Introduction à la traduction automatique

Karën Fort

Karen.Fort[at]loria.fr
LORIA, Equipes Calligramme/TALARIS

Support de cours : sauvons la forêt amazonienne !

Le cours est au format pdf, sur ma page Web :
http ://www.loria.fr/~fortkare/

Avertissement : esprit critique es-tu là ?

J'ai fait de mon mieux pour citer mes sources, les croiser, vérifier les informations présentées, mais je suis loin d'être infaillible ou de tout savoir.
Je n'aurai sûrement pas toutes les réponses à vos questions, mais j'espère vous donner les moyens et l'envie de les chercher par vous-même.

Brève présentation

- Traductrice qui a mal tourné en découvrant le Traitement Automatique des Langues (TAL).
- Double compétence : presque linguiste, presque informaticien.
- 9 ans d'expérience en TAL, principalement en tant que gestionnaire de ressources multilingues.
- Actuellement ingénieur spécialiste au LORIA (Nancy).

Et vous ?

- Vous venez d'où (Master actuel, formation antérieure) ?

Et vous ?

- Vous venez d'où (Master actuel, formation antérieure) ?
- Projet sur lequel vous travaillez ?

Et vous ?

- Vous venez d'où (Master actuel, formation antérieure) ?
- Projet sur lequel vous travaillez ?
- Que connaissez-vous de la traduction automatique ?

Et vous ?

- Vous venez d'où (Master actuel, formation antérieure) ?
- Projet sur lequel vous travaillez ?
- Que connaissez-vous de la traduction automatique ?
- Que connaissez-vous de la traduction assistée par ordinateur ?

Et vous ?

- Vous venez d'où (Master actuel, formation antérieure) ?
- Projet sur lequel vous travaillez ?
- Que connaissez-vous de la traduction automatique ?
- Que connaissez-vous de la traduction assistée par ordinateur ?
- Que connaissez-vous des autres applications en traitement automatique de la langue ?

Et vous ?

- Vous venez d'où (Master actuel, formation antérieure) ?
- Projet sur lequel vous travaillez ?
- Que connaissez-vous de la traduction automatique ?
- Que connaissez-vous de la traduction assistée par ordinateur ?
- Que connaissez-vous des autres applications en traitement automatique de la langue ?
- Qu'attendez-vous de moi ?

Plan

- Bref historique de la traduction automatique
- Principales approches :
  - systèmes à base de règles
  - systèmes basés sur des données

Plan

- Bref historique de la traduction automatique
- Principales approches :
  - systèmes à base de règles
  - systèmes basés sur des données

1946-1949 : les prémisses

- 1946 : premières calculatrices électroniques
- 1946 : A. Booth, életronicien, demande des fonds à W. Weaver, mathématicien et vice-président de la fondation Rockefeller pour construire le premier ordinateur britannique
- 1948 : premières expériences de Booth et Weaver
- 1949 : se basant sur les méthodes de décryptage des codes secrets employées durant la deuxième guerre mondiale, Booth et Weaver suggèrent qu'il serait possible de traduire automatiquement grâce à l'ordinateur (memorandum de Weaver) :

"I have a text in front of me which is written in Russian but I am going to pretend that it is really written in English and that it has been coded in some strange symbols. All I need to do is strip off the code in order to retrieve the information contained in the text."

1954-1960 : la période faste

- Janvier 1954 : des chercheurs de l'Université de Georgetown et d'IBM démontrent la faisabilité de la traduction automatique (TA) en traduisant une soixantaine de phrases russes en anglais (vocabulaire de 250 mots, six règles de syntaxe).
- S'ensuit une période faste pour la recherche en TA, financée notamment par les militaires et les services de renseignements.
- Les soviétiques (1956), les Japonais (1956), les Chinois (1958-59), les Italiens (1959), les Français (1959) et les Belges (1961) s'engagent dans des recherches.
- La plupart des idées qui présideront au développement du traitement automatique de la langue (TAL) datent de cette première période : méthodes probabilistes, langues intermédiaires sémantiques, méthodes sur corpus, dictionnaires électroniques, ...

1960-1966 : l'analyse syntaxique

- 1959 : le philosophe-mathématicien-linguiste Bar-Hillel, précurseur de la TA, affirme dans un rapport qu'une traduction totalement automatique de qualité (FAHQMT) est impossible, non seulement techniquement, mais sur le principe même.
- A partir de 1960, c'est l'analyse syntaxique qui est mise en avant comme seule voie possible pour la TA (grammaire catégorielle de Bar-Hillel, grammaire générative de Chomsky).

1966 : le rapport ALPAC

Une commission (pas tout à fait) indépendante (Automatic Language Processing Advisory Committee) statue que la TA est plus chère, plus lente et moins bonne que la traduction humaine, et que la recherche a peu de chances de conduire à des résultats satisfaisants.
La linguistique computationnelle tire son épingle du jeu...

La survie de la recherche en TA s'organise...

- France : le CETA/GETA poursuit le développement d'Ariane (1971), grâce au CNRS
- Allemagne : Susy est élaborée à l'Université de Sarrebruck
- USA : les Mormons travaillent sur la traduction automatique de la bible (Weidner, ALPS)
- Canada : des chercheurs de l'Université de Montréal mettent au point TAUM-METEO (1975) un système spécialisé pour traduire les rapports météorologiques du ministère canadien de l'environnement (sous-langage).
- Europe : Eurotra (1977-1994), système de traduction multilingue pour la CE

... alors que la “force brute” se développe

- 1969 : un système de traduction entre le russe et l’anglais est mis à l’essai dans les quartiers généraux de la US Air Force.
- En 1975, une version anglais-français du système est mise au service de la communauté européenne.
- Depuis 1997, Systran alimente le service BabelFish et autres portails sur le Web.

1980-1990 : le “tournant” japonais

- Projet “5e génération“ soutenu par le MITI
- Développement de la TAO (traduction assistée par ordinateur)
- Développement de la TA basée sur l’exemple : système de TA qui “apprendrait“ à traduire à partir d’exemples.

Depuis 1990 : le renouveau

- 1989 : Les laboratoires d'IBM commencent à mettre au point un système de traduction "statistique" (Candide). Ce système ne repose sur aucune connaissance linguistique a priori. Il se nourrit exclusivement de (grandes quantités de) traductions existantes.
- 2002 : Language Weaver est la première entreprise privée à offrir une technologie de traduction automatique statistique.
- Depuis une dizaine d'années, les méthodes statistiques dominent la recherche. . . mais on assiste à un retour en douce des approches syntaxiques...

Plan

- Bref historique de la traduction automatique
- Principales approches :
  - systèmes à base de règles
  - systèmes basés sur des données

Quelles différences ?

- systèmes à base de règles :

Quelles différences ?

- systèmes à base de règles :
  - linguistes

Quelles différences ?

- systèmes à base de règles :
  - linguistes
  - ressources

Quelles différences ?

- systèmes à base de règles :
  - linguistes
  - ressources
  - temps

Quelles différences ?

- systèmes à base de règles :
  - linguistes
  - ressources
  - temps
- systèmes basés sur des données :

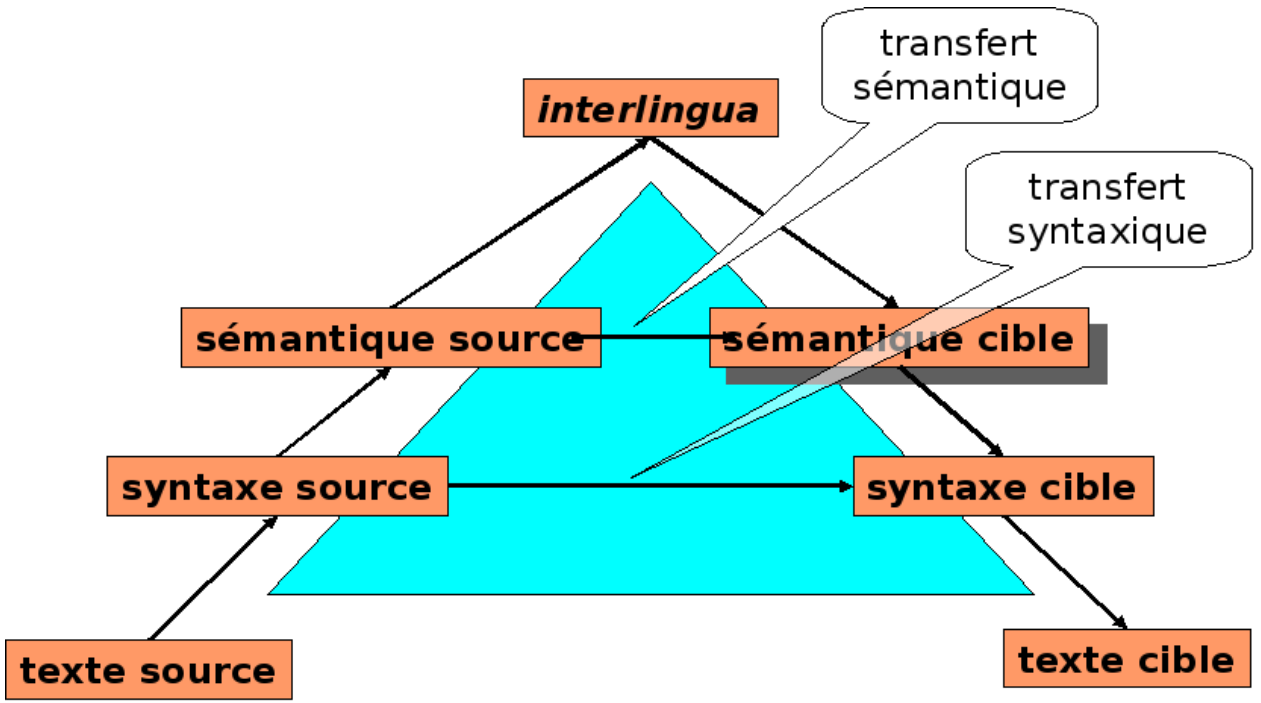
Quelles différences ?

- systèmes à base de règles :
  - linguistes
  - ressources
  - temps
- systèmes basés sur des données :
  - corpus bilingue parallèle

Quelles différences ?

- systèmes à base de règles :
  - linguistes
  - ressources
  - temps
- systèmes basés sur des données :
  - corpus bilingue parallèle
  - méthode d'apprentissage

Un exemple de TA à base de règles



Etapes de l'analyse : découpage en "mots"

- l'arbre

Etapes de l'analyse : découpage en "mots"

- l'arbre
- aujourd'hui

Etapes de l'analyse : découpage en "mots"

- l'arbre
- aujourd'hui

$\Rightarrow$ tokenization

Etapes de l'analyse : analyse des "mots"

- porte +Nf + Sg

Etapes de l'analyse : analyse des "mots"

- porte +Nf + Sg
- porte +VT + 1/3P + Sg

Etapes de l'analyse : analyse des "mots"

- porte +Nf + Sg
- porte +VT + 1/3P + Sg

$\Rightarrow$ analyse morphologique

Etapes de l'analyse : analyse des "mots" dans la phrase

- Jean regarde un homme sur la colline avec un télescope.

Etapes de l'analyse : analyse des "mots" dans la phrase

- Jean regarde un homme sur la colline avec un télescope.
- *Qui est sur la colline ?*

Etapes de l'analyse : analyse des "mots" dans la phrase

- Jean regarde un homme sur la colline avec un télescope.
- *Qui est sur la colline ?*
- *Qui a un télescope ?*

Etapes de l'analyse : analyse des "mots" dans la phrase

- Jean regarde un homme sur la colline avec un télescope.
- *Qui est sur la colline ?*
- *Qui a un télescope ?*

$\Rightarrow$ analyse syntaxique

Etapes de l'analyse : analyse des "mots" dans la phrase

- Tous les hommes aiment une femme.

Etapes de l'analyse : analyse des "mots" dans la phrase

- Tous les hommes aiment une femme.
- *Chaque homme aime une femme ou tous les hommes aiment la même femme ?*

Etapes de l'analyse : analyse des "mots" dans la phrase

- Tous les hommes aiment une femme.
- *Chaque homme aime une femme ou tous les hommes aiment la même femme ?*

$\Rightarrow$ analyse sémantique

Etapes de l'analyse : le mythe de l'interlingua

- troisième langue qui relie la langue source à la langue cible

Etapes de l'analyse : le mythe de l'interlingua

- troisième langue qui relie la langue source à la langue cible
- exemple : UNL (Universal networking Language)

Etapes de l'analyse : le mythe de l'interlingua

- troisième langue qui relie la langue source à la langue cible
- exemple : UNL (Universal networking Language)
- représentation abstraite universelle valable pour toutes les langues ? ?

Etapes de l'analyse : le mythe de l'interlingua

- troisième langue qui relie la langue source à la langue cible
- exemple : UNL (Universal networking Language)
- représentation abstraite universelle valable pour toutes les langues ? ?

$\Rightarrow$ transfert lexical et adaptation de la structure

Conclusion sur les systèmes à base de règles

- Les systèmes à base de règles incorporent des connaissances linguistiques approfondies.
- Ils requièrent peu de ressources informatiques (comparés aux méthodes statistiques).
- Ils peuvent traduire au niveau du paragraphe, voire de la page (Ariane-G5).
- MAIS
- Ils sont fragiles
- dispendieux à transférer à d'autres domaines ou paires de langues
- génèrent typiquement une seule traduction par phrase

Plan

- Bref historique de la traduction automatique
- Principales approches :
  - systèmes à base de règles
  - systèmes basés sur des données

Deux types de systèmes basés sur des données

- Systèmes statistiques purs
- Traduction par l'exemple

Les systèmes basés sur des données

Hypothèse : il n'y a pas de "bonne" réponse...

- S : le chat pourchasse la souris
- T1 : the cat chases the mouse around P = 0.22
- T2 : the cat is running after the mouse P = 0.08
- ...
- Tn : I will not buy this record, it is scratched P = 0.0000000001

...mais certaines réponses sont plus probables que d'autres !

Les systèmes statistiques purs

- Basés sur une théorie mathématique (Jelinek, Brown).

Les systèmes statistiques purs

- Basés sur une théorie mathématique (Jelinek, Brown).
- Modèle probabiliste de traduction à partir d'un texte bilingue.

Les systèmes statistiques purs

- Basés sur une théorie mathématique (Jelinek, Brown).
- Modèle probabiliste de traduction à partir d'un texte bilingue.
- Modèle probabiliste de la langue cible à partir d'un texte monolingue.

Les systèmes statistiques purs

- Basés sur une théorie mathématique (Jelinek, Brown).
- Modèle probabiliste de traduction à partir d'un texte bilingue.
- Modèle probabiliste de la langue cible à partir d'un texte monolingue.
- Traduction cible générée à partir de traduction(s) de mots individuels.

La traduction par l'exemple

- La phrase est l'unité de traduction.

La traduction par l'exemple

- La phrase est l'unité de traduction.
- Recherche des meilleurs exemples de ref. dans une base, puis adaptation.

La traduction par l'exemple

- La phrase est l'unité de traduction.
- Recherche des meilleurs exemples de ref. dans une base, puis adaptation.
- Possibilité d'ajout de règles (système hybride).

Conclusion sur les systèmes basés sur des données

- Faciles à entretenir.
- Faciles à adapter à de nouveaux domaines ou paires de langues – dans la mesure où des données sont disponibles.
- Pour une phrase source, peuvent produire plusieurs traductions, avec une mesure de confiance.
- MAIS
- Nécessitent des ressources informatiques lourdes (processus "gourmands").
- Difficiles à faire évoluer.

Remarques sur Systran et Reverso (?)

- Jean Véronis (01/2006) : http ://aixtal.blogspot.com/2006/01/traduction-systran-ou-reverso.html
- Systran : gros dicos + règles simples
- Reverso : idem mais intervention possible (?)

Les langages contrôlés

- Boeing : Simplified Technical English (STE).
- Dassault Aerospace : Français Rationalisé.
- Caterpillar : Caterpillar Technical English (CTE), Caterpillar Fundamental English (CFE).
- Nortel : Nortel Standard English (NSE).
- Scania : Scania Swedish.
- Sun Microsystems : Sun Controlled English.
- Xerox : Xerox Multilingual Customized English

De la TA à la TAO

- Utilisation de mémoires de traductions :
  - Sous Windows (payantes) : Trados Workbench, DéjàVuX, SDLX, Star Transit, Similis, etc
  - Multiplateformes (gratuites) : OmegaT, Open Language Tools

$\Rightarrow$ Traduction Assistée par Ordinateur (TAO)

De la TA au TAL

- Détection de la langue
- Fouille de texte
- Aide au terminologue
- Aide à la rédaction
- Moteur de recherche
- etc !

⇒ http ://rali.iro.umontreal.ca/

Quelques définitions (G. Perrier)

- morphologie : concerne la combinaison des signes minimaux d'une langue, ses morphèmes, pour former des mots.
- syntaxe : touche à la combinaison des mots pour former des phrases.
- sémantique : touche au sens des énoncés.

Je leur ai tout piqué !

- John Chandioux : http ://w3.gril.univ-tlse2.fr/TAL/TRAD/TRADAUTO1.htm.
- Michel Simard, du Conseil national de recherche du Canada (Technologies langagières interactives) : La traduction automatique et vous...
- Philippe Langlais, RALI, Université de Montréal. (http ://www.iro.umontreal.ca/ felipe/IFT6010-Automne2006/.
- Jacqueline Léon : Le CNRS et les débuts de la traduction automatique en France.
- Guy Perrier, Professeur à Nancy II : définitions.
- Machine translation : An Introductory Guide (http ://www.essex.ac.uk/linguistics/clmt/MTbook/).

Copyright et al

- Ce cours a été réalisé en LaTex Beamer.
- Il est disponible sous licence Creative Commons.