Classification automatique de textes basée sur une ontologie normée

Application du Extensible Business Reporting Language (XBRL) au Reuters Corpus Volume 1 (RCV1)

Stéphane Gagnon, Sadia Messaoudi, Alain Charbonneau

Université du Québec en Outaouais
283, boulevard Alexandre-Taché
C.P. 1250, succursale Hull
Gatineau (Québec) Canada J8X 3X7
stephane.gagnon@uqo.ca

RÉSUMÉ. *Nous démontrons que l'utilisation d'une ontologie normée selon le domaine d'application permet d'améliorer significativement la Classification automatique de textes (CAT). Nous utilisons le Extensible Business Reporting Language (XBRL) pour définir une ontologie normée et comparons la performance d'un engin de CAT (IBM Classification Module v.8.6) face à 2 autres listes de concepts, soient simple et hiérarchique. Notre échantillon de nouvelles financières est tiré du Reuters Corpus Volume 1 (RCV1), où 2 experts en finance nous aident à coder 1 000 des 45 000 nouvelles portant sur les fusions et acquisitions. Nous rapportons le rappel, la précision, la mesure F, et en plus une mesure dite hiérarchique ajustée pour la pertinence de classification au niveau des classes parents, ainsi qu'une mesure plus détaillée évaluant l'amélioration de la classification au niveau de chaque texte.*

ABSTRACT. *We demonstrate that applying a domain-specific ontology standard significantly improves Automated Text Classification (ATC). We use the Extensible Business Reporting Language (XBRL) to define a standard ontology and compare the performance of an ACT engine (IBM Classification Module v.8.6) against 2 other list of concepts, namely simple and hierarchical. Our sample of financial news is extracted from the Reuters Corpus Volume 1 (RCV1), where 2 experts in finance help us code 1000 of the 45000 news dealing with mergers and acquisitions. We report recall, precision, the F measure, and in addition a hierarchical measure adjusted for classification relevance in parent classes, as well as a more detailed measure evaluating the classification improvements at the level of each text.*

MOTS-CLÉS : *Classification automatique de textes, Nouvelles financières, Reuters Corpus Volume 1 (RCV1), Ontologie, Extensible Business Reporting Language (XBRL)*

KEYWORDS: *Automated Text Classification, Financial News, Reuters Corpus Volume 1 (RCV1), Ontology, Extensible Business Reporting Language (XBRL)*

1. Introduction

La Classification automatique de textes (CAT) est souvent requise en gestion des documents numériques, en particulier la classification hiérarchique selon une taxonomie et/ou ontologie du domaine d'application. Nous proposons d'améliorer la performance de ce type de classification via l'utilisation d'une ontologie normée.

Nous utilisons le Extensible Business Reporting Language (XBRL) comme ontologie normée et comparons la performance d'un engin de CAT (IBM Classification Module v.8.6) face à 2 autres listes de concepts, soient simple et hiérarchique. On l'utilise comme ontologie dans le sens que les interrelations entre les concepts ne sont pas uniques et linéaires comme dans le cas d'une taxonomie. Notre échantillon de nouvelles financières est tiré du Reuters Corpus Volume 1 (RCV1), où 2 experts en finance nous aident à coder 1 000 des 45 000 nouvelles portant sur les fusions et acquisitions. Nous rapportons le rappel, la précision, la mesure F, et en plus une mesure dite hiérarchique ajustée pour la pertinence de classification au niveau des classes parents, ainsi qu'une mesure plus détaillée évaluant l'amélioration de la classification au niveau de chaque texte.

2. Fondements

La classification de textes selon une hiérarchie de classes ou taxonomie n'a été formalisée que très récemment (Koller and Sahami 1997). Les mesures appliquées aux classifieurs plats telles la précision et le rappel, ne conviennent pas à une classification hiérarchique car elles ne prennent pas en considération les types d'erreurs liées à la mauvaise classification (Kiritchenko, Matwin et al. 2006).

Dans les tâches de classification hiérarchique, il est important de considérer la pertinence d'un document non seulement par rapport à sa classe mais aussi par rapport à la classe parent (Sokolova and Lapalme 2009). Ceci est surtout dû au fait qu'une classe parente représente des sujets plus généraux que ceux des classes enfants (Yi 2006).

Pour surmonter ces défis, nous utilisons la mesure hF, basés sur les ancêtres pour évaluer la classification (Kiritchenko, Matwin et al. 2004). Formellement, en considérant une classification hiérarchique multi-étiquettes, on peut définir la mesure d'évaluation hF de la façon suivante (Kiritchenko, Matwin et al. 2006).

Pour toute instance (d_i, C_i) classifiée sous le sous-ensemble $C_i^{'}$ avec $C_i^{'} \subseteq$ C, $d_i \in D$, $C_i \subseteq C$, on aura Les micro-moyennes hP (Précision) et hR (Rappel) telles que :

$$hP = \frac{\sum_i \left| \text{Ancêtre}(C_i) \cap \textit{Ancêtre}(C_i^{'}) \right|}{\sum_i \left| \text{Ancêtre}(C_i^{'}) \right|} \qquad [1]$$

$$hR = \frac{\sum_i \left|\text{Ancêtre}(C_i) \cap \textit{Ancêtre}(C_i')\right|}{\sum_i \left|\text{Ancêtre}(C_i)\right|}$$

La combinaison des deux valeurs hP et hR permet de calculer la F-Score (hF) :

$$hF_\beta = \frac{(\beta^2 + 1)hP.hR}{\beta^2 hP + hR}, \quad [2]$$

$\beta \in [0,+\infty]$. Afin de donner le même poids à la précision et au rappel, on utilise $\beta = 1$.

3. Méthodologie

Notre étude se concentre sur un problème particulier, soit d'évaluer si l'utilisation d'une ontologie normée aidera à améliorer la classification hiérarchique de textes. Nous voulons comparer la performance de ce type de hiérarchie par rapport à la classification non-normée, telle qu'une simple liste de sujets ou une liste de sujets avec hiérarchie limitée. Nous utiliserons un classifieur commercial, IBM Classification Module (ICM) v.8.6, sans regard aux algorithmes utilisés. Nous utilisons une méthode à 4 étapes répétée pour 3 listes différentes :

1. Développement d'une liste de sujets (mots clés) et des ontologies.
2. Échantillonnage des nouvelles.
3. Évaluation du classifieur sur les échantillons.
4. Comparaison et interprétation des résultats des divers tests.

Les sujets pour les 3 types de classifications ont été sélectionnés sur un sous-sujet d'un corpus de nouvelles financières, soit sur les fusions et acquisitions (Haleblian, Devers et al. 2009), pour produire 3 listes développées suite à notre recherche :

1. Simple : sujets choisis parmi la littérature académique.
2. Hiérarchique : sujets choisis selon la littérature, regroupés par facteur principal.
3. Normée : sujets tirés d'une norme comptable internationale.

Pour construire la liste normée, nous avons utilisé le schéma du Extensible Business Reporting Language (XBRL) v.2.1, selon le International Financial Reporting Standards (IFRS) (IASB 2009). Nous utilisons en particulier 2 normes :

1. International Accounting Standard 1 (IAS 1) pour la présentation des états financiers :
   1.1. [310005] Income statement, by function of expense - Separate financial statements.
   1.2. [220005] Statement of financial position, order of liquidity - Separate financial statements.

2. IFRS 3 pour les Notes aux états financiers pour les combinaisons d'entreprises :
   2.1. [817000] Notes - Business combinations.

La base de données utilisée pour nos tests est le Reuters Corpus Volume 1 (RCV1) (Lewis, Yang et al. 2004a). Nous utilisons seulement les 42 890 nouvelles liées au code C181, Mergers and Acquisitions, appartenant au code C18, Ownership Changes. Le nombre de nouvelles a été réduit pour nettoyer la base des nouvelles incomplètes et produire le RCV1v2 (Lewis, Yang et al. 2004b).

Grâce à une petite application en Visual Basic, un certain nombre de nouvelles aléatoires est extrait en vue de les traiter dans les prochains processus de classification supervisée. Il s'agit d'une interface permettant aux experts d'étiqueter les nouvelles selon les concepts de l'ontologie offerte par la norme XBRL. Une fois l'échantillonnage des 1000 nouvelles finalisé, on procède à la classification manuelle des données ciblées avec l'aide de deux experts du domaine. Les deux experts en finance, recruté parmi les étudiants du MBA de l'Université du Québec, travaillent indépendamment l'un de l'autre et classifient chacun une copie de l'échantillon sur la base des sujets hiérarchiques normés choisis. Une application développée sous Access est utilisée afin de faciliter l'analyse de chacune des nouvelles, leur classification, et leur récupération en vue du prochain processus.

ICM a été entrainé sur la base de 120 nouvelles extraites semi-aléatoirement de l'échantillon de 1000 nouvelles codées. Toutefois, le choix de ces 120 nouvelles se base sur une liste de classes dominantes. La liste des classes dominantes a été choisie sur la base de la comparaison de la classification des 2 experts en utilisant le tableau de contingence. Les classes au TP élevé étaient alors candidates à la sélection. Cette méthode de travail contient un certain nombre d'anomalies qu'on a tenté de corriger par une réduction des nouvelles et classes.

Afin de faire une analyse riche de la dispersion et du poids de chaque sujet ou classe et de chaque nouvelle, on s'est appuyé sur l'expertise récupérée de la classification manuelle. Les observations suivantes ont été faites :

1. Des classes sont dominantes telles que ***Acquisition***, ***Sales*** et ***Merger***.
2. D'autres classes sont insignifiantes pour les 2 experts, telles que ***Gross***, ***Depreciation***, ***Other***, ***Impairment***, et ***Inventory***.
3. En éliminant les nouvelles contradictoires par rapport aux experts (classifications totalement différentes), on obtient un nombre de nouvelles utilisable pour l'étude équivalent à 779 nouvelles (car il y a 221 nouvelles contradictoires) parmi lesquelles se trouvent les 81 nouvelles compatibles.

Dans le but d'assurer des mesures non-biaisées par la présence d'un trop grand nombre de classes non-utilisées, on a fait une nouvelle sélection de nouvelles classées de façon compatibles. Notre approche vise ainsi à exploiter les classes les plus pertinentes pour réduire la propagation des erreurs dues aux mauvais

classements à des niveaux inférieurs de la hiérarchie (Bennett and Nguyen 2009). On a alors choisi 402 nouvelles basées sur l'utilisation du tiers des classes dominantes présentées au Tableau 2, pour des échantillons d'entraînement et de classification égaux de 201 nouvelles. Il montre qu'en comparant l'expert1 à l'expert2, on remarque que certaines classes ont été privilégiées par les 2. Ainsi, si on estime que le nombre 17 est satisfaisant et que le nombre de classes intéressantes est 14.

Tableau 2. *Identification des classes dominantes parmi les 3 listes*

Classe	Fréquence du choix des 2 experts	Classe	Fréquence du choix des 2 experts
Acquisition	447	Investment	40
Merger	198	Debt	38
Sales	126	Costs	30
Cash	94	Property	26
Price	62	Value	24
Earnings	52	Taxes	17
Administrative	49	Control	17

Une fois la classification des 402 nouvelles finalisée sur la base d'une liste normée de 14 classes, on a analysé dans le détail le choix des classes par ICM et chaque expert, et on a conclu que l'erreur se trouvait dans le fait que 8 des classes feuilles de la liste normée touchaient une partie des nouvelles et non toutes les nouvelles car n'apparaissant pas dans les listes simple et hiérarchique. En fait on avait comparé en usant d'une probabilité différente qui ne fournissait donc pas le bon résultat.

La probabilité concernant le fait qu'une nouvelle quelconque soit affectée à l'une des classes dominantes est de 1/6 (il y a 6 classes feuilles) dans la liste simple et dans la liste hiérarchique. La même nouvelle a une probabilité moins importante face à une liste normée dont les feuilles ne correspondent pas totalement à celles des listes simple et hiérarchique (probabilité de 1/14).

Afin de corriger l'anomalie des résultats non totalement probants, une nouvelle liste d'entrainement contenant 203 nouvelles basée sur 6 classes dominantes a été choisie. Un nouvel échantillon de classification a également été sélectionné pour 462 nouvelles. La liste normée a été réduite sur les classes apparaissant dans les listes simple et hiérarchique : ***Merger, Acquisition, Price, Control, Debt, Value***.

4. Résultats

Les résultats de la classification automatique par rapport aux 2 experts sont rapportés selon les mesures classiques de la précision, du rappel, et de la mesure F. On rapporte également la mesure hF de Kiritchenko et al. On exécute ICM sur

l'échantillon de 462 nouvelles sur 6 classes dominantes, où chaque nouvelle reçoit un nombre variable de classes pertinentes.

On remarque au Tableau 3 que la liste normée améliore significativement toutes les mesures classiques. Ce résultat n'est cependant pas fiable car il faut également évaluer la performance de chaque liste selon les relations parent-enfant des classes.

Le Tableau 4 montre les résultats qui permettent de mieux compléter les mesures classiques. En plus de la micro et la macro F-Mesures, la mesure hF de Kiritchenko et al. est présentée. Les 2 mesures F enregistrent des résultats plus probants par rapport à la mesure hF. Cela n'est pas basé sur le fait que les résultats de la mesure hF aient baissé mais plutôt que la micro et macro mesure aient augmenté. Cette augmentation nous met dans l'obligation de trouver des explications dans une autre forme d'analyse qui va se concentrer sur le raisonnement du classifieur plutôt que sur des calculs qui pourraient mettre de côté la valeur d'une classification améliorée et/ou enrichie pour une liste normée par rapport aux listes simple et hiérarchique.

Tableau 3. *Résultats du ICM sur les mesures de base en comparaison aux 2 experts*

Expert 1	Liste Simple	Liste Hiérarchique	Liste Normée
Précision	0,5870	0,7156	0,8172
Rappel	0,8104	0,8414	0,8407
F-Mesure	0,6808	0,7734	0,8288

Expert 2	Liste Simple	Liste Hiérarchique	Liste Normée
Précision	0,6684	0,7350	0,7700
Rappel	0,7473	0,8260	0,8728
F-Mesure	0,7057	0,7779	0,8182

Tableau 4. *Résultats du ICM sur les mesures F et hF en comparaison aux 2 experts*

Expert 1	Liste Simple	Liste Hiérarchique	Liste Normée
Macro-F-Mesure	0,5165	0,5056	0,4950
Micro-F-Mesure	0,6809	0,7734	0,8288
Kiritchenko-hF-Mesure	0,6809	0,8397	0,7828

Expert 2	Liste Simple	Liste Hiérarchique	Liste Normée
Macro-F-Mesure	0,4417	0,5159	0,5664
Micro-F-Mesure	0,7057	0,7779	0,8182
Kiritchenko-hF-Mesure	0,7057	0,8593	0,8521

Dans le but de mieux comprendre l'origine des résultats sur la mesure hF, et possiblement de bien démontrer si la liste normée donne une performance supérieure, nous proposons une méthode d'analyse des améliorations de la

classification entre les 3 listes. On compare, pour chaque nouvelle, le nombre de classes choisies par le classifieur par rapport aux choix de l'expert :

1. **Amélioration :** Le classifieur a choisi les mêmes classes choisies par l'expert en plus d'autres classes que l'expert n'a pas trouvé.
2. **Stabilité :** Le classifieur a choisi les mêmes classes choisies par l'expert.
3. **Diminution :** Une classe en moins a été trouvée par le classifieur correspondant au choix de l'expert, ou aucune des classes choisies par l'expert n'a été trouvée par le classifieur.

On compare ensuite le nombre de classes qui changent de qualité entre les 3 listes. Le Tableau 5 rapporte le nombre de classes correspondant aux 3 qualités du classement. Le nombre de classes rapportées nous permet alors de démontrer si la liste normée donne des résultats améliorés, stables, ou diminués, par rapport aux 2 autres listes.

On constate tout d'abord que, pour les 2 experts, la liste normée permet une plus grande stabilité du nombre de classes bien classifiées. De plus, dans le cas de l'expert 2, la liste normée permet une moins lourde diminution de la performance de la classification. Elle ne permet pas cependant d'améliorations des classifications.

En résumé, lorsqu'on compare les résultats des Tableaux 3, 4 et 5, on peut conclure qu'une liste normée améliore les mesures classiques, n'a pas d'effet particulier face aux mesures hiérarchiques, et permet une classification plus fiable par rapport aux autres listes non-normées.

Tableau 5. *Comparaison de la qualité du classement entre les 3 listes*

Expert 1	Simple > Hiérarchique	Simple > Normée	Hiérarchique > Normée
Amélioration	235	230	222
Stabilité	221	219	231
Diminution	6	13	9

Expert 2	Simple > Hiérarchique	Simple > Normée	Hiérarchique > Normée
Amélioration	267	299	202
Stabilité	186	159	257
Diminution	9	4	3

5. Conclusion

Nous avons démontré que l'utilisation d'une ontologie normée permet d'améliorer significativement la performance d'un engin de CAT. Nous avons élaboré une méthodologie utilisant un classifieur commercial, et avons classifié, selon 3 listes de sujets ou classes, un échantillon de 1000 nouvelles du RCV1 codé

par 2 experts en finance. Nous avons enfin évalué la performance selon des mesures classiques, une mesure hiérarchique, et nouvelle méthode pour évaluer l'amélioration de la classification.

Au plan théorique, notre étude a permis de déterminer la valeur relative des ontologies normées pour alimenter d'autres pistes de recherches prioritaires. Elle pourrait être utile aux chercheurs désireux de réduire la complexité de la base de connaissance utilisée. Au plan des applications, nos résultats devraient servir à améliorer la performance liée au secteur des finances. Nous envisageons également des systèmes d'aide à la décision plus complexes, tels qu'un système de surveillance des marchés financiers permettant d'interpréter divers évènements affectant les sociétés cotées en bourse, dans le but de lier ces évènements à des prévisions des cours boursiers.

6. Références

Bennett, P. N. and N. Nguyen (2009). "Refined experts: Improving classification in large taxonomies", 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Boston, MA.

Haleblian, J., C. E. Devers, et al. (2009). "Taking stock of what we know about mergers and acquisitions: A review and research agenda", Journal of Management 35(3): 469-502.

IASB. (2009). "International Financial Reporting Standards - The IFRS XBRL Taxonomy Illustrated".

Kiritchenko, S., S. Matwin, et al. (2004). "Hierarchical text categorization as a tool of associating genes with gene ontology codes", The 2nd European Workshop on Data Mining & Text Mining for Bioinformatics: 26–30.

Kiritchenko, S., S. Matwin, et al. (2006). "Learning and Evaluation in the Presence of Class Hierarchies: Application to Text Categorization", Lecture Notes in Computer Science - LNCS - Advances in Artificial Intelligence. Berlin, Springer. 4013: 395-406.

Koller, D. and M. Sahami (1997). "Hierarchically classifying documents using very few words", Stanford InfoLab.

Lewis, D. D., Y. Yang, et al. (2004a). "RCV1: A new benchmark collection for text categorization research", Journal of Machine Learning Research 5(December): 361 - 397.

Lewis, D. D., Y. Yang, et al. (2004b). "RCV1-v2/LYRL2004: The LYRL2004 Distribution of the RCV1-v2 Text Categorization Test Collection", 12-Apr-2004 Version.

Sokolova, M. and G. Lapalme (2009). "A systematic analysis of performance measures for classification tasks", Information Processing and Management 45(4): 427-437.

Yi, K. (2006). "Les défis de la catégorisation automatique utilisant les systèmes de classification de bibliothèque", World Library and Information Congress (WIIC) 72nd IFLA General Conference and Council.