Extracting correspondences between terminologies for an easier access to biomedical information

Adila Merabti✉, Lina F Soualmia, Stéfan J Darmoni
TIBS LITIS laboratory EA 4108, Rouen University Hospital, France

Motivation and Objectives

Biomedical terminologies play important roles in clinical data capture, annotation, reporting, information integration, indexing and retrieval. More particularly, genomic terminologies and ontologies are very useful for indexing genomic information. Several sources of information and terminologies have already been developed. For instance, the Gene Ontology (GO, http://www.geneontology.org/, last accessed on July 17, 2012), which is a controlled vocabulary widely used for the annotation of gene products; the Human Phenotype Ontology (HPO, http://www.human-phenotype-ontology.org/, last accessed on July 17, 2012) in which terms describe phenotypic abnormalities encountered in human disease, such as "atrial septal defect"; and ORPHANET, http://www.orpha.net/consor/www/cgi-bin/index.php?lng=FR, last accessed on July 17, 2012) the portal for rare diseases and orphan drugs. These knowledge sources have mostly different formats and purposes. For example, ORPHANET is a rare disease database whereas HPO is an ontology which supports the description of phenotypic information. Faced with this reality and the need to allow cooperation between various health actors and their related health information systems, it appeared necessary to link these terminologies by developing a semantic repository to integrate them. The most known repository is the Unified Medical Language System (UMLS) (Lindberg et al., 1993). Several works were based on the UMLS to align terminologies in French (Merabti et al., 2012) and in English (Bodenreider et al., 1998; Milicic Brandt et al., 2011; Mougin et al., 2011). However, HPO and ORPHANET are not yet included in the UMLS. Thus, another solution is to find correspondences between these terminologies in French and in English using automatic methods. In (Merabti et al., 2012) we have proposed a lexical method to map biomedical terminologies either included or not into the UMLS. Nevertheless, these methods remain very dependent on the terminologies languages since they used NLP tools such as stemming or normalization. We propose in this study a string-based method to find correspondences between a subset of terminologies for an easier access to biomedical information. It is based on the combination of several string metrics and it is neither based on the UMLS, nor language dependent. Mixed with lexical or conceptual approaches developed in previous studies (Merabti et al., 2012), it could improve the number of correspondences between terminologies with a high precision. Semantic methods are also an envisaged issue to complete this study.

Methods

To map biomedical terminologies, we used string matching methods where concept names, terms and their labels are considered as sequences of characters. A string distance is determined to compute a similarity degree. Some of these methods can skip the order of characters. In this paper, the union of three metrics was used (i) Dice (Dice, 1945), (ii) Levenshtein (Levenshtein, 1965) and (iii) Stoilos (Stoilos et al., 2005).

The Dice's coefficient calculates the ratio between the number of bigrams of characters in-common to both the strings x and y and the total number of bigrams for two strings defined by the following equation where nb-big(x) is the number of bigrams of x:

$$Dice(x,y) = \frac{2 \times number\ of\ common\ bigrams}{nb\text{-}big(x) + nb\text{-}big(y)}$$

The Levenshtein distance between two strings x and y is defined as the minimum number of elementary operations that is required to pass from a string x to a string y. There are three possible transactions: replacing a character with another, deleting a character and adding a character. This measure takes its values in the interval [0, ∞ [. The Normalized Levenshtein (Yujian and Bo, 2007) (LevNorm) in the range [0, 1] is obtained by dividing the distance of Levenshtein Lev(x, y) by the size of the longest string and it is defined by:

$$LevNorm(x,y) = 1 - \frac{Lev(x,y)}{Max(|x|,|y|)}$$

LevNorm (x,y) is element of [0,1] as Lev(x,y) < Max(|x|,|y|). |x| is the length of the string x.

The Stoilos distance has been specifically developed for strings that are labels of concepts in ontologies. It is based on the idea that the similarity between two entities is related to their commonalities as well as their differences. Thus, the similarity should be a function of both these features. It is defined by:

$$Sim(x,y) = Comm(x,y) - Diff(x,y) + winkler(x,y)$$

Where Comm(x,y) stands for the commonality between the strings x and y, Diff(x,y) for the difference between x and y, and Winkler(x,y) for the improvement of the result using the method introduced by Winkler in (Winkler, 1999). The function of commonality is determined by the substring function. The biggest common substring between two strings (MaxComSubString) is computed. This process is further extended by removing the common substring and by searching again for the next biggest substring until none can be identified. The function of commonality is given by the equation:

$$Comm(x,y) = \frac{2 \times \sum_i |Max\ Com\ Sub\ String_i|}{|x| + |y|}$$

The function of Difference is defined in the following equation where p is element of [0, ∞ [(usually p= 0.6), |ux| and |uy| represent the length of the unmatched substring from the strings x and y scaled respectively by their length:

$$Diff(x,y) = \frac{|u_x| \times |u_y|}{p + (1-p) \times (|u_x| + |u_y| - |u_x| \times |u_y|)}$$

The Winkler parameter Winkler(x,y) is defined by the equation:

$$Winkler(x,y) = L \times P \times (1 - Comm(x,y))$$

where L is the length of common prefix between the strings x and y at the start of the string up to a maximum of 4 characters and P is a constant scaling factor for how much the score is adjusted upwards for having common prefixes. The standard value for this constant in Winkler's work is P=0.1. To evaluate the correspondences between the terminologies found using the proposed method we have calculated the precision on a sample set evaluated manually and defined as:

$$Precision = \frac{|\{Correct\ correspondences\}|}{|\{total\ correspondences\}|}$$

Results and Discussion

In this study we presented a combination of tree string matching methods to align several biomedical terminologies. The results showed that combining these methods on general terminologies such as MeSH and SNOMED provided more correspondences than only one method and with good results (with a precision>99%). Aligning genomic terminologies provided also good results with high precision. However, we evaluated

	Dice	Levenshtein	Stoilos	Combination
MeSH with SNOMED INT	NB _ align=75,176 P=99.82 % CI95%=[99.79-99.85]	NB _ align=64,657 P=99.80% CI95%=[99.77-99.83]	NB _ align=133,419 P=99.75% CI95%=[99.72-99.78]	NB _ align=156,877 P=99.78% CI95%=[99.76-99.80]
HPO with GO (EN)	NB _ align=161	NB _ align=49	NB _ align=207	NB _ align=291
HPO with GO (FR)	NB _ align=10 P=75.00%	NB _ align=7 P=83.00%	NB _ align=9 P=80.00%	NB _ align=11 P=72.22%
HPO with ORPHANET (EN)	NB _ align=2,593	NB _ align=1,506	NB _ align=3,718	NB _ align=4,237
HPO with ORPHANET (FR)	NB _ align=3,506 P=97.18% CI95%=[96.63-97.73]	NB _ align=2,246 P=94.14% CI95%=[93.17-95.11]	NB _ align=5,405 P=94.87% CI95%=[94.28-95.46]	NB _ align=6,040 P=96,49% CI95%=[96.03-96.95]

Table 1: Total number of correspondences (NB _ align) with a threshold of 0.8 and their associated precision (P%) according to each method. Only the correspondences in French were evaluated. We evaluated a sample of 100 correspondences.

here only "exact" correspondences and rated them as "correct" or "not correct". Indeed, correspondences such as "broader–narrower" or "sibling" relations between terms were not considered. For example, when a correspondence is founded between two terms which one string is included in another one in most cases it is more general than the second, and a "broader-narrower" correspondence could exist (for example, correspondence between "insuffisance surrenale" term (Adrenal insufficiency) and all the terms such as "insuffisance surrenale aigue" (Acute Adrenal insufficiency), "insuffisance surrenale primaire" (Primary adrenal insufficiency)). These preliminary good results encouraged us to apply the combination of these string matching methods on other health terminologies. The correspondences found between two terminologies in their French version may be projected on their versions in other languages. As perspectives of this study, these methods will be completed with normalization techniques and the validation of the correspondences, manual here, will be done according to the UMLS semantic types for the terminologies included in it such as in (Mougin et al, 2011).

References

1. Bodenreider O, Nelson SJ, et al. (1998) Beyond synonymy: exploiting the UMLS semantics in mapping vocabularies. In Proc. AMIA Symp. 1998, pp.815–819.
2. Dice LR (1945). Measures of the amount of ecologic association between species. Ecology 26, pp.297–302.
3. Levenshtein VI (1965) Binary codes capable of correcting deletions, insertions and reversals. Soviet Physics Dokl.10, pp.707–10.
4. Lindberg DA, Humphreys BL, et al. (1993) The Unified Medical Language System, Methods Inf Med 32(4): 281–291.
5. Merabti T, Soualmia LF, et al. (2012) Aligning Biomedical Terminologies in French: Towards Semantic Interoperability in Medical Applications. In Book Medical informatics, InTech, pp.41–68.
6. Milicic Brandt M, Rath A, et al. (2011) Mapping Orphanet terminology to UMLS. In Proc. AIME, LNAI 6747, pp.194–203.
7. Mougin F, Dupuch M, et al. (2011) Improving the mapping between MedDRA and SNOMED CT. In Proc. AIME. LNAI 6747, pp. 220-224.
8. Stoilos G, Stamou G, et al. (2005) A string Metric for Ontology Alignment. In Proc. ISWC, pp.624–37.
9. Winkler W (1999) The state record linkage and current research problems. Technical report: Statistics of Income Division, Internal Revenue Service Publication.
10. Yujian L, Bo L (2007) A normalized Levenshtein distance metric. IEEE Transactions on Pattern Analysis and Machine Intelligence, 29(6):1091–1095.