Alignement de mots par apprentissage artificiel de règles de propagation syntaxique en corpus de taille restreinte

Sylwia Ozdowska (1), Vincent Claveau (2)

(1) ERSS - Université de Toulouse le Mirail
5 allées Antonio Machado
31058 Toulouse Cedex 1
ozdowska@univ-tlse2.fr

(2) OLST - Université de Montréal
CP 6128 succ. Centre-Ville
Montréal, QC, H3C 3J7, Canada
vincent.claveau@umontreal.ca

Mots-clefs : alignement de mots, corpus alignés, apprentissage artificiel, programmation logique inductive, analyse syntaxique

Keywords: word alignment, aligned corpus, machine learning, inductive logic programming, syntactic analysis

Résumé Cet article présente et évalue une approche originale et efficace permettant d'aligner automatiquement un bitexte au niveau des mots. Pour cela, cette approche tire parti d'une analyse syntaxique en dépendances des bitextes effectuée par les outils SYNTEX et utilise une technique d'apprentissage artificiel, la programmation logique inductive, pour apprendre automatiquement des règles dites de propagation. Celles-ci se basent sur les informations syntaxiques connues pour ensuite aligner les mots avec une grande précision. La méthode est entièrement automatique, et les résultats évalués sur les données de la campagne d'alignement HLT montrent qu'elle se compare aux meilleures techniques existantes. De plus, alors que ces dernières nécessitent plusieurs millions de phrases pour s'entraîner, notre approche n'en requiert que quelques centaines. Enfin, l'examen des règles de propagation inférées permet d'identifier facilement les cas d'isomorphismes et de non-isomorphismes syntaxiques entre les deux langues traitées.

Abstract This paper presents and evaluates an effective yet original approach to automatically align bitexts at the word level. This approach relies on a syntactic dependency analysis of the texts provided by the tools SYNTEX and uses a machine-learning technique, namely inductive logic programming, to automatically infer rules called propagation rules. These rules make the most of the syntactic information to precisely align words. This approach is entirely automatic, and results obtained on the data of the HLT evaluation campaign rival the ones of the best existing alignment systems. Moreover, our system uses very few training data: only hundreds of sentences compared to millions for the existing systems. Furthermore, syntactic isomorphisms between the two spotted languages are easily identified through a linguistic examination of the inferred propagation rules.

1 Introduction

L'enjeu que représente l'alignement des corpus parallèles au niveau des mots n'est plus à démontrer : ce dernier trouve ses applications dans des tâches telles que la traduction automatique ou encore la construction de ressources lexicales bi ou multilingues (Véronis, 2000). Il existe principalement deux types d'approches pour aligner des mots : celles à dominante statistique qui s'appuient notamment sur les modèles IBM (Brown *et al.*, 1993), et celles qui tendent à combiner calculs statistiques simples et différentes sources d'information linguistique (Ahrenberg *et al.*, 2000 ; Barbu, 2004). Destinés principalement à la traduction automatique, les systèmes purement statistiques se sont progressivement enrichis en incorporant des données linguistiques issues de l'analyse syntaxique (Lin & Cherry, 2003 ; Ding & Palmer, 2004) et ce afin de mieux prendre en compte les variations systématiques entre les langues impliquées dans le processus de traduction (Dorr, 1994 ; Fox, 2002). L'alignement sur des bases purement syntaxiques a également fait l'objet de travaux : D. Wu (2000) a par exemple proposé une méthode basée sur une analyse en constituants ; S. Ozdowska (2004), dont nous reprenons le cadre expérimental, utilise quant à elle une analyse en dépendances dans le but de proposer une étude contrastive fine des divergences syntaxiques entre le français et l'anglais. Sa démarche a consisté à définir manuellement des règles d'alignement, dites de propagation, qui exploitent les relations de dépendance mises en évidence dans chaque partie d'un corpus parallèle.

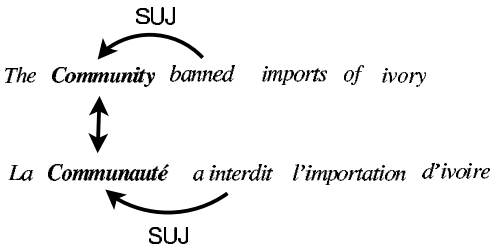
Cet article présente une technique d'alignement proche de cette dernière. Cependant, l'originalité de notre démarche réside dans le fait que les règles de propagation sont acquises de manière automatique en corpus par une technique d'apprentissage artificiel, la programmation logique inductive. Ces règles de propagation, exploitant des informations syntaxiques issues des analyseurs SYNTEX, sont automatiquement inférées à partir d'exemples d'alignements valides. L'objectif de cet article est d'une part de montrer que, contrairement aux approches statistiques, notre technique ne nécessite que très peu de données d'apprentissage. D'autre part, on se propose de vérifier si les règles obtenues et les alignements qu'elles produisent varient en fonction du type de corpus d'apprentissage.

Pour ce faire, nous exposons tout d'abord le cadre méthodologique dans lequel nous avons mené nos travaux. Puis, nous décrivons la technique d'apprentissage automatique des règles de propagation en section 3. Enfin, nous présentons et discutons les résultats obtenus en section 4 avant d'indiquer les perspectives de poursuite de ce travail.

2 Contexte d'expérimentation

2.1 Alignement de mots par propagation syntaxique

L'utilisation de règles de propagation pour aligner des bitextes au niveau des mots a déjà fait l'objet de plusieurs travaux. Ainsi, S. Ozdowska (2004) exploite les relations de dépendance syntaxique dans le processus d'alignement. Elle utilise des règles de propagation syntaxique définies à la main qui, étant donnés deux mots en relation d'équivalence dans un couple de phrases alignées, appelés *couple amorce*, permettent de propager le lien d'alignement à d'autres mots en suivant les relations de dépendance syntaxique connues. Dans l'exemple suivant, il est ainsi possible d'aligner *ban* et *interdire* en exploitant la relation sujet portant sur le couple amorce. Dans cet exemple et les suivants, les couples amorces sont notés en gras.



Chaque règle de propagation est donc décrite en fonction de la relation syntaxique qui sert de base à la propagation et de la direction dans laquelle s'effectue la propagation (et éventuellement des restrictions portant sur les parties du discours des mots concernés). Si nous reprenons l'exemple précédent, la règle de propagation anglais/français utilisée est :
V $\xrightarrow{\text{SUJ}}$ **Nom** / V $\xrightarrow{\text{SUJ}}$ **Nom**
Elle indique que la propagation se fait à partir d'un couple amorce de noms régis (*Community / Communauté)* vers un couple de verbes recteurs (*ban / interdire*) par la relation sujet. Une autre règle de propagation possible est celle qui va du couple amorce de noms régis (*ivory / ivoire*) au couple de recteurs (*imports / importation*) par la relation de préposition :
Nom $\xrightarrow{\text{PREP}}$ **Nom** / Nom $\xrightarrow{\text{PREP}}$ **Nom**

La plupart des règles utilisées dans ce type d'approche ont été définies en accord avec l'hypothèse d'isomorphisme direct entre les langues selon laquelle les structures syntaxiques seraient conservées lors de la traduction, comme dans l'exemple précédent (Hwa *et al.*, 2002). Cependant quelques unes traitent des cas de non-isomorphisme, comme l'alignement de *tax* et *fiscales* dans la biphrase : *tax* ***expenditures*** *have been (...) / les* ***dépenses*** *fiscales demeurent (...)*. Si l'on part du couple amorce de noms recteurs (*expenditures / dépenses*), les structures syntaxiques qui se correspondent dans les deux langues sont (NN représente la dépendance entre deux noms et MOD la dépendance générique tête-modifieur, ici nom-adjectif) :
Nom $\xrightarrow{\text{NN}}$ Nom/**Nom** $\xrightarrow{\text{MOD}}$ Adj

Ce type d'approche permet d'obtenir des alignements qui offrent en général une bonne précision, le rappel se révélant cependant de moins bonne qualité. En effet, le principe d'isomorphisme permet de générer des alignements corrects dans la plupart des cas où il s'applique mais il semble, dans certains cas, trop contraignant. Par ailleurs, ces approches nécessitent une expertise humaine pour écrire ces règles de propagation, ce qui peut se révéler coûteux. C'est ce dernier point que nous proposons de contourner en utilisant une technique d'apprentissage artificiel pour inférer automatiquement des règles de propagation.

2.2 Données d'apprentissage et d'évaluation

Nous avons choisi comme données de référence celles mises à disposition dans le cadre d'une campagne d'évaluation des systèmes d'alignement au niveau des mots notamment pour les paires de langues anglais/français (Mihalcea & Pederson, 2003). En voici la description (Och & Ney, 2000) :

- corpus d'entraînement anglais/français, issu du HANSARD (débats parlementaires), comptant 1.3 million de biphrases. Pour les expériences que nous reportons en section 4, nous n'avons utilisé qu'une portion variant de 10 à 1000 couples de phrases alignées de ce corpus.
- corpus de test constitué de 447 phrases alignées extraites d'une partie différente du HANSARD.

– le jeu de référence contient les alignements effectués par deux annotateurs sur le corpus de test. Chaque lien d'appariement établi s'est vu attribuer la valeur S, s'il s'agissait d'un lien considéré comme non ambigu, ou P dans le cas contraire. La valeur P est choisie en présence d'expressions figées ou de traductions libres. Dans le jeu de référence final, la valeur S est conservée pour les alignements pour lesquels il y a accord inter-annotateurs ; la valeur P est attribuée dans tous les autres cas. La figure 1 présente un exemple de phrase annotée ; les alignements S sont en traits pleins et les P en pointillés. Dans un premier temps, pour évaluer notre approche, nous nous focalisons sur les alignements 1-1 entre mots lexicaux ; les expérimentations décrites en section 4 ne portent donc que sur les annotations S entre mots lexicaux.

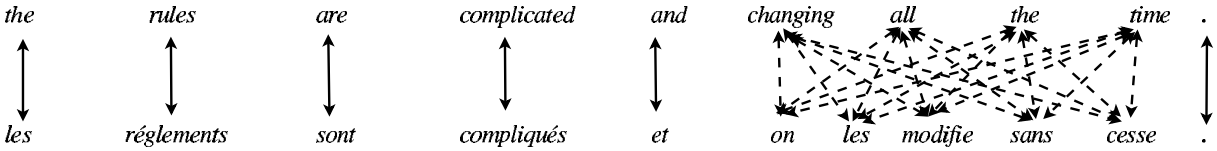


FIG. 1 – Annotation pour la campagne d'alignement HLT

Comme nous l'avons dit précédemment, en plus du HANSARD, les expériences d'inférence de règles de propagation que nous présentons en section 4 sont effectuées sur deux autres corpus. Le premier est un extrait du corpus INRA[1]. Il s'agit d'un corpus spécialisé anglais/français du domaine de la recherche agronomique de 1000 biphrases. Le second est un corpus fourni dans le cadre de la campagne d'évaluation ARCADE (Véronis & Langlais, 2000). Il est constitué de questions-réponses traitées à la Commission Européenne. Là encore, nous n'avons retenu que 1000 biphrases.

Le repérage des relations de dépendance syntaxique dans les trois corpus d'entrainement est effectué indépendamment pour chacune des deux langues par les analyseurs SYNTEX français et anglais (Bourigault & Fabre, 2000). Ces derniers prennent en entrée un texte étiqueté et identifient, pour chaque phrase, des relations telles que sujet, objet direct et indirect, modifieur... Les deux outils sont conçus suivant la même architecture et mettent en oeuvre les mêmes procédures de repérage des relations de dépendance. Par ailleurs, les relations identifiées ainsi que leur représentation sont globalement les mêmes d'une langue à l'autre.

3 Alignement par apprentissage artificiel

Comme nous l'avons déjà dit, l'originalité de notre approche tient au fait que contrairement aux travaux précédemment exposés (Ozdowska, 2004), les règles de propagation ne sont pas données manuellement mais inférées automatiquement. Les deux sous-sections suivantes présentent la technique d'apprentissage artificiel et son utilisation pour inférer ces règles. La technique d'amorçage fournissant automatiquement les exemples nécessaires à cette technique supervisée est décrite en sous-section 3.3.

[1]Nous remercions A. Lacombe, INRA, de nous avoir permis d'utiliser ce corpus.

3.1 Programmation logique inductive

Le principe de notre approche est le suivant : à partir d'exemples de propagations valides au sein de deux phrases alignées, on tente d'apprendre des règles qui les définissent. Pour ce faire, nous utilisons une technique d'apprentissage artificiel supervisée, la programmation logique inductive (PLI). Une présentation approfondie de cette méthode d'apprentissage peut être trouvée dans (Muggleton & De Raedt, 1994), nous n'en donnons ici que les grandes lignes. La PLI permet d'inférer des règles générales (des clauses de Horn) décrivant un concept à partir d'un jeu d'exemples de ce concept E^+ (avec éventuellement des contre-exemples E^-) et un ensemble d'informations externes B, appelées *Background Knowledge*. L'ensemble de règles inférées, appelé classifieur et noté H par la suite, est obtenu en généralisant les exemples en fonction de B.

Quelques conditions imposées à cette tâche d'apprentissage forment le cadre logique de la PLI ($\square$ signifie faux et $\models$ représente l'implication logique) :
- la consistance *a posteriori* impose qu'aucune contradiction n'existe entre B, H et E^+ : $B \wedge H \wedge E^+ \not\models \square$;
- la complétude assure que tous les exemples positifs sont expliqués avec H et les informations du *Background Knowledge*, soit $B \wedge H \models E^+$.

En pratique, les règles composant H sont recherchées à travers un espace d'hypothèses regroupant toutes les règles possibles. Cet espace est organisé hiérarchiquement, ce qui permet de le parcourir efficacement. Une règle de cet espace est retenue si elle maximise un score, généralement défini en fonction du nombre d'exemples (et éventuellement de contre-exemples) qu'elle couvre. La PLI, de par son expressivité (exemples et règles sont exprimés en logique des prédicats), a été utilisée pour de nombreuses tâches d'apprentissage, et notamment en TAL (Cussens & Džeroski, 2000).

3.2 Apprentissage de règles de propagation

Dans notre cas, les règles recherchées sont des propagations et les exemples que nous utilisons sont des phrases alignées comportant des alignements valides ; nous n'utilisons pas de contre-exemples. L'algorithme de PLI que nous utilisons est ALEPH[2]. Dans B sont stockées toutes les informations concernant les dépendances syntaxiques entre mots des phrases exemples et les couples amorces connus. Le formalisme logique de la PLI permet d'encoder facilement ces informations relationnelles. Ainsi, si l'on sait que *companies/entreprises* peuvent être alignés dans l'extrait de biphrase suivant (l'identifiant de chaque mot est noté après les barres obliques) :
... private/id_1_en ***sector****/id_2_en companies/id_3_en*
... les/id_1_fr entreprises/id_2_fr du/id_3_fr ***secteur****/id_4_fr privé/id_5_fr*
on ajoute à E^+ : alignement(id_3_en,id_2_fr). et à B (le nom du prédicat représente le nom de la relation syntaxique, le premier argument représente le recteur et le second le régi) :

```
determinant(id_2_fr,id_1_fr).   prep_de(id_2_fr,id_3_fr).   adjectif(id_2_en,id_1_en).
preposition(id_3_fr,id_4_fr).   adjectif(id_4_fr,id_5_fr).   nom_nom(id_3_en,id_2_en).
amorce(id_2_en,id_4_fr).
```

Une règle qui peut être inférée à partir de cet exemple est :

```
alignement(M_Ang,M_Fr) :- nom_nom(M_Ang,A1), prep_de(M_Fr,F1), preposition(F1,F2),
```

[2] ALEPH est développé par A. Srinivasan et disponible à http://web.comlab.ox.ac.uk/oucl/research/areas/machlearn/Aleph.

amorce(A1,F2).
Avec les notations précédentes, cette règle s'écrit :
M_Ang $\xrightarrow{\text{NN}}$ **A1** / M_Fr $\xrightarrow{\text{PREP\_DE}}$ F1 $\xrightarrow{\text{PREP}}$ **F2.**
Elle souligne l'équivalence des structures Nom-Nom en anglais avec Nom de Nom en français ; tout couple apparaissant dans une biphrase avec cette structure peut ainsi être aligné.

3.3 Amorçage

Des exemples d'alignements valides sont nécessaires à notre technique d'apprentissage. Cette phase de supervision, pénible si elle était conduite manuellement, est dans notre cas automatisée par une technique dite d'amorçage. Notre approche d'inférence de règles de propagation ne requiert donc finalement aucune intervention humaine ; elle est dite semi-supervisée.

Pour générer ces alignements exemples, ou couples amorces, nous utilisons deux approches complémentaires. Il s'agit d'une part d'une technique statistique classique et d'autre part d'une recherche de cognats. En ce qui concerne la méthode statistique, nous considérons comme couples amorces les paires de mots (anglais/français) apparaissant ensemble dans des phrases alignées de manière statistiquement significative (Ahrenberg *et al.*, 2000) ; la force du lien entre deux mots est calculée par un Jaccard sur les fréquences d'apparition conjointe des deux mots (Ozdowska, 2004). Pour le repérage de cognats, c'est-à-dire de chaînes de caractères identiques ou proches dans les deux langues, la méthode mise au point est similaire à celle décrite dans (Fluhr *et al.*, 2000). Elle consiste à identifier la sous-chaîne maximale commune à deux mots qui cooccurrent dans un couple de phrases alignées.

Par la conjonction de ces deux méthodes, ce sont ainsi en moyenne entre 4 et 6 couples amorces (selon les corpus) par phrase qui sont détectés. Environ 5% des couples amorces se révèlent erronés (*i.e.* mots ne devant pas être alignés) ; ce faible taux ne devrait donc pas gêner le processus d'apprentissage. Chaque couple amorce, allié aux deux phrases alignées dont il est tiré, peut ainsi servir d'exemple pour notre technique d'apprentissage de règles de propagation. Une phrase permet donc de produire autant d'exemples qu'elle comporte de couples amorces.

À partir des exemples obtenus par cette technique, il nous est donc possible d'inférer des règles de propagation à partir de nos trois corpus d'entraînement. Ces règles peuvent ensuite être appliquées à de nouvelles données dans lesquelles on aura préalablement repéré des couples amorces.

4 Résultats

Cette section présente tout d'abord les résultats obtenus par notre approche sur le jeu d'évaluation HLT. Nous décrivons ensuite quelques causes d'erreurs récurrentes et examinons enfin certaines des règles inférées.

4.1 Performances d'alignement

Les trois systèmes d'alignement (*i.e.* les trois ensembles de règles inférées à partir de nos corpus) sont évalués à l'aide des données de la campagne HLT. Leurs performances sont présentées

de manière classique en termes de taux de rappel, taux de précision et f-mesure.

La table 1 présente les résultats des trois systèmes. Pour ces expériences, la phase d'apprentissage a été menée sur 1000 phrases de chaque corpus. À des fins de comparaison, nous indiquons les résultats obtenus par les meilleurs systèmes d'alignement – en terme de f-mesure – ayant participé à la compétition HLT ; il s'agit de Ralign (Simard & Langlais, 2003), XRCE (basé sur GIZA++) (Dejean *et al.*, 2003) et BiBr (Simard & Vogel, 2003), tous les trois utilisant principalement des approches statistiques. Nous indiquons aussi les résultats du système de S. Ozdowska (2004) dans lequel les règles de propagation sont définies manuellement.

Système	HANSARD	ARCADE	INRA	Ozdowska	Ralign	XRCE	BiBr
Précision	**88.51%**	82.65%	86.15%	81.59%	72.54%	55.54%	63.03%
Rappel	60.03%	60.25%	60.73%	58.43%	80.61%	**93.46%**	74.59%
F-mesure	71.54%	69.69%	71.24%	68.10%	**76.36%**	69.68%	68.32%

TAB. 1 – Performances des systèmes d'alignement sur les données HLT

À l'examen de ce tableau, on remarque que les résultats de nos systèmes varient peu en fonction du corpus d'entraînement. D'autre part, les performances de nos trois systèmes sont de niveau comparable à celles des autres. Ils se classent en effet deuxième derrière le système Ralign en terme de f-mesure. Ils jouissent par ailleurs d'une précision très supérieure aux autres systèmes, mais d'un rappel relativement plus bas. Ce rappel s'explique en partie par l'insuffisance de couples amorces et la couverture imparfaite de l'étiquetage syntaxique, ce qui rend certains couples inaccessibles à nos règles de propagation.

On s'intéresse dans un second temps à l'évolution des performances selon la taille des corpus d'entraînement. Pour cela, on fait varier le nombre de phrases servant à produire les exemples pour l'apprentissage. La figure 2 présente les taux de rappel, de précision et la f-mesure obtenus selon le nombre de phrases à partir du corpus HANSARD. Les résultats sont très éloquents : il n'y

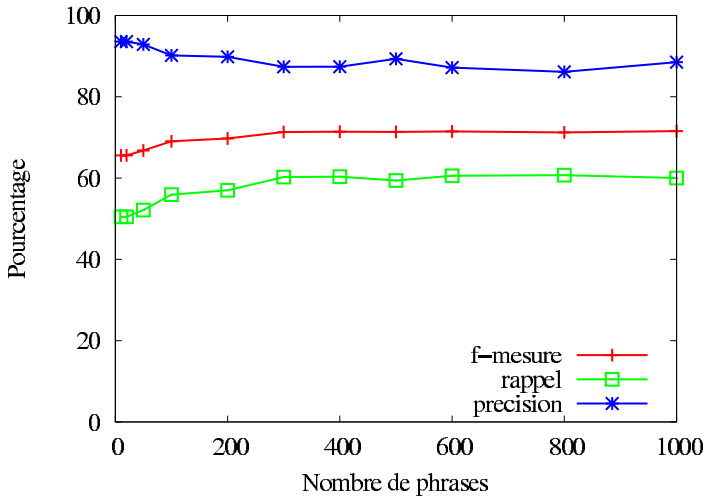


FIG. 2 – Variation des performances selon le nombre de phrases utilisées à l'apprentissage

a quasiment aucune variation de rappel et de précision de 300 à 1000 phrases. En dessous de 300 phrases, la précision augmente sensiblement alors que le rappel décroît. Cela s'explique par le

fait que seules quelques règles de propagation, parmi les plus sûres, sont trouvées. On remarque enfin qu'avec 10 phrases seulement, notre algorithme d'apprentissage est capable d'inférer des règles suffisamment pertinentes pour mener à une f-mesure de 65%. Ces résultats sont donc très positifs, notamment en regard des tailles très restreintes de nos corpus d'entraînement. À titre de comparaison, les systèmes Ralign, XRCE et BiBr utilisent 1.3 million de phrases pour s'entraîner.

4.2 Examen des résultats

Les erreurs d'alignement les plus courantes faites par nos systèmes peuvent se classer en plusieurs grandes catégories. Comme nous l'avons dit précédemment, une grande part des faux négatifs (*i.e.* des alignements non détectés) est due à une trop faible densité de couples amorces en plus d'absences de dépendances au sein de certaines phrases.

En ce qui concerne les faux positifs (*i.e.* des alignements détectés à tort), certains viennent simplement d'erreurs d'étiquetage de SYNTEX (elles-mêmes parfois causées par des erreurs de l'étiqueteur catégoriel utilisé en amont). Par exemple, dans la biphrase *federal government carpenters get $ 6.42/Les menuisiers du gouvernement fédéral touchent $ 6.42*, l'adjectif *federal* a incorrectement été rattaché à *carpenters*, ce qui a provoqué l'alignement incorrect de *carpenter/gouvernement*, tous deux notés recteurs du couple amorce *federal/fédéral*. D'autres erreurs de ce type sont causées par certaines des règles inférées qui ne sont pas assez spécifiques pour éviter de ramener du bruit. C'est notamment le cas des règles manipulant les dépendances objet ou sujet qui, à cause du manque d'informations dont dispose l'algorithme d'apprentissage, ne font pas de différence entre les voix actives et passives. Ainsi, à partir du couple amorce *bring/apporter* dans la biphrase *good legislation has been brought in by Liberal governments / les gouvernements libéraux ont apporté de bonnes mesures législatives*, *gouvernement* et *legislation* ont été alignés à tort. Enfin, des phénomènes de reformulations plus ou moins fidèles lors de la traduction perturbent parfois nos tentatives d'alignement. Ainsi, dans la phrase *the Government must implement the recommendations of the Commissioner of Official Languages/le gouvernement se doit de respecter les recommandations du Commissaire aux langues officielles*, *implement* et *respecter* ont été alignés alors que ce couple n'est pas noté valide dans le jeu de test HLT.

4.3 Règles obtenues

Environ une trentaine de règles de propagation sont obtenues pour chacun des corpus d'entraînement avec 1000 phrases. Il y a peu de différences entre ces règles dans les trois corpus, ce qui explique la proximité des performances observée en section 4.1. Elles sont, pour leur quasi totalité, très similaires à celles proposées par S. Ozdowska. Notons cependant que des règles, non retenues par S. Ozdowska, comme celles exploitant la coordination ou la relation attribut, se révèlent en pratique très productives et expliquent la différence de résultats avec notre approche par apprentissage.

La plupart des règles mettent donc en exergue des isomorphismes connus entre la syntaxe anglaise et française, comme l'alignement des adjectifs modifiant deux noms alignés, ou l'alignement des compléments d'objet direct de deux verbes alignés :

```
alignement(M_Ang,M_Fr) :- adjectif(C,M_Ang), adjectif(D,M_Fr), amorce(C,D).
```

alignement(M_Ang,M_Fr) :- objet(C,M_Ang), objet(D,M_Fr), amorce(C,D).
Ces cas d'isomorphismes parfaits représentent près de 50% des règles de propagation. Certains cas de non-isomorphisme syntaxique sont également trouvés, comme par exemple la construction standard des syntagmes nominaux Nom Nom en anglais et Nom de Nom en français (*cf.* section 3.2). D'autres types de non-isomorphismes peuvent même mener à l'alignement de mots ayant des parties du discours différentes, comme par exemple des noms et des adjectifs :
alignement(M_Ang,M_Fr) :- nom_nom(C,M_Ang), adjective(D,M_Fr), amorce(C,D).

D'une manière générale, il ressort de l'examen de ces règles que la plupart d'entre elles sont des règles de propagation que l'on peut qualifier de génériques. Elles sont effectivement pour une grande partie similaires à celles trouvées manuellement par S. Ozdowska (2004), ce qui confirme la validité de notre processus d'apprentissage. Cependant quelques règles inférées sont plus inattendues – et leur validité peut-être discutée – comme par exemple :
alignement(M_Ang,M_Fr) :- adjectif(M_Fr,C), nom_nom(D,M_Ang), adjectif(D,E), amorce(E,C).
qui permet d'aligner *bargaining* et *négociation* dans la biphrase ... *to have some hang-up with regard to the* ***collective*** *bargaining process*/... *éprouver certains complexes à l'égard de la négociation* ***collective***.

5 Conclusion et perspectives

Nous avons présenté une méthode originale d'alignement de mots basée sur la syntaxe et sur une technique d'apprentissage semi-supervisée. Celle-ci permet d'apprendre automatiquement des règles de propagation à partir d'exemples de couples de mots alignés. Ces exemples sont par ailleurs fournis à l'aide d'une procédure d'amorçage qui confère à notre approche une complète autonomie. Les résultats d'alignement obtenus sont bons et comparables aux meilleurs systèmes d'alignement actuels. De plus, et c'est l'originalité de ce travail, contrairement aux systèmes existants, très peu de données sont nécessaires pour entraîner notre système. On a montré également que les règles de propagation sont relativement génériques et changent peu d'un bitexte à un autre.

Plusieurs perspectives sont ouvertes par ce travail. Concernant la technique d'apprentissage, nous prévoyons d'intégrer les informations catégorielles pour permettre d'inférer des règles ne portant plus seulement sur les dépendances syntaxiques mais aussi sur les parties du discours. Cela permettra d'éviter certaines fausses détections reportées précédemment. L'utilisation d'exemples négatifs, qui permettraient d'empêcher des généralisations excessives et donc des règles de propagation pas assez précises, est également à l'étude. D'un point de vue applicatif, notre méthode étant entièrement automatique, elle peut aisément être adaptée à d'autres paires de langues, pourvu que celles-ci soient suffisamment proches d'un point de vue syntaxique et qu'un analyseur en dépendances existe pour chacune d'elles. Des expériences dans ce sens permettraient d'intéressantes études des cas d'isomorphismes et de non-isomorphismes syntaxiques dans les phrases alignées à travers l'étude des règles de propagation inférées.

Références

AHRENBERG L., ANDERSSON M. & MERKEL M. (2000). *A knowledge-lite approach to word alignment*, In *Parallel Text Processing: Alignment and Use of Translatin Corpora*, chapitre 5, p. 97–138. Kluwer Academic Publishers : Dordrecht.

BARBU A. M. (2004). Simple linguistic methods for improving a word alignment algorithm. In *7th International Conference on the Statistical Analysis of Textual Data, JADT'04*, Louvain-la-Neuve, Belgique.

BOURIGAULT D. & FABRE C. (2000). Approche linguistique pour l'analyse syntaxique de corpus. *Cahiers de Grammaire*, 25, 131–151. Université Toulouse le Mirail.

BROWN P. F., PIETRA S. A. D., PIETRA V. J. D. & MERCER R. L. (1993). The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, 19(2), 263–311.

J. CUSSENS & S. DŽEROSKI, Eds. (2000). *Learning Language in Logic*. Lecture Notes in Artificial Intelligence. Sringer Verlag.

DEJEAN H., GAUSSIER E., GOUTTE C. & YAMADA K. (2003). Reducing parameter space for word alignment. In *HLT-NAACL 2003 Workshop: Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, Edmonton, Alberta, Canada.

DING Y. & PALMER M. (2004). Automatic learning of parallel dependency treelet pairs. In *1st International Joint Conference on Natural Language Processing*, Sanya City, Chine.

DORR B. (1994). Machine translation divergences: A formal description and proposed solution. *Computational Linguistics*, 20(4), 597–633.

FLUHR C., BISSON F. & ELKATEB F. (2000). *Parallel Text Alignment Using Crosslingual Information Retrieval Techniques*, chapitre 9. In (Véronis, 2000).

FOX H. J. (2002). Phrasal cohesion and statistical machine translation. In *Empirical Methods in Natural Language Processing, EMNLP'02*, Philadelphia, PA, États-Unis.

HWA R., RESNIK P., WEINBERG A. & KOLAK O. (2002). Evaluating translational correspondence using annotation projection. In *40th Annual Conference of the Association for Computational Linguistics*, Philadelphia, PA, États-Unis.

LIN D. & CHERRY C. (2003). Proalign: Shared task system description. In *HLT-NAACL 2003 Workshop: Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, Edmonton, Alberta, Canada.

MIHALCEA R. & PEDERSON T. (2003). An evaluation exercise for word alignment. In *HLT-NAACL 2003 Workshop: Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, Edmonton, Alberta, Canada.

MUGGLETON S. & DE RAEDT L. (1994). Inductive Logic Programming: Theory and Methods. *Journal of Logic Programming*, 19-20, 629–679.

OCH F. J. & NEY H. (2000). Improved statistical alignment models. In *38th Annual Conference of the Association for Computational Linguistics*, Hong Kong.

OZDOWSKA S. (2004). Appariement bilingue de mots par propagation syntaxique à partir de corpus français/anglais alignés. In *conférence RECITAL'04*, Fès, Maroc.

SIMARD B. & VOGEL S. (2003). Word alignment based on bilingual bracketing. In *HLT-NAACL 2003 Workshop: Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, Edmonton, Alberta, Canada.

SIMARD M. & LANGLAIS P. (2003). Statistical translation alignment with compositionality constraints. In *HLT-NAACL 2003 Workshop: Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, Edmonton, Alberta, Canada.

J. VÉRONIS, Ed. (2000). *Parallel Text Processing : Alignment and Use of Translation Corpora*. Dordrecht : Kluwer Academic Publishers.

VÉRONIS J. & LANGLAIS P. (2000). *Evaluation of Parallel Text Alignment Systems. The ARCADE Project*, chapitre 19. In (Véronis, 2000).

WU D. (2000). *Bracketing and Aligning Words and Constituents in Parallel Text using Stochastic Inversion Transduction Grammars*, chapitre 7. In (Véronis, 2000).