Constitution et exploitation d'un corpus parallèle issu du web pour l'extension d'une terminologie multilingue

Louise Deléger[1] Pierre Zweigenbaum[1,2,3]

[1]Inserm, U729 ; [2]AP-HP ; [3]Inalco, CRIM

Journée d'étude ATALA
Le Web comme ressource pour le TAL
11 mars 2006

Inserm
French Institute of Health and Medical Research

Introduction

- Contexte : projet VUMeF.
- Objectif : acquérir des traductions françaises de termes médicaux anglais pour enrichir les terminologies médicales.
- Méthode : alignement des mots d'un corpus médical parallèle français/anglais.

Inserm
French Institute of Health and Medical Research

Inserm
French Institute of Health and Medical Research

Inserm
French Institute of Health and Medical Research

Pages web parallèles

Santé Canada | Health Canada — Canada

English | Contactez-nous | Aide | Recherche | Site du Canada
Vie saine | Soins de santé | Maladies et affections | Protection de la santé | Salle des médias

Santé Canada
ON VOUS INFORME — En direct

Santé Canada Accueil
Communiqué
- Retour aux communiqués

Information
- Retour au réaction au rapport du vérificateur général

novembre 1999

Information

Réponse de Santé Canada au rapport du vérificateur général

Gestion des poussées d'intoxication alimentaire

Rôles et responsabilités

Quand il s'agit de salubrité des aliments, Santé Canada travaille en étroite collaboration avec l'Agence canadienne d'inspection des aliments (ACIA). Par le truchement du Programme des produits alimentaires, Santé Canada établit les politiques et les normes relatives à la sécurité et à la qualité nutritive des aliments vendus au Canada. Santé Canada a aussi pour mandat d'évaluer l'efficacité des activités de l'ACIA relatives à la sécurité des aliments.

Lorsque survient une poussée d'intoxication d'origine alimentaire, le Laboratoire de lutte contre la maladie de Santé Canada coopère avec l'ACIA et d'autres ordres de gouvernement aux activités d'enquête et de circonscription rapide de la poussée. L'ACIA est chargée de l'application des mesures se rapportant aux situations d'urgence liées aux aliments et joue un rôle prépondérant dans les

Health Canada | Santé Canada — Canada

Français | Contact Us | Help | Search | Canada Site
Healthy Living | Health Care | Diseases & Conditions | Health Protection | Media Room

Health Canada
KEEPING YOU INFORMED — Online

Health Canada Home
News Release
- Back to Release

November 1999

Health Canada's response to the Auditor General's Report

Managing Food-Borne Disease Outbreaks

Roles and Responsibilities

When it comes to food safety, Health Canada works closely with the Canadian Food Inspection Agency (CFIA). Through the Food Program, Health Canada establishes the policies and standards for the safety and nutritional quality of food sold in Canada. It is also responsible for assessing the effectiveness of CFIA's activities related to food safety.

In the event of a food-borne disease outbreak, Health Canada's Laboratory Centre for Disease Control collaborates with CFIA and other levels of government to investigate and control the outbreak in a timely way. The CFIA is responsible for enforcement actions in food-related emergencies and the Agency takes a lead role in investigations and coordination of food safety emergency responses. POI Provincial and local medical officers of health have a legislative mandate to investigate disease outbreaks, and provincial laboratories provide laboratory services

Repérage de pages parallèles

Quelques travaux sur ce thème :

- Strand (Resnik, 1998-1999 ; Resnik & Smith, 2003)
- PTMiner (Chen & Nie, 2000 ; Kraaij, Nie & Simard, 2003)
- BITS (Ma & Liberman, 1999)
- Paradocs (Patry & Langlais, 2005)

Nos travaux :

- Constitution du corpus parallèle Santé Canada (Papin, CRIM 2004 ; Deléger, CRIM 2005 ; Deléger et al., 2006)

Inserm
French Institute of Health and Medical Research

Méthode générale

Critères d'inclusion : collecte de couples de documents candidats

- Identification d'un site web / d'une page de départ
  - Où trouver une page / un site qui possède une traduction ?
- Identification de couples de pages candidats
  - Une page étant donnée, où se trouve sa traduction ?

Critères d'exclusion : filtrage des couples obtenus

- Un couple de pages étant donné, quels indices permettent de décider qu'elles sont en relation de traduction ?

En pratique, filtrages successifs : finalement, uniquement critères d'exclusion ?

Inserm
French Institute of Health and Medical Research

Indices de parallélisme

- Métainformations
  - Faire partie du même site (!)
  - Noms de fichiers (URL)
  - Liens entre documents (hyperliens)
    - page « parente », page « sœur »
- Être écrit dans deux langues différentes
- Similarité du contenu
  - Longueur des fichiers ; nombre de paragraphes
  - Similarité de la structure
    - Séquence des balises principales
    - Séquence des longueurs des phrases
  - Similarité des mots
    - En direct : « cognats »
    - À travers un lexique bilingue : proportion de mots traduisibles
  - Qualité de l'alignement des phrases

Inserm
French Institute of Health and Medical Research

Inserm
French Institute of Health and Medical Research

Constitution du corpus parallèle Santé Canada

- DESS d'ingénierie multilingue d'Annie Papin (2004)
  - Débroussaillage du sujet
  - Alignement des phrases
- DESS d'ingénierie multilingue de Louise Deléger (2005)
  - Alignement des mots
  - Extension d'une terminologie médicale bilingue

Inserm
French Institute of Health and Medical Research

Le site Santé Canada

- Site gouvernemental québequois
- Site quasi-entièrement bilingue français-anglais
- Plus de 100 000 pages selon Google début 2005
- Formé d'une série de « sous-sites »
  - Le parallélisme des noms n'est pas toujours simple

Inserm
French Institute of Health and Medical Research

Méthode générale

- Téléchargement du site « entier »
  - `wget -m -l10 -Dwww.hc-sc.gc.ca -T20 -t1 -w10 --user-agent=toto -X http ://www.hc-sc.gc.ca/`
  - `wget --force-directories --input-file=santecanada2.liens.manquants`
- Repérage de couples de pages HTML par leurs liens (pages « sœurs »)
  - Note : n'a pas exploité les documents PDF
- Vérifications :
  - langue, taille du texte
  - *noms de fichiers* (langue)
  - *bijectivité*
  - qualité de l'alignement des phrases

Inserm
French Institute of Health and Medical Research

Repérage de couples de pages : liens entre pages « sœurs »

- Structure extrêmement régulière :
  - Un lien sur chaque page pointe vers la page correspondante dans l'autre langue
- Ce lien est étiqueté par un texte ou une image indiquant la langue cible
  - « Français », « English »
  - `<a href="A"><img src="X" alt="Y"></a>`
  - `<a href="A">Z</a>`
  - avec *X* ou *Y* ou *Z* contient « Français » ou « English »
  - fournit l'URL *A* du document cible

```
<a href="../../../english/reports/2002/Acc_Ass_NCRC.htm">
  <img src="../../images/b-english.gif" width="90" height="19" border="0"
alt="English">
</a>
```

Repérage de couples de pages : implémentation

- Implémentation
  1. (Surface : expressions régulières en perl)
  2. Structure : module perl HTML : :Parser
  3. (voir plus bas : xslt)
- Le document cible est téléchargé s'il ne l'était pas encore
- Résultat : 12 549 paires de documents

Inserm
French Institute of Health and Medical Research

Filtrage des documents (1)

- Méta-informations : balise <meta> avec la langue du document ;

  *<meta name="dc.language" scheme="ISO639-2" content="**eng**">*

- Noms de fichiers : ne doivent pas contenir « français » ou « english ».

Filtrage des documents (2)

- Critères de cohérence globale (bijectivité) : suppression des documents en double et des documents à plusieurs correspondants ;
- Critères de taille : calcul du rapport entre les tailles de deux documents ;
- Critère spécifique au traitement à accomplir : évaluation de la qualité de l'alignement.

Inserm
French Institute of Health and Medical Research

Problèmes relevés

- Erreurs dans le site
  - Un lien « français » ne pointe pas toujours sur une page en français
    - → intérêt de vérifier la langue du document
  - Balises « meta » : une page étiquetée « français » n'est pas toujours en français
- Erreurs potentielles dans le site
  - Documents avec plusieurs correspondants (→ bijectivité)
  - Documents à la fois dans la liste anglaise et la liste française
- Pages « sœurs » non parallèles : index, glossaire

→ évaluation a posteriori de la qualité de l'alignement des phrases

Inserm
French Institute of Health and Medical Research

Inserm
French Institute of Health and Medical Research

Quantité et qualité

- 11 041 paires de documents ;
- 27,7 millions de mots ;
- grande quantité : traitements statistiques (calculs de co-occurrences) plus efficaces ;
- bruitage : fautes d'orthographes, espaces insérées ou oubliées... Dans quelle mesure cela affecte-t-il la tâche à accomplir ?

Inserm
French Institute of Health and Medical Research

Encodage

- windows 1252 ;
- problème de portabilité : caractères de position 128 à 160 ;

→ conversion en UTF-8.

Langue

- bilingue anglais/français ;
- français du Québec vs français de France :
  - vocabulaire parfois différent : *“poser des gestes”, “recherchistes” ;*
  - peut être considéré comme du bruit à la validation.

Public et Domaines

- **général** vs **spécialisé** : grand public et professionnels de la santé.
  - Piste : caractérisation du contenu.
- division en “sous-sites” traitant de divers sujets ;
  - Piste : faire émerger les grands pôles thématiques (cf Delbecque) pour obtenir des corpus plus homogènes.

Inserm
French Institute of Health and Medical Research

Inserm
French Institute of Health and Medical Research

Préparation du corpus : conversion au format texte

Méthodes successives

1. (Débalisage — par exemple, Perl)
   - Attention aux « contenus » qui n'en sont pas (scripts)
   - Difficile à faire proprement sur du HTML non normalisé
2. (Module Perl HTML : :FormatText)
   - Passe par une analyse de l'arbre d'éléments HTML : plus propre
   - Problèmes de codage de caractères, passage intempestif en utf-8
3. Passage par XML (cf Kraiij et al., 2003)
   - Normalisation et conversion en XHTML (`tidy`)
   - Extraction des parties utiles (transformation `xslt`)

Inserm
French Institute of Health and Medical Research

Questions restantes

- Quel marquage conserver ?
  - Paragraphes, titres, ancres ?
  - Faire en fonction des besoins des traitements suivants
    - Segmentation en phrases
    - Alignement des phrases
    - Alignement des mots
- Que faire de la « décoration » d'une page ?
  - En-tête
  - Pied de page
  - Menus en tout genre

Inserm
French Institute of Health and Medical Research

Préparation du corpus : segmentation en phrases

- nécessaire pour l'alignement en phrase ;
- programme PERL ;
- aidé par le marquage HTML : fin de paragraphe, balise <br>... constituent des fins de phrases ;

Inserm
French Institute of Health and Medical Research

Alignement : phrases

- alignement de phrases : GMA[1] (Dan Melamed) ;
  - conservation de marquage HTML : balises de paragraphes, titres et liens comme points de correspondance ;
  - évaluation de la qualité de l'alignement pour détecter des documents non parallèles : score à chaque alignement, élimination au-dessous d'un certain seuil ;

[1] http ://nlp.cs.nyu.edu/GMA/

Inserm
French Institute of Health and Medical Research

Alignement : mots

- alignement de mots :
  - pré-traitement :
    - étiquetage (Treetagger, français et anglais)
    - analyse syntaxique (Syntex, français et anglais) ;
  - I*Tools (Linköping, Suède) ;
  - alignement des mots simples et expressions complexes.

Inserm
French Institute of Health and Medical Research

Sélection : termes médicaux

Projection

- de termes médicaux anglais connus (entrées des terminologies MeSH, SNOMED)
- sur la partie anglaise des couples alignés (mots simples ou expressions complexes)

→ propositions de traductions françaises pour ces termes médicaux anglais

Inserm
French Institute of Health and Medical Research

Inserm
French Institute of Health and Medical Research

Résultats

- conversion en texte et alignement de phrases : corpus entier ;
- alignement de mots : 540 paires de documents ;
- 9860 termes médicaux anglais avec leur traduction française ;
- évaluation sur un échantillon de 145 traductions de termes MeSH :
  - 66 déjà connues et 79 nouvelles ;
  - soumission des nouvelles à un expert : 64 valides (**81 %**).

Inserm
French Institute of Health and Medical Research

Analyse des « nouvelles » traductions

Chaque terme MeSH anglais possède déjà au moins une traduction dans le MeSH français. L'alignement réalisé fournit :

– – des traductions d'un autre sens du terme anglais
– des variantes morphologiques (pluriel, féminin...)
\+ des variantes morphosyntaxiques (adjectif → SP)
++ des synonymes

Résultats : quelques exemples

Anglais	**Français**	**Terme(s) du MeSH français**	**Type**	**Valide**
reproduction rights	droits de reproduction	droits en matière de reproduction	*autre sens*	non
adipose tissue	tissus adipeux	tissu adipeux	variante morphologique	oui
health priorities	priorités sanitaires	priorités en santé	variante morphosyntaxique	oui
bone cancer	cancer des os	tumeur des os / tumeurs osseuses	synonyme	oui

Inserm
French Institute of Health and Medical Research

Inserm
French Institute of Health and Medical Research

Conclusion

- mise en place d'une stratégie d'acquisition de corpus combinant plusieurs méthodes (techniques d'inclusion et d'exclusion) ;
- acquisition de termes médicaux français non recensés dans le MeSH ;
- permet de mieux prendre en compte les usages ;
- perspectives : découper en sous-corpus homogènes selon ces usages
  - « vocabulaire patient » vs. vocabulaire « professionnel » ;
  - différents sous-domaines.

Inserm
French Institute of Health and Medical Research


Jiang Chen and J-Y. Nie.
Parallel web text mining for cross-language IR.
In *Proceedings of RIAO 2000 : Content-Based Multimedia Information Access*, volume 1, pages 62–78, Paris, France, April 2000. C.I.D.


Louise Deléger.
Alignement de mots dans un corpus parallèle pour l'enrichissement de la terminologie médicale.
DESS d'ingénierie multilingue, Institut National des Langues et Civilisations Orientales, 2005.


Alexandre Patry and Philippe Langlais.
Paradocs : un systèeme d'identification automatique de documents parallèles.
In Michèle Jardino, editor, *Proceedings of TALN 2005 (Traitement automatique des langues naturelles)*, pages 223–232, Dourdan, juin 2005. ATALA, LIMSI.


Philip Resnik and N.A. Smith.
The Web as a parallel corpus.
Computational Linguistics, 29 :349–380, 2003.
Special Issue on the Web as a Corpus.

Inserm
French Institute of Health and Medical Research