

Minimalisme et question-réponse : le système Œdipe

Antonio Balvet, Mehdi Embarek et Olivier Ferret

CEA – LIST/LIC2M
92265 Fontenay-aux-Roses Cedex
{balveta, embarekm, ferreto}@zoe.cea.fr

Résumé - Abstract

Cet article présente le système Œdipe, développé par le LIC2M pour sa participation à la campagne d'évaluation EQUER. Ce système se caractérise par une approche minimaliste au niveau des moyens utilisés. L'article détaille en particulier la stratégie adoptée pour la constitution de patrons d'analyse des questions et en évalue l'efficacité pour EQUER.

This article presents the Œdipe system, which was developed by the LIC2M for participating to the EQUER evaluation campaign and is characterized as relying on minimum means. This article gives more particularly details about the strategy we used for building the patterns for finding the type of questions and evaluates its efficiency for the EQUER questions.

1 Introduction

La campagne EQUER d'évaluation des systèmes de question-réponse en français a été l'occasion pour nous de développer un premier système de question/réponse, le système Œdipe, en s'appuyant sur les outils d'analyse linguistique qui avaient été développés au LIC2M durant les deux années précédentes. Ce développement s'inscrivait donc comme un point de départ et nos ambitions, tant en termes de degré de sophistication du système conçu qu'en termes de résultats attendus, étaient donc modestes. Dans sa forme actuelle, Œdipe peut ainsi être considéré comme un système de question/réponse minimaliste, fournissant en tant que réponse des passages de taille fixe à partir d'un ensemble de documents sélectionnés par un moteur de recherche.

2 Description du système Œdipe

L'architecture du système Œdipe est tout à fait classique. Elle s'appuie à la base sur l'analyseur linguistique LIMA (LIC2M Multilingual Analyzer) (Besançon et al., 2004) qui permet d'une part, de normaliser les mots apparaissant dans les documents et dans les questions et d'autre part, d'en extraire des entités nommées de type MUC (personnes, lieux, organisations, dates et unités de mesure ainsi que produits). La normalisation des mots est réalisée par une analyse morphologique et un étiquetage morfo-syntaxique. La reconnaissance des entités nommées est effectuée quant à elle par une série d'automates appliqués au résultat de l'analyse linguistique. Chaque question posée est analysée afin de déterminer si la réponse attendue est une entité nommée et le cas échéant, le type d'entité

nommée concerné. Cette analyse repose sur un ensemble d'environ 150 automates du même type que ceux définis pour la reconnaissance des entités nommées. Les mots pleins de la question sont par ailleurs pondérés afin de caractériser leur importance *a priori*.

Les documents font quant à eux l'objet d'un traitement en deux temps. Un premier traitement permet de localiser les extraits en relation directe avec la question en s'appuyant sur les zones de forte densité en mots de celle-ci. Chaque extrait se voit attribuer un poids en fonction des mots de la question qu'il contient et éventuellement, de la présence d'entités nommées correspondant au type attendu de réponse. Les extraits sont ensuite ordonnés suivant leur score et les N (égal à 20 pour EQUER) premiers sont sélectionnés. Le second module est chargé de localiser la réponse à la question dans les extraits retenus. Si la réponse attendue est une entité nommée, Édipe retient comme réponse possible la partie de l'extrait considéré centrée sur une entité nommée du type attendu et qui présente le score le plus élevé, score comparable à celui calculé pour les extraits. Si la réponse attendue n'est pas une entité nommée, Édipe applique une fenêtre glissante (égale à la taille souhaitée de la réponse) sur l'extrait en calculant pour chacune de ses positions un score comparable à celui de l'extrait. Il retient ensuite comme réponse possible la partie de l'extrait dans laquelle ce score est maximal. Un ensemble de réponses possibles dotées chacune d'un score est ainsi constitué. La liste finale des réponses est obtenue en ordonnant cet ensemble et en le tronquant en fonction du nombre de réponses désiré. Si le score de la meilleure réponse est trop faible, Édipe suppose qu'aucune réponse n'existe dans les documents. Cette même heuristique est exploitée pour le traitement des questions polaires : une réponse négative est donnée lorsque le score de la meilleure réponse est trop faible. Édipe prend également en compte les questions de type liste, sans autre particularité que de rechercher le nombre possible de réponses attendues lors du traitement de la question en prenant comme référence la première entité nommée numérique trouvée.

3 Une stratégie de découverte de patrons de questions

Une des parties importantes d'un système de question-réponse est l'analyse des questions. Dans le cas d'Édipe, cette analyse repose sur l'application d'un ensemble de patrons morpho-syntaxiques. La stratégie adoptée pour leur découverte s'inspire de stratégies d'extraction de patrons couramment employées en extraction d'information (Riloff, 1994) et de travaux dans le domaine de l'apprentissage dit « par alignement » (Van Zaanen, 2001). La justification linguistique de cette approche est à chercher du côté des opérations de segmentation-commutation employées en linguistique structurale pour l'étude des régularités linguistiques en corpus, ainsi que du côté des stratégies d'analyse « naïve » mises en œuvre par les locuteurs d'une langue et qualifiées par Saussure de « fausse analogie ». Concrètement, cette stratégie de co-analyse suit la procédure schématisée ci-dessous :

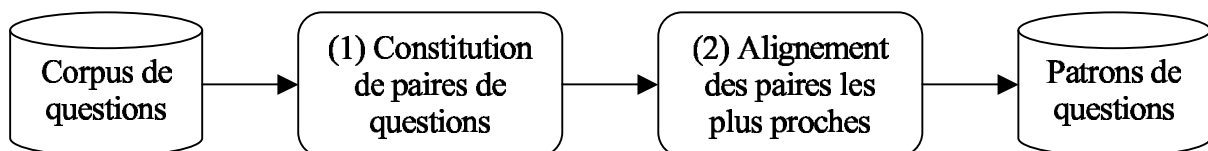


Figure 1 : Étapes pour la constitution d'une base de données de patrons de questions

L'étape (1) se base sur la mesure de la distance d'édition entre deux chaînes de caractères, dite distance de Levenstein, calculée à partir d'opérations d'insertion, d'élimination et de

déplacement. Elle aboutit à une liste de paires de questions associées à un score de distance d'édition. L'étape (2) cherche, pour toutes les paires, la plus longue sous-chaîne commune de mots en s'inspirant de l'algorithme Longest Common Substring. Le résultat de ces deux étapes est une liste de paires de questions, chaque paire étant caractérisée par un score de distance d'édition, ainsi que des scores dérivés de celui-ci, et par la plus longue sous-chaîne de mots commune aux deux questions de la paire. Par exemple, la recherche de la plus longue sous-chaîne commune de mots pour la paire de questions ci-dessous donne le patron suivant, où les '_' marquent des positions possibles dans la séquence de mots analysée¹ :

Quelle est la capitale de la Yougoslavie ?
Quelle est la capitale de Madagascar ? → *Quelle est la capitale de __ ?*

Le patron extrait est ensuite traduit sous la forme d'une expression régulière typée qui est utilisée lors de l'identification du type d'une question. Par ailleurs, les mots non alignés peuvent être considérés comme les membres d'un même paradigme (*i.e.* des noms de pays pour *Yougoslavie* et *Madagascar*).

L'approche adoptée ici, concrétisée par la plate-forme CoPT², est donc une approche de surface, ne mettant en œuvre aucune connaissance linguistique explicite (*i.e.* morphologique, syntaxique ou sémantique) autre que des récurrences de chaînes de caractères et des coïncidences de position pour ces chaînes de caractères. De ce fait, elle est susceptible d'être appliquée sur tout type de corpus de spécialité, y compris dans un contexte multilingue : la stratégie adoptée est en effet indépendante du jeu de caractères employé. Elle requiert simplement une certaine stabilité dans les patrons morpho-syntaxiques employés³, conformément aux présupposés de l'approche structuraliste. Il reste cependant à améliorer l'algorithme d'appariement utilisé car certains appariements sont impossibles dans la version décrite ici : pour les chaînes « A B C » et « A X B X C », où « X » représente une insertion quelconque (un mot), le patron extrait est « A _ _ _ ». Autrement dit, les insertions en position centrale sont mal détectées. Ce problème est néanmoins en voie de résolution.

4 Résultats et analyse

Dans le cadre d'EQUER, le LIC2M a participé à la fois à la tâche générale et à la tâche médicale, en utilisant dans les deux cas exactement le même système et en traitant tous les types de questions (le lecteur pourra se reporter à (Ayache, 2005) pour plus de détails concernant la description des tâches et des métriques utilisées). En revanche, seules des réponses prenant la forme de passages de 250 caractères ont été renvoyées. La Table 1 synthétise les résultats du système Œdipe. Pour la tâche générale, la MRR globale (moyenne de l'inverse des rangs de la première bonne réponse) se situe à 0,7 pour le meilleur système et aux alentours de 0,3 pour la majorité des systèmes, à comparer à 0,5 et 0,1 pour la tâche médicale. Comparativement, les résultats d'Œdipe sont donc faibles et même très faibles pour le domaine médical. Ceci peut s'expliquer bien sûr par le minimaliste d'Œdipe mais la Table 1 montre également un nombre anormalement élevé de questions jugées sans réponse dans le corpus d'évaluation par le système (ce qui explique d'ailleurs le relatif bon score obtenu pour les questions polaires). Il est donc probable qu'une part de nos faibles résultats

¹ La plus longue sous-chaîne commune est alignée sur la plus longue des séquences traitées.

² Corpus Processing Tools, disponible à l'adresse <http://copt.sourceforge.net>

³ Nous avons pu constater que cette stabilité est plus rare dans des corpus littéraires par exemple.

Tâche	Passages corrects / # questions	MRR Sauf listes	MRR Sauf listes et polaires	MRR Polaires	Précision moyenne Listes	Détection d'absence de réponse (#)
générale	113 / 464	0,18	0,17	0,38	0,13	236 précision : 0 rappel : 0,4
médicale	7 / 175	0,02	0,02	0	0	n/a

Table 1 : Résultats du système Édipe pour l'évaluation EQUER

puisse s'expliquer par des valeurs de seuil inadéquates à ce niveau. Afin d'éclaircir l'origine des insuffisances d'Édipe, nous avons mené une analyse manuelle concernant les performances du typage des questions, analyse dont les résultats sont donnés dans la Table 2. Celle-ci laisse apparaître que le typage des questions effectué par le système Édipe se révèle assez efficace. Si l'on prend en compte à la fois les cas dans lesquels Édipe trouve le type d'entité nommée attendue comme réponse (première ligne) et les cas dans lesquels il considère que la réponse n'est pas une entité nommée (seconde ligne), on constate qu'il se trompe dans 28,6% des cas pour les 469 questions analysées (hors questions polaires) du domaine général et dans 10,8% des cas pour les 176 questions analysées du domaine médical.

Typage	Jugement manuel	Général	Médical
Type identifié par Édipe	correct	215 / 254 (45,8%)	17 / 29 (9,7%)
	incorrect	39 / 254 (8,3%)	12 / 29 (6,8%)
Type non identifié par Édipe ≡ réponse non factuelle	incorrect	95 / 215 (20,3%)	7 / 147 (4%)
	correct	120 / 215 (25,6%)	140 / 147 (79,5%)

Table 2 : Résultats de l'analyse manuelle du typage des questions par Édipe

Un travail d'analyse plus poussé reste donc à mener pour déterminer à quel point le minimalisme du système Édipe est à l'origine de ses faibles performances et contribuer ainsi sur un plan plus général à éclaircir le rapport, pour un système de question-réponse, entre les moyens engagés et les performances à en attendre.

Références

AYACHE C. (2005), *Campagne EVALDA/EQUER : Evaluation en Question-Réponse*, rapport final de la campagne EVALDA/EQUER (<http://www.technolangu.net/article61.html>).

BESANÇON R., DE CHALENDAR G., FERRET O., FLUHR C., MESNARD O., NAETS H. (2004), *Concept-Based Searching and Merging for Multilingual Information Retrieval: First Experiments at CLEF 2003*, LNCS 3237, Springer, pp. 174-184.

RILOFF E. (1994), *Information Extraction as a Basis for Portable Text Classification Systems*, Ph.D. de l'Université du Massachusetts Amherst.

VAN ZAAENEN M. (2001), *Bootstrapping Structure into Language: Alignment-Based Learning*, Ph.D. de l'Université de Leeds.