A General Method for Sifting Linguistic Knowledge from Structured Terminologies

N. Grabar, M.Sc., P. Zweigenbaum, Ph.D.

DIAM — Service d'Informatique Médicale, DSI, Assistance Publique – Paris Hospitals &
Département de Biomathématiques, Université Paris 6, Paris, France
`{ngr,pz}@biomath.jussieu.fr http://www.biomath.jussieu.fr/`

Morphological knowledge is useful for medical language processing, information retrieval and terminology or ontology development. We show how a large volume of morphological associations between words can be learnt from existing medical terminologies by taking advantage of the semantic relations already encoded between terms in these terminologies: synonymy, hierarchy and transversal relations. The method proposed relies on no a priori linguistic knowledge. Since it can work with different relations between terms, it can be applied to any structured terminology. Tested on SNOMED *and ICD in French and English, it proves to identify fairly reliable morphological relations (precision $>$ 90%) with a good coverage (over 88% compared to the UMLS lexical variant generation program). For English words with a stem longer than 3 characters, recall reaches 98.8% for inflection and 94.7% for derivation.*

INTRODUCTION

Morphological knowledge has long been acknowledged as an important area of medical language processing, medical information indexing and ontology development[1,2,3,4]. *Inflection* (the various forms of a same word according to grammatical features, *e.g.*, plural in *artery*, *arteries*), *derivation* (adding affixes to a base word form to produce new words, *e.g.*, *infection*, *infectious*) and *compounding* (combining several radicals to obtain complex words, *e.g.*, *hyper-* + *calc-* + *-emia* yields *hypercalcemia*) produce morphologically and semantically related word forms. Being able to uncover these relations is a need for language-based medical information processing tasks, for instance terminology servers.

Although large medical terminologies such as the International Classification of Diseases (ICD) and Medical Subject Headings exist in many languages, morphological resources for medical language processing are scarce — with the exception of that included in the UMLS Specialist Lexicon, directly exploitable with the Lexical Variant Generation (`lvg`) tool[5] — and their manual production for each language is a labor-intensive enterprise. For instance, project MorTAL[6] has just started the construction of a morphological knowledge base for general French derivation.

We have shown in previous work[7] that when terminologies contain synonym terms (as is the case, *e.g.*, with the Systematised Nomenclature of Medicine, SNOMED International), algorithms can take advantage of these synonyms to learn morphological knowledge about their words, without any initial knowledge of the input language, and to generalize this knowledge to other words of the same language. Related methods have been independently proposed to collect derivational knowledge from a (non-derivational) electronic lexicon[8] or to infer the base or stem of an unknown word given examples from a derivational lexicon[9]. However, requiring the presence of synonym terms results in two limitations. One is that terminologies with no synonymy relation are excluded from this method. The other is that terminologies with both the synonymy relation and other relations may not be exploited in full.

We test here a generalization of this method to any semantic relation between terms. The additional relations we examine are (i) the hierarchical relations that provide the basic structure of most medical terminologies ("is-a", "part-of"), and (ii) transversal relations across this basic structure, such as the cross-reference relations that exist in SNOMED.

Taking as the baseline the results obtained with synonyms[7] on SNOMED and its French Microglossary for Pathology, we examine whether we can improve over them on the one hand with other relations in SNOMED, and on the other hand with the ICD (analytical volume) hierarchical relations. For French, we observe that even though SNOMED Microglossary synonym terms are a very rich source for learning morphological word variation, the other SNOMED relations do provide additional knowledge; and that the ICD hierarchy is also a valuable source of precise morphological knowledge, although in slightly lesser volume. For English, the hierarchy relation in SNOMED provides almost as good results as synonymy for inflection, whereas synonyms are unbeaten for derivational knowledge. ICD (9 or 10) does not yield as much knowledge, which is not surprising since ICD-9-CM is included in SNOMED, and because of the much larger size of SNOMED. We finally discuss further lines of improvement of the current method.

MATERIAL

We use two kinds of input material: a thesaurus with some internal structure and a list of word forms. By "internal structure", we mean that the thesaurus is not a "flat" list of terms, but includes relations between some terms. The relations we consider are (see examples (1) below) *synonymy* (two terms express the same "concept"), *hierarchy* (two terms are associated with concepts where one is broader than the other) and *cross-reference* (two terms are associated with concepts where one "references" the other). In this paper, we report on experiments with the French version of the SNOMED Microglossary for Pathology[10] (henceforth μG) and of the ICD-10 main terms (analytical volume); and with the English versions of SNOMED V3.5, ICD-9-CM and ICD-10 as obtained from the 1999 UMLS Metathesaurus. The latter were extracted by selecting (from files MRSO + MRCON) the strings (SUI) with source vocabulary (SAB) SNMI98, ICD99 and ICD10 respectively. A reference list of word forms was obtained by collecting all word forms in the SNOMED and ICD: for French, the μG and ICD-10, for English the full SNOMED and ICD-9-CM. For evaluation, we ran the UMLS `lvg` tool with options (i) `lvg -m -fi` to produce inflections and (ii) `lvg -m -fRf` for derivations.

METHODS

Our basic method is divided in two steps. The first step, starting from semantically related terms, identifies an initial list of pairs of (hypothetically) morphologically related word forms; the second step induces morphological rules from this initial sample and applies them to the full reference list of word forms. We extend here this method to take into account not only synonym terms[7], but also terms related through other relations: hierarchical and cross-reference relations. We explain here how the method is generalized, and how we compare the results obtained with different relations or different input terminologies.

Production of series of semantically related terms

We first collect pairs of semantically related terms in the input thesaurus. For synonyms, for each concept with n terms ($n > 1$), we take the $n(n-1)/2$ pairs of synonym terms built from these terms. For hierarchical relations, for each concept with descendants, we take the pairs of terms where one term expresses the concept and the other term expresses a descendant. There are $n \times m$ pairs for a concept and a descendant if n synonym terms are associated with the concept and m with the descendant. The same goes for each concept with cross-reference relations to other concepts.

Synonym terms were obtained from SNOMED (and μG). Terms in hierarchical relations were computed from SNOMED and from ICD based on their alphanumerical codes (a more complete set of relations would be obtained by using the explicit hierarchical relations present in the UMLS Metathesaurus). Cross-reference relations were found in the table of the μG^{10}. As a reviewer pointed out, yet another source could be the dagger/asterix relations in ICD.

Alignment of morphologically linked word forms

The general principle is that if (i) two terms T_1 and T_2 are semantically related and (ii) they include word forms $w_1^i \in T_1$ and $w_2^j \in T_2$ that are morphologically similar, there is a high chance that these two word forms are morphologically related, *i.e.*, that they are built from a common stem. For instance, we identify the following pairs of word forms:

(1) <u>*Synonymy:*</u> F-C0000*: "immunity, NOS"; "immune state, NOS"* → *{immunity, immune}*
<u>*Hierarchy:*</u> F-C0000*: "immune state, NOS"* > F-C1000*: "immunogen, NOS"*
→ *{immune, immunogen}*
<u>*Cross-reference:*</u> D2-01100*: "sinusite, SAI"* ↦ T-22000*: "sinus paranasal, SAI"*
→ *{sinusite, sinus}*

Our specific algorithm requires that two such word forms share an initial substring of length $l \geq \lambda$. The parameter λ is set to 4 in this series of experiments. The set of different words pairs obtained at this step constitutes initial examples of morphologically related word forms.

Generalisation to reference word list

We then induce morphological rules from this initial set of examples. For each word pair, we hypothesize a rule that takes into account the minimal difference between the two word forms: their differing final substrings. For instance,

(2) *{immune, immunogen} induces rule e|ogen*

These rules are then applied to the reference list of word forms to identify pairs of word forms that may entertain the same morphological relations as the initial examples. For instance, since *goitre* and *goitrogen* occur in the reference list,

(3) *rule e|ogen identifies {goitre, goitrogen}*

Assuming the transitivity of morphological relatedness, series of word pairs are joined into morphological families (their transitive closure). For instance,

(4) *goitre, goiter, goitrogen, goitrogens*

Ideally, all pairs of word forms in such a morphological family are constructed from a common stem through inflection, derivation or compounding operations. Note that because the alignment step is based on common initial substrings, the common stem is necessarily in initial position. The complete set of such word pairs can also be enumerated from a morphological family. It will be useful for the evaluation of recall.

Evaluation of precision and recall

The issue then arises of evaluating the results obtained. A standard measure of quality is *precision*. It was evaluated by a human review of word pairs and morphological families; for English data, because of its size, only samples were reviewed. Precision is defined as the proportion of word pairs (resp. morphological families) found by the algorithm that were considered correct by human judgment. Let us mention that *specificity* is not useful here: since true negatives are several orders of magnitude over false positives, specificity would be extremely close to 1.

Recall measures the quantity of data collected. It could be evaluated for English data by taking as gold standard the pairs of words of the reference word list where one is an inflected (resp. derived) form of the other, as identified by `lvg` (see Material above). It is defined as the proportion of word pairs identified by `lvg` that were also identified by the above algorithm[7]. We compute here three measures of recall: (i) WP-recall is computed with the word pairs directly identified by the morphological rules; (ii) F-recall is computed with the full word pairs enumerated by transitivity from the morphological families; and (iii) λ-recall: since our method, by construction, can only produce word pairs with a common prefix longer than λ characters, we also compute the recall with a gold standard with the same restriction; we actually measure λF-recall, *i.e.*, F-recall with a λ-restricted gold standard.

Comparison of morphological families

`lvg` only covers part of the kind of knowledge produced by our method: compound forms are not included in the above gold standard. Therefore, we also examine the "productivity" of a given method: a normalized amount of word pairs generated by our algorithm. Combined with the precision, it can give an idea of the global amount of correct knowledge extracted. We measure the (minimal) number of word pairs that would be necessary and sufficient to obtain, by transitivity, the final morphological families[11]. We use the same principle to examine the additional knowledge provided by an experiment (*i.e.,* a given terminology and semantic relation) over another experiment: we compute the minimal number of word pairs that should be added to the first set of morphological families to cover all the word pairs derivable from the second set of morphological families. This gives a measure of the added value of a given method over another. To take into account the relative sizes of the input terminologies, we try to normalize it with the number of input term pairs that have been examined: the "relative productivity" is the number of final word pairs divided by the number of initial term pairs.

RESULTS

We performed experiments, in French and in English, using different semantic relations between terms and different input terminologies. The reference list was kept constant for each language. For French, we examined the results for the relation of hierarchy (table 1) with the μG, keeping both preferred and synonyms terms (μGh+) or keeping only preferred terms (μGh–), and with ICD-10; the synonyms (μGs) and cross-reference (μGr) methods were also run on the μG. Precision, shown for final word pairs and families, is excellent (97∼98%) for all methods. The bottom of the table displays the number of minimal word pairs (productivity) for each method and the new word pairs from one method to the other. μGs is the most productive (3,486 minimal word pairs, involving 5,164 words out of 8,874, with 0.67 word pairs per term pair). The other methods nevertheless do identify some additional word pairs: +270 (+7.7%) with μGh+, +183 (+5.2%) with μGr+ and +199 (+5.7%) with ICD-10.

For English (table 2), we compared the knowledge obtained from the full SNOMED (hierarchy; with and without synonym terms) with that obtained with ICD-9-CM and ICD-10 (hierarchy). The base experiment with SNOMED synonyms obtains a precision of 92.5±1.3 % (not in the table: final word pairs, 1/15th sample) and 91.9±1.5 % (final families, 1/5th sample; confidence intervals were computed with α=0.05). The hierarchy experiment Sh+ obtains the highest productivity (17,130 minimal word pairs). The "relative productivity" is however maximal for ICD-10 (1.79 word pairs per term pair) and minimal for Sh+ and Sh–

Table 1: Comparing results for the hierarchy relation in French ICD-10, in SNOMED μG with (μGh+) and w/o (μGh–) synonyms, and for the cross-reference (μGr+) and synonymy (μGs) relations in μG (λ=4).

	ICD-10	μGh–	μGh+	μGr+	μGs
Terms	10,800	9,098	12,555	12,555	12,555
Series	1,554	1,949	1,949	2,082	2,344
Term pairs	9,412	24,261	66,318	13,089	5,204
Word pairs	3,571	1,562	3,831	2,316	1,572
Unique w.p.	776	554	1,121	867	1,086
Families	485	327	436	334	623
Rules	273	295	612	447	567
Reference list	8,874	8,874	8,874	8,874	8,874
Word pairs	3,389	3,566	4,735	4,307	4,573
Precision	98.5%	98.2%	97.7%	na	98.3%
Families	1,510	1,520	1,606	1,625	1,678
Precision	98.1%	97.6%	97.3%	na	97.3%
Words	4,139	4,451	5,006	4,785	5,164
W. per family	2.74	2.93	3.12	2.94	3.08
Minimal # w.p.	2,629	2,931	3,400	3,160	3,486
Relative prod.	0.28	0.12	0.05	0.24	0.67
New % ICD-10	–	+619	+936	+728	+1,056
New % μGh–	+317	–	+470	+510	+762
New % μGh+	+165	0	–	+207	+356
New % μGr+	+197	+281	+447	–	+509
New % μGs	+199	+207	+270	+183	–

(1 word pair per 100 term pairs).

WP-recall for inflection word pairs is stable around 85% for the three SNOMED experiments. ICD-10 and ICD-99 work on a much smaller input volume than SNOMED; the output for ICD-10 is much below the rest; although low, the output for ICD-99 is sufficient to obtain a good inflection recall. Taking into account the transitive closure of morphologically related word pairs (F-recall) consistently increases recall by a few percents. Focussing on word pairs with a common prefix $\geq \lambda$=4 (λF-recall), either method on SNOMED obtains a very high recall. Combining the results of the hierarchy and synonyms method on SNOMED (Sh+ $\cup$ Ss) again increases λF-recall to 98.9%, with only 27 word pairs not found out of 2417.

Derivation recall is more contrasted, and the synonyms method again obtains the best figure (75.8%). Transitivity (F-recall) adds more than 10% of word pairs, so that 86.5% of the derivational word pairs found by lvg can be discovered, or 92.9% of those with a common prefix $\geq \lambda$ (λF-recall). The combination Sh+ $\cup$ Ss produces 19,120 minimal word pairs with an F-recall of 88.2% (+1.7%) and a λF-recall of 94.7% (+2.2%: 148 word pairs not found out of 2769).

Table 2: Comparing results for the hierarchy relation in English ICD-10, ICD-9-CM (ICD-99) and SNOMED with (Sh+) and w/o (Sh–) synonyms, and for the synonymy relation in SNOMED (Ss) (λ=4).

	ICD-10	ICD-99	Sh–	Sh+	Ss
Terms	4,911	17,660	93,750	128,855	128,855
Series	557	2,933	18,038	18,215	26,295
Term pairs	2,170	34,996	1,044,986	1,754,449	64,718
Word pairs	381	5,923	72,846	95,208	15,549
Unique w.p.	149	705	5,522	8,916	6,556
Families	116	379	1,867	2,353	3,188
Rules	89	419	3,473	5,529	3,039
Reference list	49,627	49,627	49,627	49,627	49,627
Word pairs	4,062	8,782	20,885	25,970	22,372
Families	3,173	4,391	6,000	6,086	6,550
Words	7,068	11,931	21,143	23,216	22,794
W. per family	2.23	2.72	3.52	3.81	3.48
Minimal # w.p.	3,895	7,540	15,143	17,130	16,244
Relative prod.	1.79	0.56	0.01	0.01	0.25
New % ICD-10	–	+4,104	+11,300	+13,272	+12,394
New % ICD-99	+459	–	+7,737	+9,685	+8,911
New % Sh–	+52	+134	–	+1,988	+3,460
New % Sh+	+37	+95	0	–	+1,991
New % Ss	+45	+207	+2,359	+2,877	–
Infl. WP-recall	51.7%	80.3%	84.8%	85.0%	85.3%
Infl. F-recall	52.0%	84.4%	87.4%	87.8%	88.0%
Infl. λF-recall	58.0%	94.2%	97.6%	98.0%	98.2%
Deriv. WP-recall	16.6%	46.9%	68.7%	72.1%	75.8%
Deriv. F-recall	17.4%	54.7%	80.0%	83.3%	86.5%
Deriv. λF-recall	18.7%	58.7%	85.9%	89.4%	92.9%

Examples of word pairs found by Sh+ (SNOMED, hierarchy) and not by Ss (SNOMED, synonyms) include {*goitre*, *goitrogen*}, {*immune*, *immunogen*} (rule *e*|*ogen*) and {*therapy*, *therapist*} (rule *y*|*ist*). Note that these rules also identify {*triple*, *triplogen*}, {*chrome*, *chromogen*}, and {*rhine*, *rhinogen*}, not found by lvg, and {*radiotherapy*, *radiotherapist*} (found by lvg) and {*physiotherapy*, *physiotherapist*} (not found by lvg).

DISCUSSION AND PERSPECTIVES

The different terminologies and relations examined may identify different morphological relations (rules), so that their combination yields a higher recall as measured against lvg. This was observed with the {*goitre*, *goitrogen*} and {*immune*, *immunogen*} example. Similar effects are found in French, where, *e.g.*, the hierarchy relation in μGh+ finds a series of 36 *-ite* (*-itis*) words not identified by μGs: {*péricarde*, *péricardite*}, {*valvule*, *valvulite*}, {*méat*, *méatite*}, etc., because the rules *e*|*ite* (learnt on {*vagin*, *vaginite*}), ϵ|*ite* (learnt on {*sinus*, *sinusite*}), etc., were not identified. The figures for relative productivity, as defined above, are contrasted. It is probably the case however that as more term pairs are examined, the rate of new word pairs identified decreases, so that smaller input sets have a higher marginal productivity than larger input sets. A deeper examination of this dependence is necessary.

Using `lvg` as the gold standard for evaluation was very convenient, but probably not an ideal setting. On the one hand, it does not handle compound words; on the other hand, a derivation lexicon is necessarily incomplete, since "constructed words" constitute an open set. In `lvg`, morphological knowledge is encoded in part as rules and in part as "facts": the part encoded as facts is liable to contain omissions. The word pairs and families produced by our algorithm may constitute a reservoir of knowledge for lexicographers wishing to extend these rules and facts. Besides, as a reviewer pointed out, an evaluation of `lvg` itself would be useful.

All word pairs identified by our morphological rules are currently examined with the same attention; it might be useful to propose some confidence rating with each rule and with each word pair it identifies, *e.g.*, based on the number of examples on which they are grounded. This kind of rating is often used in similar work[8,9] and more generally in data mining[12]. It might also help process some shorter words without loosing too much precision. Besides, the transitivity hypothesis may not always be correct because of ambiguous words (*homographs*) and may propagate erroneous word pairs. Complementary methods[8] can help subclassify overmerged families such as *ischial*, *ischium* . . . *ischemia*, *ischemic* (Sh+).

The generalization of this algorithm to different kinds of term relations present in a terminology makes it a very powerful method, applicable to any structured terminology. The choice of ICD-10 after SNOMED was reasonable for French since only the μG was available, its size is comparable to that of ICD-10, and ICD-10 is not included in SNOMED. For English, there is a too large intersection between the two; the method should therefore be tested on more independent terminologies (*e.g.*, on the MeSH). Working on the full UMLS Metathesaurus, with its different kinds of relations, may also be an interesting path to explore. The expectable explosion of data would probably strengthen the need for a confidence rating.

ACKNOWLEDGMENTS

We wish to thank Dr. RA Côté for graciously providing us a pre-commercial copy of the French version of the SNOMED Microglossary for Pathology, and the NLM for the UMLS resources.

References

1. Pacak MG, Norton LM, and Dunham GS. Morphosemantic analysis of -ITIS forms in medical language. *Methods Inf Med* 1980;19:99–105.
2. Wolff S. Automatic coding of medical vocabulary. In: Sager N, Friedman C, and Lyman MS, eds, *Medical Information Processing - Computer Management of Narrative Data*. Addison Wesley, Reading Mass, 1987:145–62.
3. Lovis C, Baud R, Rassinoux AM, Michel PA, and Scherrer JR. Medical dictionaries for patient encoding systems: a methodology. *Artif Intell Med* 1998;14:201–14.
4. Webber B, Markert K, Hardiker N, and Rauch B. Towards consistent, minimal terminologies. In: Chute CG, ed, Proc Conference on Natural Language Processing and Medical Concept Representation, Phoenix, Az. IMIA WG6, 1999:131–41.
5. McCray AT, Srinivasan S, and Browne AC. Lexical methods for managing variation in biomedical terminologies. In: Proc Eighteenth Annu Symp Comput Appl Med Care, Washington. Mc Graw Hill, 1994:235–9.
6. Dal G, Namer F, and Hathout N. Construire un lexique dérivationnel : théorie et réalisations. In: Amsili P, ed, Actes de TALN 1999, Cargèse. July 1999.
7. Grabar N and Zweigenbaum P. Language-independent automatic acquisition of morphological knowledge from synonym pairs. *J Am Med Inform Assoc* 1999;6(suppl):77–81.
8. Gaussier E. Unsupervised learning of derivational morphology from inflectional lexicons. In: Kehler A and Stolcke A, eds, ACL workshop on Unsupervised Methods in Natural Language Learning, College Park, Md. June 1999.
9. Pirrelli V and Yvon F. The hidden dimension: a paradigmatic view of data-driven NLP. *J Expt Theor Artif Intell* 1999;11:391–408.
10. Côté RA. Répertoire d'anatomopathologie de la SNOMED internationale, v3.4. Université de Sherbrooke, Sherbrooke, Québec, 1996.
11. Zweigenbaum P and Grabar N. Liens morphologiques et structuration de terminologie. In: IC 2000 : Ingénierie des connaissances, 2000:325–34.
12. Agrawal R and Srikant R. Fast algorithms for mining association rules. In: Proc. of the 20th Int'l Conf. on Very Large Databases, Santiago. 1994.