Extraction et identification d'entités complexes à partir de textes biomédicaux

Julien Lorec*,**, Gérard Ramstein**, Yannick Jacques*

*INSERM U601, Département de Cancérologie, Équipe 3: cytokines et récepteurs
{julien.lorec,yjacques}@nantes.inserm.fr
**LINA, Équipe C.O.D, École polytechnique de l'université de Nantes
gerard.ramstein@polytech.univ-nantes.fr

Résumé. Nous présentons ici un système d'extraction et d'identification d'entités nommées complexes à l'intention des corpus de spécialité biomédicale. Nous avons développé une méthode qui repose sur une approche mixte à base d'ensemble de règles a priori et de dictionnaires contrôlés. Cet article expose les techniques que nous avons mises en place pour éviter ou minimiser les problèmes de synonymie, de variabilité des termes et pour limiter la présence de noms ambigus. Nous décrivons l'intégration de ces méthodes au sein du processus de reconnaissance des entités nommées. L'intérêt de cet outil réside dans la complexité et l'hétérogénéité des entités extraites. Cette méthode ne se limite pas à la détection des noms des gènes ou des protéines, mais s'adapte à d'autres descripteurs biomédicaux. Nous avons expérimenté cette approche en mesurant les performances obtenues sur le corpus de référence GENIA.

1 Introduction

A cette date, de nombreuses méthodes d'étiquettage d'entités biologiques pour les corpus de spécialité ont été proposées ; quelles soient à base de règles (Fukuda et al. (1998)) ou encore réposant sur des techniques d'apprentissage (Collier et al. (2000)). Néanmoins, la simple détection de la présence d'une entité nommée dans un texte ne suffit pas pour l'identifier et l'associer à une instance d'entité biologique particulière. Le couplage des méthodes d'extraction des entités nommées avec l'utilisation de dictionnaires semble être une solution particulièrement adaptée à ce type de problématique (Koike et al. (2003)). De plus, la majorité de ces techniques d'extraction d'entités nommées ont été développées dans le but de ne détecter que quelques types particuliers et spécifiques d'objets biologiques, notamment les gènes et les protéines, et ne peuvent être facilement adaptés à d'autres contextes.
Il existe trois principales difficultés à prendre en compte lors d'une recherche à base de dictionnaire :

- la présence de termes synonymes et la résolution des différentes abréviations et acronymes,
- la variabilité des mots tant au niveau de l'orthographe que de la morphologie et de la syntaxe mais aussi d'un point de vue lexico-sémantique, de la présence d'insertions/déletions et permutations,

– la présence de noms ambigus que se soit entre des entités de même nature, entre des entités de natures différentes ou des collisions avec le dictionnaire anglais standard.

Ces différents points restent particulièrement difficiles à traiter dans les textes de biologie et de médecine. Les problèmes d'ambiguïté dans les corpus biomédicaux sont résumés dans (Tuason et al. (2004)).

2 Description des dictionnaires

Nous avons construit différents dictionnaires regoupant des descripteurs biologiques de nature diverse et variée : d'une part les *gènes et protéines humaines*, certaines zones particulières de l'ADN (les *sites de liaison aux facteurs de transcription*) humain, les différentes *souches de cellules* humaines, les *tissus* et *organes* humains mais aussi les *protocoles expérimentaux et techniques ainsi que les appareillages utilisés au cours d'expériences biologiques*. Ces dictionnaires proviennent de l'assemblage de plusieurs bases de données publiques : LocusLink, Hugo, GDB, OMIM, TRRDSITE, TRRDFACTORS, TFD, COMPEL, TFFACTOR et de différentes sources du Metathesaurus UMLS (Lindberg et al. (1993)).

Nous stockons dans nos dictionnaires, non pas les noms bruts issus des alias et des acronymes de chaque base de données, mais les formes variantes (orthographiques, morphologiques, lexico-sémantiques, etc) de chaque alias et symbole que nous normalisons.

Nous utilisons aussi les informations issues des nomenclatures spécifiques de chaque base de données afin de générer de nouveaux alias d'une entité. Ces formes inédites peuvent être retrouvées dans les publications scientifiques alors qu'elles sont absentes des bases de données. Par exemple, un effort important a été fourni afin de produire l'ensemble des combinaisons de noms complets/formes acronymiques potentielles d'une même entité ("chemokin like receptor 1", "CMKLR1", "CMKL receptor 1", "CMK light receptor 1", "chemokin like R 1" et "chemokine LR 1") et les multiples insertions/déletions et permutations de groupes de mots descripteurs ("class III alcohol dehydrogenase", "alcohol dehydrogenase class III" et "adenosine monophosphate deaminase 1 isoform M", "adenosine monophosphate deaminase 1").

La construction de tels dictionnaires ne sera pas décrite dans cette article.

Actuellement, les dictionnaires contiennent un total de 205 736 variants dont 49 656 entités distinctes. Les différentes molécules, cellules, organes et sites de liaison sur l'ADN répertoriés proviennent de l'humain ou à défaut de mammifères.

3 Recherche des entités nommées dans les textes

Chaque document est tout d'abord découpé en phrases grâce à des heuristiques puis à chaque mot de chaque phrase est associé son *part of speech* grâce à GENIA POS Tagger (Tsuruoka et al. (2005)). Nous n'utilisons pas de *shallow parser* mais uniquement les informations fournies par les *part of speech*. Ceci à le mérite d'alléger la procédure à condition que l'étiquettage soit correct (Amrani et al. (2005)).

Tous les syntagmes plus ou moins complexes sont extraits de chaque phrase séquentiellement, découpés en plus petites unités, grammaticalement correctes en biologie, et normalisés jusqu'à correspondance exacte avec une entité du dictionnaire. En effet, les entités nommées présentes

dans les publications biomédicales peuvent être relativement complexes et s'étendre sur plus d'un groupe nominal.

Extraction des entités nommées Les groupes de mots pouvant potentiellement représenter un ou plusieurs objets biologiques sont dégagés des textes de la façon suivante :

1. Les groupes nominaux simples correspondant aux blocs de noms propres ou communs avec les éventuels symboles, cardinaux et adjectifs associés sont extraits. Par exemple "Interleukine 2".
2. Sont rattachés aux groupes nominaux simples les verbes au gérondif ou au participe passé en suffixe, ou en préfixe si le mot précédent le verbe n'est pas un modal, un pronom ou un adverbe. Deux groupes nominaux simples sont concaténés si un de ces verbes permet d'en faire la jonction. Par exemple "Interferon regulating factor 8".
3. Deux de ces groupes nominaux étendus peuvent être ensuite réunis si certaines prépositions ou conjonctions telles que "of", "in", "at", "on", "by", "for", "to" ou "with" les séparent. Par exemple "regulator of G-protein signalling 4" ou "cell adhesion molecule regulated by oncogenes".
4. De la même manière, deux des groupes nominaux étendus provenant de l'étape précédente sont reliés entre eux s'ils sont séparés par une conjonction de coordination "and, "but" ou "or". Par exemple "Signal transducer and activator of transcription 3 interacting protein 1".

A cette étape, nous pouvons donc avoir des syntagmes de complexité très différentes à analyser. A priori, chaque syntagme représente une seule et même entité. Nous cherchons donc son occurrence au sein de nos dictionnaires après normalisation (cf paragraphe suivant). Néanmoins si la mise en correspondance exacte n'a pu être réalisée, nous considérons que le syntagme contient alors plus d'une entité, chaque entité pouvant être représenté par une portion indépendante du texte.
Nous devons donc redécouper le bloc de texte contenu dans le syntagme en sous-unités de complexité légèrement moindre. Chaque sous-unité est alors testée individuellement contre nos dictionnaires et si la mise en correspondance s'avère infructueuse, celle ci est décomposée en constituants plus simples, et ainsi de suite, jusqu'à détection positive de la présence d'une entité ou obtenir une portion de texte non résolue et atomique.
Le découpage des syntagmes est réalisé grâce à la règle contextuelle décrite ci-dessous que nous appliquons en fonction des séparateurs suivants, séquentiellement :

1. les conjonctions de coordination,
2. les prépositions (sauf s'ils sont précédés d'un verbe),
3. les gérondifs et participes passés (et l'éventuelle préposition associée).

La précédence du séparateur numéro (2) sur le séparateur numéro (3) a été décidé empiriquement en analysant la composition de nos dictionnaires.

Les différentes combinaisons de blocs de texte de part et d'autre d'un séparateur sont générées. Par exemple, l'expression "suppressor of G2 allele of SKP1 pseudogene" donne les combinaisons "suppressor of G2 allele" et "G2 allele of SKP1 pseudogene". Des nouveaux syntagmes générés, ceux possédant le plus grand nombre de séparateurs ont précédence sur

ceux en contenant moins et sont traîtés en priorité par la suite. De même, en cas de présence de prépositions, la position des entités nommées au sein des expressions est située préférentiellement à droite des séparateurs. Nous traitons donc en priorité les nouveaux syntagmes en fin de texte. Sur l'exemple précédent, l'ordre de priorité est désormais : "G2 allele of SKP1 pseudogene" puis "suppressor of G2 allele". Le syntagme original, non découpé à l'étape en cours, est utilisé par la règle suivante. En revanche, chaque nouveau bloc est de nouveau traité par la règle en cours.

Par exemple, le syntagme "modulator of G-protein signalling 4 down-regulated by oncogenes" contient une entité nommée : "G-protein signalling 4" que l'on souhaite découvrir. Les séparateurs détectés dans le texte sont "of" et "down-regulated by". Nous testons contre nos dictionnaires les blocs de texte suivants dans cet ordre : tout d'abord "modulator of G-protein signalling 4 down-regulated by oncogenes" puis

- d'une part "G-protein signalling 4 down-regulated by oncogenes" puis
  - "G-protein signalling 4" et "oncogenes" indépendamment
- et d'autre part "modulator".

Une limite principale à la stratégie actuellement mise en place consiste en l'impossibilité de retrouver des concepts basés sur des groupes verbaux. Ceci n'influe pas sur la capacité du système à détecter des noms d'objets biologiques mais réduit son abilité à reconnaître des concepts de plus haut niveau.

Identification des entités nommées Chaque bloc de texte que l'on souhaite mettre en correspondance avec les entités présentes dans nos dictionnaires sont tout d'abord normalisées (génération des variants morphologiques, suppression des déterminants, lemmatisation et passage sous la forme de *compound nouns*). Les portions de texte que nous obtenons à l'étape d'extraction sont à base de groupes nominaux plus ou moins complexes, or il est très fréquent de trouver associés aux entités nommées des noms *satellites* qui peuvent soit décrire une action dont l'objet est l'entité biologique (par exemple "assimilation", "transcription", "screening", etc) ou qualifiant l'entité (par exemple "gene", "protein", "experiment", etc). En anglais, de tels termes sont majoritairement ajoutés en suffixe au nom de l'entité. Aussi nous supprimons au fur et à mesure les noms à droite du texte avant de les soumettre à la recherche dans les dictionnaires. Cette méthode simple permet de répondre convenablement à ce type de problème. Il reste néanmoins des cas de figure non négligeables de construction de groupes nominaux où les noms d'action ou descriptifs sont retrouvés devant le nom de l'objet biologique (par exemple, "interleukine protein IL2") et qui sont non résolus automatiquement pour le moment. La présence en préfixe d'adjectifs (par exemple "ubiquitous"), cardinaux ou symboles est également prise en compte. La seule différence avec la technique précédente réside dans le sens de la réduction des termes : ici ce sont les adjectifs, cardinaux ou symboles les plus à gauche du texte qui sont enlevés.

Désambiguation Pour l'instant l'étape de désambiguation des noms est assez rudimentaire :

- Afin d'améliorer la qualité des groupes nominaux à tester, sont détectées les énumérations simples du type "interleukine 1, 2 and 3 receptors" impliquant des numériques ou des symboles/identifieurs afin d'être explicitées sous la forme "interleukine 1 receptor and interleukine 2 receptor and interleukine 3 receptor".

– Beaucoup d'auteurs d'articles en biologie définissent entre parenthèses des abréviations qu'ils utilisent tout au long du document en lieu et place de l'objet biologique tel qu'il est décrit dans nos dictionnaires. Il est très important de pouvoir les détecter et les associer correctement aux entités qu'elles représentent. Les termes entre parenthèses qui précèdent une entité nommée identifiée mais dont la nature est inconnue seront automatiquement associés à cette entité reconnue lors d'une prochaine occurrence dans le texte.

Typage Dans nos dictionnaires, nous pouvons avoir une même entité associée à différents descripteurs biologiques. Seul le contexte dans lequel l'entité a été identifiée peut permettre de clarifier sa nature. Lorsque plusieurs entités de nature différente (par exemple *facteur de transcription* ou *site de liaison à un facteur de transcription*) correspondent à un même bloc de mots, les noms éliminés à l'étape d'Identification des entités nommées puis d'Extraction des entités nommées permettent de mesurer la vraisemblance respective de chaque descripteur "contextuel" associé à l'objet biologique grâce à un lexique de mots contrôlés. Ce lexique contient un ensemble de termes qui sont associés spécifiquement à un ou plusieurs types d'entités dans les textes (par exemple, "neuropeptide" qualifie exclusivement une protéine, "transcription" un gène et "assay" un protocole expérimental). Le type prédominant parmi les mots *satellites* résolus par le lexique doit ainsi permettre d'aider à clarifier le type contextuel de l'entité. Le recours à des techniques plus perfectionnées est alors nécessaire lorsque l'entité n'est associée à aucun terme du lexique (pour le moment, l'entité reste non-résolvable) ou lorsque des contradictions apparaissent (ici l'entité est a priori considérée comme étant un faux positif).

4 Résultats et conclusion

L'outil a été développé en Java et la base de données implémentée avec PostgreSQL.

Nous avons mesuré les performances du système sur le corpus de référence GENIA (Jin-Dong et al. (2003)). La couverture de nos dictionnaires sur ce corpus ne sera pas analysée. Nous avons sélectionné aléatoirement 126 phrases référençant uniquement des objets biologiques potentiellement détectables grâce aux dictionnaires, soit 254 entités nommées annotées, dont 73 entités biologiques distinctes, toute nature confondue. Nous avons correctement identifié un total de 236 entités avec une précision de précision de 92% et un rappel de 86%. Parmi ces 236 entités, 68 sont des entités biologiques distinctes (32/35 *gènes et protéines*, 20/21 *souches de cellules*, 9/10 *protocoles expérimentaux et techniques ainsi que les appareillages utilisés au cours d'expériences biologiques*, 4/4 *sites de liaison aux facteurs de transcription* et 3/3 *tissus*). La principale source de faux positifs provient d'erreurs de typage et la majorité des faux négatifs sont issus de l'insuffisance du nombre de formes variantes dans les dictionnaires.

Nous avons présenté une méthode simple d'extraction et d'identification d'entité nommées biologiques complexes utilisant à la fois des techniques à bases de règles et de dictionnaires contrôlés.

Les approches par dictionnaire ont pour principale limitation de ne pouvoir détecter des entités encore inconnues mais restent efficaces pour permettre de caractériser les relations entre ces différents objets biologiques ou certaines de leurs propriétés en combinaison avec un système d'extraction d'information en aval. Le principal avantage des techniques mises en oeuvre dans cet article est leur relative généricité permettant de traiter des objets biologiques de natures très

différentes sans avoir à utiliser différentes méthodes complexes en parallèle. Plusieurs difficultés restent néanmoins en suspens : la principale, et la plus difficile à résoudre, est l'inévitable cas des termes au sens variable selon le contexte (et notamment comment distinguer la véritable nature d'une entité et savoir si l'on a affaire à une véritable entité biologique ou non. Ce qui se pose particulièrement pour les sites de liaison aux facteurs de transcription d'après nos résultats préliminaires). Une autre difficulté réside dans l'exhaustivité très relative des dictionnaires présentés ici et la nécessité d'avoir des sources contrôlées et vérifiées, ce qui rend les mises à jour encore assez ardues. Beaucoup d'entités présentes dans les dictionnaires sont inappropriées ou issues d'erreurs de saisie lorsque les bases de données respectent mal ou peu les nomenclatures en vigueur.

Références

Amrani, A., M. Roche, Y. Kodratoff, et O. Matte-Tailliez (2005). Inductive improvement of part-of-speech tagging and its effect on a terminology of molecular biology. *Canadian Conference on AI 2005*, 366–376.

Collier, N., C. No, et J. Tsujii (2000). Extracting the names of genes and gene products with a hidden markov model. *Proc. COLING 2000*, 201–207.

Fukuda, K., T. Tsunoda, A. Tamura, et T. Takagi (1998). Toward information extraction : Identifying protein names from biological papers. *Proc. of the Pacific Symposium on Biocomputing '98*.

Jin-Dong, K., T. Ohta, Y. Teteisi, et J. Tsujii (2003). Genia corpus - a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(suppl. 1). i180–i182.

Koike, A., Y. Kobayashi, et T. Takagi (2003). Kinase pathway database : An integrated protein-kinase and nlp-based protein-interaction resource. *Genome Res.*, 13(6A) :1231–43.

Lindberg, D., B. Humphreys, et A. McCray (1993). The unified medical language system. *Methods Inf Med.*, 32(4) :281–91.

Tsuruoka, Y., Y. Tateishi, K. Jin-Dong, T. Ohta, J. McNaught, S. Ananiadou, et J. Tsujii (2005). Developing a robust part-of-speech tagger for biomedical text. *Proceedings of the 10th Panhellenic Conference on Informatics*, (à paraître).

Tuason, O., L. Chen, H. Liu, J. Blake, et C. Friedman (2004). Biological nomenclatures : Source of lexical knowledge and ambiguity. *Proceedings of the Pacific Symposium of Biocomputing*, 9 :238–249.

Summary

We present a tool for the extraction of complex named entities dedicated to biomedical corpora. Our method relies on a hybrid approach based on controlled dictionaries and a set of a priori rules. This paper discusses some techniques to overcome or restrict the problems of synonymies, term variability and ambiguous names. Our method is not limited to gene or protein names but aims at identifying other biomedical descriptors. The GENIA corpus has been used to evaluate the performances of this tool.