Université Paris IV Master 2 Langue et Informatique

BENAISSA ELIAS

# Devoir : Plate-formes logicielles pour le TAL 3

Année 2015-2016

## Introduction

Dans le cadre de ce devoir portant sur l'extraction et le résumé des différents résultats présentés dans des articles de recherches, j'ai analysé et proposé une première solution dans le processus d'annotation.

Nous utiliserons donc le logiciel Unitex étudié dans le cadre de ce cours, qui nous permettra d'annoter et d'extraire les différents résultats possibles, le but étant d'avoir un taux de précision élevé, un taux de silence bas tout en ayant le moins de bruit possible.

Dans le cas également où le devoir devra être en équipe et donc avoir potentiellement des résultats similaires, le rappel devra également être minimisé, en se divisant astucieusement le travail de manière à ne pas avoir par exemple de graphe Unitex potentiellement similaire, sachant également que l'accumulation d'erreur se fera également par regroupement des graphes.

Ces mesures d'évaluations permettront donc de mesurer la qualité de nos annotations en plus de la F-Mesure.

## Annotation

Après avoir étudié différents articles de recherche, la tâche d'annotation dans l'extraction des informations résumant les résultats d'article de recherche, a quelques contraintes techniques à prendre en compte.

J'ai dans un premier temps remarqué une contrainte technique dans la mise en texte d'un fichier pdf dans certains cas. Même si aujourd'hui le procédé je dirais est plutôt stable, car ayant fais un stage où j'ai eu à travailler sur des fichiers textes d'origine pdf (même si dans le cadre d'un processus totalement automatisé j'ai observé effectivement des annotations bruités, généralement lorsque l'image pdf était de mauvaise qualité), dans le cadre de ce devoir j'ai pu remarquer que beaucoup d'articles présentent leurs résultats sous forme de tableau (et c'est bien logique), et j'ai pu constater qu'il est dans ce cas nécessaire de prendre en compte les dimensions du tableau afin d'annoter les bons éléments dans les bonnes cases, par exemple dans la figure 1 dire que 71.8 est le taux de rappel du LIA (CRF) étant dans les transcriptions de référence.

1 présente les résul- | Référence | ASR | | | | | | |
SER | P | R | SER | | | | |
23,9 | 86,4 | 71,8 | 43,4 | (-19,5) | | | |
30,9 | 81,1 | 70,9 | 45,3 | (-14,4) | | | |
37,1 | 80,7 | 55,4 | 54,0 | (-16,9) | | | |
33,7 | 79,3 | 65,8 | 50,7 | (-17,0) | | | |
sultats pour la transcription de ré- | LSIS (CRF) | 35 | 82,6 | 73 | 55,3 | (-20,3) | |
férence. Pour la transcription au- | Synapse (syntaxe profonde) | 9,9 | 93 | 89,3 | 44,9 | (-35,0) | |
(ASR, fournie par le | Xerox (syntaxe profonde) | 9,8 | 93,6 | 91,5 | 44,6 | (-34,8) | |
LIMSI), une approche à base d'ap- | Tableau 1 - Reconnaissance d'EN lors d'Ester2 sur les transcription de référence (SER, Précision, Rappel) et automatiques (ASR) | | | | | | |

*Figure 1 – Mise en forme du tableau pdf Figure2 en fichier .txt*

| | Référence | | | ASR |
|---|---|---|---|---|
| **Participant (approche)** | **SER** | **P** | **R** | **SER** |
| LIA (CRF) | 23,9 | 86,4 | 71,8 | 43,4 (-19,5) |
| LIMSI (syntaxe surface) | 30,9 | 81,1 | 70,9 | 45,3 (-14,4) |
| LINA (syntaxe surface) | 37,1 | 80,7 | 55,4 | 54,0 (-16,9) |
| *LI Tours (syntaxe surface)* | *33,7* | *79,3* | *65,8* | *50,7 (-17,0)* |
| LSIS (CRF) | 35 | 82,6 | 73 | 55,3 (-20,3) |
| Synapse (syntaxe profonde) | 9,9 | 93 | 89,3 | 44,9 (-35,0) |
| Xerox (syntaxe profonde) | 9,8 | 93,6 | 91,5 | 44,6 (-34,8) |

Tableau 1 – *Reconnaissance d'EN lors d'Ester2 sur les transcription de référence (SER, Précision, Rappel) et automatiques (ASR)*

*Figure 2 – Mise en forme du tableau pdf d'origine.*

Ces résultats étant tout de même primordiaux dans la tâche de ce devoir, j'imagine que nous pouvons annotés ces résultats sans tout de même pouvoir déterminer chacun de ces chiffres où alors en proposant une aide humaine qui scindera donc le fichier .txt en deux tableau un Référence et un ASR (restera également à prendre en compte que P pourra signifier la précision par exemple).

Dans ce processus de pdf to text, les images figés graphique sont également inaccessibles par la plus part des outils proposés.

Vous trouverez dans le même dossier que ce rapport, un fichier pdf annoté exactement comme le ferait un système automatique selon des règles que je vous propose ici.

Les différents résultats sont donc présentés dans des tableaux mais aussi paraphrasés dans le texte et donc plus facilement annotables par des graphes Unitex, les annotations du fichier tal-2010-court-034.pdf ont toujours une portée par phrase (une annotation est égale à une phrase). Ce choix permet dans l'article proposé d'annoter correctement les différentes

paraphrases expliquant les résultats, mais peux également créer du silence si l'explication se poursuit dans une autre phrase.

Nous pouvons désormais nous intéressé donc à ces phrases qui seront annotés, et j'ai pour cela fait des règles adaptable et largement inspiré des techniques de détection par mot-clé et de pondération d'une phrase susceptible de réellement résumé des résultats :

1 – Phrase contenant les mots : ANR, précision, rappel, F-Mesure…

2 – Phrase contenant les mots : reconnaître (lemme) + chiffre.

## Conclusion

Ces règles proposent une petite base sur laquelle nous pourront s'appuyer dans l'annotation des résultats dans les articles de recherche, cependant j'ai également pu remarquer en analysant ces règles sur d'autres articles qu'en utilisant ce système de « dictionnaire de mot pertinent » je pourrais probablement améliorer la précision du système d'annotatation en ayant des dictionnaires par domaine d'article de recherche, dans le domaine d'extraction d'entité nommé en TAL, où dans un article particulier en médecine par exemple.

Dans le cadre d'une possible application industrielle par exemple, cette possibilité permettra au programme d'interagir avec un humain qui pourra par exemple indiquer le domaine d'étude de l'article, ce procédé pouvant résoudre en partie les problèmes qu'un mot puisse signifier différents concepts selon le domaine d'étude de l'article, par exemple si l'article parle d'extraction d'entité nommé, on pourra retrouver des termes comme « loc » « pers » « date ».