Identification automatique de zones dans des documents pour la constitution d'un corpus médical en français

Louise Deléger Aurélie Névéol
LIMSI – CNRS UPR 3251, Orsay, France
louise.deleger@limsi.fr, aurelie.neveol@limsi.fr

Résumé. De nombreuses informations cliniques sont contenues dans le texte des dossiers électroniques de patients et ne sont pas directement accessibles à des fins de traitement automatique. Pour pallier cela, nous préparons un large corpus annoté de documents cliniques. Une première étape de ce travail consiste à séparer le contenu médical des documents et les informations administratives contenues dans les en-têtes et pieds de page. Nous présentons un système d'identification automatique de zones dans les documents cliniques qui offre une F-mesure de 0,97, équivalente à l'accord inter-annoteur de 0,98. Notre étude montre que le contenu médical ne représente que 60% du contenu total de notre corpus, ce qui justifie la nécessité d'une segmentation en zones. Le travail d'annotation en cours porte sur les sections médicales identifiées.

Abstract. Much clinical information is contained in the free text of Electronic Health Records (EHRs) and is not available for automatic processing. To advance Natural Language Processing of the French clinical narrative, we are building a richly annotated large-scale corpus of French clinical documents. To access the most medically relevant content of EHRs we develop an automatic system to separate the core medical content from other document sections, such as headers and footers. The performance of automatic content extraction achieves 96.6% F-measure, on par with human inter-annotator agreement of 98%. We find that medically relevant content covers only 60% of clinical documents in our corpus. Future annotation work will focus on these sections.

Mots-clés : Traitement Automatique de la Langue Biomédicale, segmentation de documents, identification de zones.

Keywords: BioNLP, Automatic document segmentation, section identification.

1 Introduction

De nombreuses informations cliniques sont contenues dans le texte des dossiers électroniques de patients et ne sont pas directement accessibles à des fins de traitement automatique. Afin de faciliter le développement d'outils et de méthodes de traitement automatique de la langue clinique, nous préparons un large corpus de documents cliniques annotés en entités biomédicales et en relations. Une première étape de ce travail consiste à séparer le contenu médical des documents et les informations administratives contenues par exemple dans les en-têtes et pieds de page. En effet, de nombreux documents de notre corpus contiennent des passages indiquant les coordonnées du service hospitalier ayant produit le document, ainsi qu'une liste des médecins susceptibles d'intervenir dans le service. Ces informations ne sont pas spécifiques au patient ni aux éléments de la prise en charge décrits dans le document. Dans cet article, nous présentons la typologie en zones de documents que nous avons établie pour notre corpus clinique ainsi qu'une méthode pour automatiquement identifier ces zones et isoler le contenu médical des documents contenus des dossiers électroniques de patients.

La structure discursive d'un document peut être très utile pour améliorer des outils de recherche d'information (Ruch *et al.*, 2007). Ainsi, de nombreux travaux ont cherché à détecter automatiquement la structure d'articles scientifiques. Certaines approches ont catégorisé les phrases de résumés scientifiques en sections telles que "Introduction", "Method", "Result", "Conclusion" (McKnight & Srinivasan, 2003; Yamamoto & Takagi, 2005; Lin *et al.*, 2006; Ruch *et al.*, 2007; Hirohata *et al.*, 2008). D'autres ont travaillé sur l'intégralité des articles en identifiant des zones argumentatives (définies par Teufel (2000)) du type "Background" (contexte), "Own" (description des travaux des auteurs), "Other" (description d'autres travaux), etc. (Teufel & Moens, 2002; Merity *et al.*, 2009). Différents types de classifieurs ont été développés pour ces tâches : Naive Bayes (Teufel & Moens, 2002; Ruch *et al.*, 2007), MaxEnt (Maximum Entropy) (Merity *et al.*, 2009), SVM (Support Vector Machines) (McKnight & Srinivasan, 2003; Yamamoto & Takagi, 2005), HMM (modèles

de Markov cachés) (Lin *et al.*, 2006), CRF (champs conditionnels aléatoires) (Hirohata *et al.*, 2008). Les performances les plus hautes atteignent des valeurs supérieures à 0.90 d'exactitude (accuracy) pour la classification des phrases, en particulier le modèle CRF de Hirohata *et al.* (2008) avec 0.943 d'exactitude.

Pour traiter les dossiers électroniques de patients, des approches ont été développées pour identifier les différentes sections qui structurent les documents cliniques, comme par exemple "Indication", "Antécédents médicaux", "Evolution dans le service", "Traitement de sortie", etc. (Denny *et al.*, 2009; Li *et al.*, 2010; Tepper *et al.*, 2012). Certaines études se sont limitées à la classification des sections, en présupposant la connaissance des frontières de sections. Par exemple, Denny *et al.* (2009); Li *et al.* (2010) utilisent des frontières de section obtenues grâce à des heuristiques établies manuellement, ce qui est peu généralisable à d'autres corpus. Dans un travail plus récent, Tepper *et al.* (2012) effectuent à la fois la détection des frontières de section et la classification du type de section, à l'aide de modèles HMM. Ils obtiennent des F-mesures proches de 0.92 sur leurs différents corpus de tests pour la détection et la classification des débuts de sections.

Enfin, les approches de segmentation en zones ont également été utilisées sur d'autres types de documents, comme par exemple le contenu textuel des e-mails (Lampert *et al.*, 2009).

Dans cette étude, nous proposons une méthode d'identification des zones de haut niveau (de type en-tête, pied de page, contenu principal) des textes cliniques, qui s'appuie sur les travaux de la littérature en détection de sections.

2 Matériel et méthodes

2.1 Documents cliniques

Pour cette étude, nous avons utilisé des documents d'un corpus de 138 000 textes cliniques issus d'un groupe d'institutions hospitalières françaises. Ce corpus contient environ 2 000 dossiers électroniques de patients. Il couvre de nombreuses spécialités médicales et plusieurs types de documents (principalement des courriers, des comptes rendus de séjour, des comptes rendus d'acte et des ordonnances). Nous avons constitué deux corpus de travail, comprenant des documents aléatoirement sélectionnés dans le corpus entier : un corpus d'entraînement et un corpus de test de 100 documents chacun. Par la suite, nous avons constitué un corpus destiné à l'annotation en entités et en relations (500 documents) à partir des documents du service "Hépato-gastro-nutrition", en nous appuyant sur le système développé. Comme nous avons pour objectif d'annoter ce corpus (ultérieurement), nous avons corrigé la segmentation effectué par notre système afin de ne manquer aucune ligne de contenu. Ceci nous a ainsi donné l'occasion d'évaluer notre système sur un corpus supplémentaire.

2.2 Accès au contenu discursif : identification de zones dans les documents

Nous avons établi un découpage en zones des documents hospitaliers et distinguons ainsi quatre types de zone :
– l'*en-tête générique* : de type « papier à lettre », avec les coordonnées du service dont provient le document. On observe le même en-tête pour tous les documents de même provenance.
– l'*en-tête spécifique* : contenant des informations utiles comme la date du jour, le lieu, le nom et la date de naissance du patient, etc. Ce type d'en-tête est spécifique à un document.
– le *contenu* principal du document
– le *pied de page* : signature du médecin, éventuellement accompagnée de formules de politesse s'il s'agit d'un courrier.

Le tableau 1 montre un exemple type de document découpé en zones. Il faut cependant remarquer que toutes les zones ne sont pas obligatoirement présentes dans chaque document ; il existe par exemple des documents sans en-tête générique ou sans pied de page. De même, certaines zones peuvent se retrouver à plusieurs endroits d'un même document. Par exemple, les « PS » apparaissant dans un courrier après la signature du médecin sont considérés comme des zones de contenu.

Afin de développer et d'évaluer notre méthode de découpage automatique en zones, le corpus d'entrainement et le corpus de test a été manuellement annoté par les deux auteurs (LD, AN). L'annotation a été effectuée avec le logiciel Brat Rapid Anotation Tool (BRAT) (Stenetorp *et al.*, 2012), en marquant le début de chacune des différentes zones selon notre typologie. Nous avons choisi ce mode d'annotation car il permet de représenter les zones simplement : une zone s'achève au début de la zone suivante ou à la fin du document. Les accords inter-annotateurs (F-mesure) pour chacun des corpus sont présentés dans le tableau 2. On observe que les accords sont hauts, avec une micro-moyenne de 0,8711 et 0,9625 respectivement pour le corpus d'entrainement et de test.

En-tête générique	Hôpital Deschamps - 15, avenue du général Leclerc 77000 Lyon |Service de Chirurgie Générale |Secrétariat (01.41.54.92.89 || |Pr Pierre LEBLOND |E-mail : Chirurgie.Digestive@deschamps.fr |Dr Marie MICHEL |
En-tête spécifique	COMPTE-RENDU : CONSULTATION du 21/07/2005 De : Marc DURAND Né(e) le : 14/04/1958 Paris le 22/07/2005 Mon cher confrère,
Contenu	Je vois ce jour en consultation, Monsieur Marc DURAND, opéré de condylômes au niveau de la marge anale et dans le canal anal le 24 novembre 2004. Les suites opératoires ont été simples. Actuellement, l'examen anal et péri-anal est normal, il n'y a plus de condylômes, les cicatrices sont propres. Dans ces conditions, il n'y a pas lieu de revoir Monsieur DURAND en consultation.
Pied de page	Bien confraternellement. J. DUPONT – Interne.

TABLE 1 – Exemple de document découpé en zones ; toutes les informations identifiantes ainsi que les dates ont été remplacées par des substituts plausibles

	Entrainement (N=100)	**Test (N=100)**
En-tête générique	0,9036	1,0000
En-tête spécifique	0,7411	0,9565
Contenu	0,9483	0,9384
Pied de page	0,9000	0,9622
Micro-moyenne	**0,8711**	**0,9625**

TABLE 2 – Accords inter-annotateurs pour le découpage en zones

Premier token de la ligne
Premier token de la ligne précédente
Premier token de la ligne suivante
Deuxième token de la ligne
Bigramme du 1er et 2è tokens de la ligne
Le premier token de la ligne est-il tout en majuscules ?
Position relative de la ligne dans le document (en cinquièmes)
Longueur de la ligne en nombre de tokens
Présence de lignes blanches avant la ligne
Présence de chiffres dans la ligne
Présence d'adresses emails dans la ligne

TABLE 3 – Ensemble de traits utilisés dans le modèle statistique. 'token' = unité de segmentation minimale

Une fois les désaccords résolus, le corpus d'entrainement a été utilisé pour mettre en place et améliorer notre système de découpage en zones. Le corpus de test a été utilisé comme jeu de test pour évaluer notre système final. Pendant la phase de développement, nous avons effectué des validations croisées (avec une partition en 10 ensembles) sur le corpus

d'entrainement afin d'optimiser l'ensemble de traits utilisés dans le modèle statistique.

Pour identifier automatiquement les zones des documents, nous avons entraîné un modèle statistique à champs conditionnels aléatoires (CRF (Lafferty *et al.*, 2001)) en utilisant l'outil Wapiti (Lavergne *et al.*, 2010). Nous classifions chaque ligne d'un document comme appartenant à l'un des quatres types de zones, en nous appuyant sur le format BIO (Beginning/Inside/Outside) pour distinguer les débuts de zones. Cette approche est similaire à celles proposées par Tepper *et al.* (2012) pour segmenter des textes cliniques en sections et Hirohata *et al.* (2008) pour identifier les sections d'articles scientifiques. L'ensemble des traits utilisés dans notre modèle CRF est détaillé dans le tableau 3.

Une fois le système développé, nous l'avons ré-entraîné sur l'ensemble des documents annotés à disposition (soit 200 documents). Nous l'avons appliqué sur le corpus entier (138 000 documents) pour sélectionner sur la base du contenu détecté automatiquement le corpus de 500 documents que nous planifions d'annoter en entités et relations. Nous avons corrigé le découpage en zones effectué par notre système, manuellement à l'aide de l'outil BRAT.

2.3 Evaluation

Nous avons évalué les performances de la méthode d'identification des zones en examinant d'abord les performances de la détection des débuts de zone, mesurées en terme de précision, rappel et F-mesure, pour chaque type de zone ainsi que la micro-moyenne. Comme nous cherchons à déterminer le contenu principal pour sélectionner des documents, il est important que les lignes de contenu soient identifiées avec un très bon rappel, et également une bonne précision. Nous avons donc également évalué les résultats pour l'ensemble des lignes, c'est-à-dire l'attribution d'une zone à chaque ligne des documents. L'évaluation de la classification des lignes a été mesurée en terme d'exactitude (accuracy) pour l'ensemble des lignes et en terme de précision, rappel et F-mesure pour chaque type de zone. L'évaluation de notre méthode a été effectuée sur deux corpus : (1) sur le corpus de test (100 documents) avec le système construit sur le corpus d'entraînement (100 documents) et (2) sur le corpus destiné à l'annotation (500 documents) avec le système construit sur la combinaison des corpus d'entraînement et de test (2x100 documents).

3 Résultats

3.1 Performances de l'identification des zones sur le corpus de test

Les performances de la détection des débuts de zone sont présentées dans le tableau 4, individuellement pour chaque type de zone ainsi que leur micro-moyenne. Les résultats sont très bons pour l'identification des débuts d'en-têtes (F-mesures supérieures à 0,94), et un peu plus bas pour les débuts de contenu et de pieds de page (F-mesures proches de 0,81). La micro-moyenne est de 0,88 de F-mesure, ce qui est légèrement supérieur à l'accord inter-annotateur sur l'échantillon 1 (0,87), mais inférieur à celui sur l'échantillon 2 (0,96). Les résultats de l'évaluation de la classification des lignes en zone sont données dans le tableau 5). L'exactitude (accuracy) est de 0,9678, et les performances sont particulièrement hautes pour les lignes de contenu et d'en-têtes.

	Précision	**Rappel**	**F-mesure**
Début d'en-tête générique	1,0000	0,9302	0,9639
Début d'en-tête spécifique	0,9697	0,9231	0,9458
Début de contenu	0,8384	0,7830	0,8098
Début de pied de page	0,8471	0,7742	0,8090
Micro-moyenne	**0,9118**	**0,8509**	**0,8803**

TABLE 4 – Performances de l'identification des débuts de zones (corpus de test)

Une analyse des erreurs montre que les frontières de section les plus difficiles à placer sont celles concernant le début du contenu et le début du pied de page. Dans les courriers (cf. table 1), la formule de politesse qui marque le début de la zone de pied de page apparait parfois directement à la suite du dernier élément de contenu (par exemple, "Je reste à votre disposition pour revoir Mr DURAND en consultation et vous addresse mes salutations confraternelles"). Une autre difficulté concerne les documents contenant plusieurs fois un même type de zone. Il s'agit par exemple de courriers avec des post-scriptum (plusieurs zones de contenu) ou de documents contenant le suivi d'un acte par plusieurs praticiens : les différents courriers échangés sont concaténés dans un seul document, qui comporte alors plusieurs zones de chaque type.

	Précision	Rappel	F-mesure
En-tête générique	1,0000	0,9868	0,9934
En-tête spécifique	0,9779	0,9705	0,9742
Contenu	0,9520	0,9806	0,9661
Pied de page	0,9187	0,7533	0,8278
Exactitude (accuracy)	**0,9678**		

TABLE 5 – Performances de la classification en zone des lignes des documents (corpus de test)

3.2 Performances de l'identification des zones sur le corpus de 500 documents

Les performances de la détection des débuts de zone sur le corpus de 500 documents sont présentées dans le tableau 6. Celles de la classification des lignes en zone sont données dans le tableau 7. On constate que les performances sont bonnes (F-mesure globale de 0,8875 pour la détection des débuts de zones et exactitude de 0,9550 pour la classification des lignes) et très similaires à celles obtenues sur le corpus de test.

	Précision	Rappel	F-mesure
Début d'en-tête générique	0,9843	0,9171	0,8875
Début d'en-tête spécifique	0,9391	0,8834	0,9104
Début de contenu	0,8834	0,7968	0,8379
Début de pied de page	0,8894	0,8446	0,8664
Micro-moyenne	**0,9210**	**0,8563**	**0,8875**

TABLE 6 – Performances de l'identification des débuts de zones (corpus de 500 documents)

	Précision	Rappel	F-mesure
En-tête générique	0,9860	0,9624	0,9741
En-tête spécifique	0,9750	0,9262	0,9500
Contenu	0,9161	0,9892	0,9513
Pied de page	0,9496	0,9009	0,9246
Exactitude (accuracy)	**0,9550**		

TABLE 7 – Performances de la classification en zone des lignes des documents (corpus de 500 documents)

3.3 Effet de l'identification du contenu principal sur la taille du corpus

Nous avons examiné la taille du corpus sélectionné (corpus de 500 documents), en calculant des stastistiques descriptives avant et après identification des zones de contenu : (1) des statistiques sur le corpus entier (avant identification des zones) et (2) des statistiques sur le contenu principal du corpus (après identification des zones). Les résultats sont présentés dans le tableau 8. La différence de taille entre le corpus brut et le corpus restreint aux zones de contenu est nette : un peu plus de 170 000 mots au total contre environ 100 000 mots de contenu. La longueur moyenne d'un document (en mots) est réduite de 41% (343 mots vs. 202 mots). En revanche, la longueur moyenne d'une phrase augmente de 22% par rapport au corpus brut. En effet, les en-têtes et pieds de pages contiennent davantage de phrases courtes par rapport au contenu principal qui comprend de longues phrases descriptives.

	Corpus brut	Zones de contenu	Différence (en %)
Nombre de mots	171 722	100 730	-41%
Nombre de phrases	18 815	9 013	-52%
Longueur moyenne d'un document (en mots)	343	202	-41%
Longueur moyenne d'un document (en phrases)	38	18	-53%
Longueur moyenne d'une phrase (en mots)	9	11	+22%

TABLE 8 – Statistiques descriptives sur le corpus de 500 documents

4 Discussion et conclusion

Les performances de notre méthode d'identification de zones sont très bonnes. En particulier, nous obtenons une précision de 0,9520 et un rappel de 0,9806 sur le corpus de test pour la classification des lignes de contenu. On constate donc que l'on perd très peu de contenu. Les performances semblent suffisamment hautes pour pouvoir baser la sélection d'un échantillon sur le contenu identifié automatiquement. Ceci est confirmé a posteriori par l'évaluation sur le corpus sélectionné de 500 documents (précision de 0,9161 et rappel de 0,9892 pour les lignes de contenu). Le corpus est sensiblement plus petit lorsque l'on se restreint au contenu principal (tableau 8). Ceci montre que les zones d'en-têtes et de pieds de pages sont très présentes dans les documents. Il est donc nécessaire de les identifier afin de travailler sur le contenu médical.

Remerciements

Ce travail a bénéficié d'une aide de l'Agence Nationale de la Recherche portant la référence ANR-13-JS02-0009-01.

Références

DENNY J. C., SPICKARD III A., JOHNSON K. B., PETERSON N. B., PETERSON J. F. & MILLER R. A. (2009). Evaluation of a method to identify and categorize section headers in clinical documents. *Journal of the American Medical Informatics Association*, **16**(6), 806–815.

HIROHATA K., OKAZAKI N., ANANIADOU S. & ISHIZUKA M. (2008). Identifying sections in scientific abstracts using conditional random fields. In *Proceedings of the IJCNLP 2008*.

LAFFERTY J. D., MCCALLUM A. & PEREIRA F. C. N. (2001). Conditional random fields : Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, p. 282–289, San Francisco, CA, USA : Morgan Kaufmann Publishers Inc.

LAMPERT A., DALE R. & PARIS C. (2009). Segmenting email message text into zones. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, EMNLP '09, p. 919–928.

LAVERGNE T., CAPPÉ O. & YVON F. (2010). Practical very large scale CRFs. In *Proceedings the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, p. 504–513 : Association for Computational Linguistics.

LI Y., LIPSKY GORMAN S. & ELHADAD N. (2010). Section classification in clinical notes using supervised hidden markov model. In *Proceedings of the 1st ACM International Health Informatics Symposium*, p. 744–750 : ACM.

LIN J., KARAKOS D., DEMNER-FUSHMAN D. & KHUDANPUR S. (2006). Generative content models for structural analysis of medical abstracts. In *Proceedings of the HLT-NAACL BioNLP Workshop on Linking Natural Language and Biology*, p. 65–72 : Association for Computational Linguistics.

MCKNIGHT L. & SRINIVASAN P. (2003). Categorization of sentence types in medical abstracts. In *AMIA Annual Symposium Proceedings*, volume 2003, p. 440 : American Medical Informatics Association.

MERITY S., MURPHY T. & CURRAN J. R. (2009). Accurate argumentative zoning with maximum entropy models. In *Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries*, p. 19–26 : ACL.

RUCH P., GEISSBÜHLER A., GOBEILL J., LISACEK F., TBAHRITI I., VEUTHEY A. & ARONSON A. R. (2007). Using discourse analysis to improve text categorization in medline. In *Stud Health Technol Inform*, volume 129, p. 710–5.

STENETORP P., PYYSALO S., TOPIĆ G., OHTA T., ANANIADOU S. & TSUJII J. (2012). BRAT : a Web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations Session at EACL 2012*, p. 102–7.

TEPPER M., CAPURRO D., XIA F., VANDERWENDE L. & YETISGEN-YILDIZ M. (2012). Statistical section segmentation in free-text clinical records. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey.

TEUFEL S. (2000). *Argumentative zoning : Information extraction from scientific text*. PhD thesis, Citeseer.

TEUFEL S. & MOENS M. (2002). Summarizing scientific articles : experiments with relevance and rhetorical status. *Computational linguistics*, **28**(4), 409–445.

YAMAMOTO Y. & TAKAGI T. (2005). A sentence classification system for multi biomedical literature summarization. In *Data Engineering Workshops, 2005. 21st International Conference on*, p. 1163–1163 : IEEE.