Fiabilité de la référence humaine dans la détection de thème

Armelle Brun, Kamel Smaïli

TALN 2004, Fès, 19–21 avril 2004

Fiabilité de la référence humaine dans la détection de thème

Armelle Brun, Kamel Smaïli
LORIA - Université Nancy2
Campus Scientifique - BP 239
54506 VANDOEUVRE-lès-NANCY
{brun,smaili}@loria.fr

Résumé - Abstract

Dans cet article, nous nous intéressons à la tâche de détection de thème dans le cadre de la reconnaissance automatique de la parole. La combinaison de plusieurs méthodes de détection montre ses limites, avec des performances de 93.1 %. Ces performances nous mènent à remettre en cause le thème de référence des paragraphes de notre corpus. Nous avons ainsi effectué une étude sur la fiabilité de ces références, en utilisant notamment les mesures Kappa et erreur de Bayes. Nous avons ainsi pu montrer que les étiquettes thématiques des paragraphes du corpus de test comportaient vraisemblablement des erreurs, les performances de détection de thème obtenues doivent donc êtres exploitées prudemment.

In this paper, topic detection is studied in the frame of automatic speech recognition. Topic detection methods combination reaches 93.1% correct detection. This rate makes us throw the reference labeling back into question. We have then studied the reliability of the topic labeling of our test corpus, by using the Kappa statistics and the Bayes error. With these measures, we show the topic label of some paragraphs may be wrong, then performance of topic detection may be carefully exploited.

Mots-clefs – Keywords

Détection de thème, Etiquetage thématique, statistique Kappa, erreur de Bayes
Topic detection, topic assignment, Kappa statistics, Bayes error

1 Introduction

Dans le cadre de la reconnaissance automatique de la parole, il a été montré que l'adaptation des modèles au signal en cours de traitement permettait une amélioration des performances. Plus concrètement, les modèles que nous évoquons ici sont tout d'abord le modèle acoustique dont l'objectif est le traitement du signal perçu. Le second modèle, quant à lui, est le modèle de langage, il représente la composante langagière d'un système de reconnaissance de la parole.

Notons A le signal acoutique perçu par le système de reconnaissance de la parole, la séquence de mots $\widehat{W}$ reconnue par ce dernier sera celle correspondant à l'équation suivante :

$$\widehat{W} = \underset{W}{argmax} \quad P(W \mid A) \tag{1}$$

Où $P(W \mid A)$ est la probabilité que la suite de mots W soit reconnue sachant que A a été perçu. En utilisant la règle de Bayes, nous pouvons obtenir :

$$P(W \mid A) = \frac{P(A \mid W) \cdot P(W)}{P(A)} \tag{2}$$

Où $P(A \mid W)$ est la probabilité que le signal A soit perçu sachant que la suite de mots W a été prononcée, elle est évaluée par le module acoustique (Haton *et al.*, 1991; Calliope, 1989). $P(W)$, quant à lui, représente la probabilité de la suite de mots composant W, elle est évaluée par le modèle de langage.

Le modèle auquel nous nous intéressons ici est le modèle de langage et plus particulièrement son adaptation au signal en cours de reconnaissance. De nombreux travaux ont porté sur l'adaptation des modèles de langage, montrant à la fois un gain en perplexité des modèles de langage mais également en taux de reconnaissance (Kuhn & De Mori, 1990; Seymore & Rosenfeld, 1997). Nous nous penchons ici sur un cas particulier de l'adaptation des modèles de langage, celui de l'adaptation au thème de la séquence en cours de reconnaissance. Cette phase d'adaptation au thème doit passer par une étape préalable de détection dudit thème.

La seconde section de cet article présente le domaine de la détection de thème : méthodes de détection et d'évaluation des performances sur des corpus textuels, puis détaille les performances effectives sur nos données. La troisième partie s'intéresse à la remise en cause de l'étiquette de référence des documents textuels sur lesquels se sont faits les tests de performance. La quatrième partie présente le gain effectif de perplexité atteint grâce à l'adaptation des modèles de langage. Une conclusion et des perspectives à ce travail sont présentées dans la dernière section.

2 La détection de thème

La détection de thème consiste, sachant un ensemble prédéfini de thèmes et un document donné, à assigner une étiquette thématique au document. Un thème dans notre cas est considéré comme étant le sujet traité par un ensemble documents. Pour permettre l'exploitation de ces thèmes, nous utilisons un corpus, composé de documents traitant des thèmes auxquels nous nous intéressons.

2.1 Les corpus

Le corpus que nous utilisons dans notre étude est issu du journal *Le Monde* des années 1987 à 1991. Les données, présentées sous forme d'articles, nous sont fournies déjà classées. Chacune de ces classes représente un secteur de rédaction du journal, et les thèmes que nous étudions seront ces mêmes secteurs de rédaction.

De ce corpus, nous devons extraire un corpus de test. Nous choisissons volontairement de ne traiter que des articles relatant d'un seul thème. Etant donné qu'il est plus probable qu'un article entier traite de plusieurs thèmes, nous choisissons de travailler au niveau du paragraphe que nous jugeons être plus probable de ne traiter que d'un seul thème.

Le corpus de test que nous avons sélectionné est donc composé de 835 paragraphes. Pour l'étiquetage des paragraphes composant ce corpus, nous avons demandé à un humain d'affecter une étiquette thématique à ces paragraphes. Ce réétiquetage des paragraphes pourra ajouter un biais à notre expérience (que nous jugeons cependant faible), les paragraphes d'apprentissage n'étant pas réétiquetés par les humains.

Le corpus d'apprentissage est composé des paragraphes ne participant pas au test.

2.2 Méthodes et performances en détection de thème

Pour pouvoir détecter le thème d'un document donné, nous devons tout d'abord "apprendre" les caractéristiques de chacun des thèmes pour ensuite les repérer dans un nouveau document. Pour cela, nous représentons chaque thème par un vecteur (Salton, 1991), où chaque dimension caractérise un mot, sa valeur représentant la plupart du temps sa fréquence dans le corpus d'apprentissage. Pour détecter le thème d'un document donné il est indispensable de le représenter également vectoriellement pour pouvoir le comparer.

L'étape suivante consiste alors à comparer la représentation vectorielle du document de test avec chacune des représentations vectorielles des thèmes. Il existe deux grandes approches pour détecter le thème d'un document sachant les représentations vectorielles, l'approche statistique et l'approche machine learning, nous présentons maintenant les principes de quelques unes des méthodes de détection de thème.

- **L'approche statistique** exploite la probabilité d'apparition des mots dans les thèmes et dans le document de test. Nous pouvons par exemple citer le *classifieur de Bayes* ou modèle unigramme (McDonough *et al.*, 1994) qui évalue la probabilité de chaque thème sachant le document de test (en exploitant la distribution de probabilité des mots dans les thèmes et le document de test). Un autre exemple de méthode probabiliste est le *classifieur TFIDF* (Salton, 1991) qui évalue la similarité existant entre le vecteur du document de test et l'ensemble des vecteurs de thème.
- **L'approche machine learning** se fonde sur l'apprentissage automatique. Les modèles les plus courants sont les réseaux de neurones (Wiener *et al.*, 1995) et les Machines à Vecteurs Supports (SVMs) (Vapnik, 1995), qui recherchent le séparateur optimal entre les différents thèmes.

L'évaluation des performances d'une méthode de détection de thème s'effectue sur le corpus de test. Nous comparons le thème proposé par la méthode étudiée avec le thème effectif du

Table 1: Performances en détection de thème

	TFIDF	Class. Bayes	Rés. neur.	SVM
Performances (%)	74.3	83.1	76.2	78.3

paragraphe (celui proposé par l'humain). Nos expérimentations on conduit aux performances maximales données dans le tableau 1.

2.3 Combinaison de méthodes de détection de thème

Les méthodes présentées ci-dessus ont été optimisées au maximum, le seul moyen d'améliorer les performances obtenues est alors de combiner les différentes méthodes. Il existe plusieurs méthodes de combinaison : combinaison linéaire, réseaux de neurones et SVM. La combinaison à l'aide des SVM nous a permis d'obtenir les plus hautes performances : celles-ci atteignent 93.1%, le gain en performances est plus que significatif (environ 10% en absolu) (Brun *et al.*, 2003).

Cependant, il reste 7% des paragraphes dont le thème n'est pas correctement détecté. Nous nous sommes alors demandés la raison pour laquelle le thème de ces paragraphes n'était pas correctement reconnu. Pour cela, nous nous sommes penchés sur l'étiquetage de référence des paragraphes de test.

3 Remise en cause de l'étiquetage de référence

Les remarques avancées précédemment nous laissent penser que potentiellement l'étiquette de référence des paragraphes utilisée n'est pas fiable et peut être remise en cause. En effet, la phase d'étiquetage des paragraphes de test a été effectuée par un humain, source potentielle d'erreurs et l'on ne peut donc pas connaître *a priori* le degré de fiabilité des étiquettes affectées aux textes. L'étude que nous présentons dans la suite a pour but de connaître le taux de confiance que nous pouvons accorder à l'étiquetage humain.

3.1 L'expérience menée

Pour évaluer le taux de confiance que nous pouvons accorder à une étiquette fournie par un humain, nous décidons d'étudier l'homogénéité de l'étiquetage fourni par un ensemble d'humains sur des paragraphes. Pour ce faire, il nous faut, pour chaque paragraphe étudié, un ensemble d'étiquettes, fournies par différents étiqueteurs (humains).

L'ensemble de paragraphes utilisé est celui que nous avons déjà traité ici : les 835 paragraphes étiquetés manuellement. Il nous faut alors recueillir plusieurs jeux d'étiquetages pour chacun de ces 835 paragraphes. Ce nombre nous a semblé trop élevé pour trouver des personnes acceptant d'étiqueter cet ensemble important de paragraphes. C'est pour cette raison que nous avons choisi de ne travailler que sur un sous-ensemble de ce dernier. Nous avons ainsi extrait, de

façon aléatoire, environ 15% des 835 paragraphes. Ensuite, pour chacun d'eux, nous avons du récolter un ensemble d'étiquetages pour pouvoir étudier l'homogénéité des étiquettes.

Pour collecter ces étiquetages, nous avons fait appel à un ensemble de personnes bénévoles. Nous leur avons tout d'abord fourni un bref descriptif de l'étude que nous menions, où nous expliquions notre objectif. Puis nous leur avons expliqué ce que nous attendions des étiqueteurs : à chacun des paragraphes fournis ils devaient donner 1 ou 2 étiquettes parmi un ensemble d'étiquettes possibles. La première étiquette est obligatoire, la seconde, quant à elle, est optionnelle. Ces étiquettes correspondent aux thèmes sur lesquels nous travaillons. Un descriptif de ce que représente chacune des étiquettes (thèmes) leur a également été fourni.

Nous avons ensuite donné à chaque étiqueteur, une ou plusieurs séries de 10 paragraphes, auxquels ils devaient affecter au moins une étiquette.

Afin de nous situer dans le cadre d'une seule étiquette par paragraphe, nous ne retenons que les paragraphes pour lesquels une seule étiquette a été donnée. A la fin de l'expérience, nous avons obtenu un ensemble de 12 étiquetages pour chacun des 89 paragraphes étudiés. Nous avons, par la suite, cherché à analyser le degré de cohésion entre les étiqueteurs afin d'en dériver le degré de confiance que l'on peut accorder aux étiquettes des 835 paragraphes.

3.2 La mesure Kappa

La statistique Kappa (Cohen, 1960), récemment exploitée par Carletta (Carletta, 1996) comme une mesure d'accord pour l'analyse de documents/discours, est un test applicable dans le cas où plusieurs sujets doivent assigner une étiquette parmi n à un texte.

Le calcul du coefficient K d'accord entre les étiqueteurs tient compte de la chance *a priori* que les étiqueteurs soient d'accord. K est indépendant du nombre d'étiqueteurs, du nombre d'éléments à classer, ainsi que du nombre d'étiquettes à affecter aux éléments.

Le coefficient K d'accord entre les étiqueteurs est défini comme suit:

$$K = \frac{P(A) - P(E)}{1 - P(E)} \tag{3}$$

où $P(A)$ est la proportion d'étiqueteurs qui sont d'accord (proposant la même étiquette) et $P(E)$ est la proportion *a priori* que les étiqueteurs soient d'accord (proportion dûe à la chance uniquement). Lorsque les étiqueteurs sont complètement d'accord, $K = 1$ et à l'inverse si les étiqueteurs ne sont pas plus d'accord que par chance, alors $K = 0$.

Selon Krippendorff (Krippendorff, 1980), une valeur de $K > 0.8$ est synonyme d'une bonne cohérence entre les annotateurs. Une valeur $0.68 < K < 0.8$ ne permet pas de rendre de décision et une valeur $K < 0.68$ montre une non cohérence entre les étiqueteurs.

3.2.1 Aperçu de l'étiquetage manuel

Le tableau TAB. 2 présente un sous-ensemble des étiquetages fournis sur les 89 paragraphes. La valeur de chaque case ($n_{i,j}$) correspond au nombre d'étiqueteurs ayant donné le thème T_j au paragraphe P_i.

Table 2: Exemple d'étiquetage de paragraphes

Paragraphe	T_1 Culture	T_2 Economie	T_3 Etranger	T_4 Histoire	T_5 Politique	T_6 Sciences	T_7 Sports	S
P_1	0	1	6	0	1	2	2	0.26
P_2	0	0	0	0	0	12	0	1
P_3	0	0	0	1	0	0	11	0.83
P_4	7	3	2	0	0	0	0	0.38
...				...				...
P_{87}	0	0	0	0	0	0	12	1
P_{88}	1	0	0	0	0	11	0	0.83
P_{89}	0	0	0	0	0	12	0	1
$N = 89$	T_1=145	T_2=56	T_3=138	T_4=33	T_5=175	T_6=271	T_7=217	Z=67.0

La dernière colonne contient les valeurs S_i, qui correspondent à l'accord entre les étiqueteurs pour le paragraphe i. S_i a une valeur de 1 lorsque tous les étiqueteurs sont d'accord et une valeur de 0 lorsqu'ils ne sont pas du tout d'accord. S_i est calculé de la manière suivante:

$$S_i = \frac{1}{C(C-1)} * \sum_{j=1}^{m} n_{ij}(n_{ij} - 1) \tag{4}$$

Avec

m le nombre de thèmes, dans notre cas $m = 7$
C le nombre d'étiqueteurs, ici $C = 12$ étiqueteurs

La valeur de K (formule 3) nécessite la connaissance de $P(A)$ et $P(E)$. $P(A)$ représente le taux d'accord entre les étiqueteurs et est calculé de la manière suivante :

$$P(A) = \frac{Z}{N} \tag{5}$$

Avec

- N le nombre de paragraphes traités, ici $N = 89$ paragraphes

- $Z = \sum_{i=1}^{N} S_i$ est la valeur correspondant à l'accord entre les étiqueteurs, tous paragraphes confondus

$P(E)$, la proportion *a priori* d'étiqueteurs d'accord sur la même étiquette, est calculée ainsi :

$$P(E) = \frac{1}{NC^2} \sum_{i=1}^{n} T_i^2 \tag{6}$$

Avec

- n, le nombre de thèmes
- NC, le nombre total d'étiquetages (= $N * C$ = 89*12)
- $T_1, T_2, \ldots, T_7$ les valeurs correspondant au nombre d'étiquettes données pour chaque thème T_j.

3.2.2 Résultats

La valeur K obtenue sur les 89 paragraphes de test est de 0.52, cette valeur nous indique que les étiqueteurs ne sont pas d'accord sur les étiquettes des paragraphes. Vu les différences au niveau de l'étiquetage entre les étiqueteurs, nous pouvons en déduire que les étiquettes que nous avons fournies à la main aux 835 paragraphes comportent probablement des erreurs.

Par conséquent, les étiquettes de référence que nous utilisons ne sont pas forcément les bonnes étiquettes. Le taux d'étiquetage correct de 93.3% s'explique donc et dans une certaine mesure représente un excellent résultat étant donnée la qualité de l'étiquetage de référence.

3.3 L'erreur de Bayes

3.3.1 Théorie

La statistique Kappa que nous venons d'étudier, ne nous permet pas de quantifier le taux de désaccord entre les étiqueteurs. L'erreur de Bayes, que nous présentons maintenant va nous permettre d'estimer le pourcentage d'erreur d'étiquetage sur les 89 paragraphes.

Nous faisons l'hypothèse que les données (d'étiquetage) résultent d'une expérience aléatoire lancée de façon itérative. Soit $\Omega = \{\omega\}$ l'ensemble fondamental de l'expérience (l'ensemble des valeurs que peut prendre le résultat de l'expérience) et $Z = (X, Y)$ la variable aléatoire telle que $X(\Omega)$ est l'ensemble des textes et $Y(\Omega)$ l'ensemble des catégories (thèmes). Chaque couple texte/thème correspond alors à un couple (x, y) dans $X(\Omega) \times Y(\Omega)$, associé à ω (*i.e.* $x = X(\omega)$ et $y = Y(\omega)$). Ces notations étant posées, on peut prendre en considération les restrictions précitées concernant l'universalité de l'étiquetage en faisant l'hypothèse qu'il n'existe pas de dépendance fonctionnelle entre x et y mais seulement une loi de probabilité jointe sur (X, Y). Dans ce cadre, il est bien connu que le classifieur ayant le plus petit taux d'erreur est celui qui implémente la règle de décision de Bayes ((Fukunaga, 1990)). Ce taux d'erreur est donné par la formule suivante :

$$R_{Bayes} = 1 - \int_{X(\Omega)} P(y_0 \mid x) dP(x) \tag{7}$$

où $y_0 = arg\ max_y P(y \mid x)$. Elle constitue une borne inférieure sur le taux d'erreur que l'on peut espérer atteindre. L'implémentation de la règle de décision de Bayes revient à choisir, pour chaque texte, la catégorie correspondant au plus grand nombre de votes. L'erreur peut ainsi être calculée de la façon suivante :

$$\widehat{R}_{Bayes} = 1 - \frac{1}{m} \sum_{i=1}^{m} max_y f(y \mid x_i) \tag{8}$$

où $f(y \mid x_i)$ est le nombre d'étiqueteurs affectant le thème y au texte x_i et m est le nombre de paragraphes. Prenons par exemple un paragraphe ayant reçu l'étiquetage suivant :
Culture : 2, Economie : 0, Etranger : 0, Histoire : 1, Politique : 1, Sciences : 8, Sports : 0. La valeur de $max_y f(y \mid x_i)$ sera alors égale à $8/12$.

3.3.2 Résultats

Sur l'ensemble des 89 paragraphes étudiés, l'erreur de Bayes est de 15.7%. Cela signifie que l'on peut supposer que quasiment 16% des paragraphes peuvent avoir une étiquette fausse. Cependant, cette valeur doit être nuancée. En effet, bien qu'un descriptif ait été fourni aux étiqueteurs avant les étiquetages, certains détails peuvent ne pas leur avoir été précisés, et un biais peut avoir été introduit dans leur étiquetage.

Ces résultats nous mènent de nouveau à penser que le fait que les performances de détection de thème ne dépassent pas 93.3% est probablement dû à la limite des performances humaines elles-mêmes.

Les résultats précédents nous indiquent donc que nous ne pouvons pas évaluer de façon fiable les performances des méthodes de détection de thème étudiées. En effet, les étiquettes de référence pouvant être remises en doute, nous n'avons aucun référentiel fiable.

Par conséquent, il nous faut trouver une alternative à l'évaluation des méthodes de détection de thème. Nous pouvons rappeler ici que nous effectuons la tâche de détection de thème dans le but d'effectuer une adaptation des modèles de langage dans les systèmes de reconnaissance de la parole. Ainsi, l'efficacité de nos méthodes de détection de thème devra être évaluée un pas plus en avant dans le processus, c'est-à-dire au niveau de l'adaptation des modèles de langage.

Plus concrètement, si nos expériences montrent que l'adaptation des modèles de langage au thème détecté n'engendre pas une amélioration des performances de reconnaissance de la parole (ou encore de perplexité), cela signifiera les méthodes de détection de thème ne sont pas efficaces. A l'opposé, si les performances s'améliorent, cela signifiera que nos méthodes sont performantes. Nous chercherons dans ce cas à maximiser le gain (en perplexité ou en reconnaissance) obtenu. Lewis, dans (Lewis, 1991), a également énoncé le problème de la non fiabilité de l'étiquetage de référence humain. Pour y faire face, il propose également d'évaluer ces performances durant une étape plus en avant dans le processus.

3.3.3 Etude plus approfondie des résultats

Nous allons maintenant étudier les performances, sur l'ensemble des 89 paragraphes, des différentes méthodes présentées dans ce document. Plus précisément, nous allons comparer les performances obtenues par nos méthodes en fonction de celles des humains (*i.e.* du degré d'accord entre les étiqueteurs). Les méthodes étudiées sont celles qui ont permis d'obtenir les meilleures performances d'étiquetage.

Le tableau TAB. 3 présente les résultats de cette étude. Nous divisons l'ensemble des 89 paragraphes en sous-ensembles, en fonction du degré d'accord des humains sur les paragraphes.

La première colonne montre parmi l'ensemble des 89 paragraphes, le nombre de paragraphes de chaque sous-ensemble étudié. Les autres colonnes montrent soit le degré d'accord, soit les performances des méthodes sur chaque ensemble.

Table 3: Comparaison du taux d'homogénéité des étiquettes humaines et des performances des méthodes de détection de thème étudiées

Nb paragraphes	Taux d'accord humain	Perfs Unigramme (Classif. Bayes)	Perfs RN	Perfs TFIDF	Perfs SVM
39	12/12	37/39	38/39	38/39	38/39
15	11/12	14/15	14/15	12/15	14/15
8	10/12	6/8	7/8	6/8	7/8
10	9/12	9/10	8/10	9/10	8/10
3	8/12	2/3	2/3	2/3	2/3
2	7/12	2/2	2/2	1/2	1/2
6	6/12	4/6	4/6	3/6	4/6
6	5/12	2/6	1/6	2/6	2/6

Il est évident que le nombre de paragraphes par thème est trop petit pour parler de pourcentage, mais nous pouvons tout de même donner quelques impressions générales sur la tendance des performances. Nous pouvons tout d'abord remarquer que sur les ensembles de paragraphes pour lesquels les humains ont tendance à être d'accord (11/12 et 12/12), les méthodes de détection de thème sont très performantes. A l'opposé, sur les paragraphes où les humains ne sont pas d'accord (doute sur le thème, taux d'accord $< 8/12$), les méthodes ont tendance à être moins performantes.

La figure FIG. 1 montre l'évolution de l'erreur en détection de thème en fonction de l'erreur de Bayes moyenne du corpus de test. A nouveau, nous pouvons constater que moins les étiqueteurs sont d'accord sur l'étiquette à assigner aux paragraphes, plus l'erreur de détection de thème augmente.

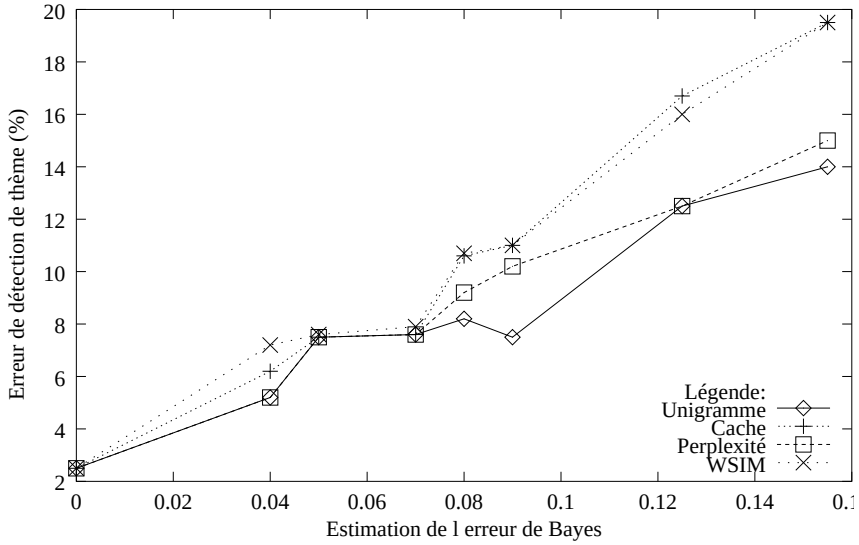

Figure 1: Evolution du taux d'erreur de détection de thème en fonction de l'erreur de Bayes du corpus de test

4 Conclusion et perspectives

Dans cet article, nous nous sommes intéressés à la tâche de détection de thème. Les performances, sur notre corpus, de plusieurs méthodes ont été présentées. Après avoir combiné les différentes méthodes de détection de thème dans un but d'améliorer les performances, nous nous sommes rendus compte que le thème de certains paragraphes n'était pas correctement reconnu. Nous avons alors étudié les raisons de ce mauvais étiquetage, et plus particulièrement nous nous sommes penchés sur le thème de référence accordé aux paragraphes de test.

L'étude que nous avons ainsi menée s'intéresse à l'homogénéité des étiquettes thématiques accordées par plusieurs humains sur des paragraphes de test. Cette étude montre que les différents étiqueteurs (humains) sont régulièrement d'accord sur l'étiquette à assigner aux paragraphes. Cependant sur certains, leur avis divergent. Nous avons ainsi étudié dans quelle proportion les avis des étiqueteurs divergaient en exploitant notamment la statistique Kappa et l'erreur de Bayes. Nous en avons ainsi dérivé le taux d'erreur sur notre corpus de test (15.7%) et nous avons pu conclure qu'une partie des erreurs d'étiquetage était probablement due aux étiquettes de référence du corpus de texte.

Références

BRUN A., SMAÏLI K. & HATON J. (2003). Nouvelle approche de la sélection de vocabulaire pour la détection de thème. In *Traitement Automatique des Langues Naturelles (TALN2003)*, p. 45–54, Nantes, France.

CALLIOPE (1989). *La parole et son traitement automatique*. Masson.

CARLETTA J. (1996). Assessing agreement on classification tasks: the kappa statistics. *Computational linguistics*, **22**(2), 249–254.

COHEN J. (1960). A coefficient of agreement for nominal scales. *PSychological easurements*.

FUKUNAGA K. (1990). *Introduction to Statistical Pattern Recognition*. Academic Press, 2nd edition.

HATON J., PIERREL J., PERENNOU G., CAELEN J. & GAUVAIN J. (1991). *Reconnaissance automatique de la parole*. DUNOD Informatique.

KRIPPENDORFF K. (1980). *Content Analysis: An introduction to its methodology*. Sage Publications.

KUHN R. & DE MORI R. (1990). A cache-based natural language model for speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **12**(6), 570–582.

LEWIS D. (1991). Evaluating Text Categorization. In *Speech and Natural Language Workshop*, p. 312–318, Asilomar.

MCDONOUGH J., NG K., JEANRENAUD P., GISH H. & ROHLICEK J. (1994). Approaches to Topic Identification On The Switchboard Corpus. In *IEEE Transactions on Acoustics, Speech, and Signal Processing*, p. 385–388.

SALTON G. (1991). Developments in Automatic Text Retrieval. *Science*, **253**, 974–979.

SEYMORE K. & ROSENFELD R. (1997). Using Story Topics for Language Model Adaptation. In *Proceeding of the European Conference on Speech Communication and Technology*.

VAPNIK V. (1995). *The Nature of Statistical Learning Theory*. Spinger, New York.

WIENER E., PEDERSEN J. & WEIGEND A. (1995). A neural network approach to topic spotting. In *Fourth Annual Symposium on Document Analysis and Information Retrieval, SDAIR-95*, p. 317–332, University of Nevada, Las Vegas.