NRC·CNRC

Institute for Information Technology

Pourquoi la traduction automatique est-elle si difficile?

Pierre Isabelle

Septembre 2010

National Research Council Canada
Conseil national de recherches Canada
Canada

NRC·CNRC
Institute for Information Technology

Traduction automatique

- Désigne tout processus par lequel une machine exécute une opération de traduction entre des textes ou des énoncés de deux langues (humaines et naturelles) différentes

  La traduction automatique coûte moins cher que la traduction humaine.

- Peut aussi désigner le résultat d'une tel processus:

  La lecture de traductions automatiques est souvent moins plaisante que celle de traductions humaines

  Pour la réflexion: une mauvaise traduction machine mérite-t-elle de s'appeler une « traduction »?

- Finalement, le terme désigne aussi l'études des méthodes et techniques qui rendent une machine capable de produire des traductions (meilleures)

  C'est ce dernier sens qui nous intéresse ici.

NRC·CNRC
Institute for Information Technology

La traduction automatique est difficile

- Dans les années 1950, certains ont cru que l'arrivée de systèmes de TA capables de bien traduire les textes techniques était imminente (« 5 ans tout au plus… »)
- Pourtant, après 55 ans d'efforts de recherche soutenus, la TA demeure encore largement confinée aux tâches de « gisting »; les traductions machine sont généralement jugées peu utiles par les grands producteurs de documents (grandes sociétés, gouvernements, etc.)
- L'optimisme exagéré des années 1950 reposait en bonne partie sur une méconnaissance de la complexité du problème

  On est parti d'une conception simpliste de la tâche à accomplir

NRC·CNRC
Institute for Information Technology

Une conception simpliste de la TA

Facile: il suffit de remplacer chaque mot du texte à traduire par l'équivalent que nous donne un dictionnaire bilingue!

- Bien sûr que non! Les choses sont très loin d'être aussi simples!
- Nous allons examiner 12 objections de plus en plus sérieuses à cette conception simpliste

NRC·CNRC
Institute for Information Technology

12 objections à la conception simpliste

1. Problèmes de segmentation du texte en mots
2. La morphologie: mot-dictionnaire (« lemme ») versus mot du texte (« forme »)
3. Les expressions idiomatiques
4. Les collocations
5. Les mots inconnus
6. L'ambiguïté lexicale dans la langue de départ
7. L'ambiguité de co-référence dans la langue de départ
8. Les divergences de structure entre les deux langues
9. Les différences d'explicitation entre les deux langues
10. L'insertion et la suppression de mots
11. La grammaire de la langue d'arrivée
12. L'ordre des mots

NRC·CNRC
Institute for Information Technology

Problème 1: Segmentation en mots

- Langues européennes: les mots sont en général séparés par des espaces
  - Mais pas toujours, e.g. les ponctuations sont souvent attachées
  - L'existence d'abréviations terminées par un « . » produit souvent des ambiguités:

    « in. » → abréviation de *inch*

    → préposition *in + point*

- Langues asiatiques: pas d'espaces → ambiguïté systématique. Imaginer en anglais:

Themanywaystowelcome…

→ Them any ways towel come…
→ Them any ways to welcome…
→ The many ways towel come…
→ The many ways to welcome…

NRC·CNRC
Institute for Information Technology

La segmentation en mots

- En conséquence…
- Un système de traduction automatique doit posséder une capacité à segmenter automatiquement le texte d'entrée en mots
- En général beaucoup de segmentations différentes sont possibles, mais une seule est correcte relativement aux intentions de l'auteur du texte
- Dans l'exemple précédent, on choisit :
  - La segmentation qui donne un phrase grammaticalement OK?
    - Oui, mais en général il peut y en avoir plusieurs
  - La segmentation la plus probable compte tenu de la grammaire, de la sémantique, du contexte, etc.

NRC·CNRC
Institute for Information Technology

Problème 2: La morphologie

- Les dictionnaires rédigés à l'intention des humains font largement abstraction des phénomènes morphologiques
  - Flexion :
    - pluriel des noms/adjectifs; conjugaison des verbes; cas (il/le/lui)
    - Ambiguïtés: leaves = *NC:leaf+S* Vs *VB:leave+S*
  - Dérivation :
    - *VB:perturber* → *AJ:pertub+able* → *AJ:im+perturbable* → *AV:imperturbable+ment*
    - Ambiguïtés: repair = *VB:repair* ('fix') Vs *re+VB:pair* ('pair again')
  - Composition :
    - En anglais/français, souvent présegmenté : *twenty-story building*
    - Mais en allemand, ambiguïtés de segmentation: Staubecken = *Stau+becken* ('reservoir') Vs *Staub+ecken* ('dusty corners')

NRC·CNRC
Institute for Information Technology

La morphologie (suite)

- En conséquence…
- Un système de TA fait souvent appel à une description explicite des phénomènes morphologiques
  - P.ex. système de règles avec exceptions comme le *Bescherelle* des verbes français
- Autrement le système de TA doit contenir une entrée séparée pour chaque forme de chaque mot
  - Beaucoup de formes différentes
  - En français, seulement pour la flexion, il faut multiplier par 7
    - Dictionnaire de 50 000 lemmes équivaut à 350 000 formes de mots différentes (pluriels, féminins, conjugaisons, etc)

NRC·CNRC
Institute for Information Technology

Problème 3: Les expressions idiomatiques

- Certaines entrées d'un dictionnaire ordinaire sont réalisées par plusieurs mots d'un texte; pas nécessairement adjacents.

  John gave up the game. → *John abandonna la partie.*

  John gave it up years ago → *Jean abandonna cela il y a des années.*

- La même suite de mots reçoit tantôt une interprétation idiomatique, tantôt une interprétation littérale:

  Jim would be prepared to give up to $200 for a copy of that book.

  → Ici « *give up* » ne reçoit pas d'interprétation idiomatique (mais « *up to* » en reçoit une)

  James kicked the bucket.

  → Deux interprétations possibles.

NRC·CNRC
Institute for Information Technology

Les expressions idiomatiques

- En conséquence…
- Un système de TA doit être capable d'enregistrer des correspondances
  - Non seulement entre des paires de mots
  - Mais plus généralement entre des paires de suites de mots

    out to lunch → dans les patates
- En outre, le système devra savoir choisir entre une interprétation littérale ou idiomatique d'une suite de mots particulière

  My boss is out to lunch → Mon patron est sorti déjeuner

  → Mon patron est dans les patates

  Ce choix devrait en principe reposer sur un examen du contexte sémantique / pragmatique

NRC·CNRC
Institute for Information Technology

Problème 4: Les collocations

- Problème apparenté à celui des expressions idiomatiques
- La traduction de certains mots dépend souvent fortement des mots avoisinants
  - *dead serious* ≠ *mort sérieux*
    = *absolument sérieux*
  - *heavy smoker* ≠ *fumeur lourd*
    = *gros fumeur*
- En conséquence, un système de TA doit souvent enregistrer un grand nombre de traductions différentes d'un modificateur particulier (p.ex. adjectif ou adverbe) selon l'élément modifié.

NRC·CNRC
Institute for Information Technology

Problème 5: Les mots inconnus

- Un système de TA rencontre souvent des mots qui ne sont pas dans son dictionnaire
  - Le dictionnaire est parfois très incomplet.
  - Même avec un dictionnaire plus complet: nouveaux noms propres, emprunts à une langue étrangère, néologismes, etc.
  - Fautes d'orthographe dans le texte à traduire.

- La présence de mots inconnus compromet sérieusement la capacité du système à traiter les phrases contenant ces mots
  - p.ex. ceci peut faire échouer une tentative de produire une analyse grammaticale de la phrase

NRC·CNRC
Institute for Information Technology

Les mots inconnus

- Les lecteurs humains possèdent une grande capacité à :
  - Reconnaître et corriger les fautes d'orthographe.
  - Deviner le rôle grammatical et le sens probable d'un mot inconnu.
- Un système de TA idéal devrait être capable de détecter les erreurs orthographiques et de les corriger (de manière virtuelle)
  - Il y a des cas faciles mais le problème général s'avère très difficile.
- Un système de TA idéal devrait aussi être capable de deviner le rôle grammatical et le sens probable d'un mot inconnu
  - Nécessaire pour traduire correctement le reste de la phrase.
- En pratique, les système de TA traitent se contentent souvent de supposer que les mots inconnus sont des noms propres
  - Truc facile, mais qui ne marche pas toujours…

NRC·CNRC
Institute for Information Technology

Problème 6: L'ambiguïté lexicale

- Au-delà de la segmentation en mots et de l'analyse morphologique, les mots d'un texte demeurent ambigus relativement aux entrées du dictionnaire.
- Ambiguïté catégorielle :
  - light, ADJ → léger ; light, N → lumière ; light, VB → allumer
- Polysémie/homographie : même catégorie, plusieurs sens
  - « Clean filter and replace it. » Vs « Discard old filter and replace it. »
    replace$_1$ → remettre en place ; *replace$_2$ → remplacer*
  - « He took his pen and wrote a poem. » Vs « Place the baby in his pen! »
    pen$_1$ → plume; *pen$_2$ → parc (d'enfant)*

NRC·CNRC
Institute for Information Technology

L'ambiguïté lexicale

- En conséquence…
- Un système de TA doit pouvoir résoudre les ambiguïtés lexicales
- Ambiguité catégorielle: analyse syntaxique de surface (« tagging ») arrive à résoudre 90-95% des cas (mais pas plus…)
- Polysémie/homographie:
  - Difficile d'imaginer des méthodes par règles réalisables. Vers 1960, Y. Bar-Hillel en tire un argument général contre la faisabilité de la TA de haute qualité.
  - En pratique, des techniques statistiques relativement simples d'association lexicale (e.g. pen: write Vs. baby) permettent de résoudre une forte proportion (~ 90%) des cas
  - Solution plus complète paraît hors d'atteinte

NRC·CNRC
Institute for Information Technology

Problème 7: La co-référence

- L'interprétation et la traduction des pronoms est souvent déterminée par leur co-référence avec un syntagme précédent ou suivant.

  Squirrels like peanuts. They are good for them.

  → Les écureuils aiment les arachides. { Elles sont bonnes pour eux. / Ils sont bons pour elles. }

  The authorities refused to grant the women a permit to demonstrate because they (feared | advocated) violence.

  → (ils | elles) selon le verbe choisi dans la subordonnée

- Le lecteur humain sélectionne instantanément l'interprétation la plus probable…)
  - Exploitation instantanée, souvent *inconsciente*, de sa connaissance des situations probables dans le monde réel.

NRC·CNRC
Institute for Information Technology

La co-référence

- En conséquence…
- Pour garantir la traduction correcte d'un pronoms, un système de TA doit en principe être capable de déterminer les relations de co-référence qu'entretiennent les pronoms dans le texte en langue de départ (pas nécessairement dans la même phrase)
- Cette capacité suppose une connaissance détaillée du monde réel : en particulier, quelles sont les situations probables ou improbables dans le monde
- Une telle capacité est présentement hors d'atteinte
- Mais on peut résoudre un bon pourcentage de cas avec des « trucs » (sans garanties!)

NRC·CNRC
Institute for Information Technology

Problème 8: Les divergences structurales

- Il est souvent impossible de traduire un mot particulier sans procéder à des ajustements de la structure grammaticale environnante
- L'existence de « trous lexicaux » dans l'une ou l'autre langue impose un changement avec effet domino
- Exemple 1: pas de nom équivalent à « *lever* » en anglais:

  Dès son lever, Max mange un croissant.

  → As soon as he gets up, Max eats a croissant.
- Exemple 2: pas de verbe équivalent à « *faillir* » en anglais:

  Anne faillit s'étrangler.

  → Anne almost choked.
- Exemple 3: Pas de construction de type « *swim across* » en français:

  Judy swam across the river.

  → Judy traversa la rivière à la nage.

NRC·CNRC
Institute for Information Technology

Les divergences structurales

- On doit souvent traduire non pas des mots mais plutôt des *constructions grammaticales*:

  swim across X → traverser X à la nage

- En conséquence…
- Un système de TA devrait être capable de:
  1. Produire une analyse grammaticale de la phrase à traduire; et
  2. Produire une traduction possédant une structure grammaticale équivalente en langue d'arrivée
- L'analyse grammaticale (« parsage ») constitue un problème très difficile, notamment à cause des *ambiguités de structure syntaxique*:

  Time flies like an arrow.

  Fruit flies like bananas.

NRC·CNRC
Institute for Information Technology

Problème 9: Insertion/suppression de mots

- Contrairement aux suppositions du « modèle simpliste », une traduction peut faire disparaître ou apparaître des mots.
- Mots grammaticaux entièrement déterminés requis par la grammaire de l'une ou l'autre langue

  Has Max seen the results?

  → Max a-t-il vu les résultats?
- Mots « semi-grammaticaux » déterminés en bonne partie par le contexte **sémantique**

  oil tank → réservoir à huile

  steel tank → réservoir en acier
- Chinois → français: le chinois ne marque pas les articles comme « un » ou « le »; marqueurs basés sur la structure discursive

  → On doit déduire tous ces éléments du contexte

NRC·CNRC
Institute for Information Technology

Insertion/suppression de mots

- En conséquence, un système de TA doit faire appel à:
  - Une grammaire détaillée des deux langues pour contrôler l'apparition ou la disparition de particules du genre de « *-t-il* »
  - Un modèle capable de reconnaître les relations sémantiques entre deux mots pour contrôler l'apparition ou la disparition de marqueurs de relations comme la préposition dans « *réservoir en acier* »
  - Dans le cas de langues éloignées comme le chinois, un modèle *capable d'accéder au contexte plus global* pour déterminer le choix des articles appropriés.

NRC·CNRC
Institute for Information Technology

Problème 10: Divergences d'explicitation

- Les langues diffèrent beaucoup à propos de ce qu'elles obligent un locuteur à rendre explicite

  Jack has invited one of his students for dinner.

  →« un de ses étudiants » OU « une de ses étudiantes » ?

  He decided to cut Milou's hair

  →« cheveux » OU « poils » ?

- Le chinois ne marque ni le genre, ni le nombre, ni le temps des verbes; tous ces éléments sont obligatoires en français.

NRC·CNRC
Institute for Information Technology

Divergences d'explicitation

- En conséquence…
- Un système de TA aura souvent besoin d'inférer beaucoup d'information à partir du contexte
  - Quel est le sexe de l'étudiant(e) que Jack a invité
  - Milou est-il un humain ou un chien?
- Le contexte pertinent n'est pas forcément local
- Le contexte pertinent n'est pas forcément linguistique
  - p. ex. la présence d'une photo de Jack et de son étudiant(e) pourrait servir à trancher la question
- Pour certains couples de langues comme le chinois → français, le recours au contexte est beaucoup plus souvent requis.

NRC·CNRC
Institute for Information Technology

Problème 11: Grammaire de la langue cible

- Contrairement aux suppositions du « modèle simpliste »,le choix de l'équivalent correct dépend souvent du *contexte en langue cible*
- Exemples:
  - *the → le | la | l' | les*

    selon le mot **français** modifié par l'article
  - know → savoir | connaître

    Selon la nature grammaticale du complément du verbe en français

    John knows that Mary is angry

    → John sait que Mary est en colère

    → John connaît la colère de Mary

NRC·CNRC
Institute for Information Technology

Grammaire de la langue cible

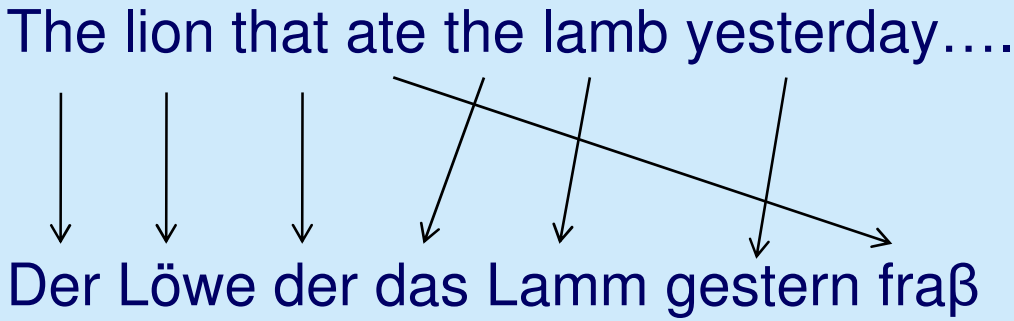
- De manière générale, un dictionnaire bilingue ne peut qu'énumérer un certain nombre de *possibilités* pour traduire les mots de la langue de départ:

  know → savoir | connaître

- Le choix final est déterminé par des considérations globales sur la structure de la phrase traduite
- En conséquence…
  - un système de TA doit en général examiner plusieurs traductions possibles pour chaque mot
  - La traduction choisie pour un mot particulier dépend souvent des traductions choisies pour les autres mots

Problème 12: Ordre des mots

- Le modèle simpliste néglige les différences parfois majeures dans l'ordre des mots des langues de départ et d'arrivée
- Par exemple, en allemand le verbe d'une proposition subordonnée se place en fin de phrase:



NRC·CNRC
Institute for Information Technology

Ordre des mots

- Comme nous l'avons déjà observé précédemment, on traduit non pas des mots isolés mais des structures grammaticales
- En conséquence…
- Un système de TA doit être capable de reconnaître la structure grammaticales des phrases en langue de départ et de lui faire correspondre une structure grammaticale équivalente en langue d'arrivée
- La correspondance en question affecte fortement l'ordre de groupes de mots (« syntagmes »)

NRC·CNRC
Institute for Information Technology

Conclusions

- Vous le saviez déjà : le modèle simpliste est… simpliste!
- Notre examen vous aura permis de réfléchir sur la nature précise de ses carences
- Il en ressort que la production de traductions :
  1. entièrement automatique;
  2. de haute qualité;
  3. de textes généraux

  s'avère plus que problématique
- Les capacités d'analyse linguistique et extra-linguistique requises dépassent de beaucoup celles des technologies actuelle
- Mais une traduction imparfaite peut s'avérer très utile à un lecteur (« gisting ») ou à un traducteur (« post-édition »)