TALN 2003, Batz-sur-Mer, 11–14 juin 2003

Introduction à la traduction guidée par l'exemple (Traduction par analogie)

Michael Carl
Institut für Angewandte Informationsforschung,
Martin-Luther-Straße 14,
66111 Saarbrücken, Germany,
carl@iai.uni-sb.de

Mots-clefs – Keywords

traduction guidée par l'exemple, traduction par analogie, traduction statistique, induction de grammaire de traduction
example-based machine translation, analogical translation, statistical machine translation, induction of translation grammar

Résumé - Abstract

Le nombre d'approches en traduction automatique s'est multiplié dans les dernières années. Il existe entre autres la traduction par règles, la traduction statistique et la traduction guidée par l'exemple. Dans cet article je decris les approches principales en traduction automatique. Je distingue les approches qui se basent sur des règles obtenues par l'inspection des approches qui se basent sur des exemples de traduction. La traduction guidée par l'exemple se caractérise par la phrase comme unité de traduction idéale. Une nouvelle traduction est génerée par analogie : seulement les parties qui changent par rapport à un ensemble de traductions connues sont adaptées, modifiées ou substituées.

Je présente quelques techniques qui ont été utilisées pour ce faire. Je discuterai un système spécifique, EDGAR, plus en detail. Je démontrerai comment des textes traduits alignés peuvent être preparés en termes de compilation pour extraire des unités de traduction sous-phrastiques. Je présente des résultats en traduction Anglais $\rightarrow$ Français produits avec le système EDGAR en les comparant avec ceux d'un système statistique.

In this paper I characterize a number of machine translation approaches: rule-based machine translation (RBMT), statistical machine translation (SMT) and example-based machine translation (EBMT). While RBMT systems make use of hand-build rules, SMT and EBMT systems explore and re-use a set of reference translations. EBMT systems are rooted in analogical reasoning, where the ideal translation unit is the sentence. Only if an identical sentence cannot be found in the reference material, EBMT systems modify, substitute and adapt sequences of the retrieved examples to generate a suitable translation.

I discuss runtime and compilation time techniques and I present a system, EDGAR, in more detail. I show how translation units are extracted off-line and how they are re-used during translation. The description of a series of experiments for the translation English $\rightarrow$ French conclude this paper. An extended bibliography provides further pointer for interested readers.

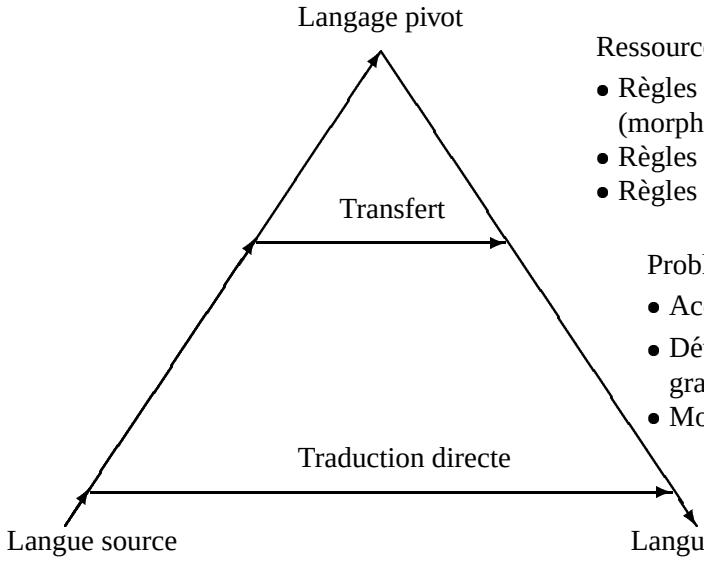
1 Contenu

- Caractérisation des approches à la traduction automatique
  — Traduction basée sur des règles, traduction statistique et traduction guidée par l'exemple
- Traduction guidée par l'exemple (EBMT)
  — Approches en termes de temps d'exécution et approches en termes de compilation
- Le système EDGAR
  — Segmentation, généralisation, spécification et raffinement
- Acquisition de grammaires de traduction
  — Propriétés de grammaires de traduction
  — Algorithme d'extraction de grammaire
- Expériences en traduction Anglais → Français
  — Acquisition des grammaires Anglais → Français à partir du Canadian Hansards
  — Echelonnement de l'approche
  — Comparaison avec *BabelFish* et traduction statistique
  — Intégration EBMT et SMT

2 Caractérisation des approches à la traduction automatique

2.1 Traduction automatique basée sur règles (RBMT)

Des approches en *traduction basée sur règles* (RBMT) sont fréquemment présentés par la pyramide de (Vauquois, 1968) (voir en dessous). Ces systèmes contiennent typiquement une série de fonctions qui analysent les phrases à traduire : analyses morphologiques, syntaxiques et/ou sémantiques, un module de transfert de la langue source en langue cible qui dépend du degré d'abstraction de la représentation du système, et une série de fonctions qui genèrent la phrase cible. Ces fonctions sont controlées par des dictionnaires et par des grammaires qui sont le plus souvent obtenues par l'inspection d'un (ou d'un groupe de) linguiste(s). Ceci a pour conséquence un développement lent du système principalement dû au problème d'acquisition de connaissances car les problèmes linguistiques de traduction doivent être d'abord complètement compris avant de les formuler en termes de règles. Mais beaucoup de problèmes en traduction automatique n'ont pas (encore) été entièrement compris ou requièrent une analyse complète sémantique et pragmatique ce qui n'est pas toujours disponible dans la plus part des cas.



Ressources :
- Règles linguistiques d'analyse (morphologique, syntaxique et/ou sémantique)
- Règles de transfert lexical et structurel
- Règles linguistiques pour la génération

Problèmes:
- Acquisition et maintenance des connaissances
- Développement et maintenance de grandes ressource lexicales et grammaticales
- Modélisation linguistique détaillé

2.2 Traduction basée sur des données

La *traduction basée sur des données* (Corpus-based Machine Translation (CBMT), mais aussi Data-driven Machine Translation) subsume un ensemble de méthodes alternatives et récentes qui visent à resoudre le problème de l'acquisition des connaissances en traduction par règles. Ces méthodes utilisent des textes bilingues traduits qui sont consultés lors de la traduction d'un texte ou d'une phrase nouvelle. Les textes bilingues sont alignés en segments de manière suivante:

Texte bilingue aligné (extrait du Canadian Hansard)

1	LA CHARTE CANADIENNE DES DROITS ET LIBERTÉS	canadian charter of rights and freedoms
2	L'hon. Benoît Bouchard (secrétaire d'État du Canada):	Hon. Benoît Bouchard (Secretary of State of Canada):
3	Monsieur le Président, je voudrais porter à l'attention de la Chambre que nous célébrons aujourd'hui, comme le savent les honorables députés, l'anniversaire de la proclamation de la Charte canadienne des droits et libertés qui a eu lieu le 17 avril 1982, ainsi que son parachèvement, il y a un an, avec l'entrée en vigueur des dispositions garantissant l'égalité à tous les membres de notre société.	Mr. Speaker, I would like to bring to the attention of the House that today, as Hon. Members are no doubt aware, we are celebrating the anniversary of the proclamation of the Canadian Charter of Rights and Freedoms which took place on April 17, 1982, and also of the coming into effect a year ago of the provisions guaranteeing equality for all members of our society.

Parmis le paradigme CBMT, deux directions principales peuvent être distinguées : la *traduction statistique* et la *traduction guidée par l'exemple*.

2.2.1 Traduction statistique (SMT)

La *traduction statistique* (SMT) se base sur la théorie mathématique de distribution et d'estimation probabiliste développée par Frederick Jelinek au *IBM* T.J. Watson Research Center et — en particulier — sur un article de (Brown et al., 1990). Les systèmes statistiques apprennent un modèle probabiliste de traduction ($Pr(t|s)$) à partir d'un texte bilingue et un modèle probabiliste de la langue cible ($Pr(t)$) à partir d'un texte monolingue. En temps d'exécution, la meilleure traduction pour une phrase nouvelle est recherchée grâce à la maximisation de ces deux modèles probabilistes.

$$arg\ max Pr(t|s) = arg\ max\{Pr(t) * Pr(s|t)\}$$

- modèle de traduction $Pr(s|t)$
- modèle de langue $Pr(t)$

- Approche d'apprentissage non-supervisée basée sur les formes fléchies.
- La traduction cible est synthétisée à partir de traduction(s) de mots individuels.
- Grande quantité de textes bilingues alignés nécessaire pour l'entraînement.

Typiquement, RBMT et SMT génèrent la phrase cible à partir des traductions de mots simples et isolés. La 'meilleure' traduction est déterminée:

en SMT par les probabilités de $Pr(s|t)$ et $Pr(t)$
en RBMT par des contraintes exprimées par des règles

2.2.2 La traduction guidée par l'exemple (EBMT) : Traduction par Analogie

La *traduction guidée par l'exemple* (Example-Based Machine Translation, EBMT) prend sa place entre la RBMT et la SMT : beaucoup d'approches intègrent des règles et des techniques statistiques. Néanmoins il y a des caractéristiques qui distinguent l'EBMT de la SMT et de la RBMT :

- La 'phrase' est l'unité de traduction idéale.
- Traduction guidée par l'exemple consiste à :
  — rechercher les meilleurs exemple(s) de référence dans une base de données.
  — substituer, modifier et adapter des séquences différentes.

Beaucoup de techniques ont été utilisées et inventées en EBMT pour substituer, modifier et adapter les séquences de mots qui diffèrent dans les exemples de la base et les nouvelles phrases à traduire. Un excellent survol de ces techniques et de leur enjeu se trouve dans (Somers, 1999; Somers, 2003). Dans mon article je présente plus en detail :

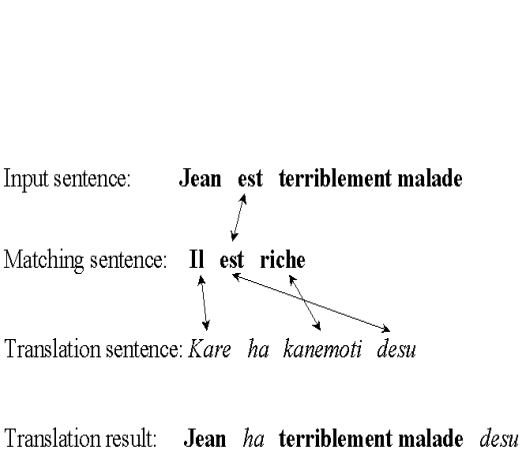
- Approches en termes de temps d'exécution
- Approches en termes de compilation
  — Représentations en schémas
  — Représentations en arbres syntaxiques

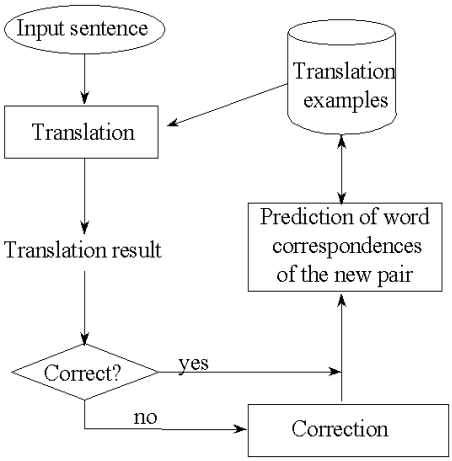
3 La traduction guidée par l'exemple (EBMT)

3.1 Approches en termes de temps d'exécution

3.1.1 Segmentation dynamique

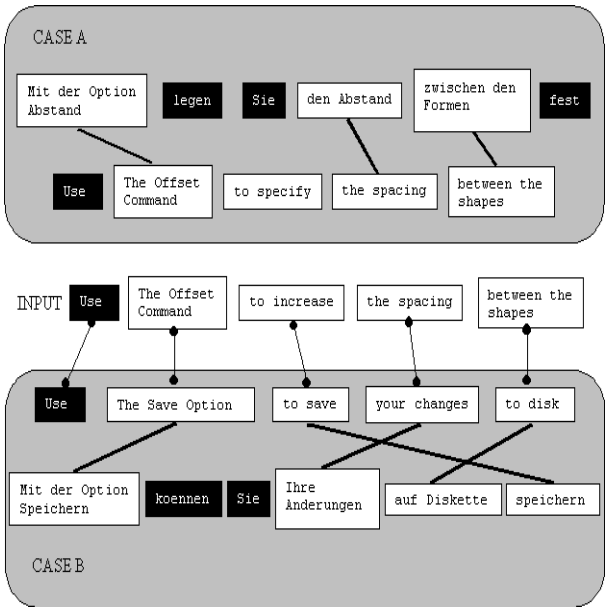
Dans l'approche proposé par (Andriamanankasina et al., 2003; Andriamanankasina et al., 1999) les exemples sont balisés et les correspondances entre les mots des deux phrases sont marqués. L'exemple le plus proche à la phrase nouvelle à traduire est rechérché et des séquences égaux sont traduites en langue cible. Ce processus est iteré jusqu'à ce que la phrase est entièrement traduite où il n'y as plus d'exemple proche disponible dans la base. La traduction peut être corrigée manuellement et inserée dans la base de manière dynamique. Andriamanankasina *et al.* montrent que ce *cycle d'apprentissage* ameliore les résultats de traduction obtenu.





3.1.2 Adaptabilité versus similarité en recherche

Dans le système de (Collins & Cunningham, 1997; Collins, 1998), voir aussi : (Collins & Somers, 2003), les exemples sont balisés et segmentés et portent l'information du rôle syntaxique. Les segments correspondants sont connectés d'une langue à l'autre. Le processus de recherche inclut une mesure d'adaptabilité qui indique la similarité de l'exemple par rapport à son contexte externe. La notion *adaptation-guided retrieval* (recherche guidée par l'adaptabilité) indique le degré auquel les exemples retrouvés sont un bon modèle pour la traduction desiré : alors que le "CASE A" est plus similaire du "INPUT", "CASE B" est le meilleur modèle pour sa traduction dû à sa meilleure adaptabilité.



3.2 Approches en termes de compilation

3.2.1 Extraction "linguistic-light"

Güvenir et Cicekli (Güvenir & Cicekli, 1998; Cicekli & Güvenir, 1996; Cicekli & Güvenir, 2003) présentent un algorithme pour l'extraction des correspondances lexicales de deux exemples de traduction : des parties du côté source doivent correspondre aux parties similaires du côté cible et des chaînes de mots différentes en langue source doivent correspondre à des chaînes de mots différentes en cible. Ces correspondances sont apprises en forme de *schémas de traduction* (translation template). Un schéma de traduction est un exemple de traduction généralisé dont certaines parties ont été remplacées par des variables liées.

Deux exemples de traduction :

I took a	ticket	from Mary	↔	Mary'den bir	bilet	aldim
I took a	pen	from Mary	↔	Mary'den bir	kalem	aldim

Généralisation de différences et extraction de correspondances lexicales :

I took a	$\mathcal{X}_1$	from Mary	↔	Mary'den bir	$\mathcal{Y}_1$	aldim

ticket	↔	bilet
pen	↔	kalem

3.2.2 Extraction “linguistic-heavy” : Microsoft Research MT (MSR-MT)

(Richardson et al., 2001; Menezes & Richardson, 2003) utilisent des règles pour obtenir les formes logiques des exemples. Ces représentations sont connectées grâce à un lexique bilingue. Ensuite des connections ambiguës sont nettoyées avec des règles de préférence. Finalement des structures de transfert de haute qualité (ce qu’ils apellent des *transfer mappings*) sont extraites. Pour chaque structure de transfert la fréquence est calculées et un contexte suffisant est gardé pour distinguer les “mappings” ambiguës pendant la traduction.

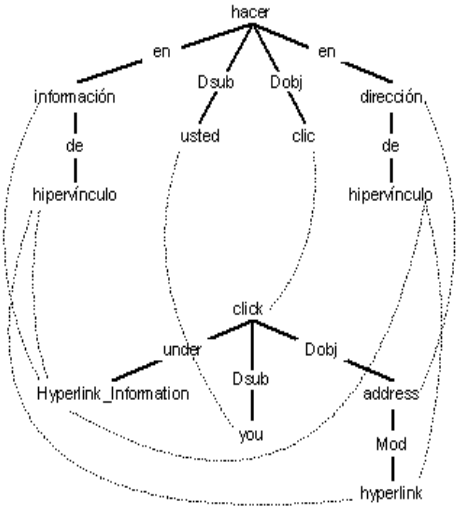
Exemple de traduction:

En Información del hipervínculo, haga clic en la dirección del hipervínculo.

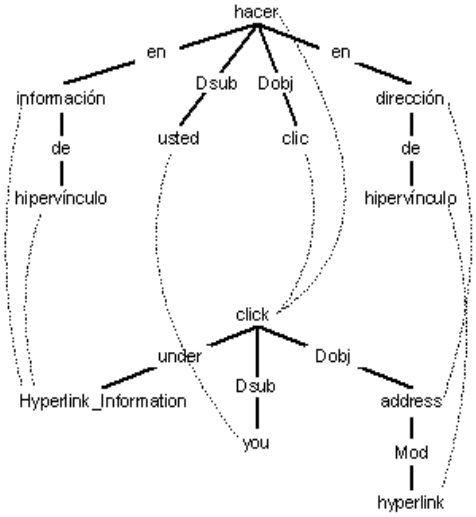
$\longleftrightarrow$

Under Hyperlink Information, click the hyperlink address.

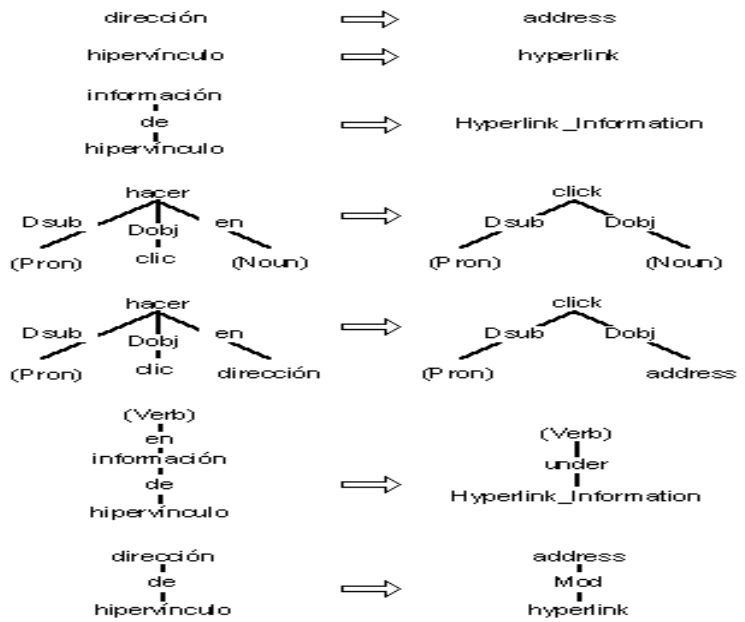
Correspondances lexicales entre les formes logiques (FL)



Alignement entre FL espagnol et anglais



Structures de transfert (transfer mappings) acquises de l’espagnol vers l’anglais



4 Présentation du système EDGAR

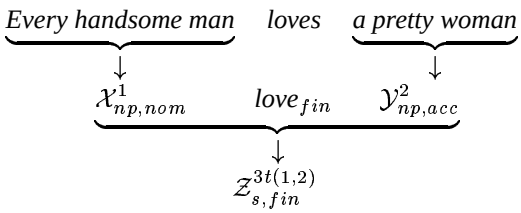
Le système EDGAR (Carl, 1999) utilise des analyseurs morphologique et syntaxique en plus des exemples de traduction. Un mécanisme d'induction généralise des exemples et produit une grammaire de traduction (Carl, 2003). La segmentation et généralisation d'une nouvelle phrase source ainsi que le raffinement de sa traduction dans la langue cible sont guidés par le contenu de la grammaire de traduction.

4.1 Segmentation, généralisation, spécification et raffinement

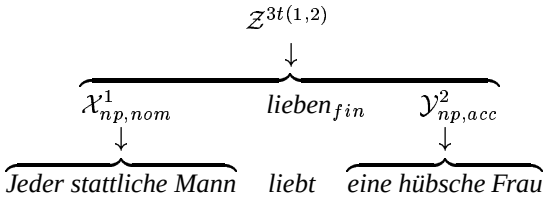
La grammaire de traduction contient des unités de traduction lexicales et des schémas de traduction variabilisés.

1 (*Every handsome man*)$_{np}$ $\longleftrightarrow$ (*Jeder stattliche Mann*)$_{np}$
2 (*a pretty woman*)$_{np}$ $\longleftrightarrow$ (*eine hübsche Frau*)$_{np}$
3 ($\mathcal{X}_{np}$ *love*$_{fin}$ $\mathcal{Y}_{np}$)$_{s}$ $\longleftrightarrow$ ($\mathcal{X}_{np}$ *lieben*$_{fin}$ $\mathcal{Y}_{np}$)$_{s}$

Segmentation et Généralisation



Spécification et Raffinement



4.2 Représentation dans le programme EDGAR

Les entrées dans la grammaire portent l'information morphologique et les lemmas sous forme d'attribut/valeur, des traits. Les traits d'une analyse d'un mot peuvent être complexe (p.ex. `agr` en bas) ou atomiques (p.ex. `lu` en bas). De plus, les traits peuvent être atomique disjonctifs (p.ex. `case=d;g`) ou comples disjonctifs. Par exemple, la représentation du mot allemand "der" porte les traits suivants:

$$\left\{\begin{array}{l}\texttt{lu=d\_art,c=w,sc=art,fu=def} \\ \texttt{agr=}\left\{\begin{array}{l}\texttt{gen=f,}\\\texttt{nb=sg,}\\\texttt{case=d;g}\end{array}\right\};\left\{\begin{array}{l}\texttt{gen=m,}\\\texttt{nb=sg,}\\\texttt{case=n}\end{array}\right\};\left\{\begin{array}{l}\texttt{nb=plu,}\\\texttt{case=g}\end{array}\right\}\end{array}\right\};\left\{\begin{array}{l}\texttt{lu=d\_rel,c=w,sc=rel,fu=np,} \\ \texttt{agr=}\left\{\begin{array}{l}\texttt{case=n,}\\\texttt{g=m,}\\\texttt{nb=sg}\end{array}\right\};\left\{\begin{array}{l}\texttt{case=g;d,}\\\texttt{nb=sg,}\\\texttt{g=f}\end{array}\right\}\end{array}\right\}$$

4.3 Percoler des traits avec des règles KURD

L'analyseur KURD (Carl & Schmidt-Wigger, 1998) sert à percoler les traits dans les arbres de dérivation et à unifier et substituer des valeurs dans les noeuds. La règle `NP` monte l'information d'accord des noeuds terminaux dans le noeud père.

```
NP  =  Aa {c=np} [
           e {c=w,sc=art,agr=_AGR},
          *a {c=adj,agr=_AGR},
          +e {c=noun,agr=_AGR}
       ]
    :  Au {agr=_AGR}
```

Les opérations de KURD

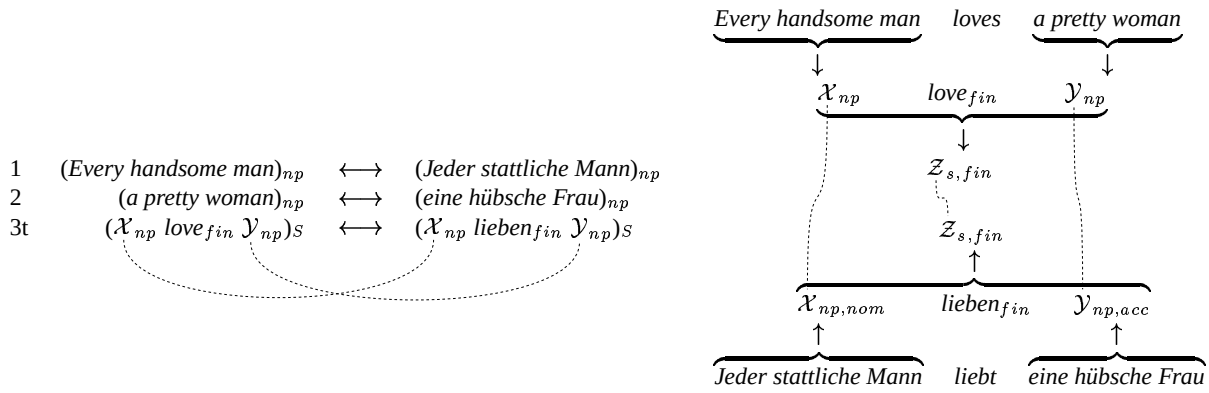
- Unification et suppression de traits.
- Concaténation et substitution de valeurs.
- Insertion et suppression de noeuds.

5 Acquisition de grammaires de traduction

5.1 Propriétés des grammaires de traduction

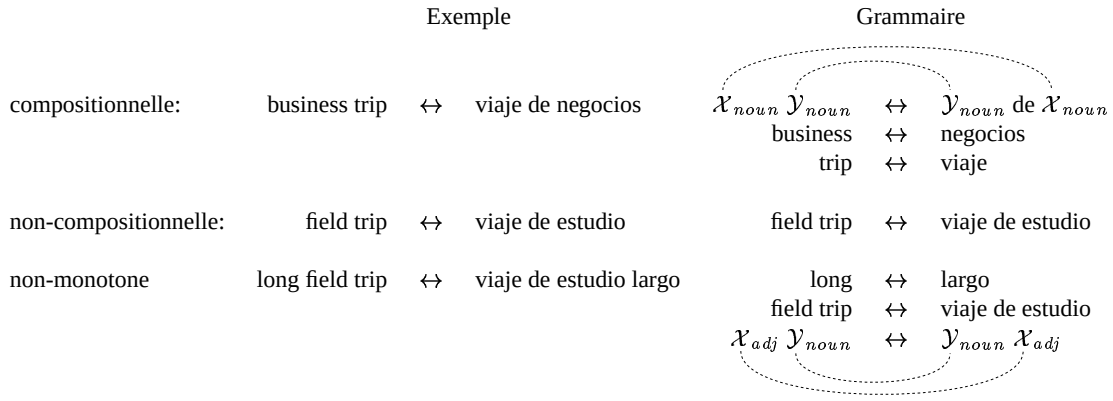
5.1.1 Grammaire de traduction homomorphe et arbres de dérivation isomorphes

Les grammaires *homomorphes* produisent des arbres *isomorphes* et rendent possible un transfert 1-à-1 de la source à la cible.



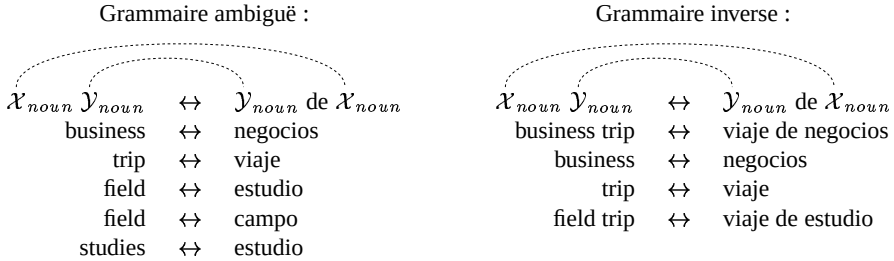
5.1.2 Traduction compositionelle versus non-monotone (partiellement compositionelle)

Les grammaires *compositionelles* segmentent la phrase source récursivement en expressions qui sont traduites indépendamment tandis que les grammaires non-monotones s'arrêtent à un certain point.



5.1.3 Grammaire ambiguë versus inverse

Les grammaires *ambiguës* permettent plusieurs traductions pour une expression source tandis que les grammaires *inverses* ne produisent qu'une seule traduction.

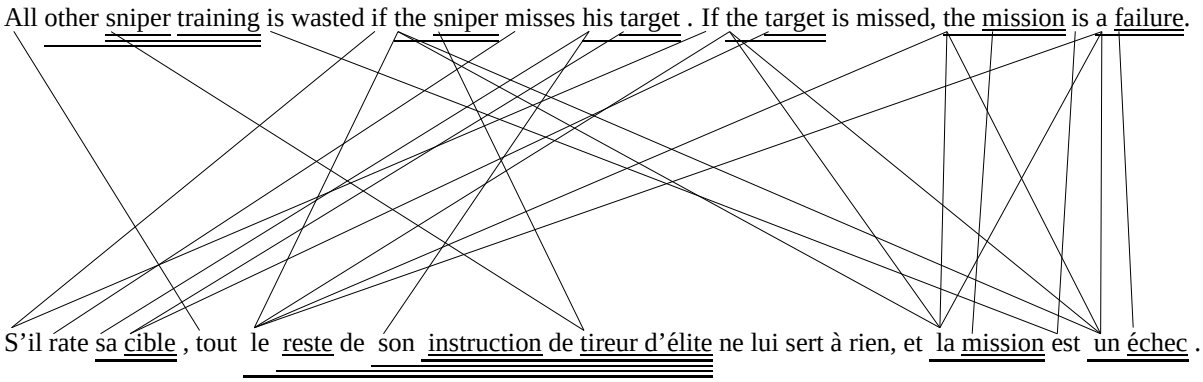


5.2 Extraction de grammaire de traduction : un algorithme

L'extraction de grammaires à partir d'exemples de traduction se pousuit en quatre étapes.

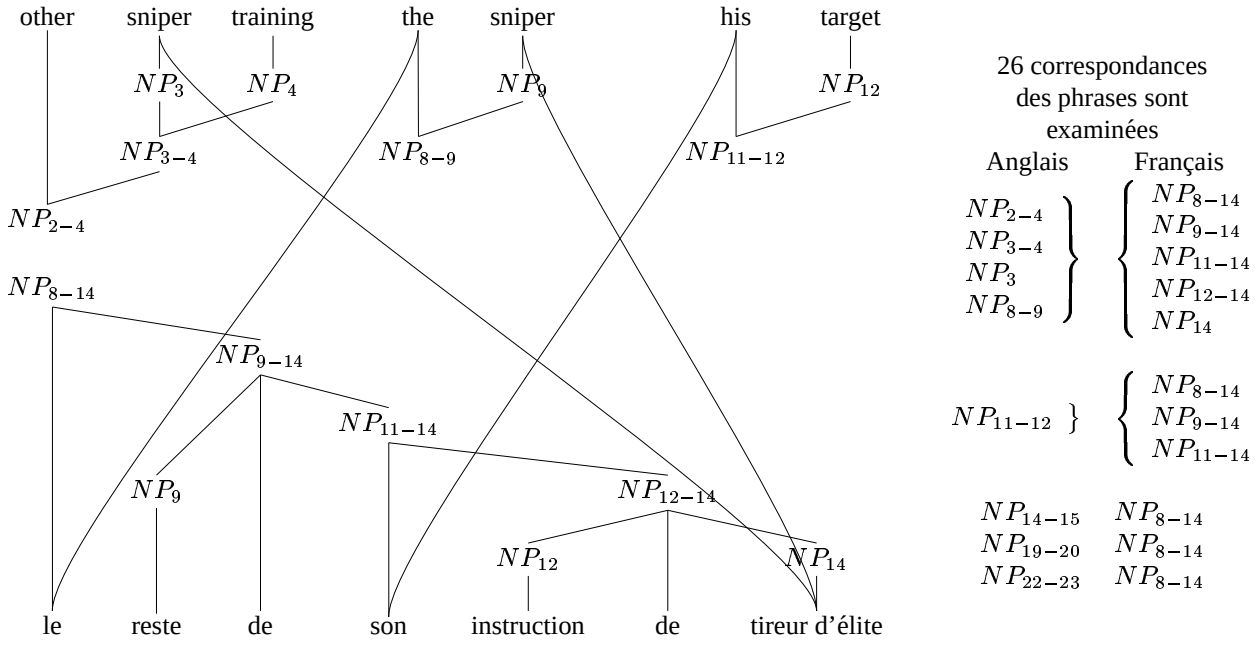
5.2.1 Alignement partiellement analysé et ancré avec un dictionnaire bilingue

D'abord l'analyse syntaxique partielle des deux côtés de l'alignement est effectuée. Des lexemes des deux côtés sont connectés grâce à un dictionnaire bilingue.



5.2.2 Détermination de correspondances phrase à phrase les plus significantes

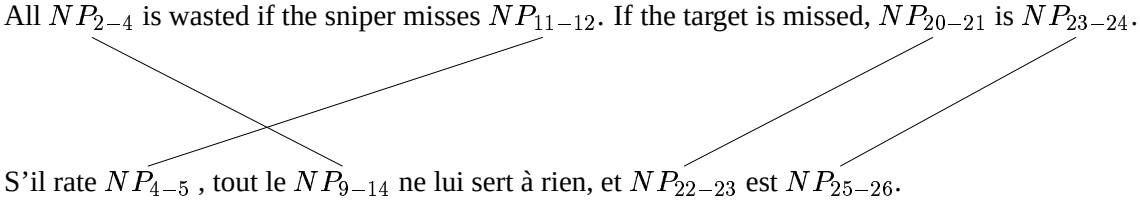
Les poids des correspondances des arbres sont calculés à partir (i) des poids et le nombre des ancres lexicales (ii) la fréquence des correspondances dans le texte et (iii) l'isomorphisme des analyses partielles (cf. (Carl, 2003))



Cet extrait d'alignement de la section 5.2.1 montre trois segments anglais differents connectés avec un segment français. Sont examinés 26 correspondances de phrases possibles dont la traduction $NP_{2-4} \leftrightarrow NP_{9-14}$: "other sniper training $\leftrightarrow$ reste de son instruction de tireur d'élite" est détectée la plus consistente dans le texte aligné.

5.2.3 Schémas de traduction générés

Les schémas sont générés par la substitution des correspondances compositionelles.



5.2.4 Grammaire de traduction générée

Une grammaire de traduction *compositionnelle* et *homomorphe* est extraite récursivement pour chaque exemple de traduction:

1	All other sniper training is wasted if the sniper misses his target. If the target is missed, the mission is a failure.	$\leftrightarrow$	S'il rate sa cible tout le reste de son instruction de tireur d'élite ne lui sert à rien, et la mission est un échec.
2	All NP^1 is wasted if the sniper misses NP^2. If the target is missed, NP^3 is NP^4.	$\leftrightarrow$	S'il rate NP^2 tout le NP_1 ne lui sert à rien, et NP^3 est NP^4.
3	other sniper training	$\leftrightarrow$	reste de son instruction de tireur d'élite
4	other NP^1 NP^2	$\leftrightarrow$	reste de son NP^2 de NP^1
5	training	$\leftrightarrow$	instruction
6	sniper	$\leftrightarrow$	tireur d'élite
7	his target	$\leftrightarrow$	sa cible
8	his NP^1	$\leftrightarrow$	sa NP^1
9	the mission	$\leftrightarrow$	la mission
10	the NP^1	$\leftrightarrow$	la NP^1
11	mission	$\leftrightarrow$	mission
12	a failure	$\leftrightarrow$	un échec
13	a NP^1	$\leftrightarrow$	un NP^1

6 Expériences en traduction guidée par l'exemple

Dans cette section nous générons des grammaires avec l'approche presenté en section 5. Un texte test est traduit avec EDGAR présenté en section 4. Une description plus étendue des expériences peut être trouvé dans (Carl & Langlais, 2003). Les expériences se basent sur le *Canadian Hansards*, texte bilingue Anglais $\leftrightarrow$ Français. Nous présentons des expériences différentes d'extraction de grammaire aussi bien en ce qui concerne le nombre d'exemples que le degré d'ambiguïté de la grammaire générée.

6.1 Extraction d'une grammaire de traduction

Les ressources utilisées pour extraire une grammaire de traduction (GT_1) Anglais $\leftrightarrow$ Français incluent un dictionnaire bilingue de 77.016 entrées, un programme de segmentation en segments et un ensemble de 50.000 exemples de traduction alignés du Canadian Hansard. Le dictionnaire couvre prèsque 3/4 des mots anglais et français du ET_1 mais contient seulement a peu près 1/3 des mots differents qui occurent dans les deux textes. Le programme de segmentation (parser partiel) génère au moyen 11 et 13 segments

rsp. par exemple de traduction pour l'anglais et le français. Plus que la moitie des segments différents (63% et 50% rsp.) font partie des règles lexicales de la grammaire inverse extraite.

	Anglais	Français
Exemples de Traduction ET_1 :		
#exemples	50.000	50.000
#mots	888.018	947.194
#mots different	17.915	23.675
#mots/exemples	17,76	18,94

	Anglais	Français
Dictionnaire bilingue (DIC) :		
#entrées du dictionnaire	77.016	77.016
%couvert en ET_1	74,77%	74,99%
#mots different	7.688	7.714
%anchors (en ET_1)	42,28%	43,53%

Grammaire de traduction extraite GT_1 :	
#règles lexicales	113.810
#schémas de traduction	70.153

Segments générés par le parser partiel :		
#segments	581,599	650,136
#segments différents	180,006	226,339

6.2 Traduction d'un texte test (TT)

Un texte test de 500 phrases est traduit de l'anglais vers le français avec les règles lexicales de GT_1. Alors que la couverture du dictionnaire bilingue (DIC) est plus grande que celle de la grammaire GT_1, la qualité de traduction, mesurée en WER[1] et en BLEU (Papineni et al., 2002), est mieux en traduction GT_1. Nous voyons ici une corrélation entre la qualité des traductions et la longueur des segments utilisés lors de la traduction.

	Texte Test (TT)	
	Anglais	Français
#exemples	500	500
#mots	8.665	9.806
#mots/exemples	17,33	19,61

	GT_1	DIC
%mots couverts	66,38%	66,99%
BLEU	0,1421	0,0573
WER	68,89%	81,68%
longueur segments ≥ 2		
#segments	966	146
#mots couverts	2,652	325
%mots couverts	30,61%	3,75%

6.3 Echelonnement de grammaires inverses et ambiguës

Dans cette expérience nous étudions (i) la capacité de l'algorithme d'utiliser un nombre différent d'exemples de traduction et (ii) l'effet de l'utilisation d'unités ambiguës. Nous comparons trois grammaires differéntes générées à partir d'ensembles différents d'exemples de traduction , tous extraites du Canadian Hansards.

	ET_0	ET_1	ET_2
#exemples de traduction	10.000	50.000	100.000
#mots anglais (En)	151.954	888.018	1.437.450
#mots français (Fr)	163.113	947.194	1.503.196
#mots différents En	7.343	17.915	22.501
#mots différents Fr	9.528	23.675	29.559

Ces trois ensembles de réference sont utilisés afin de générer deux types de grammaires différentes : des grammaires inverses GT_0, GT_1 et GT_2 (dont GT_1 est égale à celle des sections 6.1 et 6.2) et des grammaires ambiguës GT_0^a, GT_1^a et GT_2^a. Les grammaires ambiguës contiennent près que 20% plus de règles de transfert lexical, alors que le nombre de mots différents reste à peu près pareil dans les deux

[1] Les chiffres WER supérieures et chiffres BLEU inferieures indiquent le meilleur résultat de traduction.

types de grammaires. On obverve aussi que le nombre moyen de mots par règle augmente dans les grammaires plus grandes.

	Règles inverses de transfert lexical			Règles ambiguës de transfert lexical		
	GT_0	GT_1	GT_2	GT_0^a	GT_1^a	GT_2^a
#règles lexicales	23.214	113.810	180.745	28.393	146.684	220.248
#mots Anglais (En)	203.426	1.223.260	1.856.392	222.473	1,355.331	2,030.390
#mots Français (Fr)	220.273	1.314.197	2.007.322	244.615	1,491.559	2,219.455
#mots différents En	7.338	17.910	22.488	7.340	17.914	22.495
#mots différents Fr	9.520	23.659	29.523	9.521	23.670	29.542

En ce qui concerne la qualité des traductions produites, les deux types de grammaire produisent un taux de WER et BLEU à peu près égal. Ceci alors qu'un nombre considérable de segments de longueur supérieure à été utilisé pour produire la traduction avec des grammaires ambiguës. Nous concluons que les entrées ambiguës représentent pour la plupart des unités de traduction de qualité inferieure.

Qualité du texte test en grammaires inverses ...				et qualité en grammaires ambiguë		
	GT_0	GT_1	GT_2	GT_0^a	GT_1^a	GT_2^a
WER	71,91%	68,89%	66,93%	71,88%	69,75%	67,22%
BLEU	0,1365	0,1421	0,1704	0,1398	0,1519	0,1706
#segments	3.581	4.405	4.685	3.599	4.314	4.659
#segments différents	992	1.279	1.387	1.055	1.343	1.450
#mots couverts	4.611	5.752	6.146	4.680	5.752	6.228
#segments longueur ≥ 2	767	966	1.050	816	1.032	1.108
#segments different	353	519	589	380	570	646
#mots couverts	1.952	2.652	2.863	2.170	2.844	3.062

6.4 Comparaison de GT, SMT et *BabelFish*

Dans cette expérience nous comparons les résultats de traduction obtenus utilisant les grammaires GT_{0-2} avec un système statistique (Langlais, 2002) entrainé sur les mêmes exemples de traduction ET_{0-2}. Nous voyons que les résultats SMT sont inférieurs (toujours WER et BLEU) à ceux obtenus en GT. Le système SMT_3 qui a été entraîné sur un texte de 1,5 millions de exemples (15 fois plus que ET_2) obtient les meilleurs résultats.

score	GT_0	GT_1	GT_2	SMT_0	SMT_1	SMT_2	SMT_3	*BabelFish*
BLEU	0,1365	0,1421	0,1704	0,1156	0,1231	0,1378	0,2061	0,1578
WER	71,91%	68,89%	66,93%	74,72%	73,54%	71,52%	61,66%	66,03%

Le système commercial *BabelFish* obtient des résultats inférieurs à ceux de SMT_3 et GT_2. Ceci est surtout dû aux particularités du texte traduit : alors que GT et SMT apprennent les traductions particulières, *BabelFish* n'a pas pu être adapté à ce type de texte. Ainsi, la traduction : "the speaker $\leftrightarrow$ le président" a été réalisée par GT et SMT alors que *BabelFish* génère la traduction "le haut-parleur". De même : "some hon. members : oh , oh ! $\leftrightarrow$ des voix : oh , oh !" est une traduction qui se voit fréquemment en Canadian Hansards mais *BabelFish* produit "membres d'un certain hon : l' OH OH". Alors que ce sont des traductions possibles correctes dans d'autres contextes, elles sont erronées quant à la traduction du Canadian Hansards.

6.5 Intégration SMT et GT

Finalement nous essayons d'intégrer les grammaires GT et le système statistique suivant (Langlais, 2002) : quand la grammaire GT contient une entrée égale à une séquence de mots dans la phrase à traduire, le système SMT est forcé d'intégrer la traduction proposée par GT dans sa sortie. La qualité produite du système hybride est meilleure quant aux grammaires inverses (SMT_{0-2}-GT_{0-2}); pour l'intégration des grammaires ambiguës dans le systeme statistique (SMT_{0-2}-GT^a_{0-2}) une amelioration des résultats n'a pas pu être observé.

	SMT_0-GT_0	SMT_1-GT_1	SMT_2-GT_2	SMT_0-GT^a_0	SMT_1-GT^a_1	SMT_2-GT^a_2
BLEU	0.1495	0.1684	0.1789	0.1406	0.1541	0.1654
WER	71.19%	70.32%	68.94%	72.70%	72.45%	71.41%

7 Résume et conclusion

Dans cet article je présente d'approches en traduction automatique. Je fais la distinction entre la traduction par règles (RBMT), la traduction statistique (SMT) et la traduction guidée par l'exemple (EBMT). Les ressources necessaires en RBMT sont obtenues par l'inspection d'un (ou d'un groupe de) linguiste(s), tandis que les approches EBMT et SMT extraient les connaissances de traduction à partir des textes bilingues alignés. Au contraste à la SMT, la 'phrase' est l'unité de traduction idéale en EBMT. Je présente des systèmes EBMT qui extraient et acquirent ces unités en termes de temps d'exécution et en termes de temps de compilation.

En suite je discute plus en détail le système EDGAR. A partir des exemples de traduction, EDGAR produit la traduction des phrases nouvelles par analogie de manière compositionnelle et isomorphe. Je présente un algorithme pour extraire une grammaire de traduction à partir des exemples de traduction. L'article conclu avec la description d'une série de expériences en traduction guidée par l'exemple. De ces expériences nous concluons que :

- La couverture de la grammaire est fonction du nombre des exemples de référence.
- Les grammaires produisent une meilleure qualité de traduction que la traduction SMT (taille identique de référence)
- Les règles ambiguës n'améliorent pas la qualité de la traduction.
- L'intégration des techniques EBMT et SMT ameliore les résultats de la traduction.

References

Al-Adhaileh, M. H. & Tang E. K. 1999. Example-Based Machine Translation Based on the Synchronous SSTC Annotation Schema. *Machine Translation Summit VII*, Singapore, 244–249.

Andriamanankasina, T., K. Araki, & K, Tochinai. 2003. Ebmt of pos-tagged sentences with inductive learning. In *(Carl & Way, 2003)*.

Andriamanankasina, T., K. Araki & K. Tochinai. 1999. Example-Based Machine Translation of Part-Of-Speech Tagged Sentences by Recursive Division. *Machine Translation Summit VII*, Singapore, 509–517.

Brown, P. F., J. Cocke, S. A. Della Pietra, V. J. Della Pietra, F. Jelinek, J. D. Lafferty, R. L. Mercer & P. S. Roossin. 1990. A Statistical Approach to Machine Translation. *Computational Linguistics* **16**, 79–85.

Brown, R. D. 1996. Example-Based Machine Translation in the Pangloss System. *Coling* (1996), 169–174.

Brown, R. D. 1997. Automated Dictionary Extraction for “Knowledge-Free” Example-Based Translation. *TMI* (1997), 111–118.

Brown, R. D. 1999. Adding Linguistic Knowledge to a Lexical Example-based Translation System. *TMI* (1999), 22–32.

Carl, M. & A. Way, editors. 2003. *Recent Advances in Example-Based Machine Translation.* Kluwer, Academic Publisher, Boston/Dordrecht/London. in press.

Carl, M. 2003. Inducing translation grammars from bracketed alignments. In *(Carl & Way, 2003).*

Carl, M. & Langlais, P. 2003. Tuning general purpose translation knowledge to a sublanguage. In *Proceedings of EAMT/CLAW*.

Carl, M. 1999. Inducing Translation Templates for Example-Based Machine Translation. *Machine Translation Summit VII*, Singapore, 250–258.

Carl, M. & Schmidt-Wigger, A.. 1998. Shallow Postmorphological Processing with KURD. In *Proceedings of NeMLaP3/CoNLL98*, pages 257–265, Sydney.

Cicekli, I. & H.A. Güvenir. 2003. Learning Translation Templates from Bilingual Translation Examples. In *(Carl & Way, 2003).*

Cicekli, I. & H. A. Güvenir. 1996. Learning Translation Rules From A Bilingual Corpus. *NeMLaP-2: Proceedings of the Second International Conference on New Methods in Language Processing,* Ankara, Turkey, 90–97.

Collins, B. & H. Somers. 2003. EBMT Seen as Case-based Reasoning. In *(Carl & Way, 2003).*

Collins, B. 1998. *Example-Based Machine Translation: An Adaptation-Guided Retrieval Approach*. PhD thesis, Trinity College, Dublin.

Collins, B. & P. Cunningham. 1997. Adaptation Guided Retrieval: Approaching EBMT with Caution. *TMI* (1997), 119–126.

Cranias, L., H. Papageorgiou & S. Piperidis. 1994. A Matching Technique in Example-Based Machine Translation. *Coling* (1994), 100–104.

Furuse, O. & H. Iida. 1992. An Example-Based Method for Transfer-Driven Machine Translation. *TMI* (1992), 139–150.

Furuse, O. & H. Iida. 1994. Constituent Boundary Parsing for Example-Based Machine Translation. *Coling* (1994), 105–111.

Güvenir, H. A. & I. Cicekli. 1998. Learning Translation Templates from Examples. *Information Systems* **23**, 353–363.

Kaji, H., Y. Kida & Y. Morimoto. 1992. Learning Translation Templates from Bilingual Text. *Coling* (1992), 672–678.

Langlais, P. 2002. Ressources terminologiques et traduction probabiliste: premiers pas positifs vers un système adaptatif. In *TALN-2002*.

Matsumoto, Y. & M. Kitamura. 1995. Acquisition of Translation Rules from Parallel Corpora. In R. Mitkov & N. Nicolov (eds) *Recent Advances in Natural Language Processing: Selected Papers from RANLP’95,* Amsterdam: John Benjamins, 405–416.

McTait, K. & A. Trujillo. 1999. A Language-Neutral Sparse-Data Algorithm for Extracting Translation Patterns. *TMI* (1999), 98–108.

Menezes, A. & S.D. Richardson. 2003. A best-first alignment algorithm for automatic extraction of transfer mappings from bilingual corpora. In *(Carl & Way, 2003)*.

Meyers, A., R. Yangarber, R. Grishman, C. Macleod & A. Moreno-Sandeval. 1998. Deriving Transfer Rules from Dominance-Preserving Alignments. *Coling-ACL* (1998), 843–847.

Nagao, M. 1984. A Framework of a Mechanical Translation between Japanese and English by Analogy Principle. In A. Elithorn and R. Banerji (eds) *Artificial and Human Intelligence*, 173–180, Amsterdam: North-Holland.

Nirenburg, S., S. Beale & C. Domashnev. 1994. A Full-Text Experiment in Example-Based Machine Translation. *International Conference on New Methods in Language Processing (NeMLaP)*, Manchester, England, 78–87.

Papineni, K., S. Roukos, T. Ward, & W.-J. Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the ACL*, Philadelphia, Pennsylvania, USA, 311–318.

Richardson, S.D., W.B. Dolan, A. Menezes & J. Pinkham. 2001. Achieving Commercial-quality Translation with Example-based Methods. *MT Summit VIII: Machine Translation in the Information Age*, Santiago de Compostela, Spain, 293–298.

Sato, S. & M. Nagao. 1990. Toward Memory-Based Translation. *Coling* (1990), Vol. 3, 247–252.

Somers, H. 1999. Review Article: Example-based Machine Translation. *Machine Translation*, 14(2):113–157.

Somers, H. 2003. An Overview of EBMT. In *(Carl & Way, 2003)*.

Sumita, E. & H. Iida. 1991. Experiments and Prospects of Example-Based Machine Translation. *29th Annual Meeting of the Association for Computational Linguistics*, Berkeley, California, 185–192.

Sumita, E., H. Iida & H. Kohyama. 1990. Translating with Examples: A New Approach to Machine Translation. *TMI* (1990), 203–212.

Vauquois, B. 1968. A Survey of Formal Grammars and Algorithms for Recognition and Transformation in Machine Translation. *IFIP Congress-68*, Edinburgh, 254–260; reprinted in Ch. Boitet (ed.) *Bernard Vauquois et la TAO: Vingt-cinq Ans de Traduction Automatique – Analectes*, 201–213, Grenoble (1988): Association Champollion.

Veale, T. & A. Way. 1997. *Gaijin*: A Bootstrapping Approach to Example-Based Machine Translation. *International Conference, Recent Advances in Natural Language Processing*, Tzigov Chark, Bulgaria, 239–244.

Watanabe, H. 1992. A Similarity-Driven Transfer System. *Coling* (1992), 770–776.

Watanabe, H. 1993. A Method for Extracting Translation Patterns from Translation Examples. *TMI* (1993), 292–301.

Watanabe, H. & K. Takeda. 1998. A Pattern-Based Machine Translation System Extended by Example-Based Processing. *Coling-ACL* (1998), 1369–1373.

Way, A. 1999. A Hybrid Architecture for Robust MT using LFG-DOP. *Journal of Experimental and Theoretical Artificial Intelligence* **11**, 441–471.