Analogie en traitement automatique des langues. Application à la traduction automatique

Yves Lepage

ATR, Keihanna[1]
yves.lepage@atr.jp

Résumé

On se place ici dans la tendance actuelle en traitement automatique des langues, celle à base de corpus et aussi dans une perspective que l'on peut qualifier d'approche à moindre effort : il s'agit d'examiner les limites des possibilités de traitement à partir de données textuelles brutes, c'est-à-dire non pré-traitées. L'interrogation théorique présente en arrière-plan est la suivante : quelles sont les opérations fondamentales en langue ? L'analogie proportionnelle a été mentionnée par de nombreux grammairiens et linguistes. On se propose de montrer l'efficacité d'une telle opération en la testant sur une tâche dure du traitement automatique des langues : la traduction automatique. On montrera aussi les bonnes conséquences de la formalisation d'une telle opération avec des résultats théoriques en théorie des langages en relation avec leur adéquation à la description des langues. De cette façon, une opération fondamentale en langue, l'analogie proportionnelle, se verra illustrée tant par ses aspects théoriques que par ses performances en pratique.

Mots-clés : Analogie, proportion, chaîness de symboles, traduction automatique,divergences entre langues

Abstract

We position ourselves in the current trend of Natural Language Processing, *i.e.*, corpus-based approach and "least effort" approach. We shall inquire how far it is possible to go without any preprocessing of raw data. The theoretical question in the background is: which operations are fundamental in language? Proportional analogy has been mentionned by many grammarians and linguists. We shall inspect the efficiency of such an operation by testing it against a difficult task of NLP : machine translation. We shall also show some good properties brought by the formalisation of such an operation with theoretical results in formal language theory and the adequacy of analogy with the description of natural languages. To summarize, we shall illustrate an operation that is fundamental in language in its theoretical aspects as well as in its practical peformance.

Keywords: Analogy, proportion, strings of symbols, machine translation,divergences across languages

1. Introduction

Ce tutoriel se propose de présenter une opération que le traitement automatique des langues a lontemps négligée. Il s'agit de l'analogie. Une analogie met nécessairement quatre objets A, B, C et D en relation ; elle se note $A : B :: C : D$ et se lit « A est à B ce que C est à D. » Par exemple, entre mots :

français : *recevoir* : *j'ai reçu* :: *percevoir* : *j'ai perçu*

[1] Les travaux rapportés ici ont été subventionnés en partie par un contrat avec l'Institut national japonais des technologies de l'information et des communications (acronyme anglais : NiCT) sur le thème : « Étude des techniques de traduction de dialogues oraux fondées sur un grand corpus ».

arabe : *aslama* : *muslimun* :: *arsala* : *mursilun*

entre phrases :

Ces fenêtres, là, je ne vous ai pas dit de les ouvrir.	:	*Est-ce que vous pouvez m'ouvrir une fenêtre ?*	::	*Ces chèques de voyage, là, je ne vous ai pas dit de les échanger.*	:	*Est-ce que vous pouvez m'échanger ces chèques de voyage ?*

L'objet de ce tutoriel est de montrer la contribution éventuelle de cette opération au traitement automatique des langues, dans la lignée des approches « à moindre effort » et à base de corpus.

2. Introduction à la notion d'analogie

Historiquement, le mot grec *analogia* est employé par Euclide. Il ne signifie rien d'autre que l'égalité de rapports entre entiers : $\frac{4}{2} = \frac{6}{3}$. Le mot a été repris tel quel par Aristote dans l'*Éthique à Nicomaque* et dans la *Rhétorique* pour désigner toute relation du type A est à B ce que C est à D.

La confusion de l'analogie avec la simple similarité a son origine dans les élaborations scholastiques de la notion, à la suite d'Aristote et de son interrogation sur les rapports de sens du même mot dans ses différents usages. Boèce aggrave les choses en parlant de proportion pour rapport et de proportionalité pour égalité de rapports, c'est-à-dire pour analogie. L'interprétation des écrits de St Thomas d'Aquin par St Cajétan ont semble-t-il accéléré le galvaudage du terme[2].

De fait, le mot latin de proportion était une traduction du grec analogie par Varron qui l'utilisait par opposition avec *anomalie*, ou absence de régularité en conjugaison ou déclinaison. La notion a été reprise par la civilisation arabe, en droit et en théologie, cependant qu'en grammaire elle sous-tend toute la description du lexique. Au XIX^e, avec la découverte des lois phonétiques, l'analogie, en tant que créatrice de formes nouvelles, perd un temps son statut d'objet scientifique au profit des lois phonétiques, et se retrouve étiquetée comme fauteur de trouble. Mais les Néogrammairiens (Osthoff, Brugmann, Paul) la replacent à égalité avec les lois phonétiques, car, si les changements phonétiques, réguliers dans leur application, créent du désordre, l'analogie, imprévisible dans son application, rétablirait l'ordre (paradoxe de Sturtevant). Quatre chapitres du *Cours de linguistique générale* de Saussure font avancer la réflexion sur l'analogie dans un tel cadre.

Depuis les travaux de (Skousen, 1989), la notion a été remise à l'honneur en traitement automatique des langues pour différentes tâches tournant autour de la notion de mots, que ce soit la transcription graphème-phonème (Damper et Eastman, 1996), (Yvon, 1994), la dérivation morphologique ou la terminologie (Hathout, 2001), (Claveau et L'Homme, 2005). En linguistique, en morphologie, la Whole word morphology (Neuvel et Singh, in press) (Singh et Ford, 2000) remet aussi la notion d'analogie à l'honneur. Pour ce qui est de la résolution d'analogies sémantiques entre mots (tests SAT), Turney et ses collègues travaillent depuis des années sur le sujet (Turney et Littman, 2005).

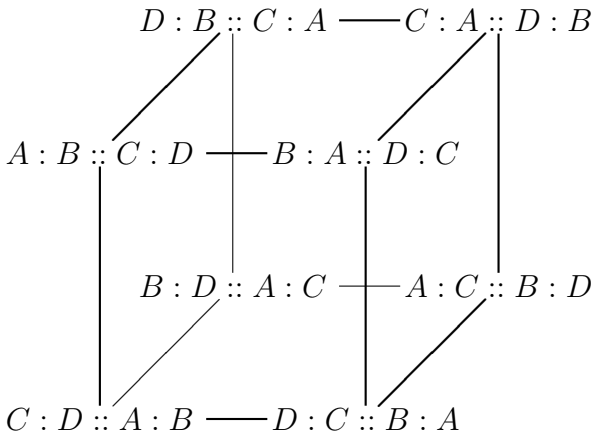
[2] Nous pensons que les travaux de Gentner sur la Structure Mapping Theory (Gentner, 1983) portent en fait sur l'explicitation des métaphores de quatrième espèce (Aristote, *Poétique*). Ces métaphores reposent sur une analogie implicite : « *un atome est comme un système solaire* » parce que (et seulement parce que) « *un electron est au noyau atomique ce qu'une planète est au soleil.* » Voir (Holyoak et Thagard, 1995).

Hermann Paul avait franchi le pas vers la syntaxe en voyant l'analogie comme le moyen privilégié de production des syntagmes nouveaux à partir du matériau linguistique déjà entendu par le locuteur (*Proportionengruppen*). Cette idée est énoncée par Bloomfield pour les phrases. Au contraire, Chomsky démontre en partie l'existence d'un module propre à la syntaxe en se basant sur le fait que les analogies entre phrases peuvent produire des phrases agrammaticales avec comme conséquence que l'analogie ne saurait être un objet d'étude en syntaxe. Mais la valeur de l'exemple toujours cité est contestée par Pullum et le débat continue. (Lavie, 2003) En traduction automatique, l'article de Nagao présentant une approche de la traduction de phrases par analogie (en fait, par similarité) est considéré comme l'un des articles fondateurs du courant de la traduction automatique par l'exemple (Nagao, 1984), (Sato, 1991).

3. Formalisation de l'analogie entre chaînes de symboles

3.1. Propriétés générales

Il existe des propriétés générales de l'analogie qui sont valables quels que soient les objets impliqués. En particulier, l'échange des moyens $A : B :: C : D \Leftrightarrow A : C :: B : D$ et des extrêmes $A : B :: C : D \Leftrightarrow D : B :: C : A$ apparaissent dans le cube des huit formes équivalentes de l'analogie donné ci-dessous.



Pour les objets particuliers que sont les chaîness de symboles, il faut bien sûr citer Copycat (Hofstadte et The Fluid Analogies Research Group, 1994), approche de la résolution de devinettes de la forme donnée ci-après par des méthodes d'intelligence articielle (modélisation du monde des lettres, etc.)

> *Supposons que la séquence de lettres* aabc *ait été changée en* aabd *; comment changer la séquence* ijkk *de la « même façon » ?*

Fondamentalement, on peut distinguer quatre espèces distinctes d'analogies qui, effectivement, ne correspondent pas aux mêmes définitions.

répétition :	(malais)	*guru* : *guru-guru* :: *perempuan* : *perempuan-perempuan*
reduplication :	(latin)	*cado* : *cecidi* :: *pago* : *pepigi*
commutation :	(arabe)	*aslama* : *muslimun* :: *arsala* : *mursilun*
miroir :		*abcde* : *xyz* :: *edcba* : *zyx*

Seules les trois premières espèces majeures d'analogie semblent pertinentes pour la linguistique. Le présent tutoriel se limitera aux analogies de commutation et seulement à celles-là. Aussi, en particulier, nous imposons un postulat de « déterminisme » qui exige que $A : A :: C : x \Rightarrow x = C$. Ce postulat nous semble inévitable, mais on verra qu'il n'est pas admis par tout le monde.

En vue d'une formalisation de la notion d'analogies, on pourrait se demander quel sens ont les signes : et ::, et on pourrait être tenté de les interpréter comme des opérations ou des relations. Il faut se garder de cette approche car on s'expose à des déceptions. Par exemple, voir la relation d'égalité dans :: est faux. En effet, en général, on n'a pas de transitivité : $A : B :: C : D$ et $C : D :: E : F$ n'impliquent pas en général $A : B :: E : F$. Soit en effet une équation analogique $A : B :: C : \mathrm{x}$ à deux solutions différentes D_1 et D_2. Par transitivité, $C : D_1 :: A : B$ et $A : B :: C : D_2$ donneraient $C : D_1 :: C : D_2$, c'est-à-dire $C : C :: D_1 : D_2$, ce qui est en contradiction avec le postulat de « déterminisme ». Nous mentionnerons d'autres résultats contraire à l'intuition.

3.2. Définitions proposées pour les chaîness de symboles

On se place ici dans une cadre minimal. Les objets intervenant dans une analogie ne sont que des chaîness de symboles sans aucune interprétation sémantique et la seule opération autorisée est la vérification de l'égalité entre deux symboles. Yvon (voir (Stroppa, 2005)) propose une formalisation élégante de l'analogie à partir de la notion de facteurs de mots (c'est-à-dire de sous-chaînes). Elle repose sur l'idée illustrée ci-dessous :

$$\begin{array}{ccccccccc} \mathrm{a} & & \mathrm{rs} & & \mathrm{a} & & \mathrm{l} & & \mathrm{a} \\ & \diagdown\!\!\diagup & & \diagdown\!\!\diagup & & \diagdown\!\!\diagup & & \diagdown\!\!\diagup & \\ \mathrm{mu} & & \mathrm{sl} & & \mathrm{i} & & \mathrm{m} & & \varepsilon \end{array} \quad \Leftrightarrow \quad \textit{aslama} : \textit{muslim} :: \textit{arsala} : \textit{mursil}$$

Ce qui s'exprime par : $A : B :: C : D \Leftrightarrow A \bullet D \cap B \bullet C \neq \emptyset$ avec $\bullet$ notant le mélange de mots. Cette approche qui s'étend fort élégamment à d'autres structures algébriques, des magmas aux groupes libre et abélien, ne respecte cependant pas le postulat de « déterminisme ». Cette formalisation est donc moins restrictive que celle, incomplète, que nous avons proposée[3] à partir de distances d'édition. C'est pourquoi, il ne s'agit que d'une implication :

$$A : B :: C : D \quad \Rightarrow \quad \left\{ \begin{array}{rcl} \mathrm{dist}(A, B) & = & \mathrm{dist}(C, D) \\ \mathrm{dist}(A, C) & = & \mathrm{dist}(B, D) \\ |A|_a + |D|_a & = & |B|_a + |C|_a, \quad \forall a \end{array} \right.$$

3.3. Définition des langages de chaînes analogiques

Grâce à l'analogie, on peut définir une famille de langages formelles. Soit $\mathcal{M}$ un ensemble de couples de chaîness de symboles (v, v') notés $v \rightarrow v'$ et apppelés *modèles*. La dérivation analogique modulo $\mathcal{M}$, notée $\vdash_{\overline{\mathcal{M}}}$, permet de produire des chaîness w' à partir de la chaînes des symboles w en « appliquant » les modèles de $\mathcal{M}$ par analogie : $w \vdash_{\overline{\mathcal{M}}} w' \Leftrightarrow \exists v \rightarrow v' \in \mathcal{M} \ / \ v : v' :: w : w'$ On peut définir, de façon standard, un « langage de chaîness analogiques » par fermeture transitive de la dérivation analogique : $\Lambda(\mathcal{A}, \mathcal{M}) = \mathcal{A} \cup \{ w' \in$

[3] Nous avons nous-même (Lepage, 2003) travaillé sur les ensembles, les multi-ensembles et les structures de traits. (Delhay et Miclet, 2004) étend la définition à un alphabet muni d'une structure plus riche.

$\mathcal{V}^*$ / $\exists w \in \mathcal{A},\ w \vdash^{+}_{\mathcal{M}} w'$ } à partir de chaîness axiomes ou *attestées* (ensemble $\mathcal{A}$) et des modèles. L'induction structurale engendre tous les éléments d'un tel langage : en partant des éléments de $\mathcal{A}$, on applique toutes les analogies possibles avec les éléments de $\mathcal{M}$ comme modèles. Réciproquement, la grammaticalité est testée simplement, *in fine*, par comparaison avec les chaîness de $\mathcal{A}$ (attestées), aprés réduction par analogie selon les modèles de $\mathcal{M}$. Ces modèles peuvent être vus comme des modèles de conjugaison, de déclinaison, de dérivation morphologique, de transformation syntaxique, de paraphrasage, etc. On remarquera, point important, que la définition des langages de chaîness analogiques ne fait pas usage de non-terminaux.

On montre aisément, par induction et par utilisation d'une hypothèse sur la concaténation d'analogies (Lepage, 2000), que les trois exemples classiques illustrant les trois grandes classes de la classification de Chomsky-Schutzenberger sont, exprimés sous forme de langages de chaîness analogiques, très simples (une seule chaînes attestée, un seul modèle) et tout à fait semblables : $\{a^n\ /n \geq 1\} = \Lambda(\{a\}, \{a \rightarrow aa\})$ (régulier), $\{a^n b^n\ /n \geq 1\} = \Lambda(\{ab\}, \{ab \rightarrow aabb\})$ (hors-contexte), $\{a^n b^n c^n\ /n \geq 1\} = \Lambda(\{abc\}, \{abc \rightarrow aabbcc\})$ (sous-contexte). Une démonstration identique permet d'obtenir que le langage qui sert de base aux contre-exemples contre l'hypothèse du hors-contexte (Shieber, 1985) est aussi un langages de chaîness analogiques : $\{a^m b^n c^m d^n\ \ / n, m \geq 1\} = \Lambda(\{abcd\}, \{abcd \rightarrow abbcdd, abcd \rightarrow aabccd\})$

3.4. Adéquation à la description des langues

Pour décrire les langues humaines, (Joshi, 1985) a proposé la notion de modérément sous-contexte, définie par a) l'inclusion des langages hors-contexte, b) le fait de rendre compte de certaines dépendances particulières, c) la propriété de croissance bornée des longueurs et d) l'analyse en temps polynomial (Joshi *et al.*, 1991).

Pour a), nous ne savons pas si les langages de chaîness analogiques incluent les langages hors-contexte. Pour b), $\{a^n b^n c^n\}$ et $\{a^m b^n c^m d^n\}$ sont des langages de chaîness analogiques. Pour c), on montre facilement que tout langage de chaîness analogiques vérifie la propriété de croissance bornée des longueurs. Pour d), on conjecture que l'analyse de certains langages de chaîness analogiques dits décroissants serait bien polynomial.

La présentation donnée plus haut des langages de chaîness analogiques permet d'entrevoir son application directe à la langue. Un ensemble de phrases (courtes) dans une langue donnée devrait être suffisant pour produire des phrases (courtes) nouvelles sans avoir besoin de patrons ni de variables. Sans aucune contrainte, ces analogies de forme produiront des phrases incorrectes ou agrammaticales (surproduction de l'analogie). Une application dans laquelle cette production serait contrainte par une représentation du sens serait préférable (Itkonen et Haukioja, 1997) ; nous verrons que c'est le cas avec la traduction automatique.

4. Analogie et corpus

Déjà, on peut se poser la question de savoir dans quelle proportion les analogies de formes repérées dans un corpus sont aussi des analogies de sens. Nous avons compté les analogies de forme (Lepage, 2004) contenues dans un corpus multilingue (anglais, japonais, chinois) très particulier, le BTEC (Takezawa *et al.*, 2002), constitué de 100 000 phrases courtes (six ou sept mots en moyenne) différentes, phrases typiques des guides pour touristes donnant une palette de variantes possibles pour une situation donnée. Les résultats sont donnés ci-dessous.

	nombre d'analogies	nombre de phrases impliquées	nombre moyen d'analogies par phrases	
			corpus entier	impliquées seulement
	(i)	(ii)	(i) / 162,318	(i) / (ii)
chinois	**1,639,068**	49,675	10.10	33.00
anglais	**2,384,202**	53,250	14.69	44.77
japonais	**1,910,065**	53,572	11.77	35.65

Les analogies de forme détectées sont-elles aussi des analogies de sens (ce qu'on appelle des « vraies analogies ») ? On peut poser que la traduction dans une autre langue n'est que l'expression du même sens dans un autre représentation. Sous cette hypothèse, l'intersection des analogies dans les trois langues est un sous-ensemble des analogies de forme et de sens. On obtient ainsi une évaluation basse de 70 000 « vraies analogies » impliquant 14 000 phrases en anglais. En « forçant » l'analogie en chinois et en japonais lorsqu'on a analogie en anglais, c'est-à-dire en essayant de produire des phrases chinoises ou japonaises, on obtient une estimation haute d'un million et demie de « vraies analogies » impliquant 50 000 phrases, soit la moitié des phrases anglaises. Enfin, une estimation par échantillonnage donne moins de 4% des analogies de forme en anglais comme « fausses » (seuil de rejet = 0,1). Voici un exemple d'analogie de forme mais pas de sens :

$$\textit{Could you tell me how to fill this from.} : \textit{Could you tell me how to fill this form.} \not:: \textit{Where is the conference centre ?} : \textit{Where is the conference center ?}$$

Il est possible de visualiser la structure analogique autour d'une phrase donnée dans une table comme ci-dessous (Lepage et Peralta, 2004). La case en haut à gauche est la phrase de départ, et toute case intérieure de la table est en relation d'analogie avec les phrases sur les bords et la phrase de départ. Ci-dessous, la phrase de départ est *I like Japanese food.* Il faut noter que la ligne avec *seafood* et les lignes avec *Chinese, Italian, …food* montrent clairement que les mots ne sont pas des frontières nécessaires pour l'analogie.

I like Japanese food.	*I prefer Japanese food.*	*I'd prefer Japanese food.*	*I feel like Japanese food.*
Do you like Italian food ?			*Do you feel like Italian food ?*
I'd like Western food.	*I'd prefer Western food.*		
I like Chinese food.	*I prefer Chinese food.*		
I like Italian food.	*I prefer Italian food.*	*(x)*	
I like Mexican food.			*I feel like Mexican food.*
I like seafood.	*I prefer seafood.*		*I feel like seafood.*
I like Western food.		*I'd prefer Western food.*	

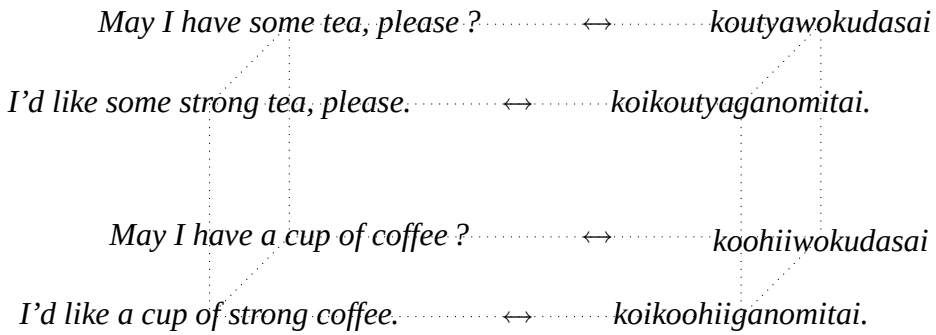
On peut ajouter des phrases dans la table, là où elles font défaut. Par exemple, la case marquée *(x)* peut recevoir la phrase *I'd prefer Italian food.* obtenue par résolution de l'équation analogique suivante, de la même manière que l'on écrit $4 : 2 = 6 : x \Rightarrow x = 3$.

$$\textit{I like Japanese food.} : \textit{I'd prefer Japanese food.} :: \textit{I like Italian food.} : x \quad \Rightarrow \quad x = \textit{I'd prefer Italian food.}$$

5. Application à la traduction automatique

5.1. Principe de base

Le problème de la contrainte de la forme par le sens, est au cœur de la traduction automatique : on peut voir la traduction d'une phrase dans une autre langue comme l'expression du même sens dans une système différent. Un corpus parallèle peut donc être vu comme un ensemble de sens (les phrases en langue source) avec leur forme ou réalisation associée (les phrases en langue cible). La technique de traduction automatique proposée repose sur la recherche d'analogies de sens (en langue source) qui sont aussi analogies de forme (en langue cible). La méthode repose donc entièrement sur la vue géométrique suivante :



Cette méthode de traduction peut s'appliquer de façon récursive. Elle s'exprime très simplement dans un langage déclaratif comme Prolog (les constantes sont en minuscules, les variables en majuscules). Il n'y a que deux prédicats : `translation` pour les paires d'exemples, et `analogy` pour la résolution des équations analogiques (avec C et $\widehat{D}$ comme inconnues sur les deux lignes). La dernière ligne ajoute une traduction a la base de faits, ce qui fait que le système apprend au fur et à mesure qu'il traduit. Ce programme montre clairement que la méthode est fondamentalement bidirectionnelle.

% base de faits :
% couples de traduction
`translation`(s_1,$\widehat{s_1}$) .
`translation`(s_2,$\widehat{s_2}$) .
⋮
`translation`(s_n,$\widehat{s_n}$) .

% moteur de traduction
`translation`(D,$\widehat{D}$) :–
`translation`(A,$\widehat{A}$),
`translation`(B,$\widehat{B}$),
`analogy`(A,B,C,D),
`translation`(C,$\widehat{C}$),
`analogy`($\widehat{A}$,$\widehat{B}$,$\widehat{C}$,$\widehat{D}$),
`assert`(`translation`(D,$\widehat{D}$)).

5.2. Amélioration du principe de base

On reste cependant confronté à la surproduction propre à l'analogie même en contraignant une langue par l'autre, mais des améliorations peuvent être apportées de façon simple. Afin d'éviter

la production de chaînes qui n'appartiendraient pas à la langue, une méthode très simple consiste à écarter toute chaînes contenant une suite de caractères de longueur fixée non-attestée dans le corpus. Ce filtre, extrêmement simple, donne d'excellents résultats dans les tâches de traduction, mais aussi dans une tâche de paraphrasage (Lepage et Denoual, 2005). La méthode est par essence non-déterministe et plusieurs candidats peuvent être produits pour une même phrase source à traduire. En plus, le même candidat peut être produit plusieurs fois. Dans l'exemple de traduction donné ci-dessous, les nombres à gauche donnent le nombre de fois où le même candidat a été produit. Intuitivement, le candidat le plus fréquent devrait être le meilleur. C'est le choix que nous avons adopté pour les différentes évaluations objectives que nous commenterons. Mais on peut montrer qu'une légère amélioration des scores peut être obtenue si un autre choix était effectué.

/i ga | itai | n desu/
'stomach NOM | painful (A) | INSIST to-be'

1744 *I have a stomach* | *ache* |.
552 *My stomach* | *hurts* |
124 *I've got a stomach* | *ache* |.
56 ~~*Do you have a stomach ache.*~~
51 ~~*Do you have a stomach ache ?*~~
50 ~~*I have a stomach ache ?*~~
2 *My stomach* | *hurts* | *me.*
1 *I have an abdominal* | *pain* | *in my stomach.*
1 *I have a* | *pain* | *in my stomach.*
1 ~~*I have a soare throat.*~~

5.3. Avantages de la méthode

La méthode, par définition d'une méthode par l'exemple, ne saurait reproduire que ce qui existe au préalable dans la corpus parallèle de départ. Un désavantage d'une approche par règles est de retomber facilement sur des procédures par défaut quand aucune règle ne s'applique, donnant par là des phrases dont la syntaxe « sent » la langue cible. Ici, fondamentalement, les divergences entre langues étant nécessairement, à la base, données dans le corpus parallèle, la méthode de traduction ne saurait faire autre chose qu'exploiter les deux systèmes des deux langues. Mais ici, le système ne fait aucune abstraction : il n'a ni règles, ni patrons, ni variable, et il n'effectue aucun prétraitement. Il n'a qu'une vue locale des choses (*cf.* (Lavie, 2003) pour une grammaire « sans catégories, sans règles, sans abstraction »).

L'exemple ci-dessus illustre le fait que des classes morpho-syntaxiques ont changé lors de la traduction entre le japonais et l'anglais : un adjectif japonais /itai/ est traduit par un verbe (*hurts*) ou par des noms (*ache, pain*). Des candidats inadmissibles ont aussi été produits par la méthode, nous les avons barrés. Les deux premiers sont intéressants car ils mettent en évidence le fait que la personne n'est effectivement pas marquée en japonais.

L'approche proposée exploite directement deux propriétés de la langue : sa systématicité et sa productivité (Lavie, 2003). Une phrase est traduite en tant qu'elle appartient au système de la langue source, et elle n'obtient de traduction que si le système de la langue cible peut créer un correspondant. La traduction est donc vue ici comme la mise en correspondance de deux

systèmes particuliers : deux espaces analogiques entre lesquels on impose la conservation des analogies (homomorphismes entre langages de chaîness analogiques). Le principe mis en œuvre est donc un principe très général décrit dans maints travaux sur l'analogie :

$$f(A) : f(B) :: f(C) : \mathsf{x} \quad \Leftrightarrow \quad A : B :: C : f^{-1}(\mathsf{x})$$

Ce scéma peut s'appliquer à différentes tâches du traitement automatique des langues, par exemple la segmentation en mots, l'analyse structurale, le paraphrasage, le rephrasage, la traduction automatique, etc.

Nous avons testé un tel système de traduction automatique sur les tâches de deux campagnes d'évaluation IWLST 2004 et IWSLT 2005 ce qui a permis de comparer ce système par l'exemple à des systèmes par approche statistique. Les résultats, obtenus avec le même moteur de traduction, montrent que l'approche obtient des résultats honorables. Si sur la tâche de 2004 nous aurions eu des scores parmi les meilleurs (ci-dessous, scores pour le japonais-anglais), l'amélioration considérable des scores des systèmes par approche statistique grâce aux méthodes d'alignement par « *phrases* », nous rejettent en fin de liste sur les tâches de 2005. Des améliorations à la méthodes sont donc absolument nécessaires pour se maintenir au niveau de l'état de l'art.

	mWER	mPER	BLEU	NIST	GTM
syst.1 hybride	**0.263**	**0.233**	0.630	10.72	0.796
syst.2 stat	0.305	0.249	0.619	**11.25**	**0.824**
syst.3 par analogie	0.324	0.300	**0.634**	9.19	0.731
syst.4 par l'ex.	0.485	0.420	0.397	7.88	0.672
syst.5 (Sytran)	0.730	0.597	0.132	5.64	0.568

La méthode est bien sûr confrontée à bien des problèmes pratiques et théoriques. Le temps de traitement est handicapant même s'il a été divisée par 10 en deux ans. Il faut encore une seconde CPU pour traduire une phrase avec entre un demi-million et un million d'équations analogiques formées en langue source. Le choix des bonnes « équations analogiques » à former reste le problème théorique le plus difficile.

6. Conclusion

Au total, l'objet de ce tutoriel est de « réhabiliter » l'analogie en traitement automatique des langues en syntaxe, puisque l'utilisation de l'analogie en morphologie n'est pas taboue. Du point de vue de la formalisation de l'opération entre chaîness de symboles, pour nous, l'héritage historique impose d'expliciter les deux notions constitutives de l'analogie, la similarité et la contiguïté. Une formalisation incomplète a été proposée, plus contrainte que celle par facteurs de chaîness. Elle permet déjà des applications pratiques.

Le schéma général est celui de l'homomorphisme entre deux espaces analogiques. La méthode peut exploiter directement des données brutes sans aucun prétraitement. En particulier, en traduction automatique il permet le traitement implicite de la divergence d'expression entre langues. Il permet aussi d'éviter une phase de segmentation des langues ne possédant pas

de séparateurs de mots (chinois, japonais, ...). Son application récursive permet l'apprentissage lors du traitement avec la constitution de données intermédiaires exploitables lors d'étapes ultérieures.

Les désavantages de la méthode de traduction automatique sont cependant nombreux : il n'y a aucun traitement du contexte, donc aucun traitement de la référence, etc. Le traitement est coûteux et une heuristique reste à trouver pour sélectionner les équations analogiques « prometteuses ».

On ne saurait prétendre que l'analogie peut tout faire en traitement automatique des langues. En particulier, les rapports entre analogie et fréquence d'emploi restent à explorer au travers des rapports de l'analogie et des modèles de n-grammes

Références

CLAVEAU V. & L'HOMME M.-C. (2005). Apprentissage par analogie pour la structuration de terminologie - utilisation comparée de ressources endogènes et exogènes. In *Actes de Terminologie et intelligence artificielle*, Université de Rouen.

DAMPER R. I. & EASTMAN J. E. (1996). Pronouncing text by analogy. In *Proceedings of COLING-96*, p. 268–269, København.

DELHAY A. & MICLET L. (2004). Analogical equations in sequences : Definition and resolution. *Lecture Notes in Computer Science*, **3264**, 127–138.

DOWTY D. R., KARTTUNEN L. & ZWICKY A. M. (1985). *Natural language processing – Psychological, computational, and theoretical perspectives*. Cambridge : Cambridge University Press.

ELITHORN A. & BANERJI R. (1984). *Artificial & Human Intelligence*. Elsevier Science.

GENTNER D. (1983). Structure mapping : A theoretical model for analogy. *Cognitive Science*, **7**(2), 155–170.

HATHOUT N. (2001). Analogies morpho-synonymiques. Une méthode d'acquisition automatique de liens morphologiques à partir d'un dictionnaire de synonymes. In *Actes de TALN-2001*, p. 223–232, Tours.

HOFFMAN R. R. (1995). Monster analogies. *AI Magazine*, **11**, 11–35.

HOFSTADTER D. & THE FLUID ANALOGIES RESEARCH GROUP (1994). *Fluid Concepts and Creative Analogies*. New York : Basic Books.

HOLYOAK K. & THAGARD P. (1995). *Mental leaps : Analogy in creative thought*. Cambridge, MA : MIT Press.

ITKONEN E. & HAUKIOJA J. (1997). *A rehabilitation of analogy in syntax (and elsewhere)*, In *In : (Kertész, 1997)*, p. 131–177. Peter Lang.

JOSHI A. (1985). *Tree adjoining grammars : How much context-sensitivity is required to provide reasonable structural description ?*, p. 206–250. In : (Dowty *et al.*, 1985).

JOSHI A., VIJAY-SHANKER K. & WEIR D. (1991). *The Convergence of Mildly Context-Sensitive Grammar Formalisms*, p. 31–81. In : (Sells *et al.*, 1991).

KERTÉSZ A. (1997). *Metalinguistik im Wandel : die kognitive Wende in Wissenschaftstheorie und Linguistik*. Frankfurt a/M : Peter Lang.

LAVIE R. J. (2003). *Le locuteur analogique ou la grammaire mise à sa place*. Thèse de doctorat, Université Paris X Nanterre.

LEPAGE Y. (2000). Languages of analogical strings. In *Proceedings of COLING-2000*, volume 1, p. 488–494, Saarbrücken.

LEPAGE Y. (2003). *De l'analogie rendant compte de la commutation en linguistique*. Mémoire

d'habilitation à diriger les recherches, Université de Grenoble.

LEPAGE Y. (2004). Lower and higher estimates of the number of "true analogies" between sentences contained in a large multilingual corpus. In *Proceedings of COLING-2004*, volume 1, p. 736–742, Genève.

LEPAGE Y. & DENOUAL E. (2005). Automatic generation of paraphrases to be used as translation references in objective evaluation measures of machine translation. In *Proceedings of the third international workshop on Paraphrasing (IWP 2005)*, p. 57–64, Jeju.

LEPAGE Y. & PERALTA G. (2004). Using paradigm tables to generate new utterances similar to those existing in linguistic resources. In *Proceedings of LREC-2004*, volume 1, p. 243–246, Lisbonne.

NAGAO M. (1984). *A Framework of a Mechanical Translation between Japanese and English by Analogy Principle*, In *(Elithorn et Banerji, 1984)*, p. 173–180.

NEUVEL S. & SINGH R. (in press). Vive la différence ! what morphology is about. *Linguistica Folia*, **35**, 9 pages.

SATO S. (1991). *Example-based Machine Translation*. Ph.d. thesis, Kyoto University.

P. SELLS, S. SHIEBER & T. WASOW, Eds. (1991). *Foundational Issues in natural language processing*. Cambridge : MIT Press.

SHIEBER S. M. (1985). Evidence against the context-freeness of natural language. *Linguistics and Philosophy*, **8**, 333–343.

R. SINGH, Ed. (2000). *The Yearbook of South Asian Languages and Linguistics-200*. Thousand Oaks : Sage.

SINGH R. & FORD A. (2000). *In praise of Sakatayana : some remarks on whole word morphology*. (Singh, 2000).

SKOUSEN R. (1989). *Analogical modeling of language*. Dordrecht : Kluwer.

STROPPA N. (2005). *Définitions et caractérisation de modèles à base d'analogies pour l'apprentissage automatique des langues naturelles*. Thèse de doctorat, École nationale supérieure des télécommunications.

TAKEZAWA T., SUMITA E., SUGAYA F., YAMAMOTO H. & YAMAMOTO S. (2002). Toward a broad coverage bilingual corpus for speech translation of travel conversation in the real world. In *Proceedings of LREC 2002*, p. 147–152, Las Palmas.

TURNEY P. D. & LITTMAN M. (2005). Corpus-based learning of analogies and semantic relations. *Machine Learning*, **60**(1–3), 251–278.

YVON F. (1994). Paradigmatic cascades : a linguistically sound model of pronunciation by analogy. In *Proceedings of ACL-EACL-97*, p. 428–435, Madrid.

Pour plus de références sur le sujet, on pourra consulter les bibliographies de (Hoffman, 1995), (Lepage, 2003) et (Stroppa, 2005).