INALCO - Institut national des langues et civilisations orientales
Département : Textes, informatique, multilinguisme (TIM)

Analyse et comparaison de deux textes – Corpus parallèles et comparables

Par

CHENG Weixuan (Université Sorbonne Nouvelle)

HUANG Yidi (Université Paris Nanterre)

Sujet choisi : Alignement de mots et syntaxe dans des corpus parallèles

Enseignant : Pierre Zweigenbaum
Master : TAL Traitement automatique des langues
Unité d'enseignement : Corpus parallèle et comparables
Année : 2022-2023

Corpus parallèles et comparables

A. Questions analytiques
A.1. Article 1

Using bilingual dependencies to align words in English/French parallel corpora

A.1.1. L'objectif principal du travail
Dans l'époque technique, l'importance de l'amélioration du système de traduction automatique devient une des tâches centrales au fil du temps. Dans ce sens, l'objectif principal du travail est de proposer une méthode d'alignement de mots et de phrases dans les corpus parallèles anglais-français en utilisant des relations de dépendance syntaxique pour optimiser la comparaison entre les ressources bilingues. Cette méthode vise à améliorer la qualité des alignements dans les corpus parallèles et à explorer s'il existe des régularités dans la traduction de structures syntaxiques spécifiques de l'anglais en français.

A.1.2. La méthode employée
La méthode employée est appelée « propagation basée sur la syntaxe ». Elle utilise l'analyseur de dépendance syntaxique profond et robuste SYNTEX pour aligner et analyser les corpus, dont l'entrée est un corpus annoté de POS utilisant le Treetagger comme étiqueteur, puis la normalisation : mesurée par JACCARD comme le « score » système pour trouver les paires de mots, considérés comme pivots pour l'alignement syntaxique ultérieur. En plus, la méthode présente une évaluation comparative par rapport à l'alignement de base statistique, fourni par les modèles IBM implémentés dans le package Giza++.

A.1.3. Les corpus employés ou visés
Ce travail est porté sur des corpus parallèles anglais-français, ce qui signifie qu'ils contiennent des textes dans les deux langues qui ont été alignés au niveau des phrases. Les corpus INRA, JOC et HLT sont considérés comme de vrais corpus parallèles car ils contiennent des textes bilingues, qui sont des textes originaux et des traductions, portant sur une thématique comme science et Commission européenne.
Concernant les autres corpus parallèles qui auraient pu être employés, on a le jeu de données de corpus parallèles de commentaires de presse (News Commentary Parallel Corpus Dataset), qui est constitué de corpus parallèles comprenant des commentaires politiques et économiques dans plusieurs langues. Et le corpus Europarl-ST, qui est une extension du corpus Europarl avec des textes alignés au niveau des phrases pour la traduction automatique.

A.1.4. L'évaluation
La méthode proposée dans ce travail a été évaluée en utilisant la mesure de précision, de rappel et de F-mesure. Les résultats obtenus ont été comparés à ceux d'autres travaux antérieurs, notamment à l'alignement giza++, qui s'est appuyé sur les méthodes statistiques. Les résultats montrent que la méthode proposée a obtenu une précision supérieure à celle des méthodes précédentes pour l'alignement de mots dans les corpus INRA et JOC, mais légèrement inférieure pour le corpus HLT. Cependant, la méthode proposée a rencontré un faible rappel, qui est inférieur d'environ 0,3 à celui du giza++. Cette baisse de performance s'explique par le manque de structure syntaxique non-isomorphe dans les sources pour l'alignement syntaxique du mot. Par conséquent, le F-mesure ne peut être comparé à celui du système d'alignement statistique. Néanmoins, les résultats de l'évaluation démontrent l'intérêt de la méthode linguistique pour cet alignement et incitent les auteurs à réfléchir à des méthodes d'amélioration, en particulier pour le rappel.

A.2. Article 2

Inferring syntactic rules for word alignment through inductive logic programming

A.2.1. L'objectif principal du travail

L'objectif principal du travail est de présenter et d'évaluer une approche pour aligner les corpus parallèles au niveau de mots. Cette méthodologie repose sur une analyse syntaxique de dépendance, générée à l'aide d'une méthode d'apprentissage automatique qui permet d'inférer des règles pour l'alignement de mots de manière semi-supervisée. Le travail constitue une évolution de la méthode linguistique pour l'alignement de mots proposée dans l'article 1.

A.2.2. La méthode employée

Inductive learning programming (ILP), une technique semi-supervisée d'apprentissage automatique symbolique, est utilisé pour inférer les règles d'alignement syntaxique en se basant sur des exemples positifs (E+) et des connaissances sur les dépendances syntaxiques (B) afin de généraliser et propager les règles (H). Avant l'apprentissage, deux outils sont employés pour le prétraitement du corpus : les dépendances syntaxiques des paires de phrases sont obtenues à l'aide de SYNTEX, et les mots d'ancrage sont automatiquement repérés par une technique de bootstrapping basée sur des fonctions de similarité et des cognates. Les paires d'ancrage sont utilisées comme exemples pour l'apprentissage automatique de l'alignement syntaxique par ILP.

ILP permet de générer des règles génériques pouvant être appliquées à d'autres corpus parallèles de deux langues comportant des alignements d'ancrage. Il permet également d'accéder à une analyse linguistique de l'isomorphisme et du non-isomorphisme, à condition que les structures syntaxiques des deux langues soient similaires ou régulières.

A.2.3. Les corpus employés ou visés

Le travail s'est basé sur trois corpus parallèles anglais-français : Canadian Hansards, INRA et JOC. Ils sont considérés comme de vrais corpus parallèles, car ils sont composés de textes et de leurs traductions, qui présentent une cohérence en termes de sujet comme la science, et une normalisation de genre et de structure.

Parmi les corpus parallèles pertinents pour cette étude, il est possible de tourner vers EUR-Lex, un site web regroupant des articles en plusieurs langues officielles de l'UE, y compris l'anglais et le français, portant sur des questions juridiques, mais qui nécessite un prétraitement avant l'alignement étant donné qu'ils sont sous format HTML et PDF. C'est le même cas pour le site des documents de l'ONU, qui contient des fichiers PDF sur différents sujets en six langues pour l'exploitation.

En plus des ressources linguistiques évoquées, les relations syntaxiques de dépendances, générées par le parser SYNTEX en font également partie.

A.2.4. L'évaluation

Les mesures : précision, rappel et F-mesure ont été utilisées pour évaluer l'efficacité de cette approche dans différents types de corpus. Lorsqu'il est appliqué à un corpus d'entraînement annoté de manière explicite, le système obtient une précision élevée, mais un rappel faible en raison du manque de paires d'ancrage ou de dépendances syntaxiques dans le corpus, ce qui conduit à un F-mesure globalement satisfaisante par rapport à la plupart des systèmes statistiques. Lorsqu'il est utilisé sur un corpus sans annotations précises, ILP obtient un F-mesure supérieur à celle du ALIBI, qui utilise une méthode manuelle pour propager les alignements syntaxiques, mais son rappel est inférieur à celui des systèmes statistiques.

Néanmoins, l'entraînement sur seulement 10 phrases permet au système d'atteindre un F-mesure de 70%. De plus, les tests effectués sur trois corpus distincts démontrent sa capacité à propager un grand nombre de règles génériques, plutôt que spécifiques au corpus. Les résultats mettent en évidence les avancées remarquables réalisées par la méthode linguistique (syntaxique) dans l'alignement de mots depuis l'approche de l'article 1.

B. Question de réflexion

B.1 les différences de deux articles en termes de méthodes, de données ou de résultats

Les deux articles présentent des méthodes d'alignement de textes parallèles anglais-français au niveau des mots. Concernant les différences en termes de méthode, le premier article utilise une méthode basée sur la syntaxe qui exploite les relations de dépendance entre les mots et un algorithme de propagation pour étendre l'alignement aux mots voisins. Comparé au premier article, le deuxième article adopte plutôt une approche innovante d'alignement automatique de textes bilingues basée sur la méthode d'apprentissage automatique symbolique appelée programmation logique inductive (ILP) qui infère les règles d'alignement à partir des exemples et une technique de bootstrapping fondée sur des fonctions de similarité et des cognates pour obtenir des mots d'ancrage. Cette méthode utilise des exemples pour induire des règles générales et inférer les règles d'alignement syntaxique à partir des données. Les deux articles évaluent leurs méthodes sur les trois corpus identiques : INRA, JOC et HLT, et les comparent à d'autres méthodes existantes.

Par rapport au résultat, les résultats obtenus dans le premier article montrent que la méthode proposée est plus précise et plus rapide que les méthodes précédentes d'alignement des mots dans les corpus INRA et JOC. Toutefois, elle est légèrement moins précise pour le corpus HLT. Cependant, elle peut aligner des textes parallèles plus longs et est efficace même lorsque les structures syntaxiques diffèrent entre les deux langues. Les auteurs poursuivent leur travail en cherchant à améliorer leur méthode et en l'étendant à d'autres langues. Notons que le premier travail n'utilise pas explicitement d'autres ressources linguistiques, mais s'appuie principalement sur les relations de dépendance syntaxique pour l'alignement. Les relations de dépendance sont des relations entre les mots dans une phrase qui reflètent la structure syntaxique de la phrase. La méthode de propagation permet de propager l'alignement à partir des mots déjà alignés vers les mots voisins. Ce qui rend la méthode efficace pour l'alignement de corpus parallèles, même lorsque les structures syntaxiques diffèrent entre les deux langues.

Alors que dans le deuxième article, à partir des conceptions de méthodes dans l'article 1, le travail utilise différentes techniques statistiques et linguistiques pour évaluer leur approche. Les résultats montrent que l'approche proposée est capable de gérer différents types de corpus, y compris les cas d'isomorphisme et de non-isomorphisme entre les langues source et cible. De plus, elle nécessite une quantité moins importante de données d'entraînement pour produire des alignements de haute qualité que les autres approches statistiques existantes.

En conclusion, ces deux travaux offrent des approches novatrices pour l'alignement automatique de textes bilingues. Le premier travail utilise une méthode basée sur la syntaxe et l'algorithme de propagation, tandis que le deuxième recourt à l'apprentissage automatique symbolique et une technique de bootstrapping simple. Les auteurs des deux travaux poursuivent leur travail en cherchant à améliorer leur méthode et en l'étendant à d'autres

langues et types de corpus. Ces approches peuvent potentiellement avoir des applications dans différents domaines tels que la traduction automatique, la recherche d'information multilingue et l'analyse de sentiments multilingue.

B.2 Continuer et améliorer le travail

B.2.1 Qu'essayez vous en premier ?

Dans le but d'améliorer la performance du système d'alignement linguistique, il est possible d'intégrer les informations de POS dans la procédure d'apprentissage automatique ILP. Cette approche permet d'obtenir une précision plus élevée en réduisant les faux alignements.
De plus, pour accroître la qualité des règles inférées, il serait judicieux d'enrichir les sources d'apprentissage avec des exemples négatifs (E-) contenant des paires fausses pour l'alignement. Cette pratique permet d'éviter une surgénéralisation des règles et d'abandonner certaines règles ambiguës qui ne sont pas assez précises ou génériques pour l'alignement syntaxique, et ainsi est produit des règles pertinentes pour différents types de corpus parallèles anglais-français. Bien que cette démarche supplémentaire puisse améliorer la performance de manière plus efficace, elle requiert des techniques totalement automatisées similaires à celles utilisées pour les exemples positifs (E+).

B.2.2 Voyez-vous d'autres pistes à explorer ?

Étant donné que l'approche est automatique, il est pratique de se concentrer sur l'évaluation de sa performance en utilisant différentes sources et données à plusieurs étapes.
Une stratégie envisageable consiste à explorer divers types d'annotations syntaxiques générées par plusieurs analyseurs syntaxiques (parsers). Cette exploration vise à déterminer dans quelle mesure l'approche dépend d'un type d'annotation spécifique. Pour cela, il convient de sélectionner plusieurs analyseurs syntaxiques automatisés tels que le Stanford Parser et Spacy, qui appliquent chacun leur propre système d'annotation et sont compétents pour détecter les relations syntaxiques dans deux langues. Ensuite, les systèmes entraînés avec différentes annotations syntaxiques seront évalués en termes de précision, rappel et F-mesure, afin de déterminer le meilleur système, mais surtout les écarts entre les scores. Ces évaluations pourraient fournir une perspective sur l'influence des types d'annotations syntaxiques, permettant ainsi de choisir les données d'entrée les plus pertinentes, mais aussi d'explorer les raisons linguistiques profondes de ce phénomène.

Une autre question repose sur les paires de langues traitées. Cette approche permet d'étudier les structures isomorphes et non-isomorphes, car les règles sont assez génériques pour l'anglais et le français. Lorsqu'il s'agit des langues qui possèdent des relations syntaxiques différentes ou complexes, le travail nécessite une étude approfondie. Par exemple, en chinois, certaines relations de dépendance sont marquées à l'aide de marqueurs grammaticaux spécifiques, qui sont placés à côté des mots pour indiquer leur rôle syntaxique. Pour aligner correctement les mots dans ce contexte, il serait préférable d'utiliser des méthodes d'apprentissage automatique pour déduire des règles supplémentaires, qui éviteraient des problèmes sur la détection des paires d'ancrages.