Traduction automatique et comparaison des synonymes français et anglais

Diana Inkpen

Université d'Ottawa

Résumé

Afin d'aider les apprenants d'une langue seconde ou étrangère à mieux distinguer les champs sémantiques de certaines mots considérés comme synonymes, nous sommes en train de concevoir un outil de traduction automatique. Une phrase traduite doit exprimer non seulement la même signification que la phrase initiale, mais aussi les mêmes nuances lexicales que celle-ci lorsqu'il faut utiliser des synonymes. Les nuances lexicales que nous distinguons dans le système que nous avons créé prennent en compte, pour ce qui est de la traduction, de l'attitude, de la stylistique et de la sémantique. Nous présentons, dans le cadre de cet article, la construction d'une petite base de connaissances de synonymes français à partir de deux dictionnaires explicatifs. Nous avons une base de connaissances de synonymes anglais et un système de génération de texte qui tient compte des nuances lexicales pour choisir les meilleurs synonymes dans le texte anglais généré. Nous avons fait appel à ce système pour trouver la meilleure traduction en anglais des synonymes français dans un ensemble de phrases choisies dans les textes des délibérations du parlement canadien. Ce corpus de textes est connu sous le nom de Hansard canadien.

Traduction automatique et comparaison des synonymes français et anglais

Comprendre le processus de l'apprentissage d'une langue seconde ou étrangère est complexe; les solutions pour y parvenir sont variées et ne sont pas toujours les mêmes. En effet, tout dépend du profil de l'apprenant, de sa langue maternelle, des autres langues connues, de son style d'apprentissage, de sa motivation, de son attitude, etc.

Ainsi, afin d'aider les apprenants de niveaux intermédiaire et avancé à mieux comprendre les champs sémantiques de certains mots synonymes, nous sommes en train de concevoir un outil de traduction automatique de synonymes français et anglais. Étant donné qu'environ 70% des mots du vocabulaire de l'anglais et du français se « ressemblent » d'une façon ou d'une autre, le problème du découpage des champs sémantiques est d'autant plus complexe. C'est dans cette perspective que nous concevons une base de connaissances des synonymes du français pour la traduction de mots à partir d'une base de connaissances des synonymes de l'anglais.

Une phrase traduite doit exprimer non seulement la même signification que la phrase d'origine mais encore, il faut choisir des synonymes exprimant les mêmes nuances lexicales. Pour respecter le plus possible la fidélité d'une traduction, le système que nous avons créé prend en compte des nuances telles que l'attitude (par exemple, le mot *alcoolique* est plus positif que le mot *ivrogne*), la stylistique (*laisser tomber* est plus familier qu'*abandonner*) et la sémantique (le mot *erreur* contrairement à son synonyme *bévue* implique l'idée d'une *irréflexion, d'une étourderie,* et souvent même *d'une bêtise*).

Afin d'identifier les mots synonymes, nous avons utilisé les entrées de plusieurs dictionnaires de synonymes. On soulignera ici que, de façon connexe, Ploux et Ji (2003) ont rassemblé, sous forme électronique, des groupes de synonymes français (environ 50000 mots),

des groupes de synonymes anglais (environ 140000 mots) et des groupes bilingues pour ces mots, à partir de divers dictionnaires. Mais ces auteurs ont seulement rassemblé des groupes de mots alors que nous avons aussi recueilli les explications à propos des différences entre les mots de chacun des groupes.

Nous avons construit, de façon automatique, une base de connaissances de synonymes anglais (5000 mots) en employant des techniques d'extraction d'informations (Inkpen et Hirst, 2001) sur le texte explicatif d'un dictionnaire selon les différences entre les synonymes anglais (Hayakawa, 1994). On pourra, à cet effet, consulter l'annexe 2 pour trouver un exemple parmi les neuf cent quatorze groupes de synonymes de cette base de connaissances.

Ainsi que nous l'avons déjà souligné, notre étude est centrée sur le processus de construction d'une petite base de connaissances des synonymes français que nous avons construite manuellement, pour cinq groupes de synonymes (un groupe de verbes, un groupe d'adjectifs et trois groupes de substantifs). Cette base de connaissances comprend présentement cinquante mots au total.

Nous avons réuni l'information provenant de deux dictionnaires explicatifs, à savoir le Bailly (1973) et le Bénac (1956). Deux entrées sont fusionnées seulement si elles ont au moins trois mots en commun.

Nous présentons, ici, deux exemples d'entrées: celui du dictionnaire des synonymes français de Bailly (1973) et celui du Bénac (1956).

Bailly (1973) :

Erreur, qui a un sens très général peut se dire de toute circonstance où l'on prend le faux pour le vrai, le mauvais pour le bon. **Méprise** suppose que l'on prend une chose pour une autre, sans que l'on puisse généralement vous en faire grief. **Bévue** implique, par contre, irréflexion, étourderie, souvent même bêtise. **Maldonne** désigne l'erreur que commet celui qui ne distribue par les cartes comme il se doit. **Aberration** ne se dit pas que d'une erreur de jugement. **Blague**, syn. d'*erreur*, de *bévue*, est familier, ainsi que **gaffe,** qui se dit d'une bévue grossière. **Boulette**, syn. de *gaffe*, est populaire.

Bénac (1956) :

Erreur: fausse opinion. *Erreur* se dit dans tous les cas où l'on prend le faux pour le vrai. **Égarement**, erreur considérable, due à une sorte d'extravagance par rapport au vrai et au bien. **Illusion**, erreur des sens ou de l'esprit, due non à eux même, mais à une fausse apparence de choses matérielles ou morales qui, en nous les faisant voir autrement qu'elles ne sont, nous induisent en erreur. **Aberration**, au contraire, se dit toujours d'une anomalie de nos fonctions, surtout intellectuelles, qui nous fait juger mal .....

Le tableau 1, qui suit, présente les cinq groupes de synonymes français utilisés dans cette étude. Le tableau 2 celui des groupes des synonymes anglais qui correspondent à ces mêmes synonymes français.

Tableau 1

Les cinq groupes des synonymes français

erreur, égarement, illusion, aberration, malentendu, mécompte, bévue, bêtise, blague, gaffe, boulette, brioche, maldonne, sophisme, lapsus, méprise, bourde
ennemi, adversaire, antagoniste, opposant, détracteur
mensonge, menterie, contrevérité, hâblerie, vanterie, fanfaronnade, craque, bourrage de crâne
abandonner, délaisser, déserter, lâcher, laisser tomber, planter là, plaquer, livrer, céder
ivrogne, alcoolique, intempérant, dipsomane, poivrot, pochard, sac à vin, soûlard, soûlographe, éthylique, boitout, imbriaque

Tableau 2

Les groupes des synonymes anglais qui correspondent aux synonymes français du Tableau 1

mistake, blooper, blunder, boner, contretemps, error, faux pas, goof, slip, solecism
opponent, adversary, antagonist, competitor, enemy, foe, rival
thin, lean, scrawny, skinny, slender, slim, spare, svelte, willowy, wiry
lie, falsehood, fib, prevarication, rationalization, untruth
leave, abandon, desert, forsake
alcoholic, boozer, drunk, drunkard, lush, sot

Les concepts implicites ou suggérés par les synonymes sont extraits de phrases françaises explicatives. Nous avons exprimé ces concepts dans le même langage de représentation que celui de la base de connaissances anglaises: soit une «interlingua» développée à l'Université de Californie du Sud, qui emploie des concepts que l'on retrouve dans l'ontologie Sensus (http://mozart.isi.edu:8003/sensus2/). Le processus comprend deux étapes. Il s'agit de:

1) trouver la meilleure traduction en anglais pour la phrase française;

2) trouver un concept ou un ensemble des concepts dans Sensus.

Par exemple, si la phrase initiale à traduire est: “aberration implique une erreur de jugement”, le mot anglais “judgment” est cherché dans Sensus. L’ensemble des concepts choisis est (P6 (c6 / |mind<view| :MOD |wrong>false|)). Le rôle MOD signifie quelque chose qui modifie le concepts principal: un adjectif ou un substantif.

Le problème des différences de **nuances lexicales entre deux langues** est complexe. En assortissant les nuances d'un synonyme français et d’un synonyme anglais, on pourra détecter leur champ commun en employant l'ontologie de Sensus. Mais il y a aussi des cas où les mots n’ont aucune partie commune. Par exemple, le synonyme français *bavure*, qui signifie une erreur faite par la police n’a pas de synonyme anglais équivalent pour traduire cette nuance particulière.

Nous avons élaboré, dans une précédente étude (Inkpen et Hirst, 2003), un système de génération de texte qui prend un ensemble de nuances d'entrées et qui choisit le meilleur synonyme dans le texte anglais produit. Ce système de génération de textes a été construit à partir du générateur des textes de Langkilde-Geary (2002), en ajoutant le traitement automatique des nuances lexicales. Dans l’étude actuelle, nous utilisons notre système avec des synonymes français, afin de trouver la meilleure traduction d'un synonyme français dans un ensemble de phrases choisies dans le Hansard canadien (http://www.isi.edu/natural-language/download/hansard/). Nous avons employé comme entrées les nuances d’un synonyme français et nous avons évalué la capacité du système à choisir le synonyme anglais adéquat.

Par exemple, voici une phrase extraite du Hansard canadien:

“Je trouvais dommage qu'on dût assister à une autre *gaffe* dans l'épopée de la privatisation de nos sociétés d'état.”

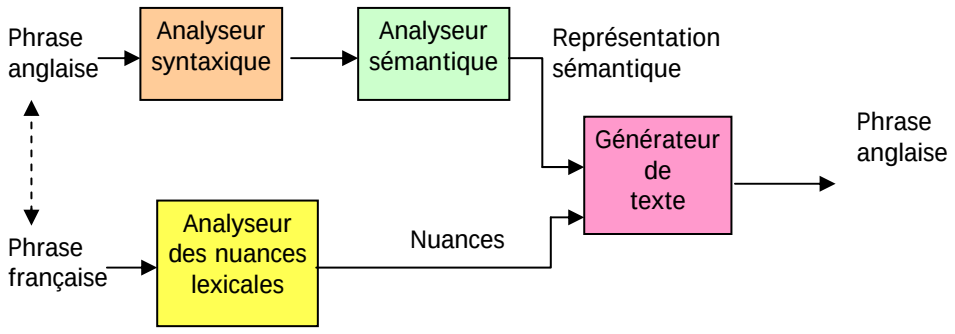
“I was not happy that we have another *blunder* in the ongoing saga of the selling off of our Crown corporations.”

Pour chaque phrase française, il y a une traduction en anglais. Ici le synonyme qui nous intéresse est *gaffe* et la meilleure traduction en anglais de ce mot est *blunder*. Peut-être existe-t-il d’autres traductions appropriées mais, selon notre méthode automatique, il n’existe qu’une seule bonne solution.

Le schéma 1 présente l’architecture de notre système. Nous avons employé comme entrées les nuances d’un synonyme français extraites à partir d’un analyseur des nuances lexicales. Pour l’analyse syntaxique et sémantique, nous avons utilisé la phrase anglaise extraite du Hansard pour obtenir la représentation sémantique. Le synonyme est remplacé par un concept général. Le but est de générer une phrase anglaise qui a le même sens que la phrase française en choisissant le meilleur synonyme.

Schéma 1

L’architecture du système



Dans ce système, nous avons besoin d'une représentation sémantique pour chaque phrase française. Nous faisons l'analyse syntaxique et sémantique de la phrase anglaise avec des outils automatiques disponibles seulement pour l'anglais. L'analyseur syntaxique est le Charniak (2000) et l'analyseur sémantique est le Langkilde-Geary (2002). Nous employons la représentation sémantique de la phrase anglaise comme approximation pour la représentation sémantique de la phrase française. Cette approximation est possible parce que les deux phrases ont la même signification, l'une étant la traduction de l'autre dans le Hansard Canadien.

Dans le tableau 3, on trouvera quelques résultats obtenus sur deux ensembles de données.

Tableau 3

Les résultats

Expérience	Nombre de phrases	Précision	
		Référence	Système proposé
Test 1	26	38%	59%
Test 2	298	71%	73%

Le premier ensemble de données comprend vingt-six phrases et l'autre, deux cent quatre-vingt dix-huit. La différence entre les deux lots testés est la manière de rassembler les phrases du corpus (le Hansard canadien). Le Test 1 contient deux phrases pour chaque paire mot français mot anglais pour assurer une représentation proportionnée de toutes les pairs. Le Test 2 contient toutes les phrases du corpus avec les paires qui nous intéressent. Le Test 2 comprend quelques paires qui ont beaucoup d'exemples et quelques paires qui ont peu d'exemples; cela reflète la

fréquence naturelle des mots dans le corpus. Toutefois, le Test 2 contient plus de cas qui sont faciles à résoudre par notre système. Dans le tableau 3, nous présentons aussi la précision d'une méthode simple; celle qui choisit toujours le synonyme le plus fréquent. Cette précision est utilisée comme référence. On pourra noter que notre méthode, qui correspond à la dernière colonne du tableau 3 du système proposé, obtient de meilleurs résultats que la méthode simple pour les deux lots testés. En somme, la nouvelle méthode permet un gain de 21% pour le Test 1 et 2% pour le Test 2. La plus petite différence pour le Test 2 est due au fait qu'il contient des phrases plus faciles à résoudre pour notre système et aussi pour la méthode simple. Le Test 1 contient plus de phrases qui posent des problèmes au système simple mais pas au nôtre .

Le problème de la **synonymie** est important également pour la traduction automatique et pour **l'apprentissage d'une langue seconde ou étrangère**. Lorsque l'étudiant doit choisir parmi des synonymes de la langue étrangère, qui ont des nuances sémantiques différentes, il va être influencé par les nuances sémantiques des synonymes dans sa langue maternelle. Il ne peut faire un choix éclairé que s'il connaît les différences. Nos bases de connaissance des synonymes anglais et français peuvent alors être utilisées comme une aide à l'apprenant en situation d'autonomie.

Notre système peut être intégré à un didacticiel de langue. Par exemple, les deux bases de connaissances peuvent être utilisées pour expliquer automatiquement à l'apprenant quelles sont les différences entre les mots synonymes. Si le français est la langue seconde ou étrangère, on utilise la base de connaissances françaises; si c'est l'anglais, on utilise la base de connaissances anglaises. Dans les deux cas, notre système peut tenir compte des nuances sémantiques d'un synonyme français avec un synonyme anglais et cela peut aider l'apprenant à choisir la meilleure traduction tout en lui expliquant pourquoi elle est meilleure que les autres. L'explication contient

la partie commune entre les nuances d'un synonyme français et d'un synonyme anglais. Par exemple, si l'apprenant doit traduire une phrase anglaise qui contient le mot *blunder* et elle choisit le mot *erreur* en français, notre système peut aider l'apprenant à comprendre pourquoi *gaffe* est un meilleure choix; c'est parce qu'il partage avec le mot *blunder* une nuance lexicale de «*bévue grossière*». Le système peut fournir des explications parce qu'il utilise l'information dans les deux bases des connaissances (voir les exemples dans les annexes 1 et 2). On note qu'une partie importante de notre travail à venir est d'enrichir le contenu de ces bases de connaissances.

En conclusion, dans cette étude, nous avons présenté la construction d'une base de connaissances des synonymes français. Nous l'avons utilisée avec notre système de génération de textes afin de trouver la meilleure traduction d'un synonyme français. Nous avons également évalué la capacité du système à choisir le synonyme anglais juste. Nous avons expliqué aussi comment nous pouvons utiliser notre système dans un didacticiel en langue.

Bibliographie

Bailly, René (1973). *Dictionnaire des Synonymes de la Langue Française*, Paris: Larousse.

Bénac, Henri (1956). *Dictionnaire des Synonymes*, Paris: Librairie Hachette

Charniak, Eugene. (2000). A Maximum-Entropy-Inspired Parser. In *Proceedings of the First Conference of the North American Chapter of the Association for Computational Linguistics and the Sixth Conference on Applied Natural Language Processing (NAACL-ANLP 2000)*, 132-139.

Hayakawa, S.I. (1994). *Choose the Right Word*, Second Edition, NewYork: HarperCollins.

Inkpen, Diana and Hirst, Graeme (2001). Building a Lexical Knowledge-Base of Near-Synonym Differences. In P*roceedings of the Workshop on WordNet and Lexical Resources, Second Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL-2001),* 47-52.

Inkpen, Diana and Hirst, Graeme (2003). Near-Synonym Choice in Natural Language Generation. In *Proceedings of the International Conference RANLP-2003 (Recent Advances in Natural Language Processing)*, 204-211.

Langkilde-Geary, Irene (2002). An Empirical Verification of Coverage and Correctness for a General-Purpose Sentence Generator, in *Proceedings of the 12th International Natural Language Generation Workshop*, 17-24.

Ploux, Sabine and Ji, Hyungsuk (2003). A Model for Matching Semantic Maps between Languages (French, English). In *Computational Linguistics*, 29(2), 155-178.

Annexe 1

Exemple d'entrée dans la base des connaissances des synonymes français

```
(defcluster generic_erreur_n
   :syns (erreur egarement illusion aberration malentendu mecompte bevue
            betise blague gaffe boulette brioche maldonne sophisme lapsus meprise bourde)
   :periph ((P1 (c1 / |take amiss| :object thing))
              (P2 (c2 / |grief,sorrow|))   (P3 (c3 / |betise|))
              (P4 (c4 / |hand,deal|  :object |card<paper|   :polarity -   :mod |the right way|))
              (P5 (c5 / (OR |voluntary>unforced| |involuntary>enforced|)))
              (P6 (c6 / |mind<view|  :mod |wrong>false|))
              (P7 (c7 / |glaring,gross|))   (P8 (c8 / |view<belief| :mod |false>untrue|))
              …
              (P18 (c18 / |action| :MOD |regretable|))  )
   :distinctions
   ((meprise usually medium Denotation P1)    ;    "prend une chose pour une autre"
    (meprise usually medium Denotation P2)    ;    "generalement faire grief"
    (bevue usually medium Implication P3)      ;    "irreflexion, etourderie, souvent meme betise"
    (maldonne usually medium Denotation P4)     ;    "ne distribue pas les cartes comme il se doit"
    (maldonne usually medium Denotation P5)  ;    "erreur voluntaire or involuntaire"
    (aberration always medium Denotation P6)  ;     "erreur de jugement"
    (blague  low Formality)    ;      "familier"
    (gaffe  low Formality)     ;      "familier"
    (gaffe usually medium Denotation P7)     ;      "bevue grossiere"
    (boulette  low Formality)    ;      "populaire"
    …
    (erreur usually medium Denotation P18)   ;      "action inconsideree, regrettable, maladroite"
  )
)
```

Annexe 2

Exemple d'entrée dans la base des connaissances des synonymes anglais

```
(defcluster generic_mistake_n
  :syns (mistake blooper blunder boner contretemps error faux_pas goof slip solecism )
  :core (ROOT GENERIC_MISTAKE (OR |fault,error| |boner| |gaffe| |slipup|))
  :periph ( (P1  (C1 / deviation))   (P2  (C1 / sin))
    (P3  (C1 / assessment :MOD (*OR* ignorant uninformed)))
    (P4  (C1 / accidental))    (P5  (C1 / careless))    (P6  (C1 / indefensible))
    (P7  (C1 / occurrence :MOD (*OR* embarrassing awkward)))
    (P8  (C1 / (*OR* action opinion judgment)))
    (P9  (C1 / (*OR* gross stupid)))   (P10  (C1 / belief))    (P11  (C1 / manners)) )
  :distinctions
  ((error usually medium Implication P1)
   (error usually medium Denotation P2)
   (blunder usually medium Implication P3)
   (slip usually high Denotation P4)
   (slip usually medium Denotation P5)
   (blooper low Formality)   (goof medium Formality)
   (goof usually medium Denotation P6)
   (contretemps usually medium Denotation P7)
   (mistake usually medium Denotation P8)
   (blunder usually medium Denotation P9)
   (error usually medium Denotation P10)
   (faux_pas usually medium Denotation P11)
   (mistake usually medium Favourable :agent)
   (blunder usually medium Favourable :agent)
    (boner usually medium Pejorative :agent)
    (contretemps usually medium Favourable :agent)  …  )  )
```

Note d'auteur

Diana Inkpen, est professeure adjointe à l'École d'ingénierie et de technologie de l'information, Université d'Ottawa.

Cette recherche a été rendue possible grâce a une subvention du Conseil de recherches en sciences naturelles et en génie du Canada (CRSNG).

La correspondance au sujet de cet article devrait être adressée à la professeure Diana Inkpen, École d'ingénierie et de technologie de l'information, Université d'Ottawa, 800 Avenue King Edward, Boîte postale 450 A, Ottawa, ON, K1N 6N5, Canada. E-mail: diana@site.uottawa.ca