Semantic distance and terminology structuring methods for the detection of semantically close terms

Marie Dupuch
CNRS UMR 8163 STL
Université Lille 1&3
59653 Villeneuve d'Ascq, France
dupuchm@hotmail.fr

Laëtitia Dupuch
Université Toulouse III Paul Sabatier
France
laetitia1dupuch@hotmail.com

Thierry Hamon
LIM&BIO (EA3969) UFR SMBH
Université Paris 13, France
thierry.hamon@univ-paris13.fr

Natalia Grabar
CNRS UMR 8163 STL
Université Lille 1&3
59653 Villeneuve d'Ascq, France
natalia.grabar@univ-lille3.fr

Abstract

The identification of semantically similar linguistic expressions despite their formal difference is an important task within NLP applications (information retrieval and extraction, terminology structuring...) We propose to detect the semantic relatedness between biomedical terms from the pharmacovigilance area. Two approaches are exploited: semantic distance within structured resources and terminology structuring methods applied to a raw list of terms. We compare these methods and study their complementarity. The results are evaluated against the reference pharmacovigilance data and manually by an expert.

1 Introduction

When an automatic system is able to identify that different linguistic expressions convey the same or similar meanings, this is a positive point for several applications. For instance, when documents referring to *muscle pain* or *cephalgia* are searched, information retrieval system can also take advantage of the synonyms, like *muscle ache* or *headache*, to return more relevant documents and in this way to increase the recall. This is also a great advantage for systems designed for instance for text mining, terminology structuring and alignment, or for more specific tasks such as pharmacovigilance.

The pharmacovigilance area covers the identification of adverse drug reactions (ADRs) in order to improve the vigilance on the health products. Pharmacovigilance reports are traditionally encoded with normalised terms from the dedicated terminologies, such as MedDRA (Medical Dictionary for Drug Regulatory Activities) (Brown et al., 1999). MedDRA is a relatively fine-grained terminology with nearly 90,000 terms. This means that a given pharmacovigilance report can be coded with different terms which have close meaning (Fescharek et al., 2004), like *muscle pain* and *muscle ache* or *headache* and *cephalgia*: although formally different the terms from these pairs have the same meaning. The difficulty is then to detect their semantic closeness. Indeed, if this semantic information is available, reports from the phramacovigilance databanks and mentionning similar adverse events can be aggregated: the safety signal is intensified and the safety regulation process is improved.

In order to aggregate the pharmacovigilance reports, several types of semantic information from MedDRA are used: (1) different hierarchical levels of MedDRA between the five levels available; (2) the SMQs (Standardized MedDRA Queries) which group together terms associated to a given medical condition such as *Acute renal failure*, *Angioedema* or *Embolic and thrombotic events*; and (3) specific resources (Bousquet et al., 2005; Iavindrasana et al., 2006; Alecu et al., 2008; Jaulent and Alecu, 2009). The SMQs are defined by groups of experts through a long and meticulous work consisting of the manual study of the MedDRA structure and of the analysis of the scientific literature (CIOMS, 2004). 84 SMQs have been created so far. They become the gold standard data of the pharmacovigilance area. However, the SMQs currently suffer from the lack of exhausitivity (Pearson et al., 2009): the set of SMQs is not exhaustive because this is an ongoing work. We assume that automatic approaches can be ex-

Proceedings of the 2012 Workshop on Biomedical Natural Language Processing (BioNLP 2012), pages 20–28,
Montréal, Canada, June 8, 2012.

ploited to systematize and accelerate the process of recruiting the semantically related MedDRA terms and to build the SMQs. We propose to exploit two approaches: methods dedicated to the terminology structuring and semantic distance approaches. We compare and combine the generated results. For the evaluation, we compare the results with the existing SMQs and also analyse them manually with an expert. Our work is different from previous work because we exploit the whole set of the available MedDRA terms, we apply several methods to cluster the terms and we perform several types of evaluation.

2 Material

We exploit two kinds of material: material issued from MedDRA and specific to the pharmacovigilance area (sections 2.1 and 2.3), and linguistic resources issued from general and biomedical languages (section 2.2). The MedDRA terms are structured into five hierarchical levels: *SOC* (*System Organ Class*) terms belong to the first and the highest level, while *LLT* (*Lowest Level Terms*) terms belong to the fifth and the lowest level. Terms from the fourth level *PT* (*Preferred Terms*) are usually exploited for the coding of the pharmacovigilance reports. They are also used for the creation of SMQs. A given PT term may belong to several SMQs.

2.1 Ontology ontoEIM

ontoEIM is an ontology of ADRs (Alecu et al., 2008) created through the projection of MedDRA to SNOMED CT (Stearns et al., 2001). This projection is performed thanks to the UMLS (NLM, 2011), where an important number of terminologies are already merged and aligned, among which MedDRA and SNOMED CT. The current rate of alignment of the *PT* MedDRA terms with SNOMED CT is weak (version 2011): 51.3% (7,629 terms). Projection of MedDRA to SNOMED CT allows to improve the representation of the MedDRA terms:

- the structure of the MedDRA terms is parallel to that of SNOMED CT, which makes it more fine-grained (Alecu et al., 2008). The number of hierarchical levels within the ontoEIM reaches 14, instead of five levels in MedDRA;
- the MedDRA terms receive formal definitions: semantic primitives which decompose the meaning. MedDRA terms can be described along up to four axes from SNOMED CT, exemplified here through the term *Arsenical keratosis*: (1) *Morphology* (type of abnormality): *Squamous cell neoplasm*; (2) *Topography* (anatomical localization): *Skin structure*; (3) *Causality* (agent or cause of the abnormality): *Arsenic AND OR arsenic compound*; and (4) *Expression* (manifestation of the abnormality): *Abnormal keratinization*. The formal definitions are not complete. For instance, only 12 terms receive formal definitions along these four axes and 435 along three axes. This is due to the incomplete alignment of the MedDRA terms and to the fact these four elements are not relevant for every term (their absence is not always problematic).

2.2 Linguistic resources

Linguistic resources provide three kinds of pairs of synonym words: (1) Medical synonyms extracted from the UMLS 2011AA (n=228,542) and then cleaned up (n=73,093); (2) Medical synonyms acquired from three biomedical terminologies thanks to the exploitation of their compositionality (Grabar and Hamon, 2010) (n=28,691); (3) Synonyms from the general language provided by WordNet (Fellbaum, 1998) (n=45,782). Among the pairs of words recorded in these resources, we can find {*accord*, *concordance*}, {*aceperone*, *acetabutone*}, {*adenazole*, *tocladesine*}, {*adrenaline*, *epinephrine*} or {*bleeding*, *hemorrhage*}. The last two pairs are provided by medical and general resources. However, the pair {*accord*, *concordance*} is provided only by medical resources.

2.3 Standardized MedDRA Queries

We exploit 84 SMQs as reference data. Among these SMQs, we distinguish 20 SMQs which are structured hierarchically. We also exploit 92 sub-SMQs, which compose these 20 hierarchical SMQs.

3 Methods

Our method consists into four main steps (figure 1): (1) computing of the semantic distance and similarity between the MedDRA terms and their clustering (section 3.1), (2) the application of the terminology structuring methods to acquire semantic re-

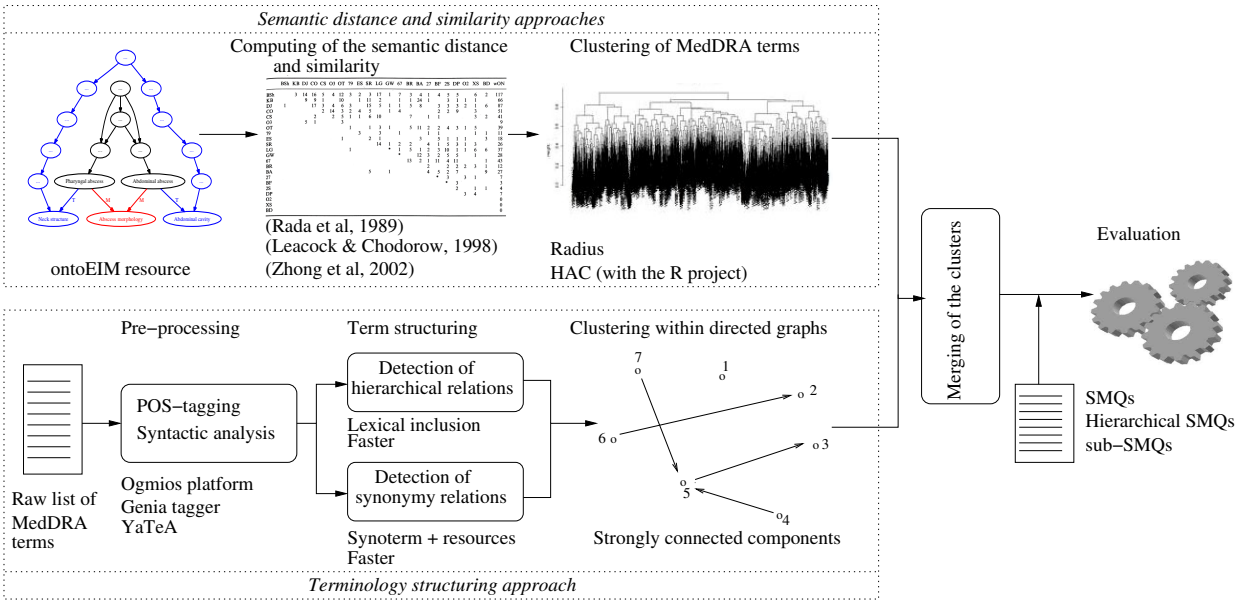


Figure 1: General schema of the experiment composed of four steps: (1) semantic distance approaches, (2) terminology structuring approaches, (3) their combination and (4) their evaluation

lations between MedDRA terms and their clustering (section 3.2), (3) the merging of these two sets of clusters (section 3.3), (4) the evaluation of the merged clusters (section 3.4). We exploit Perl language, R[1] project and several NLP tools.

3.1 Semantic distance approach

The semantic distance and similarity approach is applied to the 7,629 *PT* MedDRA terms and their formal definitions from ontoEIM. The two main steps are: computing the distance or similarity (section 3.1.1) and clustering of terms (section 3.1.2).

3.1.1 Computing the semantic distance

Because we work with a tree-structured resource, we exploit edge-based algorithms to compute the distance or similarity between two terms *t1* and *t2*: two semantic distances (*Rada* (Rada et al., 1989) and *Zhong* (Zhong et al., 2002)) and one semantic similarity (Leacock and Chodorow, 1998). In the following, we call them semantic distance algorithms. For each algorithm, three paths may be exploited: between the MedDRA terms but also between the elements of their formal definitions on two axes (morphology M and topography T often involved in diagnostics (Spackman and Campbell, 1998)). For the illustration, let's consider two MedDRA terms, *Abdominal abscess* and *Pharyngeal abscess* defined as follows:

- *Abdominal abscess*: M = *Abscess morphology*, T = *Abdominal cavity structure*
- *Pharyngeal abscess*: M = *Abscess morphology*, T = *Neck structure*

The shortest paths sp are computed between these two MedDRA terms and between their formal definitions, whose hierarchical structure is also inherited from SNOMED CT. The weight of edges is set to 1 because all the relations are of the same kind (hierarchical), and the value of each shortest path corresponds to the sum of the weights of all its edges. The semantic distance sd are then exploited to compute the unique distance between the ADR terms from MedDRA: $\frac{\sum_{i\in\{ADR,M,T\}} W_i * sd_i(t1,t2)}{\sum_{i\in\{ADR,M,T\}} W_i}$, where the three axes $\{ADR, M, T\}$ respectively correspond to terms meaning the ADR, axis Morphology M and axis Topography T; $t1$ and $t2$ are two ADR terms; W_i is the coefficient associated with each of the three axes; and sd_i is the semantic distance computed on a given axis. We carry out several ex-

[1] http://www.r-project.org

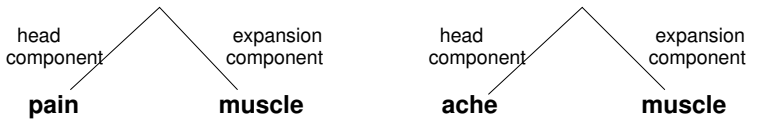


Figure 2: Syntactically analyzed terms (*muscle pain* and *muscle ache*) into their head and expansion components

periments. Semi-matrices 7629*7629 with semantic distance between the terms are built.

3.1.2 Clustering of terms

An unsupervised creation of clusters is applied to the semi-matrices. We exploit two approaches:

- R radius approach: every MedDRA term is considered a possible center of a cluster and its closest terms are clustered with it. The thresholds tested correspond to the following intervals: 2 and 3 for *Rada*, [0; 5.059] for *LCH* and [0; 0.49] for *Zhong*. The intersection of these clusters is not empty.
- *HAC* hierarchical ascendant classification is performed through the *R Project* tools (*hclust* function). Iteratively, this function chooses the best centers for terms and builds the hierarchy of terms by progressively clustering those which are closest to these centers. Then the unique cluster with all the terms is split up. Several splitting values between 100 and 7,000 are tested. These clusters are exclusive.

Clusters created with the radius approach are merged in order to eliminate smaller clusters included in bigger clusters and in order to aggregate clusters which have an important intersection between them. For the intersection, we test several intersection values within the interval [10; 90], which means that two compared clusters may have between 10% and 90% of common terms.

3.2 Terminology structuring approach

The terminology structuring methods are applied to a raw list of 18,209 MedDRA PTs. They allow the detection of semantic relations between these terms. The POS-tagging is done with Genia tagger (Tsuruoka et al., 2005) and the syntactic analysis with the YATEA parser (Aubin and Hamon, 2006). Three kinds of methods are applied for the acquisition of synonymy and hierarchical relations: lexical inclusions (section 3.2.1), morpho-syntactic variants (section 3.2.2) and compositionality (section 3.2.3). The terms are then clustered (section 3.2.4).

3.2.1 Lexical inclusion and hierarchy

The lexical inclusion hypothesis (Kleiber and Tamba, 1990), which states that when a given term is lexically included at the head syntactic position in another term there is a semantic subsumption between them, allows to identify hierarchical relations between terms. For instance, on figure 2, the short term *pain* is the hierarchical parent and the long term *muscle pain* is its hierarchical child because *pain* is the syntactic head of *muscle pain*. The lexical inclusions are computed on POS-tagged and syntactically analyzed terms. We compute two kinds of lexical inclusions:

- syntactic dependencies on minimal syntactic heads: the parent term corresponds to the shortest lexical form of the syntactic head. For instance, within the term *kaolin cephalin clotting time*, the minimal head is *time*;
- syntactic dependencies on maximal syntactic heads: the parent term is the most complete lexical form of the syntactic head. Within the same term *kaolin cephalin clotting time*, the maximal head is *cephalin clotting time*.

Parent and child terms have to be MedDRA terms.

3.2.2 Morpho-syntactic variants

We exploit Faster (Jacquemin, 1996) for the indentification of morpho-syntactic variants between the PT terms. This tool applies several transformation rules, such as insertion (*cardiac disease*/*cardiac valve disease*), morphological derivation (*artery restenosis*/*arterial restenosis*) or permutation (*aorta coarctation*/*coarctation of the aorta*). Each transformation rule is associated with hierarchical or synonymy relations: the insertion introduces a hierarchical relation (*cardiac valve disease* is more specific than *cardiac disease*), while the permutation introduces a synonymy relation. When several transformations are involved, the detected relations may

be ambiguous: *gland abscess* and *abscess of salivary gland* combines permutation (synonymy) and insertion (hierarchy) rules. In such cases the hierarchical relation prevails.

3.2.3 Compositionality and synonymy

The synonymy relations are acquired in two ways. First, the synonymy relation is established between two simple MedDRA terms if this relation is provided by the linguitistic resources. Second, the identification of synonym relations between complex terms relies on the semantic compositionality (Partee, 1984). Hence, two complex terms are considered synonyms if at least one of their components at the same syntactic position (head or expansion) are synonyms. For instance, on figure 2, given the synonymy relation between the two words *pain* and *ache*, the terms *muscle pain* and *muscle ache* are also identified as synonyms (Hamon and Nazarenko, 2001). Three transformation rules are applied: on the head component (figure 2), on the expansion component and on both of them. We perform several experiments: each medical synonymy resource is first used individually and then in combination with WordNet.

3.2.4 Clustering of terms

The sets of terms related through the lexical inclusions are considered as directed graphs: the terms are the nodes of the graph while the hierarchical relations are the directed edges. We partition these directed graphs and identify clusters of terms which could correspond to or be part of the SMQs. Among connected components and strongly connected components, we choose to generate the strongly connected components: they allow an intersection between clusters which means that a given term may belong to several clusters (this is also the case with the SMQs). Thus, within the directed graphs G we have to identify the maximal sub-graphs H of G where for each pair $\{x, y\}$ of the nodes from H, there exists a directed edge from x to y (or from y to x). To improve the coverage of the obtained clusters, we also 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. From a graph theory point of view, the initial graph is augmented with two edges going from and to the synonyms.

Methods and relationships	#relations
Hierarchical relations	
Maximal syntactic head	3,366
Minimal syntactic head	3,816
Morpho-syntactic variants	743
Medical synonyms	
3 biomedical terminologies	1,879
UMLS/Filtered UMLS	190
Morpho-syntactic variants	100
Medical synonyms and WordNet	
3 biomedical terminologies	1,939
UMLS/Filtered UMLS	227

Table 1: Hierarchical and synonymy relations generated by terminology structuring methods

3.3 Merging of clusters from two approaches

We merge the clusters generated by the two approaches. The merging is performed on the intersection between the clusters. As previously, we test intersection values within the interval [10; 90].

3.4 Evaluation

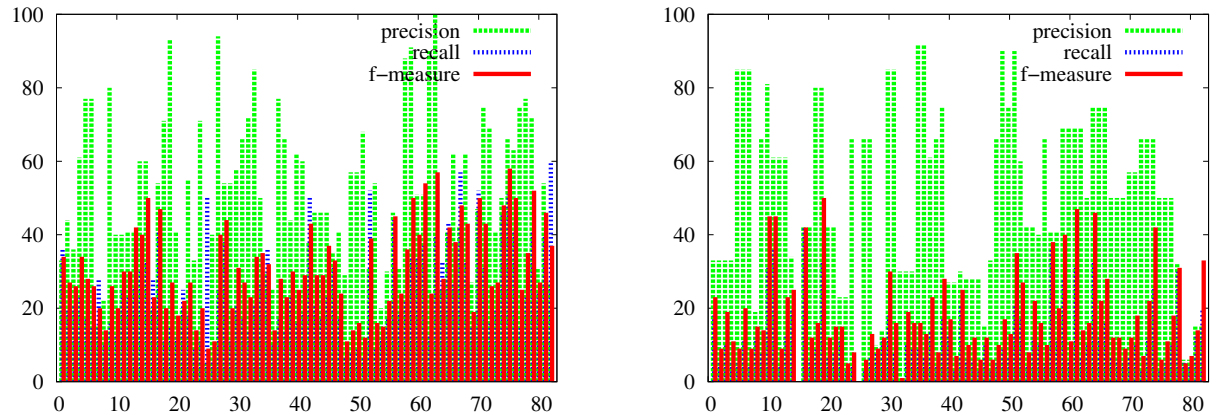
We give judgments on: (1) the correctness of the generated relations, (2) their relevance according to the reference data, (3) their relevance according to the manual evaluation by an expert. The evaluation is performed with three measures: precision P (percentage of the relevant terms clustered divided by the total number of the clustered terms), recall R (percentage of the relevant terms clustered divided by the number of terms in the corresponding SMQ) and F-measure F_1. The association between the SMQs and the clusters relies on the best F_1.

4 Results

Semantic relations acquired with terminology structuring are indicated in table 1. There is a small difference between relations acquired through maximal and minimal syntactic heads, although the influence of medical resources for the acquisition of synonymy varies according to the resources. WordNet slightly increases the number of synonyms. Faster generates a large set of hierarchical and synonymy relations. MedDRA terms have also been processed with semantic distance and clustered. The best thresholds with the radius clustering are 2 for *Rada*,

Approach	Hierarchical SMQs			SMQs and sub-SMQs		
	#clusters	interval	mean	#clusters	interval	mean
Semantic distance	2,667	[2; 1,206]	73	2,931	[2; 546]	17
Structuring (hierarchical)	690	[1; 134]	3.69	748	[1; 117]	3.43
Structuring (hierarchical+synonymy)	690	[1; 136]	4.11	748	[1; 119]	3.82
Merging (hierarchical)	2,732	[1; 1,220]	72.40	2,998	[1; 563]	24.44
Merging (hierarchical+synonymy)	2,732	[1; 1,269]	75.94	2,998	[1; 594]	26.03

Table 2: Number of clusters and their size (the interval and the mean number of terms per cluster) for individual approaches and for their merging computed for hierarchical SMQs and also for SMQs and sub-SMQs



(a) Semantic distance

(b) Terminology structuring

Figure 3: Results (precision, recall and F-measure) for semantic distance and terminology structuring approaches

4.10 for *LCH* and 0 for *Zhong*. With the HAC, the best results are obtained with 300 classes (number of terms per class is within the interval [1; 98], mean number of terms per class is 25.34). Our results show that the best parameters for the semantic distance are the Rada distance, radius approach and no formal definitions, while the best parameters for the terminology structuring are maximal syntactic head with hierarchical relations by Faster augmented with synonyms. For the merging of the clusters we apply 50% intersection for hierarchical SMQs and 80% intersection for SMQs and sub-SMQs. We exploit and discuss these results. The percentage of the MedDRA terms involved by the terminology structuring is the 32% with hierarchical relations, it reaches 40% when the synonymy is also considered. With semantic distance, all the terms from ontoEIM (51% of the MedDRA) are used.

Table 2 provides information on clusters: number of clusters, number of terms per cluster (their interval and the mean number of terms per cluster). In table 2, we first indicate the results for the individual approaches, and then when the merging of the approaches is performed. We observe that the merging has a positive effect on the number and the size of clusters: data generated by the individual approaches (and by synonymy) are complementary.

4.1 Correctness of the semantic relations

A manual analysis of the generated hierarchical relations indicates that these relations are always correct: the constraint involved through the syntactic analysis guarantees correct propositions. Nevertheless, we observed a small number of syntactic ambiguities. They appear within 144 pairs (5%) with maximal syntactic heads and correspond to pairs like: {*anticonvulsant drug level*, *drug level*}, {*blood smear test*, *smear test*}, {*eye movement disorder*, *movement disorder*}. Thus, within the first

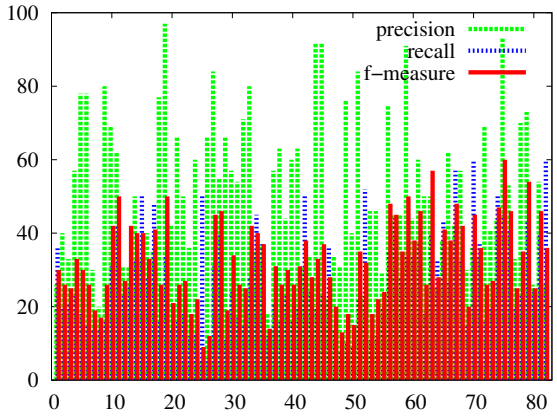


Figure 4: Results (precision, recall and F-measure) obtained when the two approaches are merged

pair, there is an ambiguity on *drug* as two dependencies seem possible: {*anticonvulsant drug level*, *drug level*} as proposed by the system and {*anticonvulsant drug level*, *level*}. But whatever the syntactic analysis performed, the semantic relations are correct.

4.2 Relevance of the generated clusters

Figures 3 and 4 provide quantitative evaluation of the clusters: semantic distance (figure 3(a)), terminology structuring (figure 3(b)), merging of these two sets (figure 4). On figure 3, we can observe that there is a great variability among the SMQs and the two approaches. The positive result is that these approaches are indeed complementary: their merging slightly increases performance. An analysis of the clusters generated with terminology structuring shows that: (1) hierarchical relations form the basis of the clusters: they correspond to 96% of the involved terms and show 69% precision. Only three clusters do not contain hierarchical relations; (2) Faster relations are involved in 50% of clusters and show precision between 75 and 85%; (3) one third of the clusters contains synonymy relations, which precision varies between 55 and 69%; (4) relations acquired with the UMLS resources are involved in 14% of clusters while their precision is only 38%.

We also performed a detailed qualitative analysis of several SMQs and clusters with an expert. Table 3 presents the analysis for three SMQs: *Angioedema*, *Embolic and thrombotic events, arterial* and *Haemodynamic oedema, effusions and fluid overload*. It indicates the number of terms in the SMQ and in the corresponding clusters *clu*, as well as the number of common terms between them *com* and the performance (precision P, recall R and F-measure F) when computed against the reference data *Reference* and also after the analysis performed by the expert *After expertise*. The results obtained with the two approaches are indicated: semantic distance *sd* and terminology structuring *struc*, as well as their merging *merg*. In the colums *Reference*, we can observe that the best F-measure values are obtained with the terminology structuring method for the SMQ *Haemodynamic oedema, effusions and fluid overload* (F=45) and with the semantic distance for the SMQ *Embolic and thrombotic events, arterial* (F=32). The merging of the two methods systematically improves the results: in the given examples, for all three SMQs.

A detailed analysis of the generated noise indicates that across the SMQs we have similar situations: we generate false positives (terms non relevant for the medical conditions, such as *Pulmonary oedema, Gestational oedema, Spinal cord oedema* for the SMQ *Angioedema*), but also the SMQs may contain non relevant terms or may miss relevant terms (thus, *Testicular oedema, Injection site urticaria, Bronchial eodema* are missing in the SMQ *Angioedema*). The expert evaluation (columns *After expertise* in table 3) attempts to analyse also the quality of the SMQs. The corrected performance of the clusters is improved in several points, which indicates that automatic approaches may provide a useful basis for the creation of SMQs.

5 Discussion

Despite the incompleteness of the ontoEIM resource, the semantic distance approach is quite efficient and provides the core terms for the building of the SMQs. Among the several algorithms tested, the most simple algorithm (Rada et al., 1989), which exploits the shortest path, leads to the best results, while the additional information on the hierarchical depth exploited by other algorithms appears non useful. The clustering method which allows the generation of non-disjoint clusters is the most efficient as MedDRA terms may belong to several SMQs.

SMQs	Number of terms			Reference			After expertise		
	SMQ	clu	com	P	R	F	P	R	F
Angioedema$_{sd}$	52	32	13	40	25	30	43	26	33
Angioedema$_{struc}$	52	31	19	61	36	45	61	36	45
Angioedema$_{merg}$	52	33	21	63	42	50	71	48	57
Embolic and thrombotic events...$_{sd}$	132	159	48	30	36	32	32	39	35.2
Embolic and thrombotic events...$_{struc}$	132	13	12	92	9	16	92	9	16
Embolic and thrombotic events...$_{merg}$	132	130	49	38	37	37.5	47	46	46.5
Haemodynamic oedema, effusions...$_{sd}$	36	22	7	32	20	24	54	33	41
Haemodynamic oedema, effusions...$_{struc}$	36	31	13	42	36	39	84	72	78
Haemodynamic oedema, effusions...$_{merg}$	36	35	16	46	44	45	86	83	84.5

Table 3: Comparison between the two approaches (semantic distance *sd* and terminology structuring *struc*) and the merging of the two approaches *merg* for three SMQs: *Angioedema*, *Embolic and thrombotic events, arterial* and *Haemodynamic oedema, effusions and fluid overload*

Traditionnal classification methods, which produce disjoint clusters, are less efficient for this task.

It has been surprising to observe that the contribution of the generated hierarchical relations is so important (table 1) and that these relations appear to be so often correct for the creation of SMQs. Indeed, because PT terms belong to the same hierarchical level of MedDRA, they should be hierarchically equivalent between them. In reality, within a cluster, we can find several hierarchical levels of the PT terms. This means that the hierarchical structure of MedDRA could be more fine-grained and that intermediate hierarchical levels could be created. As for the generated synonymy relations, their number is low and they contribute in a lesser way to the building of the clusters: this means that the PTs are semantically differentiated between them.

Finally, the merging of these two approaches is beneficial for the generation of clusters: the performance is improved, although slightly. The two approaches provide indeed complementary results. The low recall and F-measure are due to the material and methods exploited: ontoEIM contains only 51% of the MedDRA terms to be processed while the exploited terminology structuring methods are not able to detect more common features between the terms.

The difference between the results obtained against the reference data and after the expert evaluation (table 3) show that the reference data are not very precise. In previous work, it has already been observed that some important PT terms can be missing in the SMQs (Pearson et al., 2009). With the proposed automatic methods we could find some of these terms. It has been also demonstrated that the SMQs are over-inclusive (Mozzicato, 2007; Pearson et al., 2009). In the proposed analysis of the SMQs, we have also found terms which have too large meaning and which should not be included in the SMQs.

6 Conclusion and Perspectives

We have applied two different approaches to the clustering of pharmacovigilance terms with similar or close meaning. We performed a comparison of the results obtained with these two approaches and analysed their complementarity. Several experiments have been carried out in order to test different parameters which may influence the performance of the methods. 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 satisfactory, while their merging leads to the improvement of their completeness. These approaches generate complementary data and their combination provides more performant results.

Future studies will lead to the identification of other parameters which influence the quality of clusters and also other factors which may be exploited for the merging of clusters. More robust distances and clustering methods will also be used in future work, as well as approaches for a better acquisi-

tion and evaluation of the hierarchical structure of SMQs. We plan also to design corpora-based methods which may also to increase the recall of the results. We will perform an exhaustive analysis of the nature of semantic relations which can be observed within the SMQs and propose other methods to further improve the coverage of the clusters. Different filters will be tested to remove the true false positive relations between terms. The results will also be evaluation by several experts, which will allow to assess the inter-expert variation and its influence on the results. Besides, the obtained clusters will also be evaluated through their impact on the pharmacovigilance tasks and through the exploring of the pharmacovigilance databases.

References

I Alecu, C Bousquet, and MC Jaulent. 2008. A case report: using snomed ct for grouping adverse drug reactions terms. *BMC Med Inform Decis Mak*, 8(1):4–4.

S Aubin and T Hamon. 2006. Improving term extraction with terminological resources. In *FinTAL 2006*, number 4139 in LNAI, pages 380–387. Springer.

C Bousquet, C Henegar, A Lillo-Le Louët, P Degoulet, and MC Jaulent. 2005. Implementation of automated signal generation in pharmacovigilance using a knowledge-based approach. *Int J Med Inform*, 74(7-8):563–71.

EG Brown, L Wood, and S Wood. 1999. The medical dictionary for regulatory activities (MedDRA). *Drug Saf.*, 20(2):109–17.

CIOMS. 2004. Development and rational use of standardised MedDRA queries (SMQs): Retrieving adverse drug reactions with MedDRA. Technical report, CIOMS.

C Fellbaum. 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):209–220.

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

N Grabar and T Hamon. 2010. Exploitation of linguistic indicators for automatic weighting of synonyms induced within three biomedical terminologies. In *MEDINFO 2010*, pages 1015–9.

T Hamon and A Nazarenko. 2001. Detection of synonymy links between terms: experiment and results. In *Recent Advances in Computational Terminology*, pages 185–208. John Benjamins.

J Iavindrasana, C Bousquet, P Degoulet, and MC Jaulent. 2006. Clustering WHO-ART terms using semantic distance and machine algorithms. In *AMIA Annu Symp Proc*, pages 369–73.

Christian Jacquemin. 1996. A symbolic and surgical acquisition of terms through variation. In S. Wermter, E. Riloff, and G. Scheler, editors, *Connectionist, Statistical and Symbolic Approaches to Learning for Natural Language Processing*, pages 425–438, Springer.

MC Jaulent and I Alecu. 2009. Evaluation of an ontological resource for pharmacovigilance. In *Stud Health Technol Inform*, pages 522–6.

G Kleiber and I Tamba. 1990. L'hyperonymie revisitée : inclusion et hiérarchie. *Langages*, 98:7–32, juin.

C Leacock and M Chodorow, 1998. *Combining local context and WordNet similarity for word sense identification*, chapter 4, pages 305–332.

P Mozzicato. 2007. Standardised MedDRA queries: their role in signal detection. *Drug Saf*, 30(7):617–9.

NLM, 2011. *UMLS Knowledge Sources Manual*. National Library of Medicine, Bethesda, Maryland. www.nlm.nih.gov/research/umls/.

Barbara H. Partee. 1984. Compositionality. In F. Landman and F. Veltman, editors, *Varieties of formal semantics*. Foris, Dordrecht.

RK Pearson, M Hauben, DI Goldsmith, AL Gould, D Madigan, DJ O'Hara, SJ Reisinger, and AM Hochberg. 2009. Influence of the MedDRA hierarchy on pharmacovigilance data mining results. *Int J Med Inform*, 78(12):97–103.

R Rada, H Mili, E Bicknell, and M Blettner. 1989. Development and application of a metric on semantic nets. *IEEE Transactions on systems, man and cybernetics*, 19(1):17–30.

K Spackman and K Campbell. 1998. Compositional concept representation using SNOMED: Towards further convergence of clinical terminologies. In *Journal of American Medical Informatics Association (JAMIA)*, pages 740–744.

MQ Stearns, C Price, KA Spackman, and AY Wang. 2001. SNOMED clinical terms: overview of the development process and project status. In *AMIA*, pages 662–666.

Y Tsuruoka, Y Tateishi, JD Kim, T Ohta, J McNaught, S Ananiadou, and J Tsujii. 2005. Developing a robust part-of-speech tagger for biomedical text. *LNCS*, 3746:382–392.

J Zhong, H Zhu, J Li, and Y Yu. 2002. Conceptual graph matching for semantic search. In *10th International Conference on Conceptual Structures, ICCS2002, LNCS 2393, Springer Verlag*, pages 92–106.