Le défi Fouille de Textes : Quels paradigmes pour la reconnaissance automatique d'auteurs ?

Violaine Prince et Yves Kodratoff

*LIRMM-CNRS et Université Montpellier 2
161 rue Ada, 34395 Montpellier cedex 5
prince@lirmm.fr
http://www.lirmm.fr/ prince
** LRI-CNRS Université Paris Sud
kodratoff@lri.fr

Résumé. Les campagnes d'évaluation en traitement automatique du langage naturel et en informatique documentaire sont devenues un passage obligé pour la reconnaissance des différentes techniques employées. Le Défi Fouille de Texte a pour objectif de permettre aux chercheurs du monde francophone de confronter leurs travaux avec un problème, plus que de primer une équipe, une méthode, ou un outil. Dans cet article nous évoquons les diverses problématiques de la fouille de textes, à savoir la recherche d'information, l'extraction ou l'enrichissement de connaissances, la classification/catégorisation de documents, la segmentation de textes, le profilage. La reconnaissance d'auteur, objet de ce premier défi, est une tâche complexe et composite qui nécessite de traiter simultanément de la segmentation, de la catégorisation et du profilage. L'idée générale est que la mise en place des défis est un outil de cartographie des diverses avancées en fouille de textes, et également un instrument scientifique de compréhension de problèmes de nature complexe.

1 Introduction

Les campagnes d'évaluation en traitement automatique du langage naturel et en informatique documentaire sont devenues un passage obligé pour la reconnaissance de la qualité et de l'efficacité des différentes techniques employées dans ces domaines. Leurs applications, en recherche d'information, extraction de connaissances dans les textes, et fouilles de texte apparaissent naturellement comme le champ privilégié de la mise en compétition des travaux des chercheurs. Au-delà même des techniques, ce sont réellement des paradigmes scientifiques qui s'affrontent.

Si pendant plusieurs décennies les méthodes à base logique, issues de l'Intelligence Artificielle pour laquelle le langage naturel a toujours représenté le défi ultime (n'est-il pas l'élément prégnant du test de Turing ?), ont eu le vent en poupe, les campagnes d'évaluation, depuis les premiers TREC[1] et MUC[2] semblent redorer fortement le blason des méthodes à fondement statistique ou probabiliste.

[1] Text Retrieval Evaluation Conference
[2] Message Understanding Conference

Or le problème n'est pas tant de privilégier un paradigme calculatoire ou représentationnel par rapport à un autre, mais bien de réaliser au mieux une tâche donnée. La création du Défi Fouille de Textes, inspiré au départ de la tâche Novelty de TREC, a en réalité pour objectif principal de permettre aux chercheurs du monde francophone de confronter leurs travaux avec un problème, beaucoup plus que de primer une équipe, une méthode, ou un outil. La notion de défi est totalement différente de celle de compétition. Si compétition il y a, cette dernière doit être vue comme un moyen pour les chercheurs de mesurer l'adéquation de leurs productions scientifiques avec les objectifs d'une tâche précise. La question n'est justement pas de faire un classement pour exhiber les meilleurs mais bien d'éviter les écueils du "Je suis le meilleur" (parce que je suis bien placé auprès des autorités compétentes) ou bien "mais on m'ignore injustement".

Cela doit relativiser complètement la notion même d'évaluation : les mesures n'ont rien de définitif. Elles indiquent seulement une aptitude relative à traiter un type donné de problème, pour une technique donnée et à une étape donnée de la maturité de celle-ci. Que chacun se compare aux autres sur une épreuve commune permet de "remettre les pendules à l'heure", c'est-à-dire de comparer défauts et qualités de la recherche universitaire et des produits industriels, des approches statistiques et de approches symboliques, de ceux qui s'attachent à l'analyse superficielle et de ceux qui s'attachent au sens de la phrase.

Ce premier objectif, le positionnement "relatif", destiné aux chercheurs, est complété par un objectif plus important aux yeux des organisateurs, celui de la cartographie des algorithmes, outils, techniques dans un espace dessiné par des tâches ou des problèmes. Cet objectif est également assorti d'une question à laquelle nous pensons que des éditions successives du Défi pourraient répondre : quels sont les domaines de validité (car nous pensons profondément qu'aucune méthode n'est universelle en soi) des différents paradigmes ? Où sont-ils les plus efficaces ? Quelles sont leurs limites et y a-t-il moyen de les dépasser ?

2 Les différentes tâches de fouilles de texte

Quand on parle de fouille de textes, on utilise un terme générique, traduction approximative de l'anglais 'text mining', et l'interprétation la plus immédiate pourrait se référer à la recherche d'information, ou à l'extraction de connaissances. C'est en effet dans ces thématiques que la fouille de texte a pris naissance. Cependant, avec le succès grandissant de ce domaine scientifique, et surtout ses possibilités d'applications concrètes relativement spectaculaires, d'autres thématiques sont venues compléter ce premier ensemble. A notre avis, l'inventaire actuel des tâches relevant de la fouille de texte, et donc de DEFT, fait l'objet des différents paragraphes de cette section.

2.1 La recherche d'information classique

Elle traite des méthodes susceptibles de retrouver des textes ou des extraits de textes pertinents par rapport à une requête, cette dernière étant exprimée à l'aide du même matériau que ces textes, c'est-à-dire avec des mots. Cette tâche est en effet la plus classique, et la plus ancienne de la fouille de textes. Les principaux ateliers de TREC lui sont consacrés. Les systèmes dit de question-réponse en relèvent également . En France, la campagne d'évaluation EQUER[3]

[3]Evaluation des systèmes de QUEstion-Réponse. Un atelier de la conférence TALN 2005 a été consacré à EQUER.

a proposé un protocole sur ce sujet, et de nombreux compétiteurs ont pu confronter leurs idées aussi bien que leurs résultats.

Au delà des moteurs de recherche qui font de l'appariement de motifs plus ou moins complets, des travaux ont montré que cette tâche pouvait être mieux réalisée dans deux cas : premièrement quand on cherche une relation sémantique (calculable) entre les mots du texte (Yang and Chute 1992) et deuxièmement, lorsque l'on adjoint au moteur des ontologies, ou des structures de concepts permettant de capturer des textes pertinents. Cela, grâce à des relations lexicales comme la synonymie ou l'hyperonymie (Miller and Fellbaum 1991), mais également grâce à des relations sémantiques d'appartenance, de sous-catégorisation (Lee et alii 1993), de participation ou d'attribution, relations présentes dans des réseaux de concepts, ou dans des arborescences de connaissances. Des réseaux tels que WordNet (Fellbaum 1998) mélangent à la fois les relations lexicales et sémantiques. Ils permettent de marquer les mots aussi bien comme matériau (linguistique) que comme idées. Cette double étiquette a des avantages et des inconvénients. La polysémie (multiplicité des sens) est parfois difficile à suivre, et peut provoquer une surcharge dans la recherche et abaisser la précision des systèmes qui s'y fient. A partir du moment où la requête est décomposée en ses principaux constituants, elle perd de sa capacité de détermination sémantique. Des techniques telles que la cohésion lexicale (Morris and Hirst 1991) viennent parfois se substituer à cette perte : cela peut s'avérer suffisant pour des requêtes simples ou courtes. Mais si la requête est elle-même un texte d'une certaine taille, il peut y avoir quelques divergences.

2.2 L'extraction de connaissances nouvelles ou la vérification d'un réseau ou arbre de connaissances existant

Les textes étant par définition des réservoirs de connaissances, ces derniers ont jusqu'à il y a environ quinze ans, été assez peu exploités tels quels. La médiation de l'esprit humain a toujours été requise pour la construction de ces réseaux, arbres ou ontologies, nécessaires aux systèmes à base de connaissances, et plus pragmatiquement, aux moteurs d'inférence un tant soit peu évolués. Si les règles de production, qui sont des modes d'emploi de la mise en relation des différents savoirs, devaient a priori provenir majoritairement de l'expertise des individus, les connaissances factuelles et classificatoires pouvaient être automatisées. C'est pourquoi, des branches thématiques issues d'une part de la représentation des connaissances, et d'autre part des bases de données, se sont penchées sur le texte comme source de connaissances modélisables, et les années 2000 ont vu l'explosion de l'acquisition des ontologies à partir de textes (Buitelaar et alii 2005).

Pour ces deux communautés, cette source est dite non structurée, ou semi-structurée. En effet, il n'y pas eu d'intervention complémentaire de la part de l'Homme pour rajouter des éléments forçant les classifications. Néanmoins, pour les spécialistes de langage naturel, *un texte est forcément une source de données structurées*, puisqu'il est composé de phrases, de paragraphes et de sections éventuellement. Les phrases sont par définition des constructions élaborées visant à la fois à exprimer des idées mises en contexte, et à restreindre les possibilités sémantiques des mots en fonction d'un fil conducteur guidant la mise en oeuvre du texte. Un texte en langage naturel, pour un lecteur humain, est un ensemble parfaitement construit : ponctuation, contextualisation sont des éléments de forme et de fond déterminants pour la compréhension. Or pour que cette compréhension puisse être automatisée, il aurait fallu disposer

d'algorithmes et de logiciels capables de détecter ses structures à l'instar du lecteur. Cela signifie disposer d'un analyseur non seulement morphologique (lemmatiseur ou 'tagger') capable d'affecter les catégories grammaticales, mais également d'un véritable analyseur syntaxique, susceptible d'isoler les constituants et de détecter les dépendances. Constituants et surtout dépendances sont les véritables clés de la structuration de la phrase, indiquant au lecteur quels sont les éléments moteurs (agents, gouverneurs) et quels sont ceux qui sont compléments et dans quelle mesure ces derniers sont modifieurs. Certes cette structuration semble autre que celle requise pour la constitution de taxonomies ou d'ontologies, où les relations entre concepts sont les éléments à détecter de façon primordiale, mais elle existe, et surtout, il semble difficile de l'écarter, justement lorsqu'il faut relier lesdits concepts. Voici quelques remarques de fond liées à ce problème.

- Considérer un texte comme un sac de mots revient à perdre de vue la mise en perspective de l'importance relative des éléments langagiers : le langage naturel est non commutatif dans ses relations de dépendance, il est relativement ordonné, et de nombreuses dépendances de type "complément" peuvent être des indices forts de relation d'attribution (attributs de concepts). Les exemples auxquels sont sensibles actuellement les chercheurs appartiennent à l'ordonnancement des groupes nominaux prépositionnels ou adjectivaux. Ainsi "*voile de bateau*" est une sous-classe du concept "voile", alors que "*bateau à voile*" est une sous-classe du concept "bateau". De la même manière, "*moyenne pondérée*" et "*poids moyen*" faisant l'un et l'autre référence aux concepts de "moyenne" et de "poids" mettent une relation d'attribution différente dans les deux cas. Dans le premier exemple, c'est le poids qui est un attribut de la moyenne, pour générer éventuellement une sous-classe de ce concept, et dans le second, la moyenne est un attribut permettant de spécifier un type de poids particulier.
- Les relations de référence attestée par anaphore sont des indicateurs d'hyperonymie ou de synonymie lorsque lesdites anaphores ne sont pas pronominales. Exemple : *Le journaliste a interrogé* ***Clinton*** *sur la guerre en Irak.* ***L'ancien président*** *s'est montré très réservé.*
- La reconnaissance de la gouvernance (sujet ou prédicat) permet de replacer correctement les relations entre concepts tels que décrits par le texte. Prenons la phrase suivante : *Une embolie pulmonaire est définie comme une oblitération totale ou partielle du réseau artériel pulmonaire par un ou plusieurs caillots de sang*. Le sujet est ici l'objet défini. Le prédicat gouverneur est la définition. Cette structure tend a montrer que la partie "comme une oblitération totale ou partielle du réseau artériel pulmonaire par un ou plusieurs caillots de sang" est un modifieur de la définition. Extraite d'un autre corpus, la définition suivante met en avant d'autres gouverneurs : *Lorsque ce caillot se décroche, il va migrer vers le cœur dans la circulation veineuse pour atteindre le cœur droit puis, après passage dans le ventricule droit, les artères pulmonaires : c'est l'embolie pulmonaire*. Ici, c'est le caillot le sujet, sa migration, le prédicat principal, et la ponctuation (les :) indique une équivalence entre les deux propositions. Cette deuxième structure est beaucoup plus explicite que la précédente en ce qui concerne les relations entre concepts. Mais si on ne regarde que les mots, on tombe sur le même ensemble que précédemment.

La perte d'information en transformant le texte en ensembles non ordonnés de mots ou de termes, même complexe, est donc, pour cette tâche, relativement importante. Les quelques résultats actuellement obtenus sont plus satisfaisants pour les esprits susceptibles de compléter

les manques observés, par leur propre action, que si on devait considérer ces résultats de la manière la plus détachée et la plus objective. D'ailleurs, à ce sujet, certains auteurs se posent la question de la valeur des objets extraits (Roche et alii 2004) par rapport à leur problématique d'une part, et par rapport aux exigences réelles de la tâche d'autre part. Extraire des connaissances pour enrichir une base existante, ou pour vérifier les éléments de cette dernière, suppose de récupérer véritablement l'ensemble des relations, car les dépendances ont un rôle sémantique et font l'articulation entre syntaxe (construction) et sémantique (interprétation).

Les travaux qui ont été réalisés jusqu'à présent dans ce domaine se sont fondés sur les particularités des domaines dont il fallait extraire des connaissances. Plutôt que des cadres généraux, les recherches se sont focalisées sur des domaines techniques ou scientifiques relativement limités, à vocabulaire restreint. Les connaissances majoritairement extraites sont factuelles et classificatoires, et donc peuvent se confondre avec la terminologie. Une première observation permet d'associer la structuration linguistique en groupes nominaux et/ou prépositionnels avec les concepts du domaine. L'extraction de ces groupes permet de baliser l'univers conceptuel de l'ontologie à construire ou à compléter. Quant à la découverte des liens entre concepts, pour l'instant, c'est le repérage des règles d'association qui semble primer. Si *embolie pulmonaire* est liée avec *phlébite* ou *thrombose* dans le corpus, une association entre ces concepts est détectable à l'aide de diverses techniques. Mais la nature précise de l'association, portée par le prédicat ou par la sémantique de la phrase (et non pas seulement des mots) échappe encore aux extracteurs de connaissance qui font l'impasse sur la compréhension syntaxique et sémantique du texte en langage naturel.

2.3 La classification de documents

La classification de documents s'est imposée à partir de la remise au goût du jour de la classification de données (numériques, génomiques, etc.). Issue majoritairement de la statistique, la méthodologie scientifique vise essentiellement à récupérer des groupes relativement homogénéisés auxquels on attribuera le vocable de classe (si ces groupes ont des propriétés relativement fortes) ou catégories (si la notion d'appartenance est plus faible, ou plus disparate).

La classification automatique apparaît soit comme un processus supervisé, c'est-à-dire avec un partitionnement préalable des documents en catégories ou classes, réalisé en général par un ou plusieurs experts humains servant de référence, soit comme un processus non supervisé, et là, l'apprentissage a toute latitude de faire émerger des regroupements à partir de calcul de proximité, ou d'algorithmes dits de *clustering* [4].

De façon massive, la classification automatique de textes est un domaine où la notion de **fréquence d'occurrence de mots**, et celle **d'apprentissage** sont tellement prépondérantes que les chercheurs n'envisagent très souvent pas qu'il puisse en être autrement. Que les résultats produits soient directement liés à des calculs de fréquence d'occurrence de termes extraits (Salton et al 1983), ou à des justifications à fondement psycho-cognitif comme dans LSA (Matveeva et al. 2005), l'approche par mot est pratiquement la seule à tenir le haut du pavé.

Très clairement également, les méthodes d'apprentissage sont prégnantes ; elles vont des modèles de régression, de l'approche $k - nn$ (Yang and Liu 1999), des modélisations bayésiennes naïves, ou adjointes à des arbres de décision (Lewis and Ringuetee 1994), à SVM

[4] Bien qu'il puisse apparaître dans la littérature une définition différenciée : le terme de classification est réservé au non supervisé et celui de catégorisation au processus supervisé.

(Joachims 2002), à HMM (Charlet et al. 2000), modèles dont les résultats sont, pour l'instant, parmi les meilleurs.

Le problème de la classification est similaire à celui des autres domaines de la fouille de texte. Tout dépend de la tâche sous-jacente et de ses objectifs.

Ici plusieurs types de buts peuvent apparaître :

- optimiser la classification d'un corpus donné : ici les méthodes les plus dépendantes des spécificités du corpus sont les plus adaptées à ce genre de tâche. Le calibrage de la méthode par les données s'impose. Qu'il soit supervisé ou non, le processus de classification est forcément lié à un apprentissage, si ce n'est de toutes les nuances du corpus, du moins de ses paramètres principaux.
- Catégoriser de grosses masses de documents en cherchant à s'adapter au mieux à des catégories existantes : les classes dominent ici les données et les méthodes devront plutôt chercher à reconnaître au mieux les caractéristiques de ces catégories dans l'ensemble des données fournies. Dans ce cas les méthodes supervisées sont plus adaptées et on cherchera surtout à optimiser la fonction de représentation des catégories.
- Découvrir de nouvelles catégories à partir d'un ensemble de documents : cet objectif rejoint d'une certaine façon les prémisses de l'extraction de connaissances. Ainsi, une connaissance régulièrement découverte dans une masse de données peut servir de catégorie pour classer, selon un point de vue, des textes. Dans ce cadre, les méthodes à apprentissage peuvent jouer un rôle de "découvreur" de catégories, et une situation semi-supervisée s'impose, car il faut en effet connaître les catégories existantes et leurs propriétés pour décider du caractère innovant de tel regroupement possible de données.

Dans de nombreux cas, les travaux de la littérature ont du mal à faire la différence entre la tâche de catégorisation de documents et la tâche de segmentation (voir le prochain paragraphe). En effet, les méthodes de *clustering* en particulier (recherche d'agrégation) traitent les ensembles de textes comme des "méta-textes" (un grand texte général issu de la concaténation) (Zhao and Karypis 2005). Classer des textes revient alors à segmenter ce méta-texte en zones thématiques indexables par des catégories existantes ou émergentes.

Les défauts que l'on peut trouver à l'ensemble de ces approches relèvent essentiellement de l'ignorance, délibérée ou non, des qualités langagières que possède un texte. Ce dernier, comme nous l'avons dit, ne se comporte pas uniquement comme un ensemble de données. Un texte surimpose données et connaissances, signal et raisonnement. C'est ce qui fait qu'un texte relève d'un discours, et non pas de la concaténation de chaînes lexicales. Un texte est animé par une *intention* particulière. Dans ces conditions,réduire le texte à ses termes constituants, et plus particulièrement à ses constituants dits significatifs (noms, verbes, adjectifs, adverbes dans les meilleurs des cas), c'est perdre une grande partie de l'intention communicative du discours. Les méthodes à forte dominance lexicale pourront effectivement classer un texte dans une catégorie, à condition dans ce cas que les catégories soient suffisamment disjointes lexicalement pour que le classement se fasse. Sans supervision, il sera difficile de mesurer les véritables performances de ce classement.

2.4 La segmentation de textes

La segmentation de texte est une tâche de reconnaissance thématique qui peut s'apparenter à une forme d'indexation. L'idée est de dégager des parties de textes offrant une certaine cohérence, et de les distinguer les unes des autres soit en les nommant (et du coup, cela indexe les

parties en question) soit en en délimitant les contours. Dans ce dernier cas, la segmentation de texte est assimilable à la détection des ruptures thématiques.

Certains travaux, comme ceux de Reynar (Reynar 1998) , puis de Ji et Zha (Ji et Zha 2003) ont largement exploité l'analogie que l'on pouvait tirer de la métaphore de la reconnaissance d'image. Segmenter un texte ou reconnaître une image dans un ensemble complexe seraient des tâches proches. Deux attitudes sont alors possibles : soit on reconnaît ce qui est caractéristique de l'image (son centre, sa zone la plus typique, la zone la plus dense), soit on en détecte les contours.

Ainsi, on peut segmenter un texte par extraction de parties typiques et, comme le font certains auteurs, mise en marge d'un index dans la marge. En revanche, cette structuration est floue sur les frontières. Les algorithmes de pavage de texte à la Hearst (Hearst 1997) sont obligés de calculer des degrés de cohésion pour limiter l'extension des pavés ainsi reconnus. On pourrait leur opposer des algorithmes de recherche des ruptures, qui tendraient à s'apparenter davantage à des algorithmes de contour.

Ce que pourraient avoir en commun toutes les recherches sur la segmentation de texte ce sont les caractéristiques suivantes :

1. La détection de la cohésion (thématique, lexicale) dans un texte
2. La définition de la limite de segment lorsqu'il y a rupture de cohésion : soit par changement lexical (Choi et alii 2001), soit par un éloignement sémantique autrement détecté
3. La capacité à présumer de l'unité du segment par rapport à une unité connexe et cohérente : ainsi, des segments seront constitués de plusieurs phrases adjacentes si toutefois ces dernières maintiennent la cohésion choisie. Des phrases séparées par plusieurs autres phrases ne pourront pas relever d'un même segment, sauf à considérer les phrases intermédiaires comme une forme remplissage (*filler*) qui ne rompt pas la chaîne (thématique, lexicale) ainsi créée.

Toujours est-il que les tâches de segmentation dépendent fortement de ce pourquoi elles sont réalisées. Elles peuvent être par exemple associées :

1. à une tâche de recherche d'information, dans laquelle on cherchera à fournir en réponse non seulement un texte (issu d'une URL par exemple) mais plutôt, dans ce texte, le ou les fragments les plus véritablement compatibles avec la question posée (Llopis et al. 2002).
2. A une tâche d'indexation d'un texte pour des buts de création de méta-données à usage pédagogique ou documentaire (Yang and Li 2005).
3. A une tâche de résumé automatique ou semi-automatique, dirigé par le thème, où le résumé se fait par extraction des segments les plus appropriés à un thème donné, et création d'un nouveau document
4. A des travaux d'extraction du plan ou de la structure du document pour diverses fonctions ultérieures.

2.5 Le profilage

Le profilage consiste à donner des contours lexicaux, sémantiques ou rhétoriques à un ensemble de textes ou de fragments de textes dans le but de :

- reconnaître un auteur particulier ou une période donnée dans une masse de documents non datés (Swan and Jensen 2000) ou mélangés ; ces recherches d'identité ou de signature font du profilage une tâche particulièrement utile à des fins historiques ou culturelles
- à l'inverse, étant donné un profil fait de préférences fournies par des utilisateurs, rechercher l'ensemble des textes obéissant à ce profil : cela fait du profilage une forme intéressante de veille technologique, stratégique ou scientifique
- détecter des tendances ou des opinions dans des discours (Kontostathis et al. 2004) : la notion de jugement (favorable, neutre, défavorable) peut largement servir aux tâches prédictives du marketing ou des instituts de sondage. Mais également, des tendances plus complexes, de type attitudes majoritaires ou minoritaires, peuvent être l'objet d'une recherche de profil.

Dans beaucoup de cas, les méthodes ou les processus du profilage peuvent fortement ressembler à ceux de la catégorisation supervisée (si le profil est fourni) ou de la segmentation thématique. Le profil est récupéré par la détection de la rupture plus que par celle de l'émergence, cette dernière s'inscrivant en négatif par rapport aux critères de rupture. La différence est surtout dans l'intention : la catégorisation consiste souvent à se limiter à la relation d'appartenance. La segmentation thématique est dépendante de la notion de thème (*ce dont on parle*). Le profilage est plus complexe car il introduit l'identification par la *manière dont on parle*. Cette manière définit la tendance parfois plus que l'objet même du discours.

3 La reconnaissance d'auteurs : une tâche évocatrice

Le Défi Fouille de Textes 2005 a porté sur la reconnaissance d'auteurs. Le profil n'existait pas *a priori*. Il devait éventuellement être défini par le corpus d'apprentissage fourni par les organisateurs du défi. La reconnaissance des auteurs, ou la capacité à répondre à la question ; à qui appartient la phrase numéro X du corpus ? Est-ce du Mitterrand ou du Chirac ? est donc une tâche supposant les actions suivantes compte tenu du fait que les discours était mélangés :

- créer des catégories à partir du corpus d'apprentissage où les fragments de texte sont indexés par leurs auteurs
- dans le corpus de test, être capable de segmenter les textes en cherchant à classer, segment par segment (dont la taille varie selon les méthodes) l'auteur servant de catégorie
- éventuellement , profiler un texte : guider la classification par ce que l'on peut savoir du profil de l'auteur.

Par conséquent, la tâche de reconnaissance d'auteur est une structure composite entre de la catégorisation ou classification (selon que l'on utilise ou non une méthode supervisée, bien qu'ici, la structuration du Défi pousse davantage vers la catégorisation), de la segmentation liée à la reconnaissance d'une catégorie. Comme il s'agit d'auteurs, des idiosyncrasies peuvent être également utilisées pour affiner les catégories, auquel cas, on a affaire à une tâche de profilage.

Pour faire de la classification il faut supposer l'existence de différences significatives entre les objets que l'on souhaite classer. Les discours politiques étant relativement semblables (certains esprits chagrins pourraient y trouver un effet secondaire de la légendaire langue de bois), il est apparu nécessaire aux organisateurs d'avoir une *différence thématique* forte pour permettre de dégager des contours distincts. C'est pourquoi, dans le texte du Défi, les organisateurs ont signalé que les discours des deux auteurs traiteraient l'un de politique nationale ou intérieure, et l'autre de politique internationale.

Cette différence aurait pu être suffisante, et du coup, la reconnaissance d'auteur aurait pu se ramener à "International donc X" et "National donc Y", soit en une division bi-partie. Il aurait fallu en ce sens caractériser par divers éléments textuels une catégorie *International* et une catégorie *National*. Mieux encore, il aurait suffit de caractériser seulement l'une d'entre elles, puisque tout ce qui n'appartiendrait pas à l'une appartiendrait forcément à l'autre. Malheureusement, le matériau textuel étant ce qu'il est, ramener cette tâche à celle de l'apprentissage d'une catégorie et à sa reconnaissance, aurait été méconnaître les problèmes suivants :

- Si la catégorie est caractérisée par des fréquences lexicales, toute phrase P apparaissant dans le texte peut très bien appartenir à un discours relevant globalement d'une catégorie donnée, tout en ne possédant aucune des caractéristiques lexicales de sa catégorie. Par conséquent une catégorie utilisée comme filtre segmenteur peut ne pas rendre de résultat précis sur le point de rupture
- Les caractérisations des catégories dépendant du corpus d'apprentissage, si on choisit de mettre par défaut dans l'autre catégorie ce qui n'est pas classable dans l'une, on risque d'avoir des phénomènes de mauvaise attribution due au silence. Après tout, il peut très bien arriver que le corpus d'apprentissage ne contienne pas l'ensemble des mots attribuables à une thématique donnée, en particulier ceux relatifs aux entités nommées.
- Même si à première vue les catégories thématiques choisies semblent relativement disjointes, en réalité elles ont une intersection non nulle : beaucoup de termes fréquents appartiennent à l'une et à l'autre, il faut donc décider de ne s'appuyer que sur une terminologie discriminante, qui se trouvera essentiellement cantonnée aux noms de pays ou de personnes. Cela revient donc aussi faire une catégorisation par noms propres, avec probablement quelque part, une modélisation "du monde" sous-jacente (de type taxonomie ou ontologie). Si cette dernière n'est issue que du corpus d'apprentissage on a les problèmes énoncés dans le point précédent. Cela voudrait dire que, implicitement, les méthodes avec "ontologie" pourraient être favorisées ici.
- Enfin, pire que cela, même si on a une représentation du monde, il se peut très bien qu'en discours, il y ait des usages "digressifs". Ainsi lorsque l'un des auteurs parle de l'Europe, il y fait allusion concernant des régulations et directives européennes contraignant la politique nationale, alors que l'on peut estimer a priori que le mot Europe fait appel directement à la thématique internationale.

Pour toutes ces raisons, il est clair que la catégorisation, avec ou sans ontologie, et la segmentation thématique guidée par elle, ne sont pas suffisantes. Les participants qui se sont intéressés au profilage comme élément complémentaire de discrimination ont eu bien raison. En effet, les structures stylistiques, le "comment dire", sont manifestement discriminantes entre les deux discours. Le profilage ici est rhétorique en partie, et temporel pour une autre. En effet, les discours ne sont pas tout à fait de la même période, et cela peut servir d'indice de différenciation en cas de doute sur certaines phrases. De plus, les rédacteurs des discours des deux présidents n'ont pas du tout la même manière d'écrire. L'un utilise certains adjectifs, et une certaine quantité d'entre eux, et l'autre est plus porté sur les formes verbales, et la structuration hachée. La différence de style est une façon utile de profiler. Elle peut ne pas être notable sur des phrases de transition, mais elle peut également lever le doute sur des phrases où la catégorisation thématique est plus incertaine.

L'examen des résultats du Défi est en ce sens probant.

- Les équipes qui ont appliqué une méthodologie générale, se préoccupant de "ce qui est

dit", c'est-à-dire le véritable contenu thématique n'ont pas été bien placées. En effet, ce contenu n'est pas discriminant. Cela revient presque à dire que les deux candidats avaient, dans le fond, des discours fort peu différents.
- Les équipes qui se sont plus attachées à la réalisation lexicale ont eu de meilleurs résultats car manifestement, les mots sont plus discriminants que leur sens. Il y a les "mots de Mitterand" et "les mots de Chirac".
- Les équipes qui avaient un traitement meilleur des entités nommées et/ou une ontologie du domaine ont vu leurs résultats améliorés.
- Enfin les équipes qui ont rajouté des éléments de profilage à toutes les autres méthodes de discrimination se sont classées en tête du peloton.

Au-delà de la technique et de son degré d'élaboration, il est clair que les approches "dépendantes du corpus" se sont beaucoup mieux placées que celles qui étaient indépendantes du corpus, c'est-à-dire celles qui ne rajoutaient que très peu d'information contextuelle et relative à la nature même des données. Les techniques "adaptées" sont forcément meilleures que les génériques. Le problème est donc : est-ce que l'on cherche à résoudre le problème instancié dans le Défi ou est-ce que l'on cherche à évaluer ses propres méthodes dans la résolution d'un problème ? Les tenants de la première attitude se sont placés sur le podium, les autres ont pu mesurer les limites adaptatives de leur démarche. Dans l'absolu, les deux attitudes sont nécessaires, car elles permettent justement d'évaluer le coup de la résolution du problème donné. Si on devait, par un principe abductif commun, extraire les caractéristiques générales de la classe des problème dont le Défi 2005 n'est qu'une instance, alors on pourrait dire que, pour faire de la reconnaissance d'auteurs dans un ensemble de textes mélangés, il faut à la fois profiler fortement les auteurs en question, et, en toile de fond, appliquer des techniques de segmentation thématique.

4 Les défis ultérieurs : pourra-t-on cartographier l'ensemble des problèmes de la fouille de texte ?

Il est clair que DEFT05 n'est que le premier d'une série de défis à relever. Ses résultats, l'implication des participants, sont des indices forts de son utilité scientifique et épistémologique. En faisant un point sur la reconnaissance d'auteurs, DEFT05 a déblayé le terrain pour au moins deux tâches qui ont ensuite été retenues pour les éditions ultérieures [5].

- Le problème de la segmentation de textes : que signifie la rupture de la cohérence inhérente à la définition d'une unité thématique ? Dans un texte relevant d'un même domaine, la rupture thématique ne peut pas juste être caractérisée par une distinction lexicale forte, comme on pourrait le penser si la catégorisation devait servir de filtre pour la segmentation. Même des distinctions plus légères, comme *national* et *international* sont trop grossières pour déterminer les changements thématiques au sein du même discours. Le profilage stylistique ne serait d'aucun secours car il s'agit ici du même auteur, ou du même groupe d'auteurs s'il n'y a pas de distinction individuelle. Par conséquent, la problématique de la segmentation thématique, ou la détection du changement est un des

[5] A l'heure où nous écrivons, DEFT06 a eu lieu sur la segmentation thématique de textes et DEFT07 a retenu le thème du discours d'opinion et de la détection de tendance

domaines de la fouille de textes qui pourrait faire l'objet d'une confrontation des méthodes. Là, il y a des chances pour que le "contenu" ait davantage d'importance.[6]

– Le problème du profilage, et de la détection de tendances en général : dans son ouvrage dédié à la fouille de textes, Fidelia Ibekwe-Sanjuan consacre à ce domaine un chapitre entier[7]. Des systèmes existent, d'autres sont en développement. Les laboratoires universitaires sont manifestement en avance sur les systèmes industriels, mais le profilage est un enjeu crucial et possède bon nombre d'applications grand public. Traité sous l'angle de la détection d'auteur de manière implicite dans DEFT05, le profilage pourrait être examiné en profondeur dans sa dimension discours d'opinion. Auquel cas il convient d'attribuer une notion de *valeur* aux termes et aux constructions. Se posera alors le problème du traitement des négations explicites (avec les marqueurs idoines) et implicites (usages d'antonymes) des amoindrissements (usage de comparatifs) et d'objets aussi difficiles à considérer sur le plan automatique que l'euphémisme ou l'ironie. Il y a là des risques de rentrer dans un discours lié à la subjectivité, et la mise en relation de ce qui est dit avec ce qui est estimé peut s'avérer une tâche plus ardue que prévu.
– Si la recherche d'information a été explorée dans sa partie question-réponse, d'autres sous-domaines ne l'ont pas été aussi bien, et une édition du Défi pourra éventuellement se pencher sur une tâche faisant se projetant au moins en partie sur ce domaine. Si la classification/catégorisation pourrait être un domaine quelque peu galvaudé (encore que les confrontations sur un même terrain n'aient pas été si nombreuses), la problématique de l'extraction et/ou de l'enrichissement de connaissances reste encore un lieu intéressant d'examen des différentes techniques. En particulier, on s'aperçoit que les besoins taxonomiques dans de nombreuses spécialités pourraient être en partie satisfaits par des approches automatiques ou semi-automatiques. Les domaines technologiques et scientifiques évoluent à grande vitesse, les hybridations sont multiples. Le rôle de veille technologique que peut jouer la fouille de textes est donc particulièrement mis en valeur. Dès lors, les taxonomies et les terminologies des domaines variant rapidement et nécessitant l'examen de grandes masses de textes, les travaux en extraction ou augmentation de connaissances pourraient être à juste titre d'un intérêt grandissant.

Toutes ces thématiques sont bien entendu des pistes. Certaines d'entre elles sont suivies telles quelles dans les éditions du Défi, d'autres seront à définir, voire à inventer. Il est clair que les retombées économiques, sociales et scientifiques de la fouille de textes sont telles que les communautés scientifiques, qu'elles soient universitaires ou industrielles, doivent continuer à remettre en cause leurs techniques par une confrontation saine et une exigence toujours renouvelées.

5 Conclusion

Dans cette brève introduction aux résultats du Défi Fouille de Textes 2005, nous avons cherché à montrer que ce domaine de recherche, qui s'affirme de plus en plus depuis quelques années, est relativement bien balisé dans le monde francophone. Si de nombreuses équipes

[6] Effectivement comme cela a été le thème du Défi 2006, les méthodes fondées sur le contenu se sont trouvées mieux placées que lors de la reconnaissance d'auteurs. Cependant, les méthodes génériques sont restées désavantagées par rapport aux démarches de calibrage sur le corpus d'apprentissage.

[7] À paraître aux éditions Hermès en 2007

francophones, qu'elles soient en recherche d'information, informatique documentaire, systèmes d'information, intelligence artificielle, traitement automatique du langage naturel sont concernées par la fouille de textes, et ont participé aux grandes campagnes d'obédience américaine que sont TREC ou DUC, ce Défi a été pour elles l'occasion de confronter leur démarche les unes par rapport aux autres, mais également dans le cadre des items suivants :

- la problématique de la langue : est-ce que le traitement peut être dépendant de certaines caractéristiques linguistiques et donc n'est que moyennement transposable d'une langue à l'autre ? A cela nous n'avons eu que des réponses partielles. Si les statistiques, prédominantes dans la technologie du Défi apparaissent comme indépendantes de la langue, il n'en reste pas moins que certaines heuristiques linguistiques ont permis par exemple à l'équipe gagnante de dominer les autres approches statistiques.
- Les techniques adoptées : les équipes (dont on lira les travaux dans ce numéro spécial) ont montré qu'elles étaient à la pointe du domaine et les algorithmes utilisés font partie du paradigme dominant sur la scène scientifique internationale.
- L'élaboration d'un Défi en milieu francophone : la tâche a été voulue attrayante, originale. Le travail fait par les organisateurs a été remarquable (ainsi que nous l'avons signalé en préface), mais au-delà de cela, se pose la question fondamentale des moyens à mettre en œuvre pour lancer une telle campagne d'évaluation. Les laboratoires français ont loin d'avoir les moyens du NIST qui organise TREC, dont les financements sont institutionnels et proviennent en grande partie du ministère américain de la Défense. Et pourtant, aujourd'hui, ce qui est demandé aux chercheurs de ce domaine, c'est d'afficher leurs résultats dans les diverses compétitions, et de convaincre quant à la faisabilité de leurs propositions.

Il importe donc très fort que ce genre de campagne puisse continuer d'exister, si ce n'est dans le domaine francophone, du moins dans l'espace scientifique européen. Nous sommes aujourd'hui face à deux phénomènes apparemment contradictoires : la standardisation linguistique par dominance de l'anglais, et l'explosion de la diversité linguistique par la mise en ligne de textes de différentes origines. La fouille de textes ne peut être indifférente à cela. Il faudra donc aussi bien évaluer les recherches qui cherchent à s'affranchir du matériau linguistique que celles qui, au contraire, se confrontent directement à ses spécificités. Les unes permettront une approche globale, majoritaire, et les autres une vision plus en profondeur, et plus adaptative. Les deux sont nécessaires, car la fouille de textes est sans conteste une technologie "sensible" occupant une place privilégiée dans l'appréhension de l'intelligence économique.

Références

[Charlet et al. 2000] D. Charlet, D. Jouvet and O. Collin. An alternative normalization scheme in HMM-based text-dependent speaker verification *Special Issue on Speaker Recognition and its Commercial and Forensic Applications* , vol. 31, no 2-3, pp. 113-120. 2000

[Buitelaar et alii 2005] P. Buitelaar, P. Cimiano and B. Magnini. Ontology Learning from Texts : an Overview. In P. Buitelaar ed. *Ontology Learning from Text : Methods, Evaluation and Applications*. pp 3-14 IOS Press. 2005

[Chauche et alii 2003] Chauché J., Prince V. , Jaillet S., Teisseire M. Classification Automatique de Textes à partir de leur Analyse Syntaxico-Sémantique . *TALN'03 : 10ème Confé-*

rence Internationale sur le Traitement Automatique du Langage Naturel , pp. 55-65 . 2003

[Choi et alii 2001] F. Y. Y. Choi, P. Wiemer-Hastings and J. Moore, Latent Semantic Analysis for Text Segmentation, *Proceedings of 6th EMNLP*, pp 109-117. 2001

[Fellbaum 1998] Fellbaum C. (ed). *WordNet : An Electronic Lexical Database*. MIT Press, Cambridge, Massachussetts,1998.

[Hearst 1997] M. A. Hearst. Text-tilling : segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, pp 59-66. 1997

[Ji et Zha 2003] X. Ji and H.Zha Domain-independant segmentation using anisotropic diffusion and dynamic programming, *Proceedings of the ACM/SIGIR Conference on Research and Development in Information Retrieval*2003

[Joachims 2002] Joachims T. *Learning to Classify Text Using Support Vector Machines* . Kluwer Academic Publishers, May 2002

[Kontostathis et al. 2004] Kontosthatis A, Galitsky L.M., Pottenger W.M., Soma R., Phelps D.J. A Survey of Emerging Trend Detection in Textual Data Mining, In Berry M.W (eds.), *Survey of Text Minin,* Springer,NY, 2004, 186-223.

[Lee et alii 1993] Lee J. H., M. H. Kim and Y. J. Lee. Information Retrieval based on conceptual distance in IS-A hierarchies. *Journal of Documentation*, 49(2), 188–207,1993.

[Lewis and Ringuetee 1994] Lewis D.D., Ringuetee, M.(1994) A Comparison of Two Learning Algorithms for Text Categorization. *Proc. of 3rd An. Symp.on Document Analysis and Information Retrieval* Pp 81-93.

[Llopis et al. 2002] Llopis F. , Ferrandez A.., Vicedo J. L., Gelbukh A. Text segmentation for efficient information retrieval *Proceedings of CICLing.2002, Lecture Notes in Computer Science* , vol. 2276, pp. 373-380. 2002

[Miller and Fellbaum 1991] Miller G. A. and C. Fellbaum. Semantic Networks in English. In Beth Levin and Steven Pinker (eds.) *Lexical and Conceptual Semantics* , 197–229. Elsevier, Amsterdam, 1991.

[Matveeva et al. 2005] Matveeva, I., Farahat, A. and Royer, C.. Document representation with Generalized Latent Semantic Analysis. *Proceedings of the Conference 0n Research and Development in Information Retrieval (SIGIR 2005*) 2005.

[Morris and Hirst 1991] J. Morris and G. Hirst. Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Text. *Computational Linguistics*, vol.17, N¡1., pp.20-48. 1991

[Reynar 1998] Jefrey C. Reynar 1998. *Topic Segmentation : Algorithms and Applications,* PhD thesis, University of Pennsylvania.

[Roche et alii 2004] M. Roche, J. Azé, Y.Kodratoff and M. Sebag : Learning Interestingness Measures in Terminology Extraction. A ROC-based approach. *Proceedings of ROCAI* pp 81-88. 2004

[Salton et al 1983] Salton G. , Fox E.A, Wu H. 1983 Extended Boolean Information retrieval. *Communications of the ACM* 26 (12). Pp. 1022-1036.

[Swan and Jensen 2000] Swan R and D Jensen. TimeMines : Constructing timelines with

statistical models of word usage, *Proceedings of KDD-2000 Workshop on Text Mining,* pp 73-80.2000.

[Yang and Chute 1992] Yang Y. and C.G. Chute. A Linear Least Square Fit Mapping Method for Information Retrieval from Natural Language Texts. *Proceedings of COLING92*, Pp. 358-362.1992.

[Yang and Liu 1999] Yang Y., Liu X.(1999)A Re-examination of Text Categorization Methods *Proc. of the 22nd ACM SIGIR Conference*, Pp 42-49.

[Yang and Li 2005] Yang C C ; Li K. W. A heuristic method based on a statistical approach for chinese text segmentation. Journal of the American Society for Information Science and Technology, vol. 56, no13, pp. 1438-1447 2005

[Zhao and Karypis 2005] Zhao Y., Karypis G. Hierarchical Clustering Algorithms for Document Datasets. *Data Mining and Knowledge Discovery,* Vol. 10, No. 2, pp. 141Ñ168, 2005.

Summary

Evaluation Conferences in natural language processing and document processing have become a mandatory step for acknowledging different techniques in the scientific community. The *Défi Fouille de Textes* (Text Mining Challenge) aims more at enabling different researchers in the francophone world to confront their works on a given problem, than distributing winning prizes to a team, a method or a tool. The different issues of the text mining field are described in this paper. They are information retrieval, knowledge extraction and enhancement, document classification or categorisation, text segmentation, and profiling. We underline the idea that implementing challenges is a survey and overview tool for every new outcome in text mining, and is to be used as a scientific tool in understanding complex issues.