République Algérienne Démocratique et Populaire
Ministère de l'Enseignement supérieur et de la Recherche Scientifique

Université d'Alger 2

Faculté des Lettres et des Langues - Département de Linguistique

MEMOIRE DE MAGISTER

INFORMATIQUE LINGUISTIQUE-

SCIENCES DU LANGAGE ET DE LA COMMUNICATION LINGUISTIQUE

OPTION **:** TRAITEMENT AUTOMATIQUE DU LANGAGE

Résolution d'anaphores pronominales de la langue arabe par apprentissage automatique

Présenté par : **Leïla NOURI**

Jury

M[r]	A. SALMI	Docteur à l'université d'Alger	Président
M[r]	A. GUESSOUM	Professeur à l'USTHB	Encadreur
M[me]	H. ALIANE	Chargée de Recherche au CERIST	Examinatrice
M[me]	L. MAHDAOUI	Maitre de conférences à l'USTHB	Examinatrice

Année universitaire 2011/2012

Remerciements

الحمد والشكر لله تعالى

Je tiens à exprimer ma profonde gratitude et mes sincères remerciements au Professeur Ahmed GUESSOUM pour m'avoir proposée ce sujet, pour sa disponibilité et pour ses précieux conseils.

Mes remerciements s'adressent aussi au président du jury Dr Abdelmadjid SALMI et aux membres du jury Dr Hassina ALIANE et Dr Latifa MAHDAOUI, pour leur disponibilité à évaluer le présent mémoire et leur participation au jury de soutenance.

Mes vifs remerciements vont aussi à ma famille, mes amis, et mes collègues du CERIST, qui m'ont toujours soutenue et encouragée au cours de la réalisation de ce mémoire.

ملخص البحث

العائد هو كلمة لها سابقة أو عدة سوابق في النص. حل العائد في اللغة العربية بواسطة التعلم الآلي هو إيجادا بطريقة آلية سابقة العائد باستعمال كم هائل من الأمثلة النصية للعملية.

في الواقع، تواتر تواجد العوائد في نصوص اللغة العربية كبير جدا، الأمر الذي يشكل إشكالية حقيقية للعلاج الآلي للمعطيات النصية، مثل التلخيص الآلي أو استرجاع المعلومات، أو الرد على استعلام في نظم حوار بين الإنسان والآلة، الخ.

ويعتبر التعلم التلقائي للشبكات العصبونية واحدا من الحلول المثيرة للاهتمام لـحل العائد .

في هذا العمل قمنا بدراسة مسألة حل العائد في اللغة العربية و قدمنا حلا لمعالجته حاسوبيا عن طريق تحشية مدونة و تشفيرها أليا، ثم استعمالها داخل الشبكات العصبونية باستخدام الشبكة ذات طبقات متعددة بطريقة التغذية الأمامية واستخدام خوارزمية الانتشار الخلفي (للتصحيح داخل الشبكة).

الكلمات الدالة :العائد، حل العائد، اللغة العربية، تحشية المدونة، وسم، الشبكات العصبونية، الشبكة ذات الطبقات المتعددة الأمامية

Résumé

Une anaphore est un mot qui a obligatoirement un ou plusieurs antécédents qui lui réfèrent dans un énoncé. L'objectif de la résolution d'anaphores pronominales de la langue arabe par apprentissage automatique est de trouver de façon informatisée l'antécédent d'une anaphore et ce par apprentissage sur un maximum d'exemples de textes.

En effet, la fréquence d'apparition des anaphores dans un texte arabe est très grande, ce qui pose un problème pour un traitement automatique des données textuelles, tel que l'élaboration d'un résumé de texte ou d'une recherche d'information, ou pour répondre à une requête dans les systèmes de dialogue homme-machine, etc.

L'apprentissage automatique des réseaux de neurones est considéré comme une des solutions intéressantes pour la résolution d'anaphores.

Dans ce travail, nous nous intéressons à la résolution d'anaphores de la langue arabe sur un corpus annoté et codifié automatiquement, soumis à l'apprentissage des réseaux de neurones en utilisant le perceptron multi-couches selon l'algorithme de la rétro-propagation du gradient.

***MOTS-CLÉS** : anaphores, résolution d'anaphores pronominales, langue arabe, annotation, étiquettes, corpus, réseaux de neurones, perceptron multi-couches.*

Abstract

An anaphora is a word that has one or more antecedents that refer to it in a sentence or parapgraph. The objective of pronominal anaphora resolution is to find the antecedent of an anaphora in a computerized way based on the use of as large a corpus of examples of manual anaphora resolution as possible.

Indeed, the frequency of occurrence of anaphora in arabic text is very large, which turns out to be an issue that needs to be handled before automatic of textual data, so as to develop quality automatic text summaries information retrieval, or query answers in human-machine dialogue systems.

Automatic training of neural networks is considered as one of the interesting solutions for the resolution of anaphora.

In this work, we are interested in anaphora resolution of the arabic language on an automatically annotated and coded corpora, and a neural network which is trained as a feed-forward multilayer perceptron with backpropagation.

KEYWORDS : *anaphora, anaphora resolution, corpora, arabic language, annotation, tags, neural networks, multilayer perceptron.*

Sommaire