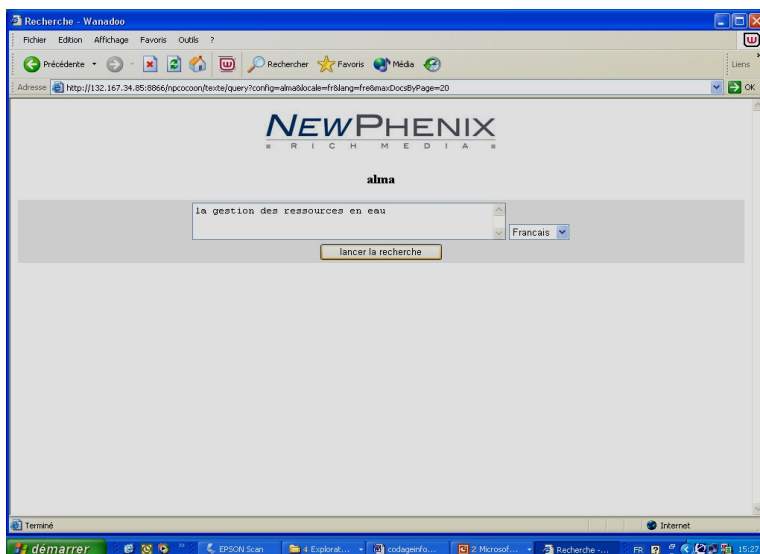
Le



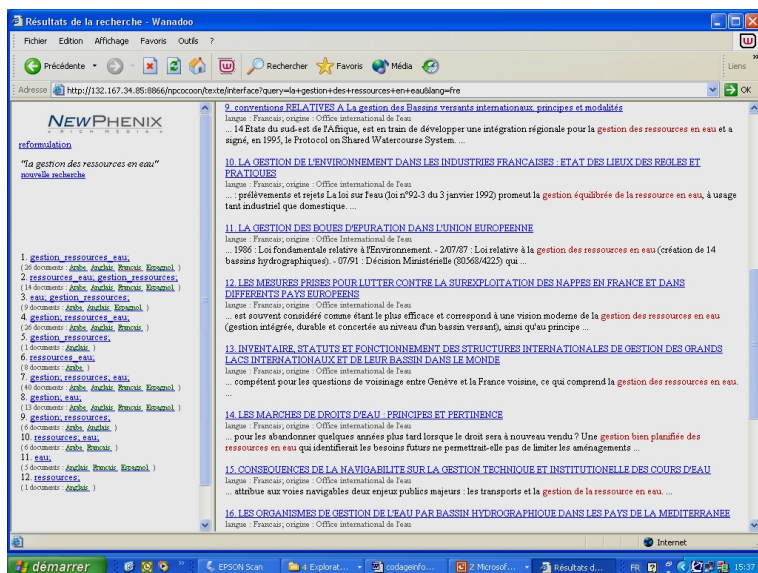
Le schéma précédent montre le fonctionnement de l'interrogation interlingue qui lance en parallèle des requêtes dans chaque langue et ensuite fusionne les résultats en s'appuyant sur les concepts communs.

Les copies d'écrans suivantes montrent un exemple d'interrogation. La requête est en français sur une base contenant des documents en français, anglais, arabe et espagnol.

De l'interrogation interlingue à la traduction automatique



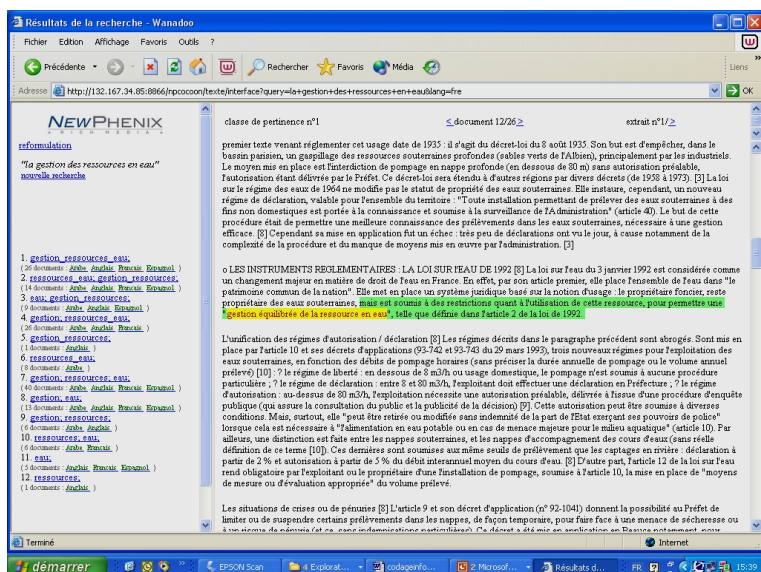
De l'interrogation interlingue à la traduction automatique



La comparaison entre requête et documents est structurale ce qui fait que le système identifie le concept de « gestion des ressources en eau » dans le texte « gestion bien planifiée des ressources en eau ».

La frame de gauche explique la nature de l'intersection sémantique qui justifie l'ordre de pertinence des documents.

De l'interrogation interlingue à la traduction automatique



Accès direct à l'information pertinente

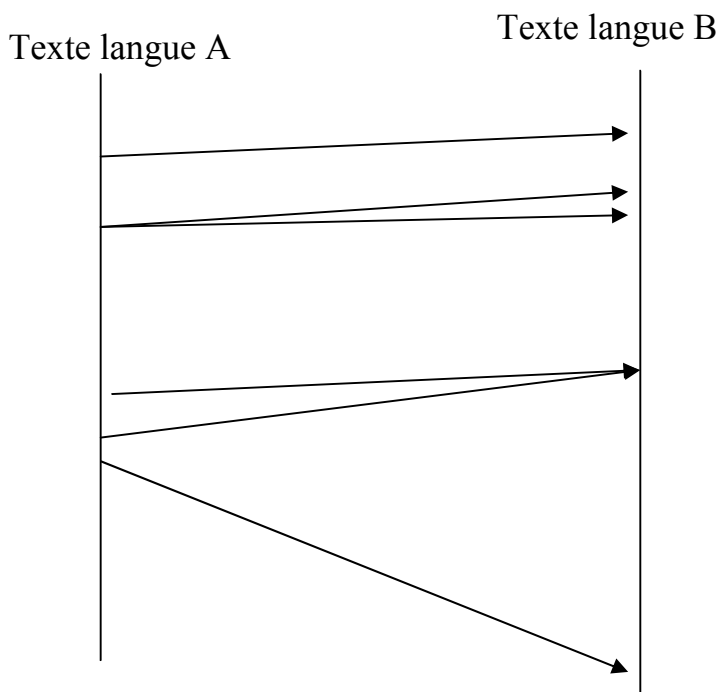
2. Alignement de textes bilingues :

L'alignement consiste à associer une phrase à sa traduction. En fait il n'y a pas toujours une association 1-1 mais à une phrase peut correspondre deux phrases ou plus dans l'autre langue. Ou pire encore à n phrases dans une langue correspondre m phrases dans l'autre.

Beaucoup de systèmes d'alignement utilisent un traitement séquentiel et des critères de longueur. Notre approche utilise nos technologies d'interrogation interlingue. En effet, si une phrase est présentée comme requête, le système va proposer les phrases qui sont le

De l'interrogation interlingue à la traduction automatique

plus proches sémantiquement. On utilisera ensuite les informations de position pour choisir parmi les phrases solutions celle qui a le plus de chance d'être bien placée pour être la traduction de la phrase requête.



Alignement de textes bilingues par utilisation de l'interrogation interlingue:

De l'interrogation interlingue à la traduction automatique

Génération d'une base de données pour chaque langue avec un document par phrase

On traite les phrases d'une langue les unes après les autres en cherchant par une interrogation interlingue les phrases de l'autre langue qui ont la plus grande intersection sémantique (en respectant dans la mesure du possible les relations de dépendances)

Pour les phrases les plus proches, on fait une interrogation interlingue inverse pour vérifier la correspondance.

On choisit parmi les meilleurs candidats celui qui optimise un critère de position (séquentialité des phrases traduites)

Cette technique s'est révélée la plus robuste dans les tests ARCADE et ARCADE II

Elle permet d'alimenter des systèmes de mémoires de traduction

Le système d'interrogation interlingue peut être utilisé en mode monolingue comme système de mémoire de traduction. Chaque phrase à traduire est confrontée aux phrases sources du corpus aligné. Les meilleures intersections sémantiques respectant la structure permettent d'accéder aux phrases cibles alignées auparavant.

Les textes bilingues alignés peuvent être aussi utilisés par les systèmes de traduction statistiques (Aachen, Google,...)

3. Traducteur basé sur un système de recherche interlingue et sur le web de la langue cible

Pourquoi aligner des textes bilingues en utilisant une interrogation interlingue et ensuite faire de la traduction statistique à partir des textes alignés alors qu'on peut directement présenter la phrase à traduire à l'interrogation interlingue et trouver la phrase traduite ou différentes parties de la phrase traduite permettant de la reconstituer.

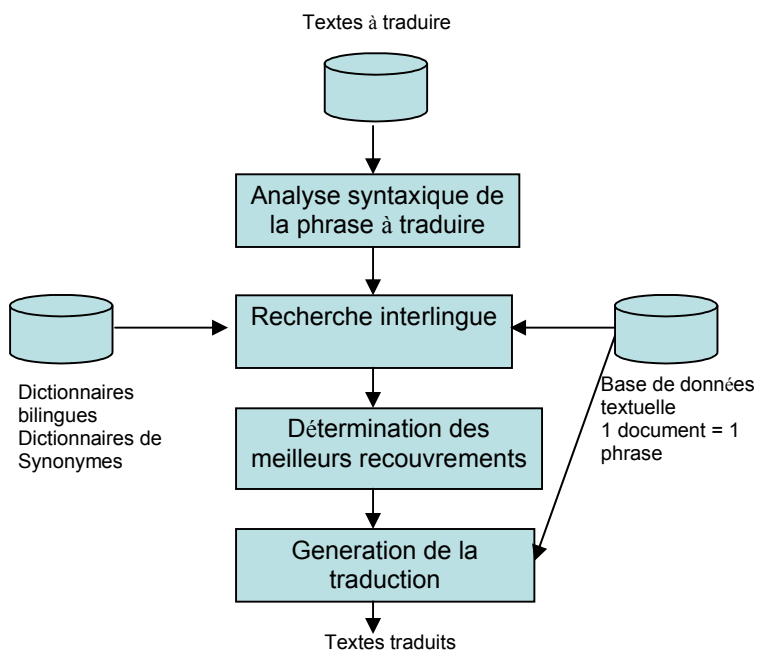
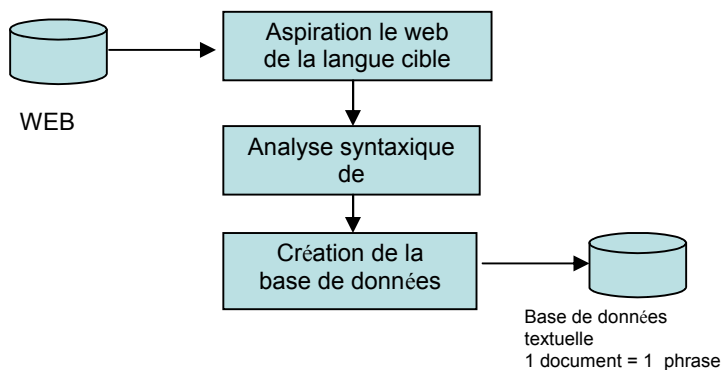
Cette approche s'inspire de la pratique des traducteurs professionnels qui utilisent de plus en plus l'interrogation sur le web pour valider les traductions.

Pour cela il « suffit » de faire une base de données de toutes les phrases de la langue cible qu'on peut trouver sur le web et dans d'autres textes que fournirait un utilisateur.

Cela donne un apprentissage infiniment plus grand que les seuls textes bilingues

Cela permet de traduire des langues plus rares pour lesquelles on ne dispose pas de beaucoup de textes bilingues. Il « suffit » dans ce cas de développer un analyseur linguistique.

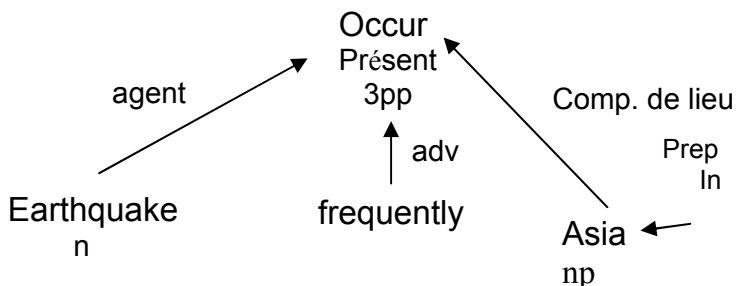
De l'interrogation interlingue à la traduction automatique



Exemple de traitement :

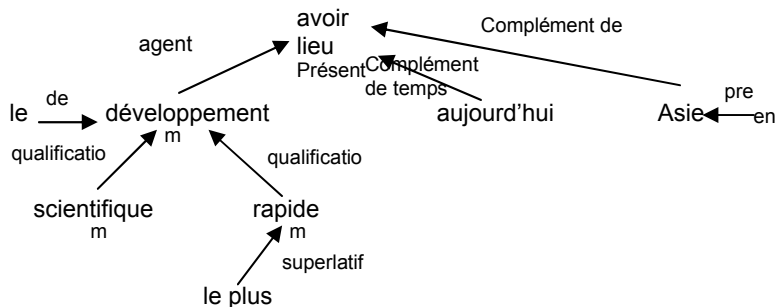
Phrase à traduire :

Earthquakes frequently occur in Asia



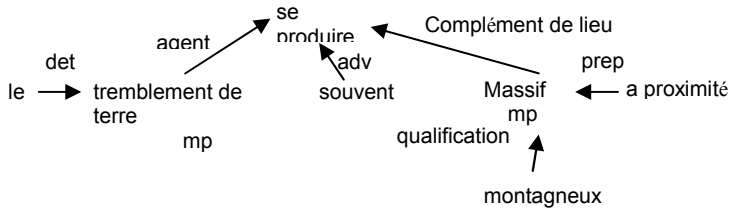
En donnant la phrase précédente au système d'interrogation interlingue, on peut récupérer les phrases suivantes qui ont une intersection sémantique partielle avec elle :

Le développement scientifique le plus rapide a aujourd'hui lieu en Asie



De l'interrogation interlingue à la traduction automatique

Les tremblements de terre se produisent souvent à proximité de massifs



Processus de recherche :

Interrogation en utilisant le dictionnaire bilingue

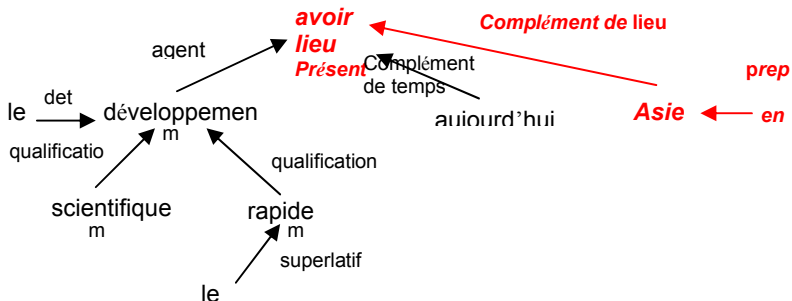
earthquake (subs) → tremblement de terre

occur (verbe) → arriver, intervenir, s'opérer, se dérouler, se produire, se présenter, se rencontrer, se trouver, survenir, avoir lieu

frequently (adv) → souvent, fréquemment

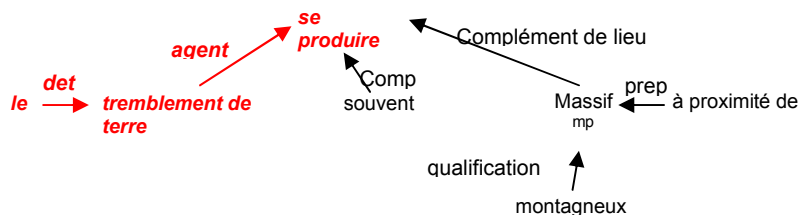
Asia (NP) → Asie

Le développement scientifique le plus rapide a aujourd'hui lieu en



De l'interrogation interlingue à la traduction automatique

Les tremblements de terre se produisent souvent à proximité de massifs montagneux



Les parties en rouge italique correspondent aux parties sémantiquement communes avec la phrase à traduire.

En recollant les parties traduites grâce à la synonymie entre « se produire » et « avoir lieu » on peut produire les phrases suivantes comme traduction de la phrase originale :

« les tremblements de terres se produisent souvent en Asie » ou « les tremblements de terres ont lieu souvent en Asie »

Tout cela suppose :

- Pour être le plus inférant possible d'assimiler la voie passive et active, de traiter les références des pronoms, éventuellement d'assimiler adjectif et complément de nom et prendre en compte les synonymes

- que les dictionnaires bilingues contiennent la plupart des traductions partielles qui ne peuvent être

traduites par parties. Il faut donc envisager d'utiliser les textes alignés existants pour améliorer les dictionnaires bilingues

- que l'on dispose d'une puissance machine suffisante

4. Conclusion :

La technologie décrite dans cet article a fait l'objet d'un dépôt de brevet. La technologie interlingue sur laquelle elle se base est aujourd'hui éprouvée. Toutefois, il reste à expérimenter la technologie en vraie grandeur. L'indexation par des technologies syntaxiques de la totalité du web d'une langue comme le français demande des moyens de calcul importants. Le laboratoire s'est doté d'un cluster de machines qui va permettre aussi bien d'indexer le web français que de réaliser l'interrogation interlingue sur la base des phrases du français.

Un prototype va être réalisé qui permettra de tester sur le couple anglais → français puis sur le couple arabe → français. D'autres couples de langues pourront ensuite être testées.

Références :

- FLUHR C., BISSON F., ELKATEB F., Mutual benefit of sentence/word alignment and crosslingual information retrieval, Chapitre dans "Parallel Text Processing", Editeur scientifique Jean Véronis, Kluwer Academic

Publishers, Text, speech and Language technologies series, 2000.

- BISSON F., FLUHR C., Sentence alignment in bilingual corpora based on crosslingual querying, Conférence RIAO2000, Paris, 12-14 avril 2000.

- LANGLAIS Ph. et SIMARD M., De la traduction probabiliste aux mémoires de traduction (ou l'inverse), TALN 2003, Batz-sur-Mer, 11–14 juin 2003

- FLUHR C., Systèmes multilingues, Recherche interlingue, Conférence CIDE'8, "Document électronique & multilinguisme", 25-28 mai 2005, Beyrouth, Liban.

- SIMARD M., CANCEDDA N., CAVESTRO B., DYMETMAN M., GAUSSIER E., Goutte C., LANGLAIS Ph., MAUSER A., YAMADA K., Traduction automatique statistique avec des segments discontinus, TALN 2005, Dourdan, 6–10 juin 2005