Présentation de DEFT 06 (DÉfi Fouille de Textes)

Thomas Heitz, Jérôme Azé, Mathieu Roche, Augusta Mela, Peter Peinl, Mezaour Amar Djalil

Présentation de DEFT'06 (DÉfi Fouille de Textes)

Les membres du Comité d'Organisation de DEFT'06 :
Jérôme Azé*, Thomas Heitz*, Augusta Mela**,
Amar-Djalil Mezaour***, Peter Peinl****, Mathieu Roche♯

*LRI, Université Paris-Sud,
{aze, heitz}@lri.fr
**LIRMM, Université Montpellier 3,
Augusta.Mela@univ-montp3.fr
***Exalead,
Amar-Djalil.Mezaour@exalead.com
****LIA, Université d'Avignon
peter.peinl@informatik.fh-fulda.de
♯LIRMM, Université Montpellier 2,
mroche@lirmm.fr

Résumé. Après le succès de DEFT'05 organisé en 2005 dans le cadre de la conférence TALN, une nouvelle édition de DEFT (DÉfi Fouille de Textes) a été mise en œuvre.

Le thème général de ce nouveau défi concerne la reconnaissance automatique de segments thématiques de textes écrits en français dans différents domaines. La segmentation thématique peut être utilisée pour différents objectifs. Elle permet, par exemple, d'isoler des zones répondant précisément à une requête. Ceci est particulièrement utile dans un système de recherche d'informations. La segmentation peut également être utilisée pour l'indexation de textes. Des méthodes de classification de documents peuvent également s'appuyer sur la segmentation de textes. Enfin, les approches de résumés de textes peuvent utiliser les informations liées à la segmentation thématique.

Cet article présente le défi dans sa globalité, les corpus utilisés et les difficultés spécifiques à chacun des corpus étudiés.

1 Introduction

Après le succès de DEFT'05 organisé l'année dernière, une nouvelle édition de DEFT (DÉfi Fouille de Textes) est mise en œuvre. Le défi est organisé dans le cadre de la Semaine du Document Numérique 2006 (SDN'06) à Fribourg en Suisse les 21 et 22 septembre 2006.

Le thème général du défi cette année concerne la reconnaissance automatique des segments thématiques de textes écrits en français dans différents domaines. La segmentation thématique peut être utilisée pour différents objectifs :

– La segmentation permet, par exemple, d'isoler des zones répondant précisément à une requête. Ceci est particulièrement utile dans un système de recherche d'informations.

– La segmentation peut également être utilisée pour l'indexation de textes.

– Les méthodes de classification de documents peuvent également s'appuyer sur la segmentation de textes.

– Les approches de résumés de textes peuvent utiliser les informations liées à la segmentation thématique.

Qu'est ce qu'un segment thématique ?

La définition générale d'un segment thématique est très problématique, c'est pourquoi nous avons choisi une définition différente pour chaque corpus. Nous avons privilégié la segmentation voulue par les auteurs des textes qui est aussi la plus simple à utiliser pour la préparation des corpus.

Pour les discours politiques, la segmentation thématique est basée sur la structure thématique des discours mis en ligne sur le site de référence. Chaque discours a été divisé en paragraphes thématiques lors de leur écriture ou lors de la constitution des corpus mis en ligne par l'organisme en charge de cette tâche.

Pour les lois de l'Union Européenne, les segments thématiques sont les lois.

Pour l'ouvrage scientifique, les segments thématiques à retrouver sont les différentes sections, à savoir les chapitres, sections, sous-sections et sous-sous-sections. Le but est donc de déterminer la première phrase de chaque section.

Dans la suite de cet article, nous présentons les tâches du défi, la préparation des données et enfin l'évaluation des résultats.

2 Tâches à réaliser pour DEFT'06

Pour DEFT'06, la segmentation thématique des textes s'est s'appuyée sur des corpus de différents domaines écrits en français :

– **Corpus 1 :** Discours politiques ;
– **Corpus 2 :** Textes juridiques ;
– **Corpus 3 :** Ouvrage scientifique ;

Le but du défi consiste à déterminer les segments thématiques de ces différents textes (c'est-à-dire, les premières phrases de chaque segment thématique).

Nous allons décrire avec plus de précision les corpus utilisés et la définition des segments thématiques pour chaque type de corpus :

Discours politiques. Le corpus est composé de discours politiques prononcés par des présidents de la république française (Valéry Giscard d'Estaing, François Mitterrand et Jacques Chirac). La segmentation thématique est basée sur la structure thématique des discours mis en ligne sur le site de référence (taille du corpus : environ 70 Mo).

Textes juridiques. Le corpus est composé d'articles de lois de l'Union Européenne. Les segments thématiques sont les articles des lois (taille du corpus : environ 110 Mo).

Ouvrage scientifique. L'ouvrage scientifique utilisé est le livre "Apprentissage Artificiel" d'Antoine Cornuéjols et Laurent Miclet (éditions Eyrolles). Avec ce corpus, les segments thématiques à retrouver sont les différentes sections (chapitres, sections, sous-sections, sous-sous-sections). Pour ce corpus, les titres des différentes sections ainsi que les figures, tableaux et les équations ont été supprimés. Le but est de déterminer la première phrase de chaque section (taille du corpus : environ 1 Mo).

L'approche retenue pour répondre au défi a été validée sur chacun de ces corpus. Les corpus d'apprentissage sont composés de 60% des corpus associés. Ces corpus contiennent les informations permettant d'identifier les segmentations thématiques. Les participants ont eu trois mois pour mettre en place leurs méthodes de segmentation sur les corpus d'apprentissage.

Les 40% des corpus restants ont été utilisés pour le test. Les participants ont eu deux jours pour appliquer, sur les corpus de test, les méthodes mises en œuvre sur les corpus d'apprentissage.

3 Préparation des données

Les principales phases de la préparation des données sont : récupération des données brutes, conversion au format texte, remplacement et suppression des parties non textuelles ou inexploitables, segmentation thématique et contrôle de la qualité.

3.1 Traitements spécifiques effectués sur chaque corpus

D'une manière générale, les références dont celles aux images et les en-têtes dont les titres ont été supprimés. Tous les corpus contiennent une phrase par ligne avec la particularité qu'un élément d'une liste constitue une phrase.

1. Pour les discours politiques, les en-têtes comprenant le titre, la date et l'orateur ont été supprimés. Il est à noter que pratiquement tous les discours de V. Giscard d'Estaing sont capitalisés, contrairement aux discours de F. Mitterrand et J. Chirac. Il existe aussi des entretiens politiques entre des journalistes et des hommes politiques dans ce corpus. De plus, les textes constituant le corpus politique sont tels qu'ils contiennent au moins un des trois intervenants suivants : V. Giscard d'Estaing, F. Mitterrand ou J. Chirac et ce pas nécessairement en tant que président en exercice.

2. Pour les lois de l'Union Européenne, les références aux images, les en-têtes, l'article final de signature et les lois de moins de 10 phrases ont été supprimés. Les références sont écrites sous la forme [REFERENCE], par exemple [EMPLACEMENT TABLE] est la plus courante. Les numéros des articles, chapitres, titres et annexes ont été remplacés par la lettre "X".
3. Pour l'ouvrage scientifique, les formules ont été partiellement converties en texte, les références aux images remplacées par [FIGURE], les citations d'autres articles remplacées par [CITATION] et les autres références remplacées par [REFERENCE].

3.2 Principales difficultés rencontrées

La phase de récupération des données brutes ne nous a posé problème que pour les discours politiques car nous avons dû écrire un programme qui envoie des requêtes sur le serveur contenant la base de données afin de récupérer un à un les fichiers.

Une difficulté qui nous a pris beaucoup de temps à résoudre concerne la conversion des fichiers au format LaTeXde l'ouvrage scientifique vers le format texte. Nous n'avons pas trouvé de logiciel libre permettant cette conversion ce qui nous a obligé à en écrire un nous même. Malheureusement, celui-ci n'atteignait pas la qualité que nous voulions à cause du manque de temps dont nous disposions pour le mettre au point. Nous avons fini par utiliser deux logiciels éprouvés, latex2html et lynx, pour convertir en deux étapes avec le langage HTML comme intermédiaire.

Les formules dans l'ouvrage scientifique ont constitué un autre problème étant donné que certaines pouvaient être converties en texte à partir de leur écriture LaTeXet d'autres non. Nous avons choisi de convertir les formules les plus simples en texte et de supprimer celles plus complexes. Le remplacement par une balise [FORMULE] est aussi envisageable.

La définition de ce qu'est une en-tête et une signature dans chaque segment thématique a été parfois difficile. Par exemple, sur les discours politiques, il n'est pas toujours facile de bien repérer la date qui termine l'en-tête ou encore la signature de l'auteur du texte. Ce sont pourtant des biais d'apprentissage qu'il faut supprimer si on ne veut pas rendre évidente la tâche.

Le petit nombre de personnes disponibles pour le contrôle des résultats et l'importante masse de données des corpus, environ 180 Mo, nous a obligé à prendre des échantillons au hasard dans les corpus pour vérifier s'ils étaient conformes à notre spécification.

Nous avons choisi de laisser les fautes d'orthographes notamment sur l'ouvrage scientifique. Ceci entraîne la perturbation de l'analyse à base de dictionnaires puisque les mots mal orthographiés ne s'y trouvent pas. Pourtant c'est une réalité que les textes comportent tous une certaine proportion de fautes d'orthographe.

4 Déroulement du défi

4.1 Corpus d'apprentissage

À partir du 13 février 2006, nous avons mis le corpus d'apprentissage à la disposition des équipes s'étant inscrites à DEFT'06.

Les trois corpus sont séparés et chaque ligne des corpus d'apprentissage est identifiée par un numéro de corpus (voir la section tâches, répertoriant les trois corpus utilisés), un numéro de ligne du corpus et un caractère "x" pour identifier la première ligne de chaque nouveau segment thématique.

Notons ci-dessous un exemple de fragment du corpus d'apprentissage de DEFT'06.

```
...
<2.22> ligne 22 du corpus 2 (textes juridiques).
<2.23x> ligne 23 du corpus 2, le "x" signifie que cette phrase correspond
au début d'un segment thématique.
<2.24> ligne 24 du corpus 2.
<2.25> ligne 25 du corpus 2.
...
```

4.2 Corpus de test

Les participants ont eu deux jours après la mise à disposition des corpus de test pour prédire les différents segments thématiques. Les corpus de test sont composés des numéros de corpus et de phrases sans indication des segments thématiques (sans présence du caractère "x" qui indiquait les premières phrases relatives aux segments thématiques du corpus d'apprentissage).

4.3 Évaluation des résultats

4.3.1 Définition du F_{score} utilisé pour le classement final

Chacune des exécutions a été évaluée en calculant le F_{score} pour chacun des corpus (avec $\beta = 1$, voir formule (1)).

$$F_{score}(\beta) = \frac{(\beta^2 + 1) \times Précision \times Rappel}{\beta^2 \times Précision + Rappel} \tag{1}$$

Dans le cadre de DEFT'06, le calcul du F_{score} pour chaque corpus peut être réécrit de la manière suivante (formule (2)) :

$$F_{score} = \frac{2 \times (nb\_phrases\_correctes\_extraites)}{nb\_total\_extraites + nb\_total\_correctes} \tag{2}$$

- $nb\_phrases\_correctes\_extraites$ correspond au nombre de phrases évaluées de manière correcte comme un début de fragment thématique données dans le fichier résultat.
- $nb\_total\_extraites$ correspond au nombre de phrases données dans le fichier résultat.
- $nb\_total\_correctes$ correspond au nombre total de phrases représentant un début de fragment thématique.

4.3.2 Définition du "F_{score} souple" utilisé pour comparer les résultats

Outre l'utilisation du F_{score}, dans le cadre de DEFT'06 nous avons introduit une mesure de "F_{score} souple". Cette mesure consiste à considérer comme correctes les phrases situées autour de la phrase caractérisant le début d'un segment thématique. Le "Fscore souple" a été évalué pour

1. une fenêtre contenant 1 phrase avant et après la phrase à prédire
2. une fenêtre contenant 2 phrases avant et après la phrase à prédire

Afin de définir le "F_{score} souple", considérons les différentes définitions suivantes utilisées dans l'algorithme propre à son calcul :

- Soit $\mathcal{V}$ l'ensemble des indices correspondant à la vérité.
- Soit $\mathcal{R}$ l'ensemble des indices des réponses d'une équipe.
- Soit t_f la taille de la fenêtre utilisée pour calculer le F_{score} souple.
  Nous considérons que $t_f \in \{0, 1, 2\}$ avec la sémantique suivante associée aux différentes valeurs de t_f :
  - $t_f = 0$: F_{score} classique. Aucune marge d'erreur n'est accordée autour de la vérité.
  - $t_f = 1$ (resp. $t_f = 2$) : F_{score} souple. Une marge d'erreur de ± 1 indice (resp. ± 2) autour de la vérité est accordée.
- Soit $\mathcal{I}$ l'ensemble contenant l'indice solution et les indices situés dans la marge d'erreur autour de l'indice solution.
- Soit $\mathcal{E}$ l'ensemble des indices appartenant aux réponses et considérés comme juste pour un indice solution donné (avec une marge d'erreur t_f).

Quelque soit la valeur de t_f, le calcul de la précision, du rappel et du F_{score} est obtenu grâce à l'algorithme 1.

Nous donnons ci-dessous une des propriétés du F_{score} souple :
$\forall t_f \; F_{score}(t_f) \leq F_{score}(t_f + 1)$

4.3.3 Algorithme utilisé pour désigner le vainqueur de DEFT'06

Les équipes sont classées en fonction des rangs obtenus sur l'ensemble des corpus et en considérant chaque soumission comme atomique. Le rang d'une soumission est donc égal à la somme des rangs associés au F_{score} de cette soumission sur chaque corpus. L'algorithme qui est utilisé est présenté ci-dessous sur la figure 2.

4.3.4 Le critère d'évaluation permet-il une marge d'erreur ?

Le critère d'évaluation pourrait être un problème sur les petits corpus tel que le corpus scientifique puisqu'il risque d'y avoir autant de chance qu'un système qui réponde avec une marge d'erreur de plus ou moins 5 phrases et un autre avec une marge d'erreur de plus ou moins 10 phrases aient les mêmes scores. Cependant, sur les plus gros corpus que sont les deux autres, cet effet est minimisé. Nous avons donc décidé d'élaborer une mesure permettant de comparer les résultats avec une marge d'erreur en plus de la mesure officielle qui reste la détection ou non de la phrase débutant le segment.

début
  pour chaque *($t_f \in \{0, 1, 2\}$)* **faire**
    prédictions(justes) = 0
    /* $\mathcal{R}'$ est une copie de $\mathcal{R}$ qui sera utilisée pour calculer le nombre de prédictions incorrectes. */
    $\mathcal{R}' = \mathcal{R}$
    pour chaque *($indice \in \mathcal{V}$)* **faire**
      /* Construction de la marge d'erreur autour de l'indice courant. */
      $\mathcal{I} = \emptyset$
      pour *($erreur$ allant de $-t_f$ à $+t_f$)* **faire** $\mathcal{I} \leftarrow \{indice + erreur\}$

      /* Recherche de l'indice (avec la marge d'erreur) dans la liste des réponses $\mathcal{R}$. */
      $\mathcal{E} = \mathcal{I} \cap \mathcal{R}$
      si *($\mathcal{E} \neq \emptyset$)* **alors**
        /* L'ensemble des réponses contient l'indice courant, avec la marge d'erreur t_f. */
        /* Toutes les réponses situées dans la marge d'erreur autour de l'indice solution sont considérées comme justes. */
        $\mathcal{R}' \leftarrow \mathcal{R}' - \mathcal{E}$
        prédictions(justes)++
      fin Si
    fin
    /* L'ensemble $\mathcal{R}'$ contient les indices qui n'ont été associés à aucune solution. */
    prédictions(incorrectes) = $|\mathcal{R}'|$
    prédictions(effectuées) = prédictions(justes) + prédictions(incorrectes)

    /* Calcul du F_{score}. */
    $précision(t_f) = \frac{prédictions(justes)}{prédictions(effectuées)}$
    $rappel(t_f) = \frac{prédictions(justes)}{|\mathcal{V}|}$
    $F_{score}(t_f) = \frac{(\beta^2+1) \times précision(t_f) \times rappel(t_f)}{\beta^2 \times précision(t_f) + rappel(t_f)}$
  fin
fin

FIG. 1 : CALCUL DU F_{score} SOUPLE

début
  pour chaque *(*$corpus \in \{discours, lois, scientifique\}$*)* **faire**
    /* Calcul des rangs pour chaque soumission de chaque équipe */
    $\mathcal{S}_{corpus}$ = Tri décroissant des soumissions selon le $Fscore_{strict}(\beta = 1)$
    /* $\mathcal{S}_{corpus}$ est un tableau trié dont les index sont les couples (equipe, soumission) */
    pour *(rang=1 ; rang*$\leq$ *dernière soumission ; rang++)* **faire**
      rangs[$\mathcal{S}_{corpus}[rang] \rightarrow equipe$] [$\mathcal{S}_{corpus}[rang] \rightarrow soumission$] [corpus] = rang;
    fin Pour
  fin
  pour tous les *(equipe ayant soumis)* **faire**
    pour tous les *(soumission de equipe)* **faire**
      rangs[equipe][soumission] = $\sum_{\forall corpus}$ rangs[equipe][soumission][corpus];
    fin
  fin
  Pour chaque équipe, la soumission ayant le rang global le plus faible est retenue
  L'équipe ayant le rang le plus faible est désignée comme vainqueur.
fin

FIG. 2 : CALCUL DU RANG DES ÉQUIPES

5 Conclusion

Cet article présente le Défi DEFT'06 consistant à déterminer les ségments thématiques dans trois corpus écrits en français qui ont des thèmes et des tailles très différentes. Outre le principe général du défi, cet article détaille la préparation des données en mettant en relief les difficultés rencontrées qui sont souvent spécifiques aux corpus traités. Enfin, les différents critères d'évaluation (F_{score} et F_{score} souple) ainsi que l'algorithme utilisé pour désigner le vainqueur de DEFT'06 sont détaillés.

Ayant présenté de manière précise les critères d'évaluation retenus pour DEFT'06, une dissussion de l'ensemble des résultats obtenus par les 7 équipes qui ont participé au défi qui sont issues de laboratoires différents pourra être menée dans le cadre de l'atelier.

6 Comités

6.1 Comité d'Organisation

Responsables : Mathieu Roche (LIRMM, TAL) et Jérôme Azé (LRI, BioInfo)

Membres :

- Thomas Heitz (LRI, I&A)
- Augusta Mela (Université Montpellier 3)
- Amar-Djalil Mezaour (Exalead)
- Peter Peinl (LIA)

6.2 Comité de Programme

Présidents : Violaine Prince (LIRMM) et Yves Kodratoff (LRI)

Membres :

- Nathalie Aussenac-Gilles (IRIT)
- Catherine Berrut (CLIPS)
- Fabrice Clérot (France Telecom)
- Guillaume Cleuziou (LIFO)
- Béatrice Daille (LINA)
- Marc El-Bèze (LIA)
- Patrick Gallinari (LIP6)
- Éric Gaussier (Xerox Research)
- Thierry Hamon (LIPN)
- Fidélia Ibekwe (URSIDOC-SII)
- Michèle Jardino (LIMSI)
- Éric Laporte (IGM-LabInfo)
- Josiane Mothe (IRIT)
- Xavier Polanco (INIST)
- Pascal Poncelet (LGI2P)
- Christian Retoré (LABRI)
- Christophe Roche (LISTIC)
- Pascale Sébillot (IRISA)
- Yannick Toussaint (LORIA)
- François Yvon (ENST)

Remerciements

Le Comité d'Organisation de DEFT'06 remercie les associations AFIA et EGC pour avoir parrainé DEFT'06. Nous remercions également le Comité d'Organisation de la Semaine du Document Numérique d'avoir accepté la mise en œuvre de DEFT'06 dans le cadre de cette manifestation d'envergure internationale.

Summary

As the previous edition of DEFT (DÉfi Fouille de Textes), organised in conjonction with the TALN conference, can be consider as a success, a new edition of DEFT has been organised in conjonction this year. The general topic of this new challenge relates to the automatic recognition of thematic part of texts, written in French, in various fields. The thematic segmentation can be used for various objectives. It allows, for example, to find part of text than answer a request. This is particularly useful in a retreival information system. The segmentation can also be used for the indexation of texts. Classification methods can also be based on texts segmentation. Finaly, text summarization can use information related to thematic segmentation.

This paper presents the corpora used and the specific difficulties of each corpus.