COLLEGE DOCTORAL PAYS DE LA LOIRE — ED MATHEMATIQUES STIC

Nantes Université

THÈSE DE DOCTORAT DE

NANTES UNIVERSITÉ

ÉCOLE DOCTORALE N° 641
Mathématiques et Sciences et Technologies du numérique, de l'Information et de la Communication
Spécialité : *Informatique*

Par
Martin LAVILLE

Évaluation en extraction de lexiques bilingues à partir de corpus comparables

Thèse présentée et soutenue à Nantes, le 1er février 2023
Unité de recherche : Laboratoire des Sciences du Numérique de Nantes (LS2N)

Rapporteurs avant soutenance :

Eric GAUSSIER	Professeur des Universités, Université Grenoble Alpes
Marianna APIDIANAKI	Chargée de Recherche, CNRS, Université de Pennsylvanie

Composition du Jury :

Président :	Prénom NOM	Fonction et établissement d'exercice *(à préciser après la soutenance)*
Examinateurs :	Eric GAUSSIER	Professeur des Universités, Université Grenoble Alpes
	Marianna APIDIANAKI	Chargée de Recherche, CNRS, Université de Pennsylvanie
	Pierre ZWEIGENBAUM	Directeur de Recherche CNRS, Université de Paris-Saclay
Dir. de thèse :	Emmanuel MORIN	Professeur des Universités, Nantes Université
Co-encadrant :	Philippe LANGLAIS	Professeur des Universités, Université de Montreal

COLLEGE ED MATHEMATIQUES
DOCTORAL STIC
PAYS DE LA LOIRE

Nantes Université

Titre : Évaluation en extraction de lexiques bilingues à partir de corpus comparables

Mot clés : plongements de mots bilingues, évaluation, extraction de lexiques bilingues

Résumé : L'extraction de lexique bilingue (BLI) a pour objectif la création, de manière automatique à partir de corpus bilingues, de lexiques entre deux langues. Le BLI est utilisé le plus souvent en domaine général, où les lexiques extraits peuvent par exemple servir en traduction automatique ou en recherche d'information. Les systèmes de BLI fonctionnent alors sur de grandes quantités de données et les résultats semblent hautement satisfaisants. Cependant, les données d'évaluation contiennent de nombreuses erreurs, ce qui pourrait conduire à une remise en question des systèmes. Un second contexte d'utilisation plus marginal du BLI est celui des domaines de spécialité, où l'objectif est l'obtention de traductions absentes des dictionnaires classiques. Les corpus spécialisés (qui ne concernent qu'un seul sujet) sont peu fournis en données et il est compliqué pour les systèmes de BLI d'obtenir d'aussi bons résultats qu'en domaine général. Il faut donc chercher à adapter les approches pour prendre en compte cette particularité. Dans cette thèse, nous améliorons les résultats obtenus en BLI en domaine de spécialité en proposant l'utilisation de techniques de sélection de données. Puis, nous nous intéressons au processus d'évaluation en domaine général et plus particulièrement à certains biais présents dans les données d'évaluation comme la surprésence de paires de mots très fréquents ou graphiquement identiques et proposons un processus d'évaluation plus précis et unifié qui prend en compte ces faiblesses dans les données.

Title: Evaluating bilingual lexicon induction using comparable corpora

Keywords: bilingual word embeddings, evaluation, bilingual lexicon induction

Abstract: Bilingual lexicon extraction (BLI) has as its objective the creation, in an automatic manner from bilingual corpora, of lexicons between two languages. It is most often used in the general domain, where the extracted lexicons can be used in machine translation or information retrieval. BLI systems work on large amounts of data and the results seem to be highly satisfactory. However, the evaluation data contains many errors, which could lead to a re-evaluation of the systems. A second and more marginal context of use of BLI systems is in specialized domains, where the objective is to obtain translations that are not available in classical dictionaries. Specialized corpora (about only one subject) are poorly supplied with data and it is complicated for BLI systems to obtain as good results as in the general domain. It is therefore necessary to adapt the approaches to take into account this particularity. In this thesis, we improve the results obtained in specialized domains by proposing the use of data selection techniques. Then, we focus on the evaluation process in general domain and more particularly on some biases present in evaluation data such as the overpresence of very frequent or graphically identical word pairs and we propose a more accurate and unified evaluation process that takes into account these weaknesses.