Quel indice pour mesurer l'efficacité en segmentation de textes?

Yves Bestgen

CECL / PSOR – Université catholique de Louvain
Place du Cardinal Mercier, 10 — B-1348 Louvain-la-Neuve — Belgique
yves.bestgen@psp.ucl.ac.be

Résumé L'évaluation de l'efficacité d'algorithmes de segmentation thématique est généralement effectuée en quantifiant le degré d'accord entre une segmentation hypothétique et une segmentation de référence. Les indices classiques de précision et de rappel étant peu adaptés à ce domaine, WindowDiff (Pevzner, Hearst, 2002) s'est imposé comme l'indice de référence. Une analyse de cet indice montre toutefois qu'il présente plusieurs limitations. L'objectif de ce rapport est d'évaluer un indice proposé par Bookstein, Kulyukin et Raita (2002), la distance de Hamming généralisée, qui est susceptible de remédier à celles-ci. Les analyses montrent que celui-ci conserve tous les avantages de WindowDiff sans les limitations. De plus, contrairement à WindowDiff, il présente une interprétation simple puisqu'il correspond à une vraie distance entre les deux segmentations à comparer.

Abstract The evaluation of thematic segmentation algorithms is generally carried out by quantifying the degree of agreement between a hypothetical segmentation and a gold standard. The traditional indices of precision and recall being little adapted to this field, WindowDiff (Pevzner, Hearst, 2002) has become the standard for this kind of assessment. An analysis of this index shows however that it presents several limitations. The objective of this report is to evaluate an index developed by Bookstein, Kulyukin and Raita (2002), the Generalized Hamming Distance, which is likely to overcome these limitations. The analyzes show that it preserves all the advantages of WindowDiff without its limitations. Moreover, contrary to WindowDiff, it presents a simple interpretation since it corresponds to a true distance between the two segmentations.

Mots-clés : Segmentation thématique, évaluation, distance de Hamming généralisée, WindowDiff

Keywords: Thematic segmentation, evaluation, generalized Hamming distance, WindowDiff

1 Évaluation en segmentation thématique

La segmentation thématique de textes a pour objectif de localiser les changements de thème dans des documents. Ce type d'informations peut permettre l'amélioration de diverses applications en traitement automatique des langues naturelles comme l'extraction d'informations, le résumé automatique ou encore la navigation à l'intérieur de longs textes. Une série de recherches ont par exemple mis en évidence l'intérêt de segmenter des textes en fonction des thèmes qu'ils abordent afin d'améliorer les résultats de procédures d'extraction d'informations (Hearst, 1997 ; Prince, Labadié, 2007). Ces dernières années, de nombreux algorithmes de segmentation thématique, basés principalement sur la cohésion lexicale, ont été proposés (p.ex., Choi, 2000 ; Ferret, 2002 ; Hearst, 1997 ; Ponte, Croft, 1997 ; Utiyama, Isahara, 2001) rendant encore plus important les problèmes que pose leur évaluation.

Si quelques recherches ont évalué les performances d'une procédure de segmentation sur la base des bénéfices qu'elle apporte à l'application pour laquelle elle a été conçue (Bellot, El-Bèze, 2001 ; Prince, Labadié, 2007), la majorité des chercheurs procèdent en comparant la segmentation postulée à une norme censée correspondre à la vraie segmentation du texte[1]. Pour déterminer cette norme, deux approches sont principalement employées. La première consiste à demander à des juges d'effectuer la même tâche que l'algorithme et donc à segmenter des textes de diverses origines (Bestgen, Piérard, 2006 ; Hearst, 1997). La seconde s'appuie sur un matériel artificiel obtenu en concaténant des textes, les changements de thème à identifier correspondant aux frontières entre ceux-ci. Cette seconde approche s'est très largement imposée en raison de l'existence d'un matériel de référence (Choi, 2000), qui permet de comparer les performances de tout nouvel algorithme à celles des algorithmes considérés comme les plus efficaces selon la littérature.

Quelle que soit l'origine de la norme, l'évaluation requiert un indice pour mesurer le degré d'accord entre la segmentation proposée par l'algorithme et la segmentation de référence. Depuis quelques années, le taux d'erreur *WindowDiff* (Pevzner, Hearst, 2002), sur la base d'une analyse critique de l'indice *Pk* (Beeferman et al., 1999), s'est imposé. Cet indice présente toutefois plusieurs faiblesses. Sa présentation et la discussion de ses limitations font l'objet des deux sections suivantes. La quatrième section présente la distance de Hamming généralisée, proposée par Bookstein et al. (2002), qui, comme l'indique la cinquième section, répond à ces limitations tout en conservant les avantages de *WindowDiff* par rapport à *Pk*.

2 Indices pour mesurer l'efficacité d'un algorithme

Dès les premières recherches en segmentation thématique, les indices classiques en extraction d'information que sont le rappel et la précision ont été critiqués parce qu'ils ne font aucune différence entre des erreurs légères, comme le fait de placer une frontière juste à côté de la position attendue, par comparaison aux erreurs plus graves, comme placer une frontière à une grande distance de cette position attendue, manquer une frontière (*faux négatif*) ou en ajouter une (*faux positif*). Pour cette raison, des indices d'efficacité spécifiques à ce champ de recherches ont été proposés. Le premier à avoir fait l'objet d'un consensus est le taux d'erreur

---

[1] Récemment, Lamprier et al. (2007) ont proposé de se passer de toute segmentation de référence en basant l'évaluation sur la stabilité de la segmentation postulée aux permutations des unités internes aux segments.

Pk (Beeferman et al., 1999). Une analyse critique par Pevzner et Hearst (2002) a souligné son intérêt par rapport aux indices classiques de précision et de rappel, mais également plusieurs de ces limitations. Afin d'y remédier, Pevzner et Hearst (2002) ont proposé une version modifiée de *Pk* qu'ils ont appelé *WindowDiff* (*WD*) et qu'ils formulent[2] comme suit

$$WD(ref, hyp) = \frac{1}{N-k}\sum_{1}^{N-k}\left(\left|b(ref_i, ref_{i+k}) - b(hyp_i, hyp_{i+k})\right| > 0\right)$$

où *b(i, j)* représente le nombre de frontières entre les positions *i* et *j*, *N* le nombre de positions, *k* correspond à la moitié de la longueur moyenne d'un segment dans l'annotation de référence[3]. On peut décrire le fonctionnement de *WD* de la manière suivante. Une fenêtre de taille *k* est déplacée tout au long des unités minimales de segmentation d'un texte (habituellement les phrases). Pour chaque position de la fenêtre, on compare le nombre de frontières de segments que celle-ci englobe selon la norme de référence au nombre de frontières détectées par l'algorithme. Celui-ci est pénalisé d'un point chaque fois que ces nombres sont différents. Le dénominateur permet d'obtenir un score *WD* compris entre 0 et 1. S'agissant d'une mesure d'erreur, plus sa valeur est proche de 0, meilleure est la performance.

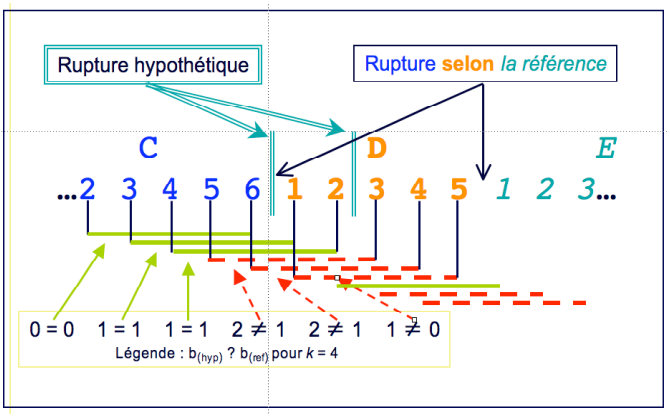

Figure 1 : Exemple de calcul de *WD* (adapté de Pevzner et Hearst, 2002).

La figure 1 illustre le fonctionnement de cet indice. Les unités minimales d'un texte y sont représentées par des chiffres qui traduisent la segmentation selon la norme de référence. La segmentation hypothétique met en évidence deux ruptures, l'une identique à une de celles mises en évidence par la norme de référence et l'autre non. Il est nécessaire de définir ici trois types d'erreurs. La non-identification par la segmentation hypothétique de la frontière entre *D* et *E* est appelée un *faux négatif*. L'ajout erroné par la segmentation hypothétique d'une frontière entre *D2* et *D3* peut être considéré de deux manières différentes. On peut le voir comme un *faux positif*, c'est-à-dire l'ajout dans la segmentation hypothétique d'une rupture qui n'existe pas dans celle de référence. On peut aussi le voir comme une *erreur légère* si on

---

[2] Une autre formule, prenant en compte le nombre de différences entre les deux segmentations et non la dichotomisation de ce nombre, est parfois (mais rarement) employée, même si elle ne correspond pas à la formule originale de Pevzner et Hearst (2002).

[3] Cette valeur a été proposée par Beeferman et al. (1999) et reprise par Pevzner et Hearst (2002) parce qu'elle permet d'attribuer des scores médiocres et relativement similaires aux algorithmes dégénérés les plus courants comme ceux qui ne segmentent jamais ou chaque fois que possible.

considère qu'il s'agit de la rupture présente entre *D* et *E* dans la segmentation de référence, mais que la segmentation hypothétique ne l'a pas placée exactement au même endroit. Cette seconde interprétation impose de revoir celle qui fait de la non-identification de la frontière entre *D* et *E* un cas de faux négatif.

Pour un paramètre *k* égal à 4, les traits horizontaux traduisent le déplacement de la fenêtre mobile. Les trois premiers traits (verts et continus) signalent des fenêtres pour lesquelles les deux segmentations sont d'accord puisqu'elles marquent le même nombre de ruptures (0 ou 1). Les traits discontinus et rouges signalent des points de désaccord : les deux extrémités de la fenêtre mobile ne sont pas séparées par le même nombre de ruptures selon les deux segmentations. Elles pénalisent donc la segmentation hypothétique.

Afin de démontrer les avantages de *WD* par rapport à *Pk*, Pevzner et Hearst (2002) ont mené une série de simulations dans lesquelles la fluctuation de la longueur des segments ainsi que la proportion et la gravité des différents types d'erreurs étaient manipulées. Il en résulte que, contrairement à *Pk*, *WD* pénalise de manière équivalente les faux positifs et les faux négatifs, qu'il pénalise moins les erreurs légères que les faux positifs de même ampleur et que, s'il reste sensible à la fluctuation de la taille des segments, il l'est nettement moins que *Pk*.

Même si de nombreux chercheurs continuent à rapporter *Pk* dans leur analyse afin de faciliter la comparaison avec les études antérieures à 2002, *WD* s'est imposé comme l'indice de référence. Son importance pour l'évaluation en segmentation thématique a encore été récemment renforcée par Artstein et Poesio (2008) qui recommandent son emploi pour mesurer l'accord entre les juges afin de prendre en compte le fait que les juges, comme les procédures automatiques, peuvent détecter les différents thèmes tout en se trompant sur la localisation exacte de leurs frontières (pour une analyse de cette recommendation, voir Bestgen, 2009).

3 Limitations de *WindowDiff*

Malgré les avantages qu'il apporte par rapport à *Pk*, *WindowDiff* a fait l'objet de plusieurs critiques. Lamprier et al. (2007) ont fait remarquer que *WD*, comme *Pk*, pénalise différemment les erreurs selon leur position dans le texte. Les erreurs qui se produisent à moins de *k* positions du début d'un texte ou de sa fin sont pénalisées moins lourdement que celles qui se produisent entre ces deux bornes. Georgescul et al. (2006) ont souligné que l'interprétation des valeurs produites par *WD* n'est pas plus évidente que celle de *Pk* parce que ces indices ne reflètent pas directement l'efficacité de l'algorithme.

WD présente deux autres problèmes, hérités de *Pk*. Tout d'abord, une erreur légère (telle que définie à la section 2) qui se produit à une distance inférieure à *k* de la vraie frontière, mais supérieure à k/2, est plus pénalisée qu'un faux positif pur (éloigné de toute segmentation dans la norme). En effet, celui-ci reçoit une pénalité de *k*, alors qu'une erreur légère reçoit une pénalité supérieure à *k* puisqu'égale au double de sa distance à la vraie frontière. Cette situation est d'autant plus problématique que *WD* a été développé dans le but de réduire les pénalités attribuées à des erreurs légères.

Un second problème, mentionné par Pevzner et Hearst (2002), est que deux ou plus faux positifs qui sont proches les uns des autres (à une distance inférieure à *k*) sont moins pénalisés que s'ils se produisent à des distances supérieures à *k*. Il en est de même de deux ou plus faux négatifs qui sont proches les uns des autres.

On notera enfin, comme le signalent Pevzner et Hearst (2002), que si *WD* est moins sensible à la fluctuation des longueurs de segments que *Pk*, il reste néanmoins affecté par celle-ci.

4 La distance de Hamming généralisée : un indice plus efficace?

L'objectif majeur de cette étude est d'évaluer un indice susceptible de répondre aux limitations présentées par *WD* tout en conservant ses avantages par rapport à *Pk* : la distance de Hamming généralisée proposée, indépendamment du développement de *Pk* et de *WD*, par Bookstein et al. (2002). Ces auteurs se sont intéressés à la mesure de la distance entre des vecteurs binaires de mêmes longueurs. Ce genre de données s'obtient en traitement du signal, mais aussi en segmentation thématique où la présence d'une frontière entre deux unités minimales peut-être codée par un "1" et l'absence de frontière par un "0". Pour ce type de données, une mesure classique est la distance de Hamming (ici appliquée à des données binaires) qui est basée sur le nombre de bits qu'il est nécessaire de modifier pour transformer une séquence en une autre. Considérée comme une forme particulière de distance d'édition, elle correspond au coût minimal des opérations nécessaires pour effectuer cette transformation lorsque les deux seules opérations autorisées sont :

– l'*opération d'insertion* qui change un 0 en 1 pour un coût $Ci = 1$;

– l'*opération de suppression* qui change un 1 en 0 pour un coût $Cs = 1$.

Les termes *insertion* et *suppression* sont, comme le soulignent Bookstein et al. (2002), employés dans un sens différent de celui utilisé habituellement dans les travaux sur la manipulation de chaînes de caractères puisque ni l'insertion, ni la suppression ne modifient la taille de la chaîne.

Dans sa version originale, la distance de Hamming présente la même limitation que les indices de précision et de rappel puisqu'elle se base exclusivement sur l'accord ou le désaccord entre deux bits. Pour dépasser cette limitation, Bookstein et al. (2002) proposent d'adjoindre une troisième opération :

– l'*opération de déplacement* qui fait glisser un "1" vers la gauche ou vers la droite de la séquence afin de le mettre en correspondance avec un "1" dans l'autre séquence. Le coût de cette opération (*Cd*) est une fonction strictement positive et monotoniquement croissante de la longueur du déplacement nécessaire.

La distance de Hamming généralisée (*DHG*) correspond au coût minimum pour transformer une séquence en l'autre au moyen de ces trois opérations. Bookstein et al. (2002) montrent que *DHG* est une vraie distance en ce sens qu'elle en présente toutes les propriétés lorsque les coûts des différentes opérations remplissent certaines conditions. C'est tout particulièrement le cas lorsque $Ci = Cs > 0$ et que le coût total d'un déplacement est proportionnel à la longueur de celui-ci. En divisant le coût minimal par la longueur des séquences, on obtient une mesure relative dont le minimum est 0 alors que le maximum dépend des coûts attribués aux différentes opérations.

La figure 2 présente l'application de *DHG* à l'exemple de segmentation déjà employé à la figure 1 sur la base des coûts suivants : $Ci = Cs = 2$ et $Cd = 1$. *Ci* et *Cs* ont une valeur égale à celle de $k/2$ de sorte que pour qu'un déplacement soit plus avantageux qu'une insertion et une suppression, sa longueur doit être inférieure à k, soit à la moitié de la longueur d'un segment dans la segmentation de référence.

Bookstein et al. (2002) proposent un algorithme qui permet de calculer *DHG* en fonction des pénalités attribuées à chaque opération. Cet algorithme a été implémenté en C++ par Vladimir Kulyukin (www.cs.usu.edu/~vkulyukin/vkweb/software/ghd/ghd.html). Jiang (2009) propose un algorithme plus rapide.

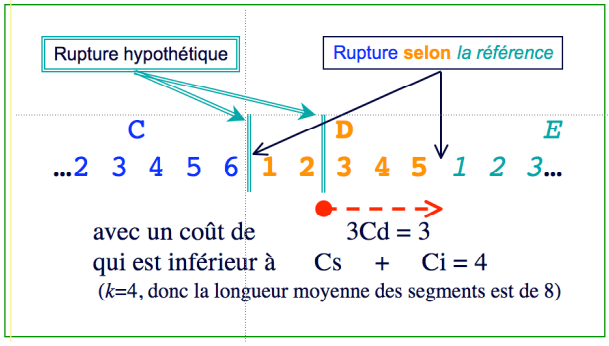

Figure 2 : Exemple de calcul de DHG.

5 Évaluation comparée de *WD* et de *DHG*

Bien que Bookstein et al. (2002) mentionnent la segmentation de texte comme un domaine d'application de leur distance, aucune étude n'a, à ma connaissance[4], comparé *DHG* à *WD*. Une analyse des propriétés de *DHG* montre qu'il est une alternative intéressante à *WD*. On notera en premier lieu qu'il présente les mêmes avantages que *WD* par rapport à *Pk*. Lorsque *Ci* = *Cs*, les faux positifs et les faux négatifs sont pénalisés de la même manière. De plus, toutes les erreurs sont pénalisées, qu'elles soient proches ou éloignées les unes des autres.

DHG répond aussi, de par sa construction, à plusieurs des critiques formulées à l'encontre de *WD*. Tout d'abord, la position d'une erreur dans la séquence, au tout début, au milieu ou à la fin, n'a aucun impact sur la pénalité encourue. Ensuite, *DHG* a une interprétation simple et directe puisqu'il s'agit d'une vraie distance qui correspond au coût minimal des opérations nécessaires pour transformer une segmentation en l'autre. Ces opérations prennent en compte la spécificité de l'évaluation en segmentation thématique en distinguant une erreur grave comme un faux négatif et un faux positif, d'une erreur plus légère comme placer une frontière à une faible distance de la position attendue.

Enfin, *DHG* répond aux deux autres critiques adressées à *WD*. Il est en effet impossible que cet indice pénalise plus une erreur légère qu'un faux positif (ou qu'un faux négatif) puisque dans ce cas, c'est l'opération de suppression (ou d'insertion) qui sera choisie afin d'obtenir le coût minimal. De même, les faux positifs reçoivent toujours la même pénalité quelle que soit la distance entre eux puisque le coût est toujours identique à *Cs*.

Afin de confirmer ces analyses, une évaluation comparative de *WD* et de *DHG*, ainsi que de *Pk*, a été réalisée au moyen des deux simulations les plus importantes décrites dans Pevzner et Hearst (2002). Pour chaque simulation, 10 segmentations de références, chacune composée de 1000 segments d'une longueur moyenne de 25 unités sont générées aléatoirement et pour

[4] Antérieurement au développement de *WD* et de *DHG*, Ponte et Croft (1997) ont employé pour évaluer leur algorithme de segmentation un indice basé sur les opérations d'édition.

chacune de celles-ci 100 segmentations hypothétiques sont générées et comparées à celle de référence au moyen des trois indices *Pk*, *WD* et *DHG*, les résultats finaux correspondants aux moyennes pour les 1000 essais.

5.1 Effet de la fluctuation de la longueur des segments

La première série de simulations a pour objectif d'évaluer l'impact de la fluctuation des longueurs de segments sur les taux d'erreurs. Dans chaque simulation, les segmentations de référence sont générées sur la base de longueurs de segments uniformément distribuées entre deux valeurs à égale distance de la longueur moyenne de 25. Ces valeurs sont [20, 30], [15, 35], [10, 40] et [5, 45]. Les segmentations hypothétiques sont quant à elles générées selon trois types de distribution d'erreurs différents :

– FN: une segmentation avec une probabilité d'occurrence d'un faux négatif de 0.5 à chaque frontière réelle.

– FP1: une segmentation avec une probabilité d'occurrence d'un faux positif de 0.5 dans chaque segment, les faux positifs étant uniformément distribués dans ceux-ci.

– FNP1: une segmentation qui combine les deux précédentes et donc les deux types d'erreurs.

Le schéma de coûts attribués aux différentes opérations pour le calcul de *DHG* correspond à celui proposé dans la section 4, sauf que les trois valeurs ont été multipliées par 2 (*Cd*=2 et *Ci=Cs=k)* de façon à faciliter la comparaison entre *WD* et *DHG* car ceci permet d'égaliser le coût pour *DHG* d'un faux positif pur (éloigné de toute segmentation dans la norme) et d'un faux négatif pur à la pénalité encourue par ces mêmes erreurs selon *WD* puisque celle-ci est de *k*. Le paramètre *k* a été fixé à 12 afin d'obtenir les valeurs les plus similaires possibles à celles rapportées par Pevzner et Hearst (2002).

	FN				FP1				FNP1			
	20-30	15-35	10-40	5-45	20-30	15-35	10-40	5-45	20-30	15-35	10-40	5-45
Pk	.240	.240	.237	.218	.128	.122	.112	.106	.314	.305	.288	.266
WD	.240	.240	.239	.233	.236	.235	.235	.232	.370	.364	.353	.339
DHG	.240	.240	.240	.240	.240	.240	.240	.240	.378	.373	.367	.356

Tableau 1 : Valeurs moyennes de *Pk*, *WD* et *DHG* pour la première série de simulations

Comme l'indique le tableau 1, *WD* est moins fortement affecté par la fluctuation des longueurs des segments que *Pk*, un résultat attendu. Plus intéressant est le fait que *DHG* est encore moins affecté par cette fluctuation. Pour quantifier objectivement cette différence, des analyses de variance à un facteur (*fluctuation*) ont été effectuées. Ces analyses permettent de déterminer la part de variance expliquée (aussi appelée R-carré) qu'apporte la connaissance des niveaux de fluctuation pour prédire les trois indices (Howell, 2008, pp. 336-338). Cette part de variance correspond au rapport entre la variabilité des valeurs moyennes d'un indice en fonction du facteur *fluctuation* (variabilité inter-niveaux) et la variabilité totale de l'indice, qui inclut la variabilité inter-niveaux et la variabilité à l'intérieur de chaque niveau de fluctuation (variabilité intra-niveau). Plus ce rapport est proche de 1, plus l'indice en question est sensible

à la fluctuation des longueurs des segments. Comme le montre le tableau 2, ces parts de variance sont à chaque fois les plus faibles pour *DHG*. Lorsqu'on compare, dans le tableau 1, les résultats pour les faux négatifs et les faux positifs, on observe que *DHG* et *WD* se comportent d'une manière similaire, mais différente de *Pk* qui sous-pénalise nettement les faux positifs comme le prédit l'analyse de Pevzner et Hearst (2002).

Pk			WD			DHG		
FN	FP1	FNP1	FN	FP1	FNP1	FN	FP1	FNP1
.58	.76	.84	.13	.03	.69	.00	.00	.48

Tableau 2 : Parts de variance expliquées par le facteur *fluctuation* pour les trois indices.

5.2 Effet du type de distribution des erreurs

La seconde série de simulations vise à évaluer l'impact de différentes distributions d'erreurs sur les indices. Pour celles-ci, Pevzner et Hearst (2002) ont choisi de n'employer qu'un seul niveau de fluctuation de la longueur moyenne des segments, celui allant de 15 à 35. Sept distributions d'erreurs ont été évaluées, dont trois sont communes avec les premières simulations (FN, FP1 et FNP1). Les quatre distributions supplémentaires sont :

– FP2: une segmentation avec une probabilité d'occurrence d'un faux positif de 0.5 dans chaque segment, les faux positifs étant distribués autour des frontières des segments selon une distribution normale avec un écart-type égal à ¼ de la longueur du segment.

– FP3: une segmentation avec des faux positifs distribués uniformément dans l'ensemble de la séquence, la probabilité d'occurrence d'un faux positif en chaque position possible étant de 0.02, ce qui correspond à une probabilité d'occurrence dans chaque segment de 0.5.

– FNP2: combine FN et FP2.

– FNP3: combine FN et FP3.

	FN	FP1	FP2	FP3	FNP1	FNP2	FNP3
Pk	.240	.122	.096	.116	.305	.268	.306
WD	.240	.235	.232	.215	.364	.340	.361
DHG	.240	.240	.240	.240	.373	.350	.385

Tableau 3 : Valeurs moyennes de *Pk*, *WD* et *DHG* pour la deuxième série de simulations

Le tableau 3 montre, comme attendu, que *Pk* se comporte très différemment de *DHG* alors que celui-ci donne lieu à des résultats très semblables à ceux de *WD*. La principale différence entre ceux-ci porte sur les valeurs obtenues pour FP3. *WD* considère que cette distribution d'erreurs est meilleure que FP2 et FP1 alors que *DHG* considère que les trois distributions FP*x* sont équivalentes. Comme l'indiquent Pevzner et Hearst (2002, p. 33), *WD* sous-pénalise FP3 parce que c'est dans ce genre de distributions que des faux positifs ont le plus de chance de se produire les uns près des autres et donc d'être sous-pénalisés. Pour *DHG*, ces trois

distributions d'erreurs donnent lieu à un même nombre d'opérations de suppression et donc à un même coût total.

Les distributions FNPx méritent aussi une attention toute particulière. FNP3 génère, par construction, les plus mauvaises segmentations. Cette distribution d'erreurs inclut, comme les deux autres, 50% de faux négatifs et un pourcentage équivalent de faux positifs, mais la distribution de ceux-ci n'est pas construite pour favoriser les erreurs légères puisque les faux positifs sont uniformément distribués. De ce point de vue, FNP2 est la moins mauvaise des distributions, puisque c'est celle qui maximise les chances que les faux positifs soient les plus proches des faux négatifs, et FNP1 est intermédiaire. Si *WD* identifie correctement l'avantage de FNP2 par rapport aux deux autres, il considère que FNP3 est meilleur que FNP1 (parce qu'il sous-pénalise FNP3 pour la même raison qu'il sous-pénalise FP3). *DHG* ne commet pas cette erreur et met nettement plus en évidence les différences.

6 Conclusion

L'objectif de cette recherche était d'évaluer les indices qui permettent de mesurer l'efficacité d'algorithmes de segmentation thématique. Tout particulièrement, la distance de Hamming généralisée (*DHG*), proposée par Bookstein et al. (2002), est décrite et comparée à l'indice le plus fréquemment employé dans ce genre de recherche, *WindowDiff* (*WD*) de Pevzner et Hearst (2002). L'analyse des propriétés de *DHG* montre qu'il conserve tous les avantages que *WD* présente par rapport à son prédécesseur, *Pk*, sans les limitations que ces deux-ci partagent. Il faut toutefois noter que, dans les simulations réalisées, les différences entre *DHG* et *WD* sont relativement faibles. Ce résultat n'est pas étonnant parce que les simulations utilisées ont été proposées par Pevzner et Hearst (2002) pour confronter *WD* à *Pk* et n'ont donc pas été conçues en fonction des différences entre *DHG* et *WD*. Des expériences complémentaires sont nécessaires. Toutefois, même si on devait juger que les améliorations apportées par *DHG* sont insuffisantes pour justifier un changement d'indice de référence, un apport de cette étude est de proposer pour *WindowDiff* une interprétation plus simple que celle donnée par Pevzner et Hearst (2002). En effet, ce raisonnement conduit à considérer que *WindowDiff* approxime une vraie distance entre les deux segmentations à comparer, distance qui correspond au coût minimal des opérations nécessaires pour transformer la segmentation proposée par l'algorithme en la segmentation de référence. Si, par contre, on considère que la distance de Hamming généralisée vaut la peine d'être adoptée, il sera nécessaire de mener des analyses approfondies de l'impact des coûts attribués aux opérations sur les valeurs obtenues, et ce, entre autres, par des algorithmes "dégénérés". Comparer dans ce type de situations *WD*, *DHG*, mais aussi l'indice de stabilité proposé par Lamprier et al. (2007), serait très informatif.

Remerciements

Yves Bestgen est chercheur qualifié du F.R.S-FNRS. Il tient à remercier les experts pour leurs commentaires.

Références

ARTSTEIN R., POESIO M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics* 34, 555–596.

BEEFERMAN D., BERGER A., LAFFERTY J. (1999). Statistical models for text segmentation, *Machine Learning* 34, 177–210.

BELLOT P., EL-BEZE M. (2001). Classification et segmentation de textes par arbres de décision. Application à la recherche documentaire. *Technique et science informatiques* 20, 107–134.

BESTGEN Y. (2009). Jugements humains et évaluation des algorithmes de segmentation thématique : application de WindowDiff. Actes de *EvalECD'09*, 15-24.

BESTGEN Y., PIERARD S. (2006). Comment évaluer les algorithmes de segmentation automatique? Essai de construction d'un matériel de référence. Actes de *TALN'06*, 407-414.

BOOKSTEIN A., KULYUKIN V.A., RAITA T. (2002). Generalized Hamming distance. *Information Retrieval* 5, 353–375.

CHOI F. (2000). Advances in domain independent linear text segmentation, Proceedings o*f NAACL-00*, 26–33.

FERRET O. (2002). Using collocations for topic segmentation and link detection. Proceedings of *COLING 2002*, 260-266.

GEORGESCUL M., CLARK A., ARMSTRONG S. (2006). An analysis of quantitative aspects in the evaluation of thematic segmentation algorithms. Proceedings of *SIGdial'06*, 144–151.

HEARST M. (1997). TextTiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics* 23, 33–64.

HOWELL D.C. (2008) *Méthodes statistiques en sciences humaines*, Bruxelles, De Bœck.

JIANG M. (2009). A linear-time algorithm for Hamming distance with shifts. *Theory of Computing Systems*, *44*, 349-355.

LAMPRIER S., AMGHAR T., LEVRAT B., SAUBION F. (2007). On evaluation methodologies for text segmentation algorithms. Proceedings of *ICTAI 2007*, 19-26.

PASSONNEAU R., LITMAN D. (1997). Discourse segmentation by human and automated means. *Computational Linguistics* 23, 103-139.

PEVZNER L., HEARST M. (2002). A critique and improvement of an evaluation metric for text segmentation, *Computational Linguistics* 28, 19-36.

PONTE J., CROFT W. (1997). Text segmentation by topic. Proceedings of *1st ECDL*, 120-129.

PRINCE V., LABADIE A. (2007). Text segmentation based on document understanding for information retrieval. Proceedings of *NLDB 2007*, 295–304.

UTIYAMA M., ISAHARA H. (2001). A Statistical model for domain-independent text segmentation. Proceedings of *ACL'2001*, 491–498.