Inferring semantic relations between pharmacovigilance terms with the NLP methods

Marie DUPUCH[a], Laëtitia DUPUCH[b], Thierry HAMON[c] and Natalia GRABAR[a]
[a]*CNRS UMR8163, Université Lille 1&3, France;*
[b]*Université Toulouse III Paul Sabatier, France;*
[c]*LIM&BIO (EA3969)Université Paris 13, Bobigny, France*

Abstract. The surveillance of the adverse drug reaction profile and the prevention of the potential or known risk related to the use of medicine products is a core activity of pharmacovigilance. Signal detection and monitoring of identified risks benefit from traditional statistical approaches and also from qualitative information on semantic relations between close adverse drug reaction terms, such as Standardized MedDRA Queries or hierarchical levels of the MedDRA terminology. Our objective is to detect the semantic relatedness between the MedDRA terms. To achieve this, we combine two terminology structuring approaches applied to a raw list of the MedDRA terms for the inferring of synonymy and hierarchical *is-a* relations. The inferred relations are considered as directed graphs and clustered within non disjoint clusters. The results are evaluated against the Standardized MedDRA Queries and with an expert.

Keywords. Adverse Drug Reaction Reporting Systems; Algorithms; Artificial Intelligence; Automatic Data Processing; Vocabulary, Controlled; Natural Language Processing

1. Introduction

The surveillance of the adverse drug reaction (ADR) profile and the prevention of the potential or known risk related to the use of medicine products is a core activity of pharmacovigilance. When detected, a new serious ADR may modify the conditions of the use of the medicinal product: reduce its use or even may result in withdrawal of the drug from the market. Safety signal detection depends on the quality and specific features of the ADR coding. Currently, the ADRs are coded with the MedDRA terminology [1] (Medical Dictionary for Drug Regulatory Activities). For the analysis of these databases and the signal detection, traditional pharmacovigilance methods [2-3] are exploited. They are currently supplemented by statistical algorithms [4-5]. To improve the signal detection, these methods benefit from groupings of related ADR terms [6], which are relevant because the structure of MedDRA is very fine-grained and closely related terms can be spread in this terminology (*hepatitis infectious*,

hepatitis infectious mononucleosis, *hepatitis viral*): the use of very specific terms for the coding of ADRs may cause a dilution of signals [7]. To remedy this effect, SMQs (Standardized MedDRA Queries) are created. They gather MedDRA terms specific to a given medical condition independent from their primary MedDRA hierarchies (or SOCs). The SMQs are defined by groups of experts through a manual study of the MedDRA's structure and the scientific literature [8], which is a tedious task. Now there are 84 SMQs that cover several important medical conditions, as for example *Acute renal failure*, *Agranulocytosis* or *Angioedema*. But several other SMQs are still to be defined and new *ad hoc* queries may be necessary as new signals emerge.
The SMQs have two specificities: (1) the variety of SOCs within SMQs varies between 4 and 25 (the full number of SOCs being 32); (2) a given term can belong to more than one SMQ. Indeed, the ADRs can appear in relation to different medical conditions. These observations show that the selection of the terms for the SMQs follows a very precise medical logic and does not respect the MedDRA hierarchy.
We propose an automatic method to assist the creation of SMQs. There are very few publications on this: grouping of the ADR terms with hierarchical subsumption [9-10] or semantic distance [11-12], or extension of Pubmed queries for pharmacovigilance [13]. Only one publication [10] has partially evaluated the results against the SMQs. In our previous work [14], we exploited the semantic similarity measures and evaluated the obtained groupings against the SMQs. The precision is usually high, although the recall remains low. In this work, we propose to tackle this problem with Natural Language Processing (NLP) methods dedicated to terminology structuring. We aim at the detection of synonym and hierarchical relations. Synonymy stands for terms which have identical or very close meanings (*hepatitis infectious*, *hepatitis viral)*. Hierarchical relations link a more general to a more specific term (*hepatitis infectious*, *hepatitis infectious mononucleosis)*. We assume these two kinds of relations may indeed help in the detection of semantically close terms and in the building of the clusters of pharmacovigilance terms. We analyze the results provided by these approaches and evaluate them against the SMQs and with an expert.

2. Material and Methods

Material. We exploit material issued from MedDRA: raw list of the 18,209 MedDRA PTs (preferred terms) and 84 SMQs (the gold standard). We exploit also three sets of linguistic resources: 1) medical synonyms directly extracted from the UMLS (n=228,542); 2) medical synonyms induced from biomedical terminologies [15] (n=28,691); 3) general language synonyms from WordNet [16] (n=45,782).

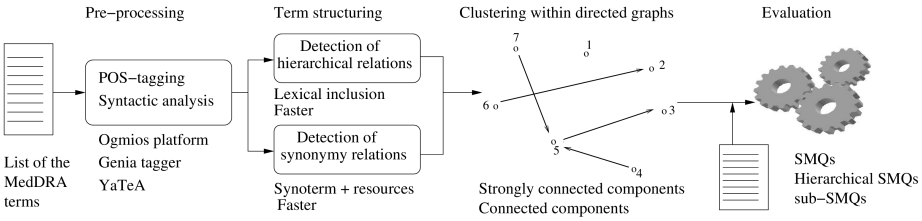

Figure 1: General schema of the approach

Methods. The proposed method consists of several steps (figure 1): 1) pre-processing of the MedDRA terms; 2) application of the terminology structuring approach to

acquire semantic relations within a raw list of the MedDRA PT terms; 3) clustering of the MedDRA terms, 2) evaluation of these clusters.

Terminology structuring methods are applied to the raw list of MedDRA terms. These are pre-processed with the POS-tagger Genia [17] and the syntactic parser YaTeA [18].

Hierarchical relations. We detect the hierarchical *is-a* relations through the lexical inclusions [19]. For instance, if one term (*hepatitis infectious*) is lexically included in another term *(hepatitis infectious mononucleosis)*, there is a hierarchical relation between them: the short term *hepatitis infectious* is the parent term and the long term *hepatitis infectious mononucleosis* is the child term.

*Synonymy relations. S*ynonymy relations are detected through their compositionality [20] which exploits the syntactic analysis of the terms and accepts the modifications in a given syntactic position, such as *infectious* and *viral* in the terms *hepatitis infectious* and *hepatitis viral* given that *infectious* and *viral* are known synonyms provided by the linguistic resources.

Morpho-syntactic variants. The detection of morpho-syntactic variants is done with the Faster [21] tool. It provides hierarchical and synonymy relations according to the transformation rules: insertion (*accessory respiratory muscles, accessory muscle*), derivation (*arterial stenosis, artery stenosis*), permutation (*eye burn, burns of eye*).

Clustering of the terms. The sets of terms related through hierarchical relations are considered as directed graphs. The graphs are partitioned into strongly connected components to obtain non disjoint clusters, which may share terms among them. To improve the coverage of the clusters, we add the synonyms: if a term has a synonymy relation with the term from a cluster then this term is also included in this cluster.

Evaluation. For the evaluation we give judgments about: (1) the correctness of the generated relations, (2) quantitative evaluation of their relevance to the creation of the SMQs through the comparison with the SMQs, (3) qualitative evaluation through a manual evaluation with an expert. The quantitative evaluation of the clusters is performed with three measures: precision P, recall R and F-measure F.

3. Results and Discussion

MedDRA terms have been processed through NLP and terminology structuring tools. The best experience is when the lexical inclusions are augmented by Faster and by compositionally computed synonyms. Manual analysis of hierarchical relations indicates that these relations are always correct: the constraint involved through the syntactic analysis guarantees correct generations. We have observed some syntactic ambiguities: *anticonvulsant drug level* may have *level* or *drug level* as hierarchical farther. But whatever the syntactic analysis, the semantic relation remains correct.

When we use only hierarchical relations (lexical inclusions and insertion rule of Faster), we obtain 748 clusters with a mean of 3.43 terms per cluster within the interval [1; 117]. When these clusters are augmented with compositionally-inferred synonyms and derivation and permutation rules of Faster, the mean size (but not the number) of clusters is increased to 3.82 terms/cluster within the interval [1; 119].

Figure 2 provides the quantitative evaluation. We can observe that there is a great variability among the SMQs and that the precision (green line) is systematically high while the recall (blue lines) is low. This means that the generated clusters are small but precise and that they cover specific aspects of SMQs. We observed that hierarchical relations form the core of the clusters (up to 96% of the involved terms) and show 69% precision. Only three clusters do not contain hierarchical relations. The Faster relations

are involved in 50% of clusters and show also a high precision between 75 and 85%. 30% of the clusters contain synonymy relations: their precision varies between 55 and 69%. Such a high rate of useful hierarchical relations seems to indicate that the processed MedDRA PT terms, although they belong to the same hierarchical level of MedDRA, have in reality not sibling but hierarchical relations among them. This aspect may be refined in the MedDRA terminology and additional hierarchical levels may be added to it. On contrary, the synonymy relations are well distinguished at this MedDRA level and very few PT terms have such synonymy relations among them.

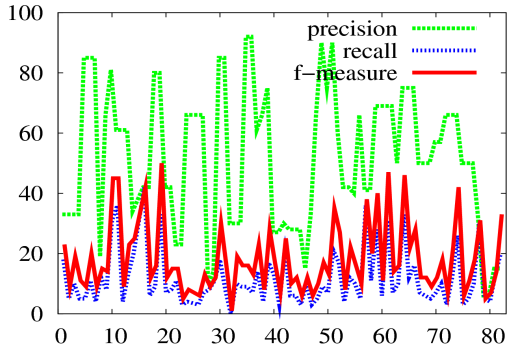

Figure 2: Precision, recall and f-measure for the terminology structuring approach

False negatives within the clusters are due to the fact that the exploited NLP methods cannot capture the lexical and semantic relations between the terms. For this, other methods should be used to increase the coverage of the clusters.

We also performed a detailed analysis of the false positives with an expert. We present the results for the *Agranulocytosis* cluster, but across the clusters we observe similar situations. Thus, it has been observed that the SMQs may miss relevant terms [22], while with our approach we have found some of the missing terms within the *Agranulocytosis* SMQ: *Herpes sepsis*, *Candida sepsis*, *Fungal sepsis* and *Anthax sepsis*. According to the expert, although these terms do not appear in the corresponding SMQ, they should be considered correct for this safety topics (in which case these four terms do not count as false positives). The other 32 terms proposed by our approach are already included in the SMQ (*i.e.*, *Listeria sepsis*, *Stenotrophomonas sepsis*, *Meningococcal sepsis*). Finally, our methods can miss the relevant terms, such as *Abscess peritonsillar*, *Acute agranulocytosis*, *Agranulocytic angina*, *Bone marrow failure* or *Catheter related septicaemia* for the SMQ *Agranulocytosis.* As a matter of fact, after the analysis of the expert, the corrected performances of the generated clusters from figure 2 may be improved by several points.

Our experiences indicate that the proposed automatic approaches may provide a useful basis for the creation of SMQs, especially because they systematically collect all the relevant terms which satisfy the given algorithmic conditions.

4. Conclusion and Perspectives

We applied terminology structuring methods for the clustering of pharmacovigilance terms. Although the automatic creation of the SMQs is a difficult task, our results seem to indicate that the automatic methods may be used as a basis for the creation of new

SMQs. The precision of the clusters is often very high. The very important amount of the hierarchical relations inferred between the PT terms suggests that the terms from this hierarchical level of the MedDRA may receive a more fine-grained organization.
Future studies will lead to the identification of other parameters which influence the quality and completeness of clusters. For instance, we plan to exploit hierarchical relations from the UMLS and to test other NLP tools and methods. We observed that the performances vary according to the SMQs and it appears that different strategies should be used for different SMQs. Different filters (i.e., lexical and hierarchical) will be tested to clean up the results and to remove the true false positive terms. Besides, the generated clusters will also be evaluated through their impact on the drug safety survey. First tests (data not presented) seem to have a positive effect. This is the main expected impact of our work.

References

[1] Brown E, Wood L & Wood S. (1999). The medical dictionary for regulatory activities (MedDRA). Drug Saf., **20**(2), 109–17.

[2] Almenoff JS, Tonning JM, Gould AL & al. (2005). Perspectives on the use of data mining in pharmacovigilance. Pharmacoepidemiol Drug Saf., **28**, 981-1007.

[3] Bailey S, Singh A, Azadian R & al. (2010). Prospective data mining of six products in the US FDA Adverse Event Reporting System: disposition of events identified and impact on product safety profiles. Pharmacoepidemiol Drug Saf., **33**(2), 139–46.

[4] Bate A, Lindquist M, Edwards I & al. (1998). A bayesian neural network method for adverse drug reaction signal generation. Eur J Clin Pharmacol, **54**(4), 315–21.

[5] Meyboom R, Lindquist M, Egberts A & Edwards I.(2002). Signal selection and follow-up in pharmacovigilance. Drug Saf, **25**(6), 459–65.

[6] Hauben M & Bate A. (2009). Decision support methods for the detection of adverse events in post-marketing data. Drug Discov Today, **14**(7-8), 343–57.

[7] Fescharek R, Kübler J, Elsasser U, Frank M & Güthlein P. (2004). Medical dictionary for regulatory activities (MedDRA): Data retrieval and presentation. Int J Pharm Med, **18**(5), 259–269.

[8] CIOMS (August 2004). Development and Rational Use of Standardised MedDRA Queries (SMQs): Retrieving Adverse Drug Reactions with MedDRA. Report of the CIOMS Working Group, CIOMS.

[9] Alecu I, Bousquet C & Jaulent MC. (2008). A case report: using SNOMED CT for grouping adverse drug reactions terms. BMC Med Inform Decis Mak, **8**(S1), 4.

[10] Jaulent MC & Alecu I. Evaluation of an ontological resource for pharmacovigilance. In Stud Health Technol Inform, pages 522--6, 2009

[11] Bousquet C, Henegar C, Lillo-Le Louët A & al.. Implementation of automated signal generation in pharmacovigilance using a knowledge-based approach. *Int J Med Inform*, 74(7-8):563--71, 2005

[12] Iavindrasana J, Bousquet C, Degoulet P & Jaulent MC. Clustering who-art terms using semantic distance and machine algorithms.In AMIA Annu Symp Proc, pages 369--73, 2006

[13] Delamarre D, Lillo-Le Louët A, Guillot L & al. Documentation in pharmacovigilance: using an ontology to extend and normalize Pubmed queries. Stud Health Technol Inform2010: 518-22

[14] Dupuch M, Bousquet C & Grabar N. Automatic creation and refinement of the clusters of pharmacovigilance terms. In ACM IHI 2012. To appear

[15] Grabar N & Hamon T. (2010). Exploitation of linguistic indicators for automatic weighting of synonyms induced within three biomedical terminologies. In MEDINFO, p. 1015–9.

[16] Fellbaum C. (1998). A semantic network of english: the mother of all WordNets. Computers and Humanities, EuroWordNet: a multilingual database with lexical semantic network, **32**(2-3), 209–20.

[17] Tsuruoka Y, Tateishi Y, Kim JD & al. (2005). Developing a robust part-of-speech tagger for biomedical text. In LNCS, p. 7746:382–92.

[18] Aubin S & Hamon T. (2006). Improving term extraction with terminological resources. In FinTAL, number 4139 in LNAI, p. 380–87.

[19] Kleiber G & Tamba I. (1990). L'hyperonymie revisitée: inclusion et hiérarchie. Langages, 98: 7–32.

[20] Partee BH. (1984). Compositionality. In Landman F. & Veltman F., editor, Varieties of formal semantics.

[21] Jacquemin C. (1996). A symbolic and surgical acquisition of terms through variation. In Connectionist, statistical and symbolic Approaches to Learning for Natural Language Processing, p. 425–38.

[22] Pearson R, Hauben M, Goldsmith D & al.(2009). Influence of the MedDRA hierarchy on pharmacovigilance data mining results. Int J Med Inform, **78**(12), 97–103.