# Vers une harmonisation automatique de la représentation de comptes rendus medicaux pour évaluer leurs similarités

Yves Parès[1], Xavier Aimé[1], Jean Charlet[1,2], Marie-Christine Jaulent[1]

[1] INSERM, U1142, LIMICS, F-75006, Paris, France;
Sorbonne Universités, UPMC UNIV PARIS 06, UMR_S 1142, LIMICS, F-75006, Paris, France;
Université Paris 13, Sorbonne Paris Cité, LIMICS, UMR_S 1142, F-93430, Villetaneuse, France.
yves.pares@etu.upmc.fr, xavier-aime@inserm.fr,
{jean.charlet,marie-christine.jaulent}@crc.jussieu.fr

[2] Assistance publique - Hôpitaux de Paris, France.

**Résumé** : De nombreux hôpitaux contiennent encore des dépôts de connaissances inexploités. Ceux-ci prennent souvent la forme d'enregistrements textuels faiblement structurés aux contenus hétérogènes, qui contiennent des cas passés. Ces enregistrements sont par exemple les comptes rendus d'examens médicaux. L'accès aux connaissances et à l'expérience qu'ils renferment pourrait aider à traiter les cas présents. Nous présentons ici une méthode pour normaliser la représentation de comptes rendus textuels en fœtopathologie de manière à constituer une base qui sera utilisée pour raisonner à partir de cas. Cette méthode se base sur la transformation des comptes rendus en arbres. Les mesures statistiques concernant le bruit et le silence générés sur 10 de nos cas sont présentées.

**Mots-clés** : Extraction d'information, normalisation de la représentation des connaissances, analyse de comptes rendus médicaux, réutilisation de données cliniques, raisonnement à partir de cas.

## 1 Introduction

L'analyse de données médicales textuelles nécessite souvent plusieurs étapes préliminaires pour rendre les données algorithmiquement gérables. Même en mettant de côté les simples fautes d'orthographe, nous devons souvent faire face dans les textes à des contextes imprécis, à des informations mal placées et à la variabilité d'expression inhérente à chaque professionnel de santé. La tâche principale du projet Accordys (Agrégation de contenus et connaissances pour raisonner à partir de cas de dysmorphologies fœtales) (Charlet, 2013) est d'aborder ces difficultés dans le domaine de la fœtopathologie. Les comptes rendus utilisés dans ce projet sont des textes écrits par différents praticiens et qui se répartissent sur une période de dix ans.

Afin de permettre aux praticiens actuels et futurs de réutiliser cette mémoire pour améliorer les diagnostics, le but d'Accordys est de développer un système d'information qui puisse (1) harmoniser la représentation des cas (homogénéiser leurs *structures*), autorisant ainsi par la suite à les comparer (identifier les similarités dans leur *contenus*) et (2) construire une base de cas grâce aux technologies du Web sémantique et à des ontologies de domaine pour améliorer l'extensibilité et les capacités d'intéropérabilité du système. Le but à long terme est d'automatiser le processus cognitif d'affinage du diagnostic d'un cas présent à partir de la recherche de cas passés décrivant des malformations similaires. Nous présentons dans cet article une méthode adaptable pour l'harmonisation de la représentation de comptes rendus qui aborde le point (1). Cette méthode se base sur la mise en correspondance d'arbres (*tree matching*) et vise à être faite en parallèle avec une chaîne de traitement automatique du langage naturel.

L'approche présentée ici opère de manière lexicale, ce qui la rend utilisable dans n'importe quel domaine n'ayant pas de ressource termino-ontologique à disposition, ce qui est le cas de la fœtopathologie. De plus les travaux effectués par d'autres membres du projet (pas encore publiés) ont montré que la couverture par MeSH de 993 études de cas en anglais de dysmorphies fœtales disponibles sur PubMed n'était pas complète. Nous voulons également être capables de

quantifier l'apport futur de l'ontologie de la fœtopathologie en cours de développement au sein d'Accordys à notre méthode.

## 2 Matériel et méthode

Notre objectif est de construire le contexte hiérarchique des observations trouvées dans un compte rendu de fœtopathologie, en plaçant ces observations dans un arbre que l'on fait correspondre à un modèle (lui aussi un arbre).

Après une description plus précise du matériel utilisé, nous exposons les approches pertinentes déjà existantes et finalement nous donnons une adaptation de l'algorithme général de *flexible tree matching* (Kumar *et al.*, 2011) afin de montrer son intérêt dans notre contexte.

### 2.1 Matériel utilisé

Nous avons pour ce travail utilisé un corpus d'environ 12 000 mots, provenant de 20 comptes rendus. Il s'agit d'un fragment du corpus final de 2000 comptes rendus qui sera utilisable par les partenaires d'Accordys une fois les accords CNIL finalisés et signés. Sur les 20 documents, 10 (aléatoirement tirés) ont servi lors du processus de réalisation du modèle de cas, et les 10 autres lors de l'évaluation de notre méthode de mise en correspondance avec ce modèle.

Les documents sont des textes Unicode avec des éléments de structure hiérarchique guidant la lecture. Dans le contexte d'Accordys, ils sont obtenus par la numérisation et l'anonymisation des comptes rendus papier de l'hôpital Trousseau à Paris. Chaque compte rendu contient au plus 9 sections principales correspondant chacune à un examen (par exemple : examen macroscopique externe du fœtus, autopsie, radiographies, examens macroscopique du placenta). Chaque section peut contenir des sous-sections ou directement de l'information sous forme soit de phrases complètes, soit de listes à puces (énumérations ou paires site anatomique/observation).

Bien qu'étant du même type, les documents n'expriment pas toujours les informations de la même manière : la plupart des comptes rendus contiennent seulement une fraction de ces sections (par exemple l'examen cytogénétique est souvent absent et l'autopsie peut avoir été refusée par les parents), pas toujours sous le même nom et pas exactement dans le même ordre. Cette variabilité sera d'autant plus présente dans le corpus final de 2000 comptes rendus. C'est pourquoi nous avons besoin d'homogénéiser leurs structures.

### 2.2 État de l'art

L'extraction du sens de chaque phrase contenue dans un rapport médical est souvent réalisée grâce à des méthodes de TAL. (Deléger *et al.*, 2014) présente celles qui sont utilisées dans le cadre d'Accordys, et (Uzuner *et al.*, 2011) catalogue les différentes approches utilisées lors de la compétition I2B2 2010, à laquelle les partenaires TAL d'Accordys ont participé. Ces méthodes visent à repérer, normaliser et typer les mots et expressions d'un corpus, ainsi qu'à dégager les relations entre eux. Ceci produit un graphe (appelé graphe de dépendance) pour chaque phrase.

Pour reconstruire le contexte de chaque phrase, des méthodes de structuration existent (Taira *et al.*, 2001) mais elles essaient souvent de faire correspondre les comptes rendus à un modèle sémantique prédéfini et encodé directement dans le logiciel. Le *tree matching* offre plus de généricité, et n'a à notre connaissance pas encore été utilisé dans le cadre de l'analyse de comptes rendus médicaux. C'est un domaine étudié depuis quelque temps (Hoffmann & O'Donnell, 1982) et notamment dans le contexte du développement et de l'exploration automatique du Web (Jindal & Liu, 2010). En fonction des contraintes, plusieurs méthodes existent et peuvent fournir soit un mapping (mise en correspondance) optimal soit une approximation. Le mapping optimal peut être trouvé en temps polynomial si on le contraint à respecter :

— la hiérarchie (si un nœud N1 du premier arbre est mis en correspondance avec un nœud N2 du deuxième arbre, alors les fils de N1 devront correspondre aux fils de N2 et inversement) ;

— l'ordre (si N1 est mis en correspondance avec N2, alors les frères suivant N1 ne pourront être mis en correspondance qu'aux frères suivant N2).

Nous avons vu que ces contraintes étaient inacceptables dans notre contexte : des titres peuvent être manquants et nous ne savons pas dans quel ordre les sections vont apparaître, de même pour les éléments d'observation présents dans chaque section.

Le *flexible tree matching* est une approche qui permet de relâcher ces contraintes. Cet algorithme utilise une méthode stochastique pour échantillonner des approximations du meilleur mapping. Un mapping est un graphe biparti dont chaque arc est valué par un coût. Chaque coût de mise en correspondance de deux nœuds (N1, N2) est la somme de trois termes : (1) la distance entre les labels de N1 et N2, (2) le nombre de fils de N1 ou N2 non mis en correspondance avec des fils de l'autre nœud et (3) le nombre de frères de N1 ou N2 non mis en correspondance avec des frères de l'autre nœud. Les termes (2) et (3) – termes d'ascendance et de fratrie – couvrent le coût structurel du mapping, en évaluant la manière dont il respecte les formes globales des arbres. Un coût est également induit par chaque nœud non mis en correspondance (ceci est symbolisé par une mise en correspondance avec un nœud vide). Chacun de ces termes est multiplié par un poids, permettant à l'utilisateur de régler l'importance de chaque terme en fonction de son domaine d'application. Voir (Kumar *et al.*, 2011) pour l'algorithme détaillé et (Kumar *et al.*, 2013) pour son application à la manipulation de pages Web.

### 2.3 Méthodologie

La première étape de la méthodologie proposée est de construire l'arbre modèle. Il a été élaboré avec la collaboration des fœtopathologistes de l'hôpital Trousseau impliqués dans Accordys. L'arbre résultant a été dérivé manuellement du gabarit originalement établi par les médecins. Des renseignements sur le type ou le contenu possible de chaque nœud présent ont été ajoutés. Ainsi, le nœud `Autopsie` sera annoté avec le type `examen` et le nœud `Capacité fonctionnelle` aura pour fils les nœuds `Normale`, `Paranormale` et `Pathologique`. Chaque nœud a donc un label et éventuellement un type.

L'étape suivante est de transformer les comptes rendus en des arbres intermédiaires que nous appelons arbres cas. Tous les arbres produits sont placés dans une base de données documentaire, dans laquelle ils sont représentés sous une forme très semblable à du JSON [1]. Un ensemble de règles simples basées sur des expressions régulières portant sur la forme de chaque ligne servent à déterminer les types des nœuds qu'elle génère et leur position par rapport au nœud précédemment généré. Par exemple, la ligne contenant juste `DISSECTION`, entièrement en capitales et entourée de lignes vides, génèrera un nœud de type `examen`, et la ligne

```
- Foie : développement normal
```

ressemblant à une paire site/observation, génèrera le nœud `Foie` et son fils `développement normal`, tous deux non typés.

Nous utilisons une représentation en arbre car celle-ci reste simple tout en épousant la structure naturelle des comptes rendus. Elle est flexible car (1) elle permet d'en capturer toutes les informations et (2) elle peut s'adapter à divers degrés de formalisation (le mapping vers l'arbre modèle étant la première étape pour monter en formalisation). Elle permettra de plus d'acceuillir les graphes de dépendance issus de la chaîne de TAL, ceux-ci remplaçant les feuilles encore sous forme de chaînes de caractèrs. On pourra également utiliser les arbres-cas homogénéisés lors de cette chaine de TAL, notamment pour aider à la désambiguisation de certains termes du corpus (e.g savoir que l'adjectif « calcanéen » utilisé seul dans la section des radiographies a de fortes chances de correspondre au « point d'ossification calcanéen »), l'arbre fournissant un contexte. Les partenaires TAL d'Accordys et nous pensons que les deux processus bénéficieront l'un de l'autre.

Une fois ces deux étapes complètes, on essaie de déterminer pour chaque arbre cas combien de ses nœuds peuvent trouver une correspondance dans l'arbre modèle. Nous utilisons

1. JavaScript Object Notation. Voir www.json.org/xml.html

le *flexible tree matching*. La tâche ici est de trouver les paramètres du processus de calcul du mapping qui donnent les meilleurs résultats dans notre contexte de comptes rendus de fœtopathologie. Comme exposé dans l'état de l'art, le *flexible tree matching* vise le mapping de moindre coût. Le coût total d'un mapping est calculé en faisant la somme des coûts de chacun de ses arcs et en la divisant par la somme de la taille des deux arbres. La fonction calculant le coût de mise en correspondance des labels est une distance lexicale : nous utilisons la distance de Levenshtein divisée par la longueur du label le plus court et multipliée par 2 si les nœuds ont des types incompatibles (ex : si deux nœuds ont des labels proches mais que l'un correspond à un titre d'examen et l'autre à une observation, alors leur distance sera artificiellement augmentée). Chaque terme structurel (termes d'ascendance et de fratrie) est multiplié par un poids. Les paramètres à trouver sont ces poids ainsi le coût d'une mise en correspondance avec le nœud vide.

## 3 Résultats et discussion

Les expérimentations montrent que le réglage des paramètres n'est pas facile. Notre arbre modèle contient pas moins de 576 nœuds et étant donné que l'algorithme (1) évalue initialement toutes les mises en correspondance possibles entre tous les nœuds de l'arbre cas et tous les nœuds de l'arbre modèle et (2) ré-évalue itérativement certains des mappings possibles, alors le processus de calcul du mapping complet d'un cas peut prendre jusqu'à une heure sur un processeur quadricore, ce qui ralentit la recherche des paramètres. Le calcul répété d'une distance de Levenshtein comme est le goulot d'étranglement. Ceci ne devrait toutefois pas être bloquant dans un système en routine car les cas de fœtopathologie sont rares, et le mapping ne devra être fait qu'au moment de l'inclusion d'un cas dans la base.

L'ensemble de paramètres qui a retenu notre attention est le suivant :

— poids du terme d'ascendance de 0,21 ;
— poids du terme de fratrie de 0,25 ;
— coût d'une mise en correspondance avec le nœud vide de 1,2.

Ces valeurs sont totalement dépendantes de la fonction choisie pour calculer la distance entre les labels. Les autres paramètres ont été laissés comme suggéré par (Kumar *et al.*, 2011). Sur une base de 10 comptes rendus (sélectionnés aléatoirement dans notre corpus), ces paramètres nous ont donné en moyenne un bruit (pourcentage de nœuds de l'arbre cas mis en correspondance avec un nœud incorrect de l'arbre modèle, créant ainsi de fausses informations) de 23.7% et un silence (pourcentage de nœuds qui ont été mis en correspondance avec le nœud vide alors qu'un nœud pertinent existait) de 8.6%.

Nous avons remarqué un comportement récurrent dans l'algorithme de flexible tree matching : il tend à commencer par mettre en correspondance les feuilles des deux arbres. Étant donné que les feuilles ne peuvent pas induire de coût d'ascendance elles semblent être les mises en correspondance les moins coûteuses au début. Le problème est que dans nos arbres cas les feuilles sont les nœuds les moins pertinents à mettre en correspondance : elles correspondent en général aux observations qui sont spécifiques au cas et qui ne peuvent que rarement ressembler à un nœud dans l'arbre (voir par exemple le nœud (1 11) en figure 1).

Ceci nous mène à penser que le fait de supprimer les feuilles de l'arbre cas ou de détailler le cas des feuilles dans le calcul de la distance entre les labels pourrait être utile, afin d'éviter d'induire en erreur l'algorithme puisque si des feuilles sont incorrectement mises en correspondance alors la mise en correspondance de leurs pères sera perturbée. Cette perturbation se voit en figure 1 au mapping du nœud (1 3 0) : ce nœud est mis en correspondance avec un nœud du modèle qui se trouve avoir le même label, mais qui ne concerne pas la configuration du placenta. Ce mapping non pertinent induira un coût d'ascendance pour son père, ce qui peut expliquer que le nœud (1 3) ne soit pas mis en correspondance par la suite. Ce problème n'est pas à imputer à l'algorithme de *flexible tree matching*, il est une conséquence du domaine d'application et des choix de modélisation lors de la construction de l'arbre modèle.

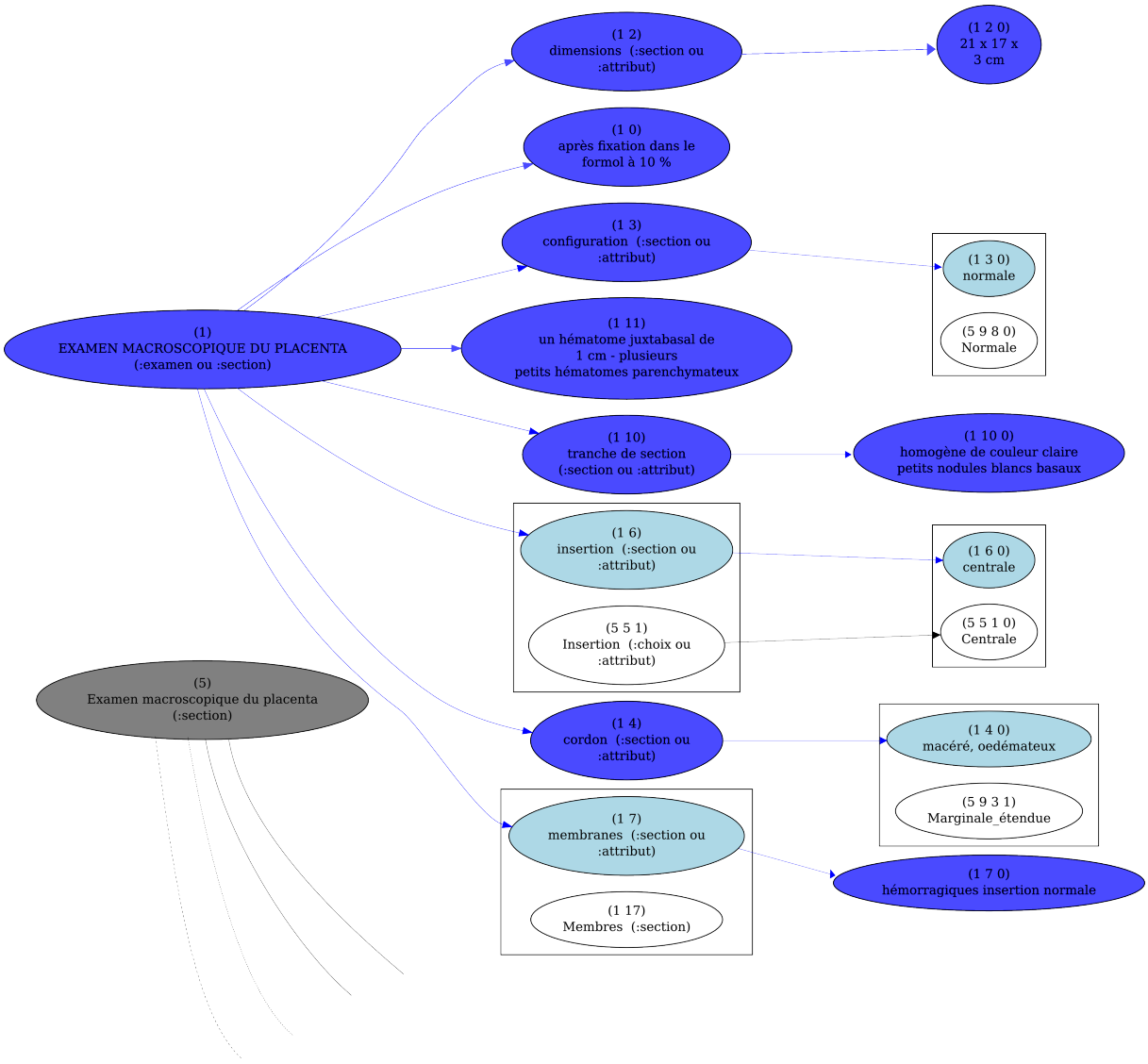


FIGURE 1 – Portion de mapping contenant des nœuds portant sur des observations portant sur le placenta. Les nœuds en bleu proviennent du compte rendu, ceux en blanc/gris du modèle. Les nœuds clairs contenus dans des boîtes sont ceux qui ont étés mis en correspondance.

## 4 Perspectives

Comme nous sommes contraints de gérer du texte libre sans structure, des méthodes statistiques pourraient nous être utiles pour segmenter le texte (Lafferty *et al.*, 2001) et construire l'arbre de façon plus fiable, nos règles simples montrant leur faiblesse pour gérer les cas d'énumérations d'observations se trouvant sur la même ligne ou présentées d'une manière inhabituelle.

Trouver les bons paramètres pour le processus de calcul du mapping pourrait aussi être fait grâce à des techniques d'apprentissage supervisé – ceci est proposé par (Kumar *et al.*, 2011) –, en donnant au système des mappings faits à la main il pourrait apprendre l'ensemble des

paramètres.

Nous sommes en train de réaliser un travail d'annotation de notre corpus à partir des ressources de CISMeF. Ceci permettra d'utiliser une mesure de similarité sémantique (Harispe *et al.*, 2013; Sánchez-Ruiz *et al.*, 2011) pour (1) remplacer la mesure lexicale lors du mapping et (2) mesurer la similarité entre deux cas. En effet, du fait du mapping, deux nœuds comparables se retrouveront au même endroit dans les arbres-cas à comparer, et nous pourrons mesurer la similarité des concepts annotant ces nœuds. Une fois la similarité entre les dysmorphies de deux cas établies, il sera possible d'adapter le diagnostic de l'ancien cas et les décisions prises à l'époque (décision de suggérer des examens génétiques aux parents, par exemple) au nouveau.

La méthode que nous présentons sera également utile pour la construction de l'ontologie du domaine : en calculant systématiquement des mappings vers un modèle commun, on peut (1) identifier des éléments de donnée communs dans les comptes rendus et de nouveaux niveaux de détail à inclure dans l'ontologie ou au moins (2) inventorier de nouveaux labels alternatifs pour les concepts existants.

## Remerciements

Cette recherche a été conduite dans le cadre d'Accordys, projet financé par l'Agence Nationale de la Recherche (ANR-12-CORD-0007).

## Références

CHARLET J. (2013). Agrégation de contenus et de connaissances pour raisonner à partir de cas de dysmorphologie foetale. In *Premier atelier du SIG IMIA francophone*.

DELÉGER L., LIGOZAT A.-L., GROUIN C., ZWEIGENBAUM P. & NÉVÉOL A. (2014). Annotation of specialized corpora using a comprehensive entity and relation scheme. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC), Reykjavik* : ELRA.

HARISPE S., SÁNCHEZ D., RANWEZ S., JANAQI S. & MONTMAIN J. (2013). A framework for unifying ontology-based semantic similarity measures : A study in the biomedical domain. *Journal of biomedical informatics*.

HOFFMANN C. M. & O'DONNELL M. J. (1982). Pattern matching in trees. *Journal of the ACM (JACM)*, **29**(1), 68–95.

JINDAL N. & LIU B. (2010). A generalized tree matching algorithm considering nested lists for web data extraction. In *Proceedings of the SIAM International Conference on Data Mining (SDM)*, p. 930–941 : SIAM.

KUMAR R., SATYANARAYAN A., TORRES C., LIM M., AHMAD S., KLEMMER S. R. & TALTON J. O. (2013). Webzeitgeist : Design mining the web. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, p. 3083–3092 : ACM.

KUMAR R., TALTON J. O., AHMAD S., ROUGHGARDEN T. & KLEMMER S. R. (2011). Flexible tree matching. In *Proceedings of the Twenty-Second international joint conference on Artificial Intelligence*, p. 2674–2679 : AAAI Press.

LAFFERTY J., MCCALLUM A. & PEREIRA F. C. (2001). Conditional random fields : Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the International Conference on Machine Learning*, volume 18, p. 282–289.

SÁNCHEZ-RUIZ A. A., ONTANÓN S., GONZÁLEZ-CALERO P. A. & PLAZA E. (2011). Measuring similarity in description logics using refinement operators. In *Case-Based Reasoning Research and Development*, p. 289–303. Springer.

TAIRA R. K., SODERLAND S. G. & JAKOBOVITS R. M. (2001). Automatic structuring of radiology free-text reports1. *RadioGraphics*, **21**(1), 237–245.

UZUNER Ö., SOUTH B. R., SHEN S. & DUVALL S. L. (2011). 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, **18**(5), 552–556.