Apprentissage Numérique pour le Résumé de Texte

Massih-Réza Amini, Patrick Gallinari

Laboratoire d'Informatique de Paris VI
8, rue du capitaine scott, 75015 Paris
{amini, gallinari}@poleia.lip6.fr
http://www-connex.lip6.fr

Abstract. Nous présentons une méthode originale d'apprentissage numérique basée sur de l' apprentissage *semi-supervisé* pour réaliser des résumés de textes par extraction de phrases pertinentes. Ce système permet d'entraîner des classifieurs en se basant sur une petite quantité de données étiquetées simultanément à une grande quantité de données non-étiquetées. Des méthodes d'apprentissage proposées jusqu'à présent, pour le résumé de textes, s'appuient sur des techniques d'apprentissage supervisé, ce qui pour apprendre nécessite l'étiquetage manuel de toutes les phrases ou les paragraphes d'une collection de documents. Ce procédé devient vite obsolète dans le cas où on disposerait d'une grande collection de documents ou bien lorsqu'on s'intéresse aux résumés *on-line* de documents issus d'un moteur de recherche. L'algorithme que nous proposons est assez générique dans le sens où il peut être utilisé avec n'importe lequel des classfieurs estimant les probabilités a posteriori de classes.

1 Introduction

Le résumé automatique permet de présenter à l'utilisateur une information de qualité qui caractérise un texte, sous une forme extrêmement condensée et qui peut être facilement appréhendée. Les résumés similaires à ceux réalisées par un humain (résumé manuel) sont néanmoins difficiles à faire sans une compréhension poussée du contenu du texte [19]. Il existe beaucoup trop de variation de styles d'écriture, de constructions syntaxiques, etc. pour pouvoir construire un système de résumé synthétique. Pour contourner cette difficulté, les systèmes de résumés proposent d'extraire des passages du texte et présentent à l'utilisateur un résumé en concaténant ces passages. Il existe deux façons d'envisager le résumé automatique de texte :

- le *résumé générique* qui résumé le contenu par rapport à l'idée principale du texte,
- le *résumé par rapport à une requête* qui résumé le texte par rapport à une requête d'utilisateur.

Le résumé automatique de textes remonte à la fin des années cinquante aux travaux de Luhn [11]. Pour extraire les phrases pertinentes nécessaires à la construction d'un résumé, il considère des caractéristiques comme la moyenne des fréquences de termes, des mots de titres et la position de la phrase. D'autres recherches sur ce sujet utilisant des approches similaires ont continué jusqu'à la fin des années quatre-vingt [6][16][18]. Avec l'avènement de l'Internet et de moteurs de recherche de plus en plus performants, l'importance d'informations condensées du type résumé a considérablement augmenté et la tâche de résumé automatique a suscité de nouvelles vocations. Beaucoup de nouvelles approches ont commencé à être explorées:

1. des approches qui tentent d'utiliser la structure du discours [12],
2. des approches linguistiques [9][13][17],
3. des approches statistiques [4][8][15],
4. une combinaison de ces deux dernières approches [3][7].

Récemment, différents auteurs ont commencé à s'intéresser aux techniques d'apprentissage pour effectuer des résumés automatiques de textes [10] [20]. Ces techniques permettent de s'adapter au corpus traité ou aux demandes particulières de l'utilisateur. Toutefois, toutes ces approches reposent sur de l'apprentissage supervisé, ce qui pour apprendre nécessite l'étiquetage manuel de toutes les entités textuelles des documents. L'étiquetage manuel au niveau entités textuelles est très coûteux en temps et infaisable pour la plupart des applications réelles.

Nous allons présenter une nouvelle approche qui facilite l'apprentissage des systèmes pour cette tâche. Cette approche est basée sur de l'apprentissage *semi-supervisé*, elle permet d'entraîner un système avec seulement une petite quantité de phrases étiquetées avec une grande quantité de phrases non-étiquetées. Ces dernières sont facilement accessibles et elles permettent d'améliorer considérablement les performances d'un système de résumé à base d'apprentissage. Nous introduirons un nouvel algorithme *semi-supervisé*, son originalité est qu'elle adopte une approche discriminante à l'apprentissage semi-supervisé plutôt qu'une approche générative classiquement employée. Cet algorithme est décrit dans le cadre de l'algorithme CEM [5] [14], il est utilisable avec n'importe lequel des classifieurs qui estiment les probabilités a posteriori de classes.

2 Modèle automatique basé sur l'apprentissage *semi-supervisé* pour le résumé

Nous avons exploré des techniques à base d'apprentissage *semi-supervisé* pour automatiser complètement les méthodes de résumé de textes [1]. Ces méthodes peuvent aussi bien s'acquitter des tâches de résumé générique que de celles basées sur les requêtes utilisateurs. Elles permettent de tirer avantage d'une large collection de documents non-étiquetés, i.e. pour lesquels on ne dispose pas de résumé.

Avec les méthodes d'apprentissage, proposées jusqu'ici, on a besoin d'une base d'apprentissage et de leur résumés associés,qui sont utilisés pour étiqueter les phrases des documents comme pertinents ou non-pertinents pour le résumé. Après le cycle d'apprentissage, ces systèmes opèrent en étiquetant les phrases des nouveaux documents par rapport à leur score de pertinence avec la requête. L'étiquetage de grandes bases au niveau des phrases est clairement prohibitif et s'applique mal au cas de requêtes génériques.

Du point de vue apprentissage, la tâche de résumé est typiquement une tâche pour laquelle il existe de grandes quantités de données non-étiquetées (tous les textes disponibles dans les différents corpus) et où les données étiquetées sont très chères. C'est par définition un cadre idéal pour l'apprentissage semi-supervisé.

Ces techniques n'ont pas été employées dans le cadre du résumé. En recherche d'Information, elles ont principalement été utilisées pour faire de la classification de texte. La plupart des travaux dans ce cadre partent d'une approche non-supervisée et proposent :

1. d'adapter l'algorithme EM pour prendre en compte des données étiquetées et non-étiquetées,
2. d'appliquer une estimation de maximum de vraisemblance.

Les modèles utilisés sont la plupart du temps des mélanges de gaussiennes ou dans le cas discret des mélanges de lois multinomiales. Différents auteurs ont observé qu'en pratique ces hypothèses sont trop restrictives et ont proposé des modèles plus complexes pour s'adapter au traitement de données réelles.

L'approche que nous proposons part de techniques d'apprentissage supervisé. Nous partons d'une partition initiale des phrases qui peut être réalisée soit par un algorithme de base soit par un algorithme de discrimination entraîné sur des données étiquetées dans le cas semi-supervisé, soit par une combinaison des deux. Ce premier étiquetage est ensuite amélioré de façon itérative en utilisant un classifieur qui utilise ses propres sorties et la confiance qui leur est associé pour re-étiqueter les phrases. [2]

L'algorithme présente les avantages des méthodes discriminantes, à savoir qu'il est juste nécessaire d'estimer des probabilités a posteriori $p(Classe \mid Forme)$ ce qui est bien moins complexe que modéliser la probabilité $p(Forme)$ comme dans les approches génératives. Les calculs sont également plus simples. Nous ne faisons pas d'hypothèses sur la forme des données, et n'importe quel classifieur qui estime les probabilités a posteriori de classes peut être utilisé.

Nous avons analysé le comportement de cet algorithme dans le cas de classifieurs linéaires et de classifieurs logistiques sur deux bases de données différentes [2]. Pour chacun des classifieurs nous avons montré sous certaines hypothèses la convergence de l'algorithme vers un maximum local et de la vraisemblance classifiante. Ce critère est beaucoup employé pour le *clustering* dans le cas d'apprentissage non-supervisé avec des modèles génératifs [5]. Nous avons montré pour les classifieurs que l'algorithme est une instance de l'algorithme *Classification Expectation Maximization* [5]. Les résultats que nous avons obtenus sur les deux bases de données `Reuters` et `Summac` permettent d'éclairer le comportement de l'algorithme dans le cas de classifieurs simples et de s'assurer de sa convergence.

3 Conclusion

Nous proposons une nouvelle technique d'apprentissage pour le résumé automatique de texte, basé sur l'extraction des phrases pertinentes d'un document. Le système proposé trie les phrases d'un document en regard de leur pertinence par rapport à une requête utilisateur. Il peut aussi donner des résumés génériques de textes. Cette technique a été testée sur deux base de `Reuters` et de `Summac`. Dans les deux cas, elle présente une augmentation des performances par rapport aux systèmes d'apprentissage classiques [10].

Toutefois, nous n'avons pas considéré le cas où les phrases aient un sens similaire (problème de doublons) qui peuvent être extraites par notre système. Ce traitement est compliqué du fait qu'il faut considérer le sens de chaque phrase dans le résumé et il nécessite un traitement plus linguistique qui sort du cadre que nous nous étions fixé.

Il y a très peu d'études sur l'apprentissage semi-supervisé en particulier à partir de modèles discriminants. Notre travail constitue une avancée dans ce domaine, mais beaucoup de questions restent ouvertes, notamment la valeur des données non-étiquetées dans le processus de l'apprentissage.

Finalement, les techniques que nous avons proposées sont générales et en particulier dans le cadre de la Recherche d'Information.

References

1. M.-R. Amini. *Apprentissage Automatique et Recherche d'Information: application à l'Extraction d'Information de surface et au Résumé de Texte*. Thèse de doctorat, Université Pierre et Marie Curie, LIP6, Juillet 2001.
2. M.-R. Amini and P. Gallinari. The use of unlabeled data to improve supervised learning for text summarization. In *Proceedings of the* 25^{th} *Annual International ACM SIGIR*, pages 105–112, 2002.
3. R. Barzilay and M. Elhadad. Using lexical chains for text summarization. In *Proceedings of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization*, pages 10–17, 1997.
4. J. G. Carbonell and J. Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the* 21^{th} *Annual International ACM SIGIR*, pages 335–335, 1998.
5. G. Celeux and G. Govaert. A classification em algorithm for clustering and two stochastic versions. *Computational Statistics and Data Analysis*, 14:315–332, 1992.
6. H.P. Edmundson. New methods in automatic abstracting. *Journal of the ACM*, 16(2):264–285, 1969.
7. J. Goldstein, M. Kantrowitz, V. Mittal, and J. Carbonell. Summarizing text documents: Sentence selection and evaluation metrics. In *Proceedings of the* 22^{th} *Annual International ACM SIGIR*, pages 121–128, 1999.

8. E. Hovy and C.Y. Lin. Automated text summarization in summarist. In *Proceedings of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization*, pages 18–24, 1997.
9. J.L. Klavans, K.R. McKeown, and M.Y. Kan. Ressources for evaluation of summarization techniques. In *Actes of First International Conference on Language Ressources and Evaluation (LREC)*, pages 899–902, 1998.
10. J. Kupiec, J. Pederson, and F. Chen. A trainable document summarizer. In *Proceedings of the* 18^{th} *Annual International ACM SIGIR*, pages 68–73, 1995.
11. P. H. Luhn. Automatic creation of litterature abstracts. *IBM Journal*, pages 159–168, 1958.
12. D. Marcu. From discourse structures to text summaries. In *Proceedings of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization*, pages 82–88, 1997.
13. K. McKeown, J. Robin, and K. Kukich. Designing and evaluating a new revision-based model for summary generation. *Information Proceedings and Management*, 3(5), 1995.
14. G. J. McLachlan. *Discrimiant Analysis and Statistical Pattern Recognition*. Wiley series in probability and mathematical statistics. Addison-WesleyJohn Wiley and Sons, INC.
15. M. Mitra, A. Singhal, and C. Buckley. Automatic text summarization by paragraph exraction. In *Proceedings of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization*, pages 31–36, 1997.
16. C.D. Paice. Constructing literature abstracts by computer: Techniques and prospects. *Information Proceedings and Management*, 26:171–186, 1990.
17. D. Radev and K. McKeown. Generating natural language summaries from multiple online sources. *Computational Linguistics*, 1998.
18. J. G. Rath, A. Resnick, and T. P. Savage. The formation of absracts by the selection of sentences. *American Documentation*, 12(2):139–143, Avril 1961.
19. K. Spark-Jones. Discourse modeling for automatic summarizing. Technical Report 29D, Computer Science Department, University of Cambridge, 1993.
20. S. Teufel and M. Moens. Sentence extraction as a classification task. In *Proceedings of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization*, pages 58–68, 1997.