Laurine Lamy
15004304

Projet TAL3
Annotation des articles de Denis Maurel

L'article que nous avons choisi s'intitule « Elaboration d'une cascade de transducteurs pour l'extraction des noms de personnes dans les textes », et est écrit par Nathalie Friburger et Denis Maurel. Nous allons établir des annotations sur cet article afin de dégager les résultats (de type scores, classements, mesures d'évaluation etc.) présentés dans celui-ci.

Avant de pouvoir faire ce travail, nous devons étudier l'article afin de dégager les indices nous permettant d'établir des règles pour une annotation automatique des résultats présents dans le texte. Pour ce faire, nous choisissons d'abord de les surligner en jaune dans notre fichier pdf. Cette première étape nous permet de construire un tableau récapitulatif des résultats trouvés (qui sont à annoter), tout en gardant les contextes gauches et droits :

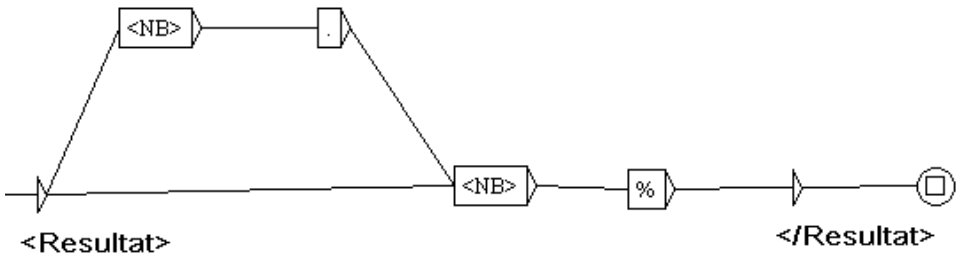
Contexte gauche	Résultat	Contexte droit
Une évaluation sur un corpus journalistique (journal Le Monde) fait apparaitre un taux de précision de	98.7%	pour un taux de rappel de 91.9%.
Une évaluation sur un corpus journalistique (journal Le Monde) fait apparaitre un taux de précision de 98.7% pour un taux de rappel de	91.9%	.
Nous pourrions utiliser un étiqueteur tel que celui de [Brill, 1992] qui donne des résultats corrects à plus de	95%	ou le système Cordial qui offre des résultats encore meilleurs.
Cas 1 :	25.9%	des noms sont précédés d'un contexte contenant un titre ou un nom de profession suivis d'un prénom et d'un patronyme.
Cas 2 :	19.1%	des noms de personnes sont précédés d'un contexte déclencheur contenant un titre ou un nom de profession suivi, soit d'un patronyme seul, soit d'un prénom inconnu (i.e. absent de notre dictionnaire des prénoms) et d'un patronyme.
Cas 3 : C'est le cas le plus fréquent.	43.4%	des noms de personnes n'ont pas de contextes descriptibles mais sont composés du prénom de la

		personne (connu de notre dictionnaire) et suivi du nom de la personne.
Cas 4 :	5.2%	des formes sont repérables grâce à la présence d'un verbe utilisé pour désigner une action mettant en jeu une personne […]
Cas 5 : Les	6.4%	des noms de personnes restants n'ont aucun contextes, même complexes qui puissent les distinguer à coup sûr d'autres noms propres.
Cependant nous avons remarqué que	49%	de ces noms de personnes restants sont détectables en réalisant une seconde passe dans laquelle on recherche les patronymes qui ont été découvert un autre endroit du texte par un des transducteurs.
Ce qui réduit à	3.3%	le nombre de formes indétectables.
Par contre, on peut remarquer que le cas 1 (nom précédé d'un titre ou d'une profession) chute de	25.9%	dans le Monde à 17.1% dans Ouest France : ceci est certainement dû à un souci de rigueur du journal Le Monde.
Par contre, on peut remarquer que le cas 1 (nom précédé d'un titre ou d'une profession) chute de 25.9% dans le Monde à	17.1%	dans Ouest France : ceci est certainement dû à un souci de rigueur du journal Le Monde.
Nous avons observé l'ordre nom-prénom dans	3%	des noms de personnes des articles étudiés dans le journal Ouest France.
Les résultats obtenus sur les trois premiers cas sont très bons ; nous obtenons	97%	de rappel et 99.2% de précision sur les noms de personnes qui sont précédés d'un déclencheur et/ou d'un prénom.
Les résultats obtenus sur les trois premiers cas sont très bons ; nous obtenons 97% de rappel et	99.2%	de précision sur les noms de personnes qui sont précédés d'un déclencheur et/ou d'un prénom.
Le cas 5 est constitué de noms de personnes sans aucun indice : seul le sens de la phrase ou la connaissance du monde d'un lecteur humain permet de déterminer qu'ils sont un nom de personne ;	48.4%	des noms de personnes du cas 5 peuvent être trouvés, grâce à leur présence sous une forme détectable ailleurs dans le texte.

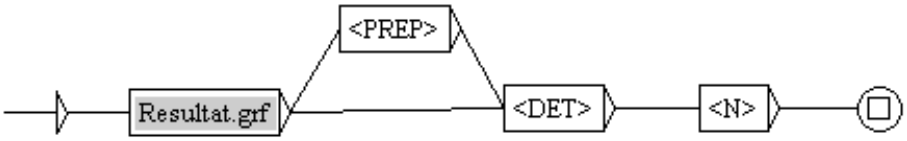
Précisons que nous n'avons pas pris en compte les données présentes dans le résumé en anglais. De même, nous n'avons pas étudié celles présentes dans les différents tableaux. D'un point de vue linguistique, cela ne semble pas possible.

Nous constatons que tous les résultats sont exprimés en pourcentage, ce qui est un point à prendre en compte. En revanche, cela ne suffit pas lui seul : il peut y avoir des pourcentages présents dans l'article, sans qu'il ne s'agisse de résultats ! Il faut voir plus loin si nous ne voulons pas obtenir du bruit lors de nos annotations automatiques. C'est ainsi que les contextes gauches et droits deviennent important pour la construction des graphes nous permettant de dégager les résultats présents dans notre article.

Tout d'abord, nous pouvons faire la proposition de créer un sous-graphe permettant de détecter tous les pourcentages présents dans l'article :



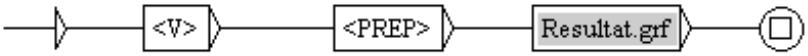
En étudiant notre tableau récapitulatif, nous pouvons émettre diverses hypothèses quant à la construction de graphes. Le premier se présenterait comme ceci :



Ainsi, ces résultats seraient dégagés :

Sous-graphe « Resultat »	Préposition*	Déterminant	Nom
25.9%		des	noms
19.1%		des	noms
43.4%		des	noms
5.2%		des	formes
6.4%		des	noms
49%	de	ces	noms
3%		des	noms
48.4%		des	noms
25.9%	dans	le	Monde
3.3%		le	nombre
97%		de	rappel
98.7%	pour	un	taux
99.2%		de	précision

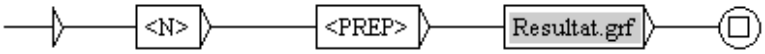
Le second se présenterait comme suit :



Ces résultats seraient dégagés :

Verbe	Préposition	Sous-graphe « Resultat »
chute	de	25.9%
réduit	à	3.3%
remarqué	que	49%

Le dernier graphe se présenterait de cette manière :



Ces résultats seraient dégagés :

Nom	Préposition	Sous-graphe « Resultat »
rappel	de	91.9%
précision	de	98.7%
Monde	à	17.1%
prénom	dans	3%

Ces graphes permettent donc d'annoter les résultats que nous avons dégagés dans notre tableau récapitulatif. En revanche, ils ne sont pas corrects !
Tout d'abord, nous constatons qu'il y a du bruit, certains de nos graphes peuvent annoter les mêmes résultats, mais de manières différentes. Par exemple :

	Verbe	Préposition	Sous-graphe « Resultat »	Préposition	Déterminant	Nom
Graphe 1			**49%**	de	ces	noms
Graphe 2	remarqué	que	**49%**			

De plus, les résultats qui peuvent se dégager de nos graphes ne correspondent pas à ce que nous attendions, notamment en ce qui concerne l'entité « Le Monde ». Il est bien sûr évident que l'ordinateur reconnaisse cela comme étant un Déterminant+Nom, mais en tant qu'humain, cela nous est assurément un nom propre, une entité nommée « Le Monde ».

Il faut donc corriger nos graphes afin qu'ils soient plus précis. En revanche, ils auraient le défaut d'être trop précis et ne s'appliqueraient pas forcément correctement à d'autres articles.

Pour cela, nous pensons à introduire un contexte négatif concernant « dans le Monde » pour alimenter le graphe 1, et crée un autre graphe pour l'insertion d'entités, si possible.
Afin de ne pas avoir de doublon comme nous l'avons expliqué précédemment, nous pouvons indiquer que les verbes ne peuvent pas être aux temps composés.
Enfin, nous ne jugeons pas correct de mélanger les pourcentages que nous pouvons calculer dans un texte et les pourcentages concernant les mesures d'évaluation des études établies lors de l'écriture de cet article. Ainsi, nous pensons également différencier ces deux types de résultats pour les annotations. Il est donc possible qu'un autre de nos graphes possède des contexte s obligatoires tels que « précision » et « rappel ».