Cross-language detection of linguistic and semantic regularities in pharmacovigilance terms

Marie Dupuch[1], Thierry Hamon[2], Natalia Grabar[1]

(1) CNRS UMR 8163 STL, Université Lille 3, 59653 Villeneuve d'Ascq, France
(2) LIM&BIO (EA3969), Université Paris 13, Sorbonne Paris Cité, France

Abstract. Our objective is the detection of semantic relations between pharmacovigilance terms. We propose to apply the Natural Language Processing methods independently in two languages (English and French) and then to combine the obtained results. The evaluation of these relations is done via their comparison with the reference data, while their complementarity is analyzed through their involvement in the clusters. Our results show that: each language contributes almost equally to the generated results; the number of common hierarchical relations is greater than the number of common synonym relations. On the whole, the obtained results point out that in a cross-language context, each language brings additional linguistic and semantic regularities. The union of the two languages is especially beneficial to the recall and the F-measure.

1 Introduction

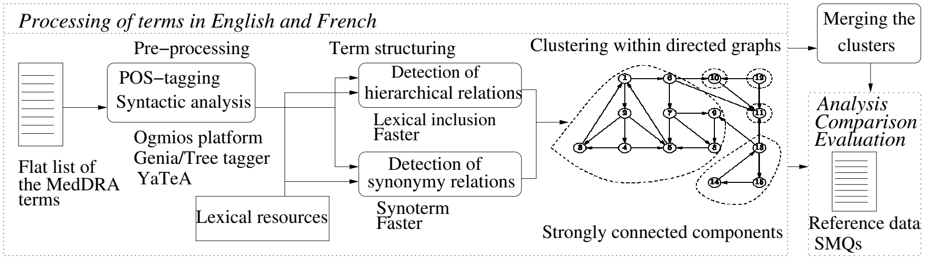
Recent research pointed out that, across languages, it is possible to find common linguistic and semantic regularities. Hence, the cross-language semantics may be exploited in several ways: (*i*) comparative studies, which allow to find interlingual universals [1, 2]; (*ii*) contrastive cross-language analysis [3]; (*iii*) transposition and adaptation of methods and resources from one language to another [4]; (*iv*) collaboration between languages [5]. Among these, we propose to work in the context of collaboration between languages. We assume the results obtained in one language can help to improve the results obtained in the other language. Our work aims at the detection of semantically and medically close terms related to the adverse drug reactions (ADRs) from the MedDRA (Medical Dictionary for Regulatory Activities) terminology [6]. This task is very useful for the generation of new alerts and for making the use of drugs more secured. Usually, the semantically close MedDRA terms are detected manually within MeDRA [7] or within specific resources [8]. We propose to use a flat list of the MedDRA terms and to detect the semantic relations among them with the Natural Language Processing (NLP) methods dedicated to the terminology structuring. The relations to be detected are morpho-syntactic variations {*artery restenosis*, *arterial restenosis*}, synonymy {*muscle ache*, *muscle pain*} and hierarchical subsumption {*renal failure*, *postoperative renal failure*}. Terms in English and French are used.

The 4th International Louhi Workshop on Health Document Text Mining and Information Analysis (Louhi 2013), edited by Hanna Suominen

Type of material	*English*	*French*
1. MedDRA Preferred Terms	18,209	18,786
2. Reference sets of relations between MedDRA terms (SMQs)	84	84
3. Linguistic resources		
UMLS synonyms [9]	227,887	126,892
Acquired biomedical synonyms [10, 11]	28,691	1,314
General language synonyms [12, 13]	50,970	115,720

Table 1. Material exploited in the two processed languages (English and French).

Fig. 1. General schema of the method.



2 Methods and Material

Material. We exploit three types of material (Table 1): (1) a flat list of the MedDRA Preferred Terms; (2) the manually detected relations between the MedDRA terms collected within 84 SMQs, or Standardised MedDRA Queries, related to various safety topics such as *Haemorrhage, Hepatic disorders, Convulsions* and (3) linguistic resources, which contain synonymy relations between simple words or terms, such as {*accord*, *concordance*}, {*aceperone*, *acetabutone*} or {*bleeding*, *hemorrhage*}. Each kind of material is exploited in English and French languages.

Methods. Figure 1 presents the general schema of the method. Terms are POS-tagged with Genia tagger [14] in English, TreeTagger [15] in French, then syntactically analyzed with the YATEA parser [16]. Three methods are applied for the acquisition of semantic relations (synonymy and hierarchical subsumption).

Morpho-syntactic variants. Identification of morpho-syntactic variants between the terms is detected with Faster [17], which applies transformation rules for processing insertion (*cardiac disease*/*cardiac valve disease*), morphological derivation (*artery restenosis*/*arterial restenosis*) or permutation (*aorta coarctation*/*coarctation of the aorta*). Insertion introduces a hierarchical relation (*cardiac valve disease* is more specific than *cardiac disease*), while permutation introduces a synonymy relation. When several rules are involved, such as in *gland abscess* and *abscess of salivary gland*, the hierarchical relation prevails.

Compositionality and synonymy. Synonymy relations are acquired in two ways: (1) synonymy relation is established between two simple terms if this relation is provided by the linguistic resources; (2) identification of synonym relations between complex terms relies on the semantic compositionality [18]. Hence, two complex terms are considered to be synonyms if at least one of their components at the same syntactic position are synonyms. For instance, given the synonymy relation between the two words *pain* and *ache*, the terms *muscle pain* and *muscle ache* are also identified as synonyms [19].

Lexical inclusion and hierarchy. According to the lexical inclusion hypothesis [20], there is hierarchical subsumption relations between two terms when one term is lexically included at a given syntactic position in another term. For instance, 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.*

Clustering of terms. The 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) and are partitioned into strongly connected components. 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 path composed of directed edges from x to y. These clusters can correspond to or be part of the reference sets (SMQs). To improve the coverage of the 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.

Evaluation and Analysis of the complementarity. The generated semantic relations are evaluated against the reference data with three measures: precision P (percentage of the relevant relations divided by the total number of the generated relations), recall R (percentage of the relevant relations divided by the number of relations in the reference SMQs) and F-measure F_1 (harmonic mean of P and R). For the analysis of the complementarity between the languages, we address issues such as: whether the relations are common or unique between the languages, whether they allow to improve the coverage or the correctness of the results, are some of the relationships more redundant between the languages.

3 Results and Discussion

In table 2, columns *# relations* show the number of the acquired relations in the two languages. We have three main observations: (1) there is more relations generated in English than in French, (2) each input resource in English contributes to the acquisition of relations, while in French the UMLS synonyms provide no results, (3) the set of the hierarchical relations induced with lexical subsumption in French (3,980) is larger than in English (3,366). Given the poor (in English) or null (in French) contribution of the UMLS synonyms (term labels which have the same concept), we show the interest to use other sources of synonyms. The following three columns (*% in clusters*) of table 2 indicate the percentage of the acquired relations in the generated clusters. We can see that the hierarchical subsumption relations bring the majority of terms into the clusters (79.57% in English and up to 96.66% in French), while the synonymy relations show but a low level of involvement (less than 1% in French, 11.81% in English).

Methods and relationships	# relations		% in the clusters		
	English	French	English	French	Union
Hierarchical rel. with lexical inclusion	3,366	3,980	79.57	96.66	89.2
Hierarchical rel. with morpho-syntactic variants	316	178	8.62	2.56	4.36
Synonymy rel. with UMLS synonyms	54	-	-	-	-
Synonymy rel. with acquired biomedical synonyms	1,110	31	-	-	-
Synonymy rel. with simple MedDRA synonyms	214	-	-	-	-
Synonymy rel. with general language synonyms	28	142	-	-	-
Total number of the acquired synonyms	1,459	164	11.81	0.78	6.44

Table 2. Relations generated in each language and their participation in the clusters.

Fig. 2. Number of common and specific relations in the two processed languages.

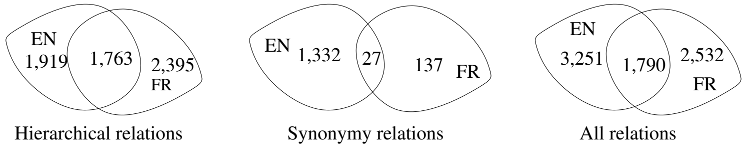


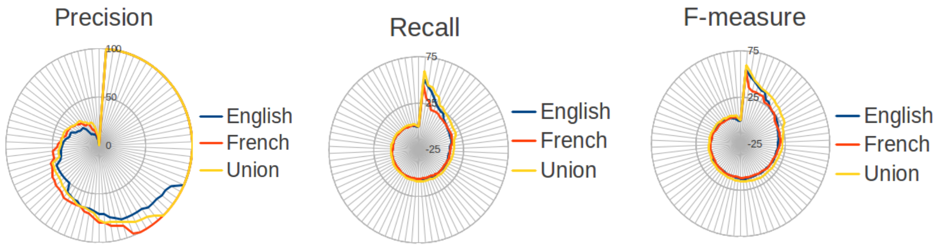
Figure 2 shows the complementarity between the two languages for each type of relations: only 27 common synonymy relations, but up to 1,763 common hierarchical relations. We also observe that the generation of the hierarchical relations seems to detect more common regularities in the two languages. Still several relations are unique to one language (*i.e.*, {*abdominal rebound tenderness*, *abdominal tenderness*} in English, {*fibrome du sein*, *tumeur du sein*} in French). In table 3, which contains data on the clusters generated with the semantic relations, we can observe that the number of clusters as well as their size intervals, are larger when the two languages are used. Thus, the two languages appear to be complementary from different points of view: within the sets of synonymy and hierarchical relations and at the level of the clusters. Moreover, their union shows that the languages contribute almost equally: 39.69% of terms unique to English, 34.03% unique to French, and 26.27% common to the two languages.

The results of the evaluation against the reference data are shown on figure 3. These are not projected on the x and y axes. Instead, the outer border of the circles indicates the reference data (84 SMQs), the radius 0-100 scale indicates the evaluation measure values. More a given line is closer to the outer border, the better the results for the corresponding method and measure. We have several observations: (1) there is an important variability between the SMQs; (2) very often, the precision is high while the recall is low (the generated clusters are smaller than the SMQs and show their different aspects), which is similar to the hierarchical subsumption obtained within MedDRA; (3) the union of the two languages has a positive effect on the Recall and F-measure. As for other automatic methods exploited within a similar task (semantic similarity, OWL queries), they tend to provide high recall but a lower precision [21, 22].

	English	French	Union
Number of clusters	965	1,133	1,571
Size of clusters (intervals)	[2; 257]	[2; 205]	[2; 301]
Size of clusters (average)	6.39	4.97	6

Table 3. Generated clusters in each language (English, French) and with their Union.

Fig. 3. Evaluation against the reference data with Precision, Recall and F-measure.



(a) Average: English 64.7%, French 70%, Union 69.5%

(b) Average: English 9.5%, French 8.5%, Union 12%

(c) Average: English 16.5, French 15, Union 20.4

4 Conclusion and Perspectives

We proposed an experiment on the exploitation of linguistic data in two languages, English and French, for the generation of semantic relations between the MedDRA terms. In this way, we can detect semantically close MedDRA terms and analyse the complementarity between the results provided in each processed language. Several analyses performed point out that two languages are better than one: we obtain more complete results and the global performance of the approach is improved when the union of the two languages is done. Each language contributes almost equally to the results. Only a small set of the synonymy relations is common to the two languages, while an important number of hierarchical relations are generated in the two languages. The hierarchical relations bring the majority of the results. We have several perspectives to this work: (1) enrich the input resources with associative relations acquired with distributional methods on corpora and terminologies; (2) exploit the compositionality-based method not only with input synonymy resources but also with input hierarchical and associative relations; (3) explore corpora and apply other methods for the automatic detection of the semantic relations between the MedDRA terms; (4) combine these results with other automatic methods [22].

References

1. Willett, T.: A cross-linguistic survey of the grammaticalization of evidentiality. Studie in Language **12** (1988) 51–97

2. Palmer, F.: Mood and Modality. Cambridge University Press, Cambridge (2001)
3. Cartoni, B., Namer, F.: Linguistique contrastive et morphologie : les noms en -iste dans une approche onomasiologique. In: CMLF. (2012) 1245–1259
4. Grabar, N., Krivine, S., Jaulent, M.C.: Classification of health webpages as expert and non expert with a reduced set of cross-language features. In: AMIA 2007, Chicago, USA (2007) 284–8
5. Banea, C., Mihalcea, R., Wiebe, J.: Multilingual sentiment and subjectivity. In: Multilingual Natural Language Processing, Prentice Hall (2011)
6. Brown, E., Wood, L., Wood, S.: The medical dictionary for regulatory activities (MedDRA). Drug Saf. **20**(2) (1999) 109–17
7. Pearson, R., Hauben, M., Goldsmith, D., Gould, A., Madigan, D., O'Hara, D., Reisinger, S., Hochberg, A.: Influence of the MedDRA hierarchy on pharmacovigilance data mining results. Int J Med Inform **78**(12) (2009) 97–103
8. Bousquet, C., Henegar, C., Lillo-Le Louët, A., Degoulet, P., Jaulent, M.: Implementation of automated signal generation in pharmacovigilance using a knowledge-based approach. Int J Med Inform **74**(7-8) (2005) 563–71
9. NLM: UMLS Knowledge Sources Manual. National Library of Medicine, Bethesda, Maryland. (2011) www.nlm.nih.gov/research/umls/.
10. Grabar, N., Hamon, T.: Exploitation of linguistic indicators for automatic weighting of synonyms induced within three biomedical terminologies. In: MEDINFO 2010. (2010) 1015–9
11. Grabar, N., Varoutas, P.C., Rizand, P., Livartowski, A., Hamon, T.: Automatic acquisition of synonym ressources and assessment of their impact on the enhanced search in EHRs. Methods of Information in Medicine **48**(2) (2009) 149–154 PMID 19283312.
12. Fellbaum, C.: A semantic network of english: the mother of all WordNets. Computers and Humanities. EuroWordNet: a multilingual database with lexical semantic network **32**(2-3) (1998) 209–220
13. Robert, L.: Le petit Robert. Le Robert, Paris (1990)
14. Tsuruoka, Y., Tateishi, Y., Kim, J., Ohta, T., McNaught, J., Ananiadou, S., Tsujii, J.: Developing a robust part-of-speech tagger for biomedical text. LNCS **3746** (2005) 382–392
15. Schmid, H.: Probabilistic part-of-speech tagging using decision trees. In: ICNMLP, Manchester, UK (1994) 44–49
16. Aubin, S., Hamon, T.: Improving term extraction with terminological resources. In: FinTAL 2006. Number 4139 in LNAI, Springer (2006) 380–387
17. Jacquemin, C.: A symbolic and surgical acquisition of terms through variation. In Wermter, S., Riloff, E., Scheler, G., eds.: Connectionist, Statistical and Symbolic Approaches to Learning for Natural Language Processing, Springer (1996) 425–438
18. Partee, B. In: Compositionality. F Landman and F Veltman (1984)
19. Hamon, T., Nazarenko, A.: Detection of synonymy links between terms: experiment and results. In: Recent Advances in Computational Terminology. John Benjamins (2001) 185–208
20. Kleiber, G., Tamba, I.: L'hyperonymie revisitée : inclusion et hiérarchie. Langages **98** (juin 1990) 7–32
21. Jaulent, M., Alecu, I.: Evaluation of an ontological resource for pharmacovigilance. In: Stud Health Technol Inform. (2009) 522–6
22. Dupuch, M., Bousquet, C., Grabar, N.: Automatic creation and refinement of the clusters of pharmacovigilance terms. In: ACM IHI. (2012) 181–190