# Deft 2011: appariements de résumés et d'articles scientifiques fondés sur des distributions de chaînes de caractères

Gaël Lejeune, Romain Brixtel, Emmanuel Giguet

# Deft 2011: Appariement de résumés et d'articles scientifiques fondé sur des distributions de chaînes de caractères

Gaël Lejeune, Romain Brixtel et Emmanuel Giguet

Université de Caen, GREYC UMR 6072, Boulevard du Maréchal Juin 14032 Caen Cedex, France
prenom.nom@unicaen.fr

**Résumé**

Nous présentons ici une expérimentation dans le cadre de la seconde tâche du défi fouille de textes (DEFT) 2011: appariement de résumés et d'articles scientifiques en français. Nous avons fondé nos travaux sur une approche à base de distribution de chaînes de caractères de manière à construire un système simple et correspondant à une conception endogène et multilingue des systèmes. Notre méthode a obtenu de très bons résultats pour la piste 1 "articles complets" (100%) mais a été moins efficace sur la piste 2 "articles sans introduction ni conclusion" (96%).

**Abstract**

We present here our work on the second task of 2011's Deft: pairing scientific articles and their abstract. Our approach is based on distribution of character strings. Our aim was not only to be efficient on that particular task on French but to build a system that can easily be used for other languages. Our method achieved very good results on track 1 "full articles" (100%) but had more problems with track 2 where introduction and conclusion were removed (96%).

**Mots-clés :** Chaînes de caractères répétées maximales, méthode endogène, approche multilingue, linguistique différentielle, algorithmique du texte

**Keywords:** Maximal repeated character strings, endogenous method, multilingual approach, differential linguistics, stringology

# 1 Introduction

Nous présentons ici une expérimentation dans le cadre de la seconde tâche du Défi Fouille de Textes 2011: appariement de résumés et d'articles scientifiques en français. Nous avons fondé nos travaux sur une approche à base de distributions de chaînes de caractères de manière à construire un système sans ressources externes d'une part et potentiellement multilingue d'autre part. Nous allons expliquer les raisons de ces choix et leurs implications.

## 1.1 Cadre de travail

L'axe de recherche "Multilinguisme, traduction, algorithmique du texte et méthodes différentielles" du laboratoire GREYC promeut depuis fort longtemps un traitement automatique des langues avec des ressources légères, dans la lignée des travaux de Jacques Vergne. Cette approche permet des traitements résolument multilingues (Lucas, 1993 ; Vergne, 2001).

Il défend la position selon laquelle l'interprétabilité du résultat ne passe pas nécessairement par l'interprétabilité des opérandes du calcul ayant produit ce résultat. Ainsi les traitements à base de ressources, si légères soit-elles dans les travaux de Vergne, ont-ils été progressivement délaissés pour laisser place à des traitements dits endogènes suite aux travaux de (Déjean, 1998). Les ressources lexicales ne sont alors pas une entrée nécessaire mais une production du calcul (Giguet & Lucas, 2004 ; Giguet & Luquet, 2006).

Le concept de mot est encore prégnant dans certains travaux du groupe qui nécessitent une segmentation classique en mots. Cette approche tend aujourd'hui à disparaitre de nos approches au profit d'un traitement basé sur les caractères (Lardilleux, 2010 ; Brixtel *et al.,* 2010 ; Lécluze, 2011). Ces opérandes souvent non interprétables par le lecteur humain, puisque pouvant débuter à n'importe quel caractère du texte pour se terminer à n'importe quel autre, ont l'intérêt de laisser envisager des traitements automatiques multilingues, incluant des langues où le mot n'est pas graphiquement délimité (Lejeune *et al*., 2010b). En effet, pour l'ordinateur, un mot est une chaîne de caractères comme une autre et effectuer une opération sur un mot n'a pas plus de sens que de l'effectuer sur une chaîne de caractère quelconque : ses capacités de calcul ne s'en trouvent pas dégradées.

Si le traitement au grain caractère a pu se développer et trouver sa pertinence, c'est certainement par le fait que parallèlement à ces réflexions sur la définition d'un grain d'analyse adéquat pour tel ou tel traitement se développait une approche guidée par le modèle, sous l'influence de Nadine Lucas (Lucas, 2004 ; Lucas, 2009a). Le traitement des langues au GREYC rencontre alors la fouille de données pour donner naissance à des travaux inter-équipes : approches inductives de la fouille de données textuelles (Turmel *et al*., 2003 ; Lucas & Crémilleux, 2004). L'approche guidée par le modèle défendue par Nadine Lucas s'oppose à la vision selon laquelle le texte serait non-structuré, une simple suite de phrases ou pire encore un sac de mots. L'introduction du modèle permet de mettre en scène le critère de position et de grain d'analyse, et ainsi d'ancrer la recherche de cooccurrences de chaînes de caractères (Lejeune *et al*., 2010a), pour produire un résultat qui fait sens pour l'utilisateur. Ainsi pourrait-on résumer le traitement des langues "à la mode de Caen" et illustrer ce positionnement dans la participation à ce Défi Fouille de Textes.

## 1.2 Stratégie de résolution

C'est par la mise en œuvre la plus simple et la plus immédiate de ces deux caractéristiques de notre méthode, à savoir traitement au caractère et approche guidée par le modèle, que nous avons choisi d'aborder le sujet et tenté de montrer la pertinence de la méthode. D'un point de vue général, nous avons souhaité une solution qui soit simple d'un point de vue calculatoire : nous n'avons donc pas cherché à maximiser une fonction de qualité globale des appariements sur la collection. Nous avons choisi au contraire un appariement séquentiel, et sans remise en cause des appariements effectués. L'hypothèse sous-jacente est qu'un résumé et un article sont en quelque sorte indissociables, de sorte qu'il n'est pas nécessaire d'envisager une quelconque ambiguïté d'appariement entre un article et plusieurs résumés ou entre plusieurs articles et un résumé.

Dans cette même recherche de solution simple d'un point de vue calculatoire, nous avons considéré qu'il était plus efficace de rechercher pour chaque document son résumé, plutôt que de rechercher pour chaque résumé son document. L'espace de recherche de la collection de résumés est en effet plus petit que l'espace de recherche de la collection d'articles. Par ailleurs on suppose qu'un article contient toutes les informations importantes disponibles dans le résumé, alors que l'inverse n'est pas vrai. Chercher à quel article correspond tel résumé serait alors potentiellement générateur d'"ambiguïtés" artificiellement engendrées par la démarche.

Notre approche prend donc à ce titre le contrepied des applications de recherche d'information classique, ou le résumé serait envisagé comme la requête posée à un moteur travaillant sur la collection d'articles indexés. Cette approche aurait été pertinente en terme de réutilisabilité de technologies disponibles, mais peut être plus discutable en terme d'adéquation au problème.

Nous n'avons pas non plus cherché a pondérer les fréquences des éléments recherchés en fonction de la collection (approche de type *tf/idf*). Nous avons opté pour une approche moins coûteuse à calculer, qui consiste a considérer que la simple cooccurrence de séquences communes au résumé et à l'article constitue un indice de corrélation suffisamment fiable pour un appariement de qualité, d'autant plus fiable qu'il est cohérent avec les positions définies dans le modèle d'article attendu.

Du point de vue linguistique, nous avons adopté une tripartition des articles : introduction, développement et conclusion. La mise en œuvre informatique consiste à calculer cette tripartition. La segmentation est déduite de la structure physique dont la trace se manifeste par la présence d'éléments XML "titre", qu'il s'agisse de titre ou de sous-titre. La segmentation ne repose donc pas sur la recherche de mots-clés comme "introduction" , "conclusion" qui induirait une dépendance à la langue ou aux variations de libellé, comme le titre "discussion" qui peut faire fonction de conclusion.

Le premier segment, du début du texte à la première balise titre est associé à l'introduction, le dernier segment, de la dernière balise titre à la fin du texte, est associé à la conclusion, et par différence, le reste est associé au développement. Cette mise en œuvre simple part de l'hypothèse que l'introduction et la conclusion sont souvent non découpées en sous-sections, contrairement au développement de l'article.

D'un point de vue pragmatique, nous supposons que le résumé contient des reprises a l'identique de l'introduction, et que le contexte, la thématique et les perspectives sont des points communs que partage le résumé avec le couple introduction-conclusion. De fait, l'implémentation traduit ces hypothèses : (1) la recherche de la plus longue séquence de caractères présente dans l'article, attendue dans l'introduction, et unique dans la collection de résumés, (2) la plus forte corrélation en terme de séquences de caractères partagées entre l'article et le résumé, attendu principalement dans l'introduction et la conclusion.

Dans la partie qui suit, nous détaillerons nos différentes expérimentations, avec différentes relaxations des contraintes.

## 2 Cadre théorique et définitions

Le fait que deux documents puissent constituer un couple résumé-article provient a priori de certaines connections que l'on peut trouver entre eux. Dans notre travail nous avons nommé ces connections, ces points communs, des affinités. Ces affinités sont des chaînes de caractères, mots ou non-mots, communes aux deux documents. Dans la terminologie que nous allons utiliser par la suite, chaque article est un célibataire qui possède un certain nombre de prétendants: les résumés. Pour former des couples nous faisons une hypothèse contrastive, parmi une collection de résumés nous recherchons celui qui possède les meilleures affinités avec un article. La proximité entre un article et un résumé ne se juge donc pas localement mais par rapport à la collection.

### 2.1 Cadre théorique

On cherchera donc à partir d'un corpus de célibataires d'une part et d'un corpus de prétendants de l'autre à obtenir le plus de couples corrects résumé-article. Le bon prétendant pour un célibataire donné sera le résumé qui partagera le plus grand nombre d'affinités avec un article.

Ces affinités seront des chaînes de caractères, mots ou non mots. Nous aurions pu utiliser simplement des mots mais dans la lignée des principes décrits plus haut nous avons souhaité:

- Ne pas nous baser sur des pré-traitements (lemmatisation par exemple) pour pouvoir effectuer certaines comparaisons (retrouver « traduction » dans « traductions » par exemple)

- Favoriser, bien que le corpus soit finalement monolingue, une méthode qui soit facilement réutilisable pour des corpus multilingues.

Plus généralement, nous n'avons stocké aucune information à l'issue de la phase d'apprentissage ni utilisé aucune ressource externe. Cette phase initiale nous a servi simplement à éprouver le système. Dans la même idée de généricité et pour faciliter le passage à l'échelle nous n'avons pas souhaité utiliser les informations concernant la revue. Le système que nous présentons va donc chercher le résumé correspondant à un article donné parmi toute la collection de prétendants sans pré-filtrage. Toutefois, et nous le verrons plus loin, une revue a posé plus de problèmes que les autres.

Nous cherchons ici à mettre en avant la généricité et la parcimonie, dans la lignée des travaux décrits plus haut. Si le but d'un concours est bien entendu de faire le meilleur score, nous nous sommes attachés dans notre démarche à ne pas créer un système trop complexe ou trop paramétrable. Au contraire c'est le même système que nous avons fait fonctionner pour le "run" de référence de chaque piste.

### 2.2 Définition des affinités

Des segments présents dans l'article sont repris par l'auteur dans l'écriture de l'*abstract*. Selon les stratégies mises en place par l'auteur pour construire son résumé, la recopie pourra être plus ou moins prononcée. Cette recopie pourra être un mot, un groupe de mots voire une phrase ou une proposition. L'utilisation des chaînes de caractères répétées maximales (rstr-max) permet de repérer des unités qui se rapprochent dans une certaine mesure des unités multi-mots (Doucet, 2006).

Pour rechercher quel résumé correspond à quel article nous recherchons donc des points d'ancrage. Ces points d'ancrage sont des rstr-max entre un résumé R et un article A. Notre hypothèse est que nous pouvons les rapprocher s'ils ont des segments en commun longs et nombreux. Nous appelons ces segments des affinités, plus un couple R-A possède d'affinités, si possible de grande taille, plus il y a de chances qu'il constitue un appariement correct.

Les chaînes de caractères constituant nos affinités sont repérées à l'aide d'une implémentation python disponible en ligne[1]. Elles sont présentes plus d'une fois (répétées) et ne sont pas strictement contenues dans une chaîne répétée plus grande de même fréquence (maximales):

***tototo*** a pour rstr-max toto (fréquence 2) et *to* (fréquence 3) mais pas *t* qui est de même fréquence que to et se trouve aux mêmes positions

Grâce à notre implémentation, en comparant un célibataire à tous ses prétendants nous obtenons une structure de données donnant pour chaque rstr-max, les documents du corpus dans lesquels elle apparaît. La fréquence de ces affinités à l'intérieur du document ne nous intéresse pas dans cette étude. D'autre part nous ne conservons que les affinités entre le célibataire et ses prétendants, les affinités entre prétendants ne sont pas prises en compte.

[1]http://code.google.com/p/py-rstr-max/

## 3 Filtrage des affinités

Les deux mesures utilisées seront la taille en caractères de la plus grande affinité (affinité-max) et le nombre total d'affinités hapax (card-affinités) pour chaque couple potentiel. Le critère affinité-max n'est pas suffisamment fiable pris isolément mais est complémentaire avec le second. Le nombre total d'affinités peut quand à lui souffrir de la sur-représentation d'affinités a priori peu significatives. En effet le nombre de rstr-max pour un document donné est potentiellement très élevé. Sur des documents en langue naturelle, ce nombre est quadratique en la taille des documents.

On filtre en ne gardant que les affinités qui sont "hapax" dans la collection de prétendants. Nous supposons que ce qui est rare peut avoir une grande valeur. En l'occurrence on ne tient compte d'une affinité entre un célibataire et un de ses prétendants que si cette affinité n'est pas partagée par d'autres prétendants, donc n'est pas banale. Ce critère d'exclusivité évite la surgénération de motifs qui pourrait intervenir avec des rstr-max.

Donc si un article a une affinité commune avec plusieurs résumés, cette affinité n'est pas considérée comme significative. De cette façon nous cherchons à ne pas tenir compte de celles qui pourraient être trop peu discriminantes. Notre hypothèse est qu'un couple réussi doit partager des affinités "originales". En pratique en plus d'éviter de prendre en compte des termes trop génériques et trop largement distribués, nous filtrons ainsi de nombreuses affinités "vides" (Figure 1).

«resse» «ymbol» «ssib» «est p» «ns et» «la mise en» «s donné» «ntifi» «à m» «qu'elle» d'une co» «e ap» «les, » «s qua» «ur l'a» «amin» «lum» «ns f»

*Figure 1: Exemples d'affinités vides*

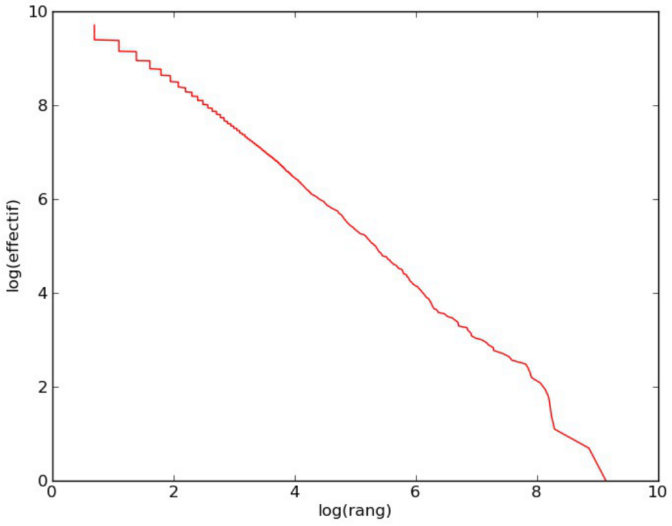


*Figure 2: Loi de Zipf sur les rstr-max*

On peut remarquer sur la figure 2 que la loi de Zipf s'applique très bien aux chaînes de caractères répétées maximales. Dès lors, on peut d'une certaine manière parvenir à caractériser des chaînes de caractères "vides". Nous utilisons ici le terme vide dans la même acception que celle qui est la sienne dans

l'opposition mot-plein/mot-vide ou terme-plein/terme-vide. Nous trouvons effectivement dans les rstr-max figurant en haut à gauche de la courbe (très courtes et très fréquentes) des affixes, des enchaînements de caractères très fréquents et des mots courts. Au contraire, les affinités rares et tout spécialement les hapax sont des chaînes auxquelles on pourrait plus facilement rattacher un sens, pour analyser des erreurs par exemple. Cette observation nous a semblé un pas intéressant vers la validation de notre hypothèse: les couples semblaient formés pour de "bonnes" raisons (Figure 3).

«a philosophie politique d» «s les organisations» «r la reconnaissance des»
«des organisations internationales» «s les années 1970» «établissements»

*Figure 3: Exemples d'affinités pleines*

La figure 4 illustre l'importance de l'utilisation du critère de fréquence. La fréquence 2 en abscisse signifie que l'affinité est présente dans l'article et dans un seul résumé, on a donc une affinité qui est hapax dans la collection de résumés. On peut voir sur cette courbe que dès que l'on relâche cette contrainte, les résultats s'en ressentent. Par exemple, tenir compte des affinités présentes dans 2 résumés (fréquence 3) fait passer les résultats sur le corpus d'entraînement de 0.97 à 0.78.

Les chaînes de caractères ci-dessus ne signifient rien en elles-mêmes. C'est au niveau de l'improbabilité relative de la présence d'une chaîne répétée que se situe le critère de décision (Church 2000). Nous pouvons voir enfin sur la figure 5 qu'un certain seuil dans la significativité de la taille des affinités peut être observé. Fixer leur taille minimale entre 7 et 12 caractères peut optimiser les résultats. En dessous de ce seuil le score reste supérieur à 0.9 donc le filtrage par les hapax est efficace. Par contre au delà, et spécialement à partir de 20 caractères, les résultats sont en chute libre : il n'y a plus assez d'affinités à observer.

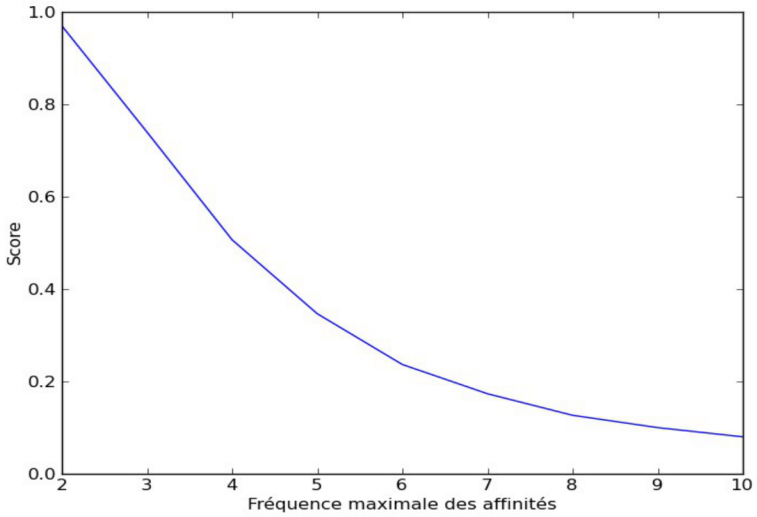


*Figure 4: Corpus d'apprentissage, évolution du score selon la fréquence maximale des affinités dans le sous corpus de prétendants*

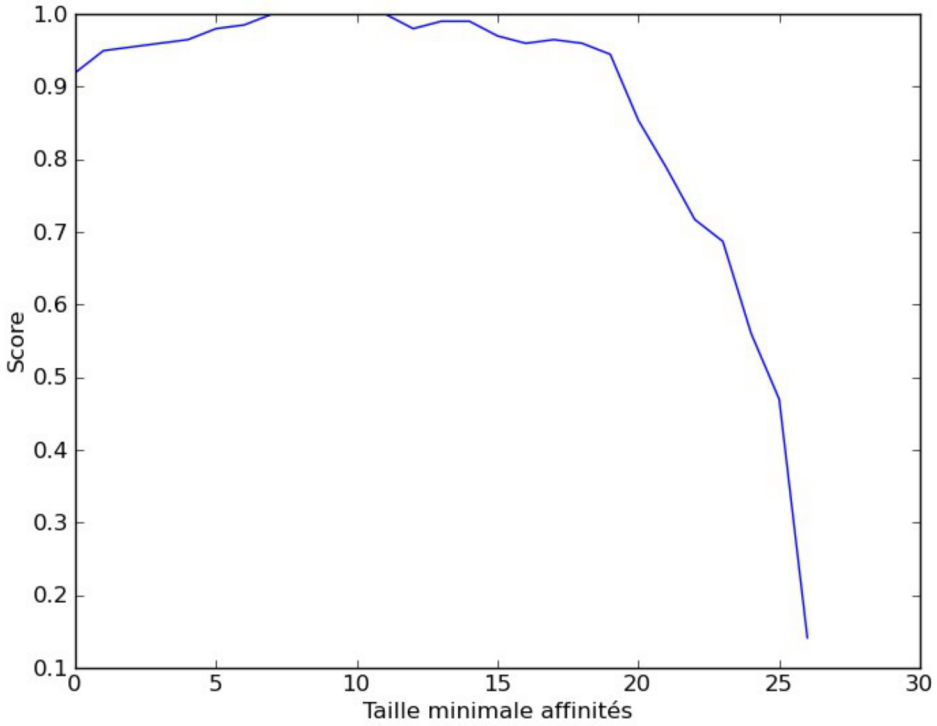


*Figure 5: Corpus de test, influence d'un seuil de taille des affinités sur le score*

## 4 Fonctionnement et résultats

Nous prenons en entrée la liste des articles et des résumés à apparier. Chaque célibataire (article à apparier) est comparé à tous ses prétendants (résumés à apparier). L'implémentation py-rstr-max calcule les chaînes de caractères répétées maximales (rstr-max). On ne garde que les affinités qui sont hapax dans le corpus de prétendants et de taille supérieure à 8. Ce seuil a été fixé pour le français de manière empirique et a été validé par le calcul mais pourrait sans doute être recalculé pour chaque collection quelle que soit la langue.

### 4.1 Description locale

On compare un article à la collection de résumés et on forme un couple chaque fois qu'il semble significativement relié par des affinités. Pour ce faire il faut qu'ils partagent l'affinité-max trouvée dans la collection de prétendants et un nombre significatif d'affinités "uniques". Si un prétendant se détache (figure 4), alors un couple est formé et le célibataire et le prétendant concernés ne seront plus confrontés aux autres.

Nous avons donc cherché comment modéliser la significativité de cette répartition. Comment juger de la significativité du nombre d'affinités d'un prétendant par rapport à un autre? Nous avons remarqué en comparant l'ensemble des affinités d'un couple correct Article 1 - Résumé 1 avec celles de tous les autres couples possibles Article 1 - Résumé "x" (Figure 6) que les affinités hapax étaient réparties en trois tiers globalement équivalents en termes de fréquence:

- Des "affinités vides" mais non filtrées par le critère d'exclusivité
- Des affinités peu significatives et non discriminantes, très proches d'un document à l'autre
- Des affinités pouvant décrire les centres d'intérêt du célibataire, les thématiques de l'article.

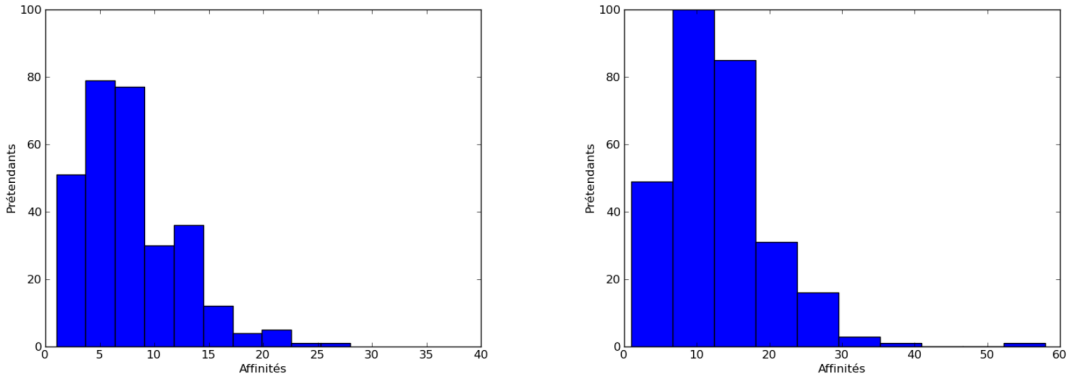


*Figure 6: Répartition des prétendants par nombre d'affinités : à gauche pas de bon couple, à droite un prétendant se détache.*

Nous avons observé que le nombre d'affinités existant entre un article et son résumé était le plus souvent au moins 1,5 fois supérieur à celui des autres prétendants. Quand ce critère n'est pas respecté on considère qu'il y a jalousie potentielle : aucun prétendant ne se détache il y a donc danger d'erreur dans la constitution du couple. Le célibataire sera alors laissé de côté et attendra une phase ultérieure pour être apparié.

On pourrait dès lors penser que l'ordre dans lequel nous traitons les célibataires introduit un biais. Le tableau suivant montre différents résultats selon l'ordre de tirage des célibataires. Nous avons fait des ordres de tirage aléatoires d'articles sur le corpus d'entraînement avec une différence peu significative. L'ordre n'introduit en fait de différences significatives que dans les dernières phases d'appariement, lorsqu'il ne reste que peu de documents à coupler (cf. Tableau 1).

| Run | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Score | 0.97 | 0.967 | 0.97 | 0.97 | 0.973 | 0.97 | 0.967 | 0.967 | 0.97 | 0.963 |

*Tableau 1: Corpus d'entrainement, tirage aléatoire de l'ordre d'apparition des célibataires dans la boucle*

Chaque fois qu'un couple est formé, on considère que le prétendant ne doit plus être présenté aux autres célibataires. De cette façon pour les célibataires ayant du mal à trouver leur résumé, la tâche est facilitée: le nombre d'affinités hapax est plus grand, le critère devient plus discriminant tout en restant très efficace. Tant qu'il reste des célibataires, on les compare aux prétendants disponibles. La constitution de tous les couples nécessite en général 5 à 6 phases. Les dernières phases sont celles où l'on voit le plus d'erreurs, soit qu'elles soient dues à des appariements erronés dans les phases précédentes (cas rare), soit que le faible nombre d'affinités en jeu rende les derniers appariements moins convaincants (cas le plus fréquent).

Nos différents tests ont montré que quels que soient les jeux de données, la première phase apparie 80% des célibataires avec une précision supérieure à 99%. Au fur et à mesure des phases la contrainte de significativité est abaissée pour faciliter l'appariement des documents restants.

### 4.2 Résultats

Nous montrerons ici les résultats obtenus sur le corpus d'apprentissage, sur le corpus de test et sur la concaténation des deux corpus. Le système est rigoureusement le même pour chacune des pistes et pour chacun des corpus. Cela bien que nous aurions pu obtenir de meilleurs résultats avec quelques heuristiques locales. On peut remarquer que nous obtenons de moins bons résultats sur les articles tronqués. C'était assez attendu puisque le phénomène de recopie que nous recherchons est moins visible dans le développement. Comme nous n'avons pas souhaité concevoir un système différent selon les jeux de données, le modèle de document attendu détériore quelque peu les résultats.

Nous avions défini une *baseline* naïve qui consistait à former des couples uniquement selon le critère affinité-max, ses résultats sont faibles. Toutefois il est intéressant de noter la complémentarité des critères affinité-max et card-affinités , affinité-max évite certaines rares mauvaises décisions basées uniquement sur card-affinités (Tableau 3).

| Critère | Affinité-max | Card-affinités | Combinaison |
|---|---|---|---|
| Piste 1: Article-résumé | 0.626 | 0.975 | 1 |
| Piste 2: Texte-résumé | 0.48 | 0.934 | 0.959 |

*Tableau 2: Corpus de test, score selon les critères utilisés*

Le tableau 4 montre les résultats par corpus. Nous tenons à préciser que nos résultats sur le corpus de test sont supérieurs à ce que nous escomptions sur la piste 1. Nous avons donc fait un test en combinant les deux corpus. Remarquons que le résultat de ce test (corpus combinés) n'est pas la simple moyenne des résultats des corpus pris isolément. Nous n'avons pu faute de place intégrer nos tests sur des corpus non symétriques (nombre différents de résumés et d'articles mais ils montraient la même robustesse.

Le système n'utilise pas l'information sur la revue mais n'en souffre pas: il apparie toujours un article de la revue X à un résumé de la même revue dans le cadre du concours. Bien que l'indépendance vis à vis de la revue soit réelle il est intéressant de noter qu'une des revues, *Meta*, a concentré la très grande majorité des erreurs d'appariements et cela quels que soient les jeux de données. Le faible nombre d'affinités hapax rencontrés dans les articles de cette revue a été un facteur déterminant. La distribution des séquences utilisées a semblé plus homogène dans les différents articles issus de cette revue et a fortement nui aux appariements.

| | Corpus d'apprentissage | Corpus de test | Corpus concaténés |
|---|---|---|---|
| Piste 1: Article-résumé | 0.97 | 1 | 0.978 |
| Piste 2: Texte-résumé | 0.96 | 0.959 | 0.96 |

*Tableau 3: Résultats selon les corpus*

Les erreurs les plus fréquentes sur la seconde tâche provenaient là aussi du plus faible nombre d'affinités détectées par le système: sans l'introduction et la conclusion, un grand nombre d'affinités hapax disparaissent (Tableau 4). La différence entre un bon couple et un couple erroné tend à s'estomper et la

qualité des résultats s'en ressent. Comme nous l'avons évoqué, quelques paramètres bien choisis auraient sans doute permis d'atteindre un meilleur score dans la piste 2 mais nous avons voulu garder la simplicité et la reproductibilité comme objectifs primordiaux. Cela corrobore l'hypothèse des linguistes que les débuts et fins de segments sont en soi intéressants à exploiter, à différents niveaux de granularité (Lucas, 2009).

| ID Résumés corpus de test | 013.res | 066.res | 073.res | 154.res | 155.res |
|---|---|---|---|---|---|
| Card-affinités du bon couple article-résumé | 58 | 54 | 76 | 71 | 49 |
| Card-affinités du bon couple texte-résumé | 42 | 42 | 49 | 33 | 37 |

*Tableau 4: Nombre d'affinités du bon couple résumé-célibataire selon la piste.*

## 5 Discussion

Nous avons présenté une méthode d'appariements d'articles et de résumés scientifiques basée sur des distributions de chaînes de caractères. Cette méthode a eu de très bons résultats sur la piste 1 qui concernait les articles complets. Le phénomène de recopie que nous cherchions à utiliser était par contre moins prégnant sur les documents de la seconde tâche, ce qui corrobore notre hypothèse fondée sur le distributionnalisme linguistique. Il nous semble que ces résultats apportent une pierre à l'édification de modèles alternatifs au "tout interprétable".

L'utilisation des chaînes de caractères mots ou non-mots revient quelque part à considérer l'espace typographique comme un caractère comme les autres et pas simplement comme une frontière immuable. Bien entendu il reste beaucoup à faire pour améliorer les résultats de ce genre d'approche mais nous pensons que la généricité multilingue en elle même, peut être exploitable dès le prochain défi fouille de textes, et qu'elle justifie les investigations dans cette voie.

Notre participation a porté sur la tâche 2, consistant à rapprocher un article de son résumé, et non sur la tâche 1, consistant à dater un extrait d'article. Ce choix n'est pas anodin puisque dans notre approche orientée modèle, « le texte est pour une linguistique évoluée l'unité *minimale*, et le corpus l'ensemble dans lequel cette unité prend son sens" (Rastier, 2002 ; Rastier, 2009). Alors que nous disposions d'un modèle de structuration des articles académiques suite aux travaux de (Lucas, 2004), et d'un modèle de structuration des articles de presse (Giguet & Lucas, 2004) dans la lignée des travaux de (Van Dijk, 88), nous ne disposions pas de modèles adaptés à la tâche 1, faute d'applications en lien avec des textes non suivis et éventuellement tronqués.

L'on pourrait bien entendu discuter du concept d'"appariement résumé/article" : en effet, un article nettoyé de son résumé reste-t-il vraiment un article académique ? L'on pourrait également s'interroger sur le fait qu'un article restructuré, voire"déstructuré", en XML, nettoyé de sa mise en page, de sa mise en forme, de ses figures ou de ses références, soit encore véritablement un article représentatif du genre académique.

On peut s'interroger plus globalement sur le fait que la déstructuration des documents soit un service rendu à la recherche en traitement des langues. Certes, cette option facilite l'entrée dans la compétition des participants mais n'a-t-elle pas une contrepartie insidieuse, un tribut peut-être un peu trop lourd à payer ? En restreignant le traitement des langues à un traitement littéral, n'est-ce pas ignorer l'importance de la sémiotique des formes non littérales dans la construction du sens ? N'est-ce pas laisser place à une vision de la langue fondamentalement ambiguë et à des traitements parfois inutilement combinatoires pour gérer ces ambiguïtés ? La plupart de ces ambiguïtés ne sont-elles pas que la résultante artificielle d'une vision peut être encore trop lexicale du traitement des documents ?

Au-delà de ces considérations épistémologiques, nous avons apprécié que des méta-informations comme le nom de la revue soient fournies. Non pas pour qu'elles soient systématiquement utilisées, mais pour que ce choix soit laissé aux participants, comme aurait pu être laissé au participant le choix de travailler soit à partir de la version XML du document, soit à partir de la version pdf texte+image. Cette option est notamment retenue par les organisateurs de l'ICDAR Booksearch Track (Doucet *et al*., 2009), a

laquelle nous participons en travaillant précisément a partir du PDF (Giguet, Baudrillart & Lucas, 2009). Dans la phase de mise au point, le nom de la revue a d'ailleurs permis de révéler la plus grande difficulté de notre approche à traiter la revue *Meta*. Nous n'avons cependant pas cherché à améliorer notre approche pour cette revue. Pour expliquer cette différence, on peut s'interroger sur la méthode de rédaction de ces résumés : sont-ils produits par l'auteur ou par le comité éditorial ? est-ce que des consignes particulières sont données, en terme de longueur ou de contenu ?

Nous avons également apprécié que la tâche choisie soit relativement simple. Les résultats des participants, homogènes en qualité, laissent finalement place à une discussion de fond sur les méthodes, sur leur légèreté, sur leur capacité à être appliquée à d'autres revues académiques, à d'autres genres où le résumé figure, à de quelconques autres langues, ou encore à des collections plus volumineuses. On constate en effet que lorsque la disparité des résultats est grande, l'attention sur la méthode est moins marquée pour les méthodes produisant des résultats dégradés, ce qui n'est bien sûr en aucun cas en lien avec la qualité des concepts sous-jacents. Typiquement, une résolution systémique nécessite un effort de conception et de développement qui n'est pas forcement valorisable par une évaluation en cours de mise au point.

## Références


Brixtel R., Fontaine M., Lesner B., Bazin C., Robbes R. (2010), Language-Independent Clone Detection Applied to Plagiarism Detection in Tenth IEEE International Working Conference on Source Code Analysis and Manipulation. IEEE Computer Society, Timişoara, Romania. Pp 77-86

Church K., (2000) Empirical estimates of adaptation : The chance of two Noriegas is closer to *p/2* than $p^2$ . in *Coling 2000 Saarbrücken*, pp.173-179

Dejean H., (1998) Concepts et algorithmes pour la découverte des structures formelles des langues *Thèse de Doctorat*, Université de Caen.

Doucet A., Ahonen-Myka H. (2006) Fast extraction of discontiguous sequences in text : a new approach based on maximal frequent sequences in *Proceedings of IS-LTC 2006, Information Society - Language Technologies Conference*, Ljubljana, Slovenia, October 9-14, 2006, pp. 186-191.

Doucet A., Kazai G., Dresevic B., Uzelac A., Radakovic B. and Todic N. (2009) ICDAR 2009 Book Structure Extraction Competition in *Proceedings of the Tenth International Conference on Document Analysis and Recognition* (ICDAR'2009), Barcelona, Spain, July 26-29, pp.1408-1412.

Giguet E., Lucas N. (2004). La détection automatique des citations et des locuteurs dans les textes informatifs. *Le discours rapporté dans tous ses états : Question de frontières,* J. M. López-Muñoz, S. Marnette, L. Rosier (eds.). Paris, l'Harmattan, 2004, pp. 410-418.

Giguet E.,Luquet P.S. (2006). Multilingual lexical database generation from parallel texts in 20 languages with endogenous resources. Actes de *Coling 2006* Sydney. pp. 271-278.

Giguet E., Baudrillart A., Lucas N. (2009) Resurgence for the Book Structure Extraction Competition. in *INEX 2009 workshop preproceedings*. Brisbane, Australia. pp. 136-142.

Lardilleux A. (2011) Contribution des basses-fréquences à l'alignement sous-phrastique multilingue. *Thèse de Doctorat* Université de Caen.

Lecluze C. (2011 ) Recherche d'une granularité optimale pour l'alignement multilingue: N-grammes de caractères ou N-grammes de mots ? in *Actes des Journées Toulousaines, JeTou 2011*, Toulouse. pp. 147-151

Lejeune G., Doucet A., Lucas N. (2010a) Tentative d'approche multilingue en Extraction d'Information in *Analyse Statistiques des Donnés textuelles, JADT 2010* Rome pp. 1259-1268

LEJEUNE G., LUCAS N., DOUCET A. YANGARBER R. (2010b). Filtering news for epidemic surveillance: towards processing more languages with fewer ressources. In *Proceedings CLIA/COLING* Beijing. pp. 3-10

LUCAS N. *ET AL.*, (1993) Discourse analysis of scientific textbooks in Japanese : a tool for producing automatic summaries, in *Department of Computer Science Tokyo Institute of Technology, Technical report* 92TR-0004.

LUCAS N. ET CREMILLEUX B. (2004). Fouille de textes hiérarchisée, appliquée à la détection de fautes. *Document numérique vol. 8 n° 3* pp.107-133.

LUCAS N. (2004). The Enunciative Structure of News Dispatches: A Contrastive Rhetorical Approach. *Language, Culture, Rhetoric : Cultural and Rhetorical Perspectives on Communication*. Cornelia Ilie, 154-64. Stockholm: ASLA, 2004.

LUCAS, N. (2009a) Discourse Processing for Text Mining. in : *Information Retrieval in Biomedicine: Natural Language Processing for Knowledge Integration,* ed. by V. Prince & M. Roche, 229 - 62 Hershey, PA, USA: Medical Information Science Reference (imprint of IGI Global).

LUCAS, N. (2009b) Etude des textes en corpus et problèmes d'échelle. in *Corpus 8*. pp.197-220.

RASTIER F. (2002) "Enjeux épistémologiques de la linguistique de corpus," in *Journées de Linguistique de Corpus*, Lorient. http://www.revue-texto.net/Inedits/Rastier/Rastier_Enjeux.html

RASTIER F. (2009). *Sémantique interprétative*, Paris PUF.

TURMEL L., LUCAS N., CREMILLEUX B. (2003). Signalling well-written academic articles in an English corpus by text-mining techniques. *UCREL technical papers Vol. 16 special issue Proceedings Corpus Linguistics*, Lancaster University. pp. 465-474

VAN DIJK T.A, (1988). *News as discourse*, Lawrence Erlbaum Associates, Hillsdale N.J,

VERGNE J. (2001) Analyse Syntaxique Automatique De Langue: Du Combinatoire Au Calculatoire. in *TALN 2001, 8e conférence sur le traitement automatique des langues naturelles*, Tours.