SANDOH, un outil pour analyser des textes hétérogènes

Vo Trung Hung

GETA-CLIPS-IMAG, Institut National Polytechnique de Grenoble
385, rue de la Bibliothèque – BP 53 – 38041 Grenoble Cedex 9 – France
Tél : +33 4 76 51 48 17 – Fax : +33 4 76 51 44 05
Hung.Vo-Trung@imag.fr

Abstract

The knowledge of the language and the coding used for the texts helps us to correctly carry out treatments on these texts as the exact search for information, the checking of orthography, the visualizing of information, the exchange of information between different systems, etc. In this article, we present the construction of a tool to analyse a heterogeneous text by method diagnoses N-grams and method of progressive segmentation to segment a heterogeneous text in homogeneous zones. This tool makes it possible to analyse a text to identify the language and coding used in this text. For the homogeneous text, the result is a couple <language, coding>, for the heterogeneous text, the result is a table of the zones with the language and corresponding coding {zone-1, language-1, coding-1}, {zone-2, language-2, coding-2}, …, {zone-n, language-n, coding-n}.

Résumé

La connaissance de la langue et du codage utilisé dans les textes aide à réaliser correctement des traitements sur ces textes comme la recherche exacte d'informations, la vérification d'orthographe dans la langue utilisée, l'affichage d'informations avec le bon jeu de caractères correspondant au codage utilisé, l'échange d'informations entre systèmes différents, etc. Dans cet article, nous présentons la construction d'un outil pour analyser un texte hétérogène selon méthode du diagnostic n-grammes et méthode de la segmentation progressive pour segmenter un texte hétérogène en zones homogènes. Le résultat de l'analyse est le nom d'un couple <langue, codage> si ce document est homogène, sinon, les zones et le couple <langue, codage> utilisé dans chaque zone {zone-1, langue-1, codage-1}, {zone-2, langue-2, codage-2}, …, {zone-n, langue-n, codage-n}.

Mots-clés : multilinguisme, texte multilingue, texte homogène, texte hétérogène, traitement de la langue naturelle, segmentation, classification, multilinguïsation.

1. Introduction

Actuellement, on peut traiter tous les systèmes d'écriture par l'ordinateur, mais on doit utiliser souvent plusieurs systèmes de codage pour présenter un système d'écriture (par exemple : ISO 8859-1, CP1252, CP 850, Macintosh, … pour le français ; ISO 8859-1, CP1252, TCVN, VNI, VPS, … pour le vietnamien ; ISO 8859-5, CP 1251, CP 886, KOI 8, Mac Cyrillic, … pour le russe) et c'est difficile de traiter ces données. Malgré Unicode, les utilisateurs rencontrent beaucoup de problèmes avec des fichiers contenant des couples de langues et de codages très différents. En pratique, on a souvent l'expérience frustrante de vouloir ouvrir une page Web ou un fichier, d'essayer tous les codages proposés, et de ne pas arriver à la lire, sans parler de l'éditer.

La connaissance de la langue et du codage utilisé dans les textes aide à réaliser correctement des traitements sur ces textes comme la recherche exacte d'informations, la vérification d'orthographe dans la langue utilisée, l'affichage d'informations avec le bon jeu de caractères correspondant au codage utilisé, l'échange d'informations entre systèmes différents. En parti-

culier, avec les documents hétérogènes (écrits en plusieurs langues et plusieurs codages), l'identification et la segmentation nous permettent traiter correctement pour chaque zone de langue et de codage différent. Par exemple, à chaque zone nous appliquons un dictionnaire propre pour vérifier les fautes d'orthographe.

Nous présentons ici l'outil SANDOH (Système d'ANalyse des DOcuments Hétérogènes) que nous avons construit pour analyser un texte hétérogène en zones homogènes. Le résultat du système est le nom d'un couple <langue, codage> si le document est homogène, sinon, les zones et le couple <langue, codage> utilisé dans chaque zone.

2. État de l'art

Cette section aborde quelques systèmes récents d'identification automatique la langue et le codage du texte homogène (écrit en seule langue et seul codage). La liste n'est pas exhaustive.

2.1. *SILC (Système d'Identification de la Langue et du Codage)*

Le logiciel SILC est développé au laboratoire RALI de l'Université de Montréal principalement par Pierre Plamodon. SILC reconnaît 28 langues encodées avec en moyenne trois encodages par langue. En totalité, il peut identifier 73 couples <L, C> différents. SILC utilise associer les unigrammes et trigrammes pour évaluer des textes. C'est un bon outil diagnostique des textes homogènes (Bouffard, 2002 ; Vo-Trung, 2003).

2.2. *TextCat*

TextCat est un système développé par Gertjan Van Noord de l'Université de Groningen (http ://odur.let.rug.nl/~vannoord/TextCat/). Ce système implante une méthode basée sur les ngrammes telle que décrite par Wiliam Cavnar et John Trenke. Il reconnaît en tous 70 langues et un total de 77 paires <L, C>.

2.3. *Xerox Mltt Language Guesser*

Ce système a été développé par Xerox et est commercialisé par Inxight Software. Il reconnaît 47 langues chacune encodée avec en moyenne 2 jeux de caractères pour un total d'exactement 93 couples <L, C> (www.xrce.xerox.com/research/mltt/tools/guesser.html).

2.4. *Lextek Language Identifier*

Lextek est un système de l'identification automatique de la langue et du codage de Lextek International. Ce système peut diagnostiquer pour 260 couples <L, C> différents (www.languageidentifie.com).

3. Objectif

Ces systèmes s'appliquent principalement à l'identification des fichiers homogènes en format texte ou HTML. L'exactitude du diagnostic varie beaucoup suivant la longueur du texte analysé. Avec les textes de plus de 50 lettres, le diagnostic est assez bon mais la qualité de diagnostic est trop mal avec les textes inférieurs à 30 lettres (Vo-Trung, 2003). Un problème posé qu'il faut chercher une nouvelle solution pour analyser des textes très courts.

Notre but est de proposer un système d'analyse des textes hétérogènes en zones homogènes sur la base du système d'identification d'un texte homogène. Nous proposons à utiliser des dictionnaires monolingues pour diagnostiquer des textes très courts.

Dans la partie suivante, nous présentons l'architecture du système et les méthodes utilisées pour ce système.

4. Système d'analyse d'un document hétérogène

Dans les systèmes multilingues, on doit souvent travailler avec des données hétérogènes qui sont écrites en plusieurs langues et peut-être sous des codages différents. Par exemple, les documents français contiennent des citations anglaises ou bien les dictionnaires contiennent plusieurs articles écrits dans des langues différentes. La connaissance de la langue et du codage utilisé dans chaque zone de ces textes aide à réaliser correctement des traitements à la suite.

4.1. *Architecture du système*

Le système d'analyse d'un texte hétérogène est organisé :

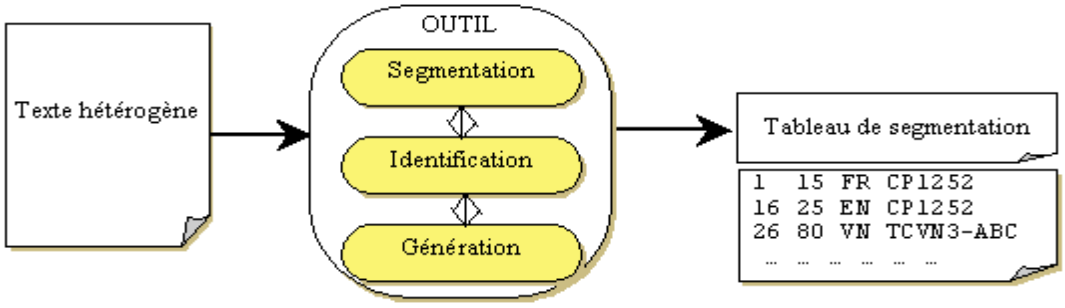

Figure 1. Outil d'analyse d'un texte hétérogène en zones homogènes

Cet outil a trois fonctions principales : la segmentation, l'identification et la génération.

- Le module de segmentation permet de segmenter un texte en zones à évaluer. Chaque zone est définie par la première et la dernière position de lettres de la zone. Par exemple, la zone 1 est définie par (p_{1z1}, p_{2z1}), la zone 2 par (p_{1z2}, p_{2z2}),…, la zone n par (p_{1zn}, p_{2zn}).
- Le module d'identification permet de diagnostiquer la zone actuelle pour évaluer si elle est hétérogène ou homogène. Si elle est homogène, il renvoie le résultat (zone, langue, codage) au module de génération et passe à une autre zone. Sinon, il renvoie cette zone au module de segmentation qui la segmentera en zones plus petites.
- Le module de génération permet d'écrire le résultat obtenu dans un fichier. Ce résultat est une table, chaque ligne de cette table présente une zone (la position de la première lettre et de la dernière lettre) avec la langue et le codage utilisé sur cette zone.

4.2. *Segmentation*

Pour segmenter un texte, nous utilisons une technique de segmentation progressive basée sur les signes de ponctuation. Normalement, les parties écrites dans des langues différentes sont séparées par des signes de ponctuation comme le point, le tiret, les deux-points, la parenthèse, le guillemet, le point d'exclamation, le point-virgule, etc. L'idée ici est qu'après avoir évalué une zone, si elle est hétérogène, on continue de séparer cette zone en zones plus petites pour l'évaluer. Mais il faut s'assurer que ces zones ne seront pas trop courtes pour l'identification.

D'abord, nous évaluons chaque paragraphe (la terminaison de paragraphe est un signe de retour à la ligne) pour diagnostiquer s'il est homogène ou hétérogène. S'il est homogène, on

passe à un autre paragraphe, sinon, il faut découper ce paragraphe en deux zones. Il suffit de séparer le texte au milieu et ensuite de tester ce qui arrive lorsque l'on se déplace d'un mot vers la gauche ou vers la droite jusqu'à obtenir que chaque zone contienne un ensemble de phrases. Nous continuons d'évaluer et de séparer cette zone en zones plus petites et à évaluer jusqu'à obtenir une zone homogène.

4.3. *Identification de la langue et du codage d'une zone homogène*

4.3.1. *Architecture générale du système d'identification de la langue et du codage*

Il y a en général deux phases pour identifier la langue et le codage : la première phase crée un modèle du couple <langue, codage> (seulement une fois pour chaque couple) et la deuxième phase identifie la langue et le codage du texte à analyser.

L'identification de la langue et du codage est réalisée comme le montre le schéma ci-dessous (Russell, 2002) :

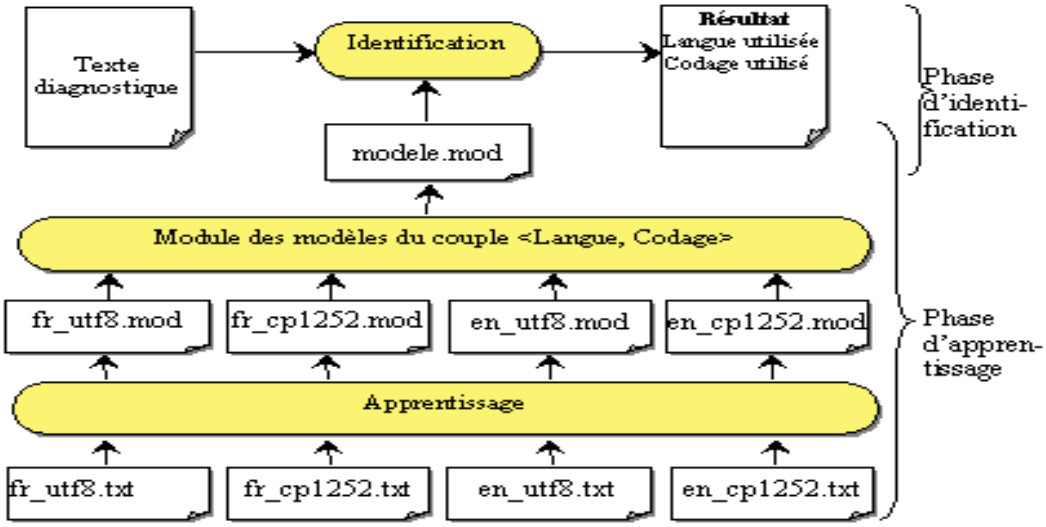

Figure 2. Architecture générale d'un identifieur de langue et de codage

La phase d'apprentissage est construite sur la base d'un modèle statistique. Au début, on doit avoir des fichiers annotés par des balises <langue, codage>. Le module d'apprentissage crée ensuite les « modèles » correspondant à chaque couple <langue, codage>. Ces modèles sont construits à partir de la fréquence des séquences qu'on a comptées dans le fichier d'apprentissage. Puis on réalise la synthèse de ces modèles pour réunir les fichiers obtenus en un fichier unique qui constitue le modèle du couple <langue, codage>. Ce fichier est utilisé pour identifier les autres textes dans le futur.

La phase d'identification permet de déterminer en quelle langue est écrit un texte et avec quel codage. Elle utilise la même méthode en comparant des segments dans le texte à analyser avec des séquences du modèle de langues pour évaluer le texte à analyser.

4.3.2. *Méthode n-grammes*

On calcule d'abord la fréquence des séquences de ***n caractères*** dans un grand fichier d'apprentissage. L'idée ici est de détecter des motifs récurrents dans des suites d'observations (Cavnar et Trenkle, 1994). Par exemple, les mots anglais se terminant par ***-ck*** sont plus nombreux que les mots français avec la même terminaison. Par contre, les mots français se termi-

nant par ***-ez*** sont plus nombreux que les mots anglais présentant la même fin. Ces motifs récurrents peuvent avoir une longueur variable.

L'algorithme d'apprentissage est un algorithme itératif sur la base d'apparition des éléments des séquences n lettres dans le fichier d'apprentissage (Beesley, 1998). Au cours de ces itérations, les segmentations du corpus évoluent, faisant émerger les séquences d'observation les plus typiques.

Après apprentissage, un modèle du couple <langue, codage> est créé, contenant les séquences sélectionnées les plus probables et leur vraisemblance.

Par exemple, avec le texte d'apprentissage « *les chiens et les chats sont des animaux* », on va obtenir les modèles de langage n-grammes suivants :

Tableau 1.
Modèles de langage avec leurs séquences et leurs fréquences

Unigrammes		**Bigrammes**		**Trigrammes**	
Séq	**Fréq**	**Séq**	**Fréq**	**Séq**	**Fréq**
_	7	s_	5	es_	3
s	6	es	3	les	2
e	5	le	2	s_c	2
a	3	_c	2		
n	3	ch	2		
t	3				

La phase d'identification consiste à calculer la probabilité sur le texte à analyser et sur les modèles du couple <langue, codage>. On choisit alors la langue et le codage correspondant au modèle du couple <langue, codage> qui est le plus probable.

On peut calculer le taux de certitude qui caractérise une langue et un codage correspondant à chaque modèle du couple <langue, codage> par la formule (Manning et Schutze, 1999) :

$$l = \sum_{i=1}^{N} P_i \qquad (1)$$

Avec :

$$P_i = \begin{cases} \log\left(\frac{f(w_i)}{N} x \frac{F(w_i)}{M}\right) & , \text{si } w_i \in \text{texte} \\ 0 & , \text{si } w_i \notin \text{texte} \end{cases}$$

N = nombre de séquences dans le modèle du couple <langue, codage> actuel

$f(w_i)$ = fréquence de la séquence w_i dans le modèle du couple <langue, codage> actuel

$F(w_i)$ = fréquence de la séquence w_i dans le texte à analyser (si $w_i \in$ texte)

M = nombre de séquences dans le texte à analyser

On conclut sur la langue et le codage appliqué sur le texte en choisissant le modèle du couple <langue, codage> qui a le taux de certitude maximum.

C'est une méthode utilisable pour tout couple <langue, codage>. La construction du modèle et l'évaluation du texte ne dépendent pas de la langue et du codage, elles consistent à analyser les segmentations des textes.

On peut changer la variable n pour obtenir des modèles différents tels qu'unigramme (n=1), bigrammes (n=2), trigrammes (n=3), tétragrammes (n=4),… et on peut aussi associer des modèles pour augmenter l'exactitude du système.

Dans notre système, nous avons utilisé un modèle du couple <langue, codage> sur des n-grammes avec n variant de 1 à 4, et chaque modèle contenant les 400 séquences de plus haute fréquence.

4.3.3. *Résultat obtenu*

Nous avons développé un outil d'identification d'un texte homogène selon la méthode ci-dessus pour plusieurs langues et codages différents. Le système actuel permet de diagnostiquer 53 langues comme l'anglais, le français, le vietnamien, le chinois, le russe, l'arabe, etc. Nous pouvons facilement ajouter d'autres langues par fourniture des fichiers d'apprentissage.

Le résultat obtenu est un tableau de scores correspondant aux modèles du couple <langue, codage> analysés et nous choisissons la langue et le codage selon le modèle du couple <langue, codage> de plus grand score.

Voici un exemple de résultat sur un texte français « *les chiens et les chats sont des animaux* » et un texte chinois 地定空厘混沌。渊面黑暗。神的灵运行在水面上 avec les 10 scores les plus hauts :

Texte français			**Texte chinois**	
Langue-codage	Score		Langue-codage	Score
Français	255.8934		Chinois-gb2312	344.3639
Roman	236.9501		Chinois-big5	234.7359
Écossais	231.5291		Coréen	222.3591
Espagnol	211.8286		Arabe-windows1256	170.5808
Latin	195.9060		Tamoul	169.7671
Slovene-iso8859_2	195.1329		Arabe-iso8859_6	140.7884
Irlandais	183.8261		Ukrainien-koi8_r	135.7941
Quechua	167.9354		Japonais-shift_jis	128.0966
Slovaque-windows1250	167.8815		Thaï	119.0824
…	…		…	…

Tableau 2. Tableau de scores de résultat diagnostique

4.4. *Évaluation des zones*

Pour chaque zone homogène identifiée, le système doit calculer le taux de certitude ou le "ratio" de chaque couple (voir la formule (1)). Comme ce ratio peut varier de façon importante en fonction du couple, on peut l'utiliser pour l'identification. Il suffit de calculer l'erreur relative entre le ratio du meilleur couple et celui du second couple :

$$q = \frac{l_1 - l_2}{l_1} \qquad (2)$$

avec :

l_1, le taux de certitude maximum calculé selon la formule (1) avec le modèle du couple <langue, codage> correspondant.

l_2, le taux de certitude secondaire calculé avec la formule (1).

L'idée ici est qu'après le calcul d'erreur relative d'une zone, si ce taux est inférieur à une valeur λ, on continue de séparer cette zone en zones plus petites à évaluer.

Exemple : avec λ=0.25 (méthode expérimentale), sur la zone actuelle, nous avons $l_1 = 0.7$, $l_2 = 0.3$, alors $q = (0.7 - 0.3)/0.7 = 0.57$ (>λ). Nous décidons que cette zone est homogène et choisissons la langue et le codage selon le modèle du couple <langue, codage> du taux l_1. Si $l_1 = 0.7$, $l_2 = 0.6$, alors $q = 0.14$ et nous décidons que cette zone est hétérogène et il faut la séparer en zones plus petites pour l'évaluer ou il faut décider la langue et le codage selon le modèle du couple <langue, codage> de l_1 si cette zone est trop petite pour continuer la séparation.

5. Expérimentation

Nous avons construit un site web pour démonstration d'outil du diagnostic et de la segmentation des textes hétérogènes. Les utilisateurs entrent un texte hétérogène ou choisissent un fichier disponible. Ce système analysera et rendra le résultat qu'il est un couple <langue, codage> utilisé si ce texte est homogène, sinon, il est un texte sous la forme XML avec des balises marquées de couple <langue, codage> d'une zone homogène. Par exemple, avec le texte "*Life is rarely as we would like it to be rather it is exactly as it is : C'est la vie!*", le résultat obtenu est un texte sous la forme XML :

```
<En-CP1252>
    Life is rarely as we would like it to be rather it is     exactly as it is :
</ En-CP1252>
<Fr-CP1252>
    C'est la vie!
</Fr-CP1252>
```

Nous avons testé cet outil avec des textes écrits en plusieurs langues et codages différents comme russe/KOI8, chinois/Big5, japonais/shift-jis...

6. Conclusion et perspective

Nous avons développé un outil en ligne pour analyser des textes homogènes et hétérogènes. Cet outil permet diagnostiquer pour 53 langues et avec plusieurs systèmes de codage différents. De plus, on peut facilement ajouter les autres langues et codages par les fichiers textes d'apprentissage. Cependant, le diagnostic des textes courts n'est pas encore bon. Avec les textes inférieurs à 30 lettres, le taux exact du diagnostic est environ 70% (Vo-Trung, 2003). Nous pensons à utiliser des dictionnaires monolingues pour évaluer des mots et des groupes de mots pour les zones très courts mais il n'est pas encore une bonne solution.

Nous sommes en train d'étudier pour appliquer cet outil dans des systèmes de traduction automatique multilingue. Il sert à segmenter et diagnostiquer la langue pour déterminer les couples langues originales et cibles de traduction de chaque zone monolingue.

Références

Beesley K. R. (1998). Language identifier : A computer program for automatic natural language identification of on-line text. In *Language at Crossroads : Proceedings of the 29th Annual Conference of the American Translators Association*.

Benny G. (2001). *Reconstruction et Utilisation de SILC*. Rapport de stage, Département d'Informatique et de recherche opérationnelle, Université de Montréal.

Bouffard V. (2002). *Évaluation de SILC*. Rapport scientifique, Département d'Informatique et de recherche opérationnelle, Université de Montréal.

Cavnar W. et Trenkle J. (1994). N-gram Based Text categorization. *Symposium On Document Analysis and Information Retrieval*, University of Nevada.

Giguet E. (1998). *Méthode pour l'analyse automatique de structures formelles sur documents multilingues*. Thèse, Université de Caen.

Grefenstette G. (1995). Comparing Two Language Identification Schemes. In *Actes des JADT 1995*.

Manning C. et Schutze H. (5-1999). *Foundations of Statistical Natural Language Processing*. MIT Press.

Russell G. (2002). *The QUE Language and Encoding Identification Package*. RALI, Université de Montréal.

Vo-Trung H. (2003). Évaluation des méthodes et des outils actuels pour identifier automatiquement la langue et le codage d'un texte homogène. In *Conférence MAJECSTIC'03*.