Language-independent Automatic Acquisition of Morphological Knowledge from Synonym Pairs

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/

Medical words exhibit a rich and productive morphology. Beyond simple inflection, derivation and composition are a common way to form new words. Morphological knowledge is therefore very important for any medical language processing application. Whereas rich morphological resources are available for the English medical language with the UMLS Specialist Lexicon, no such resources are publicly available for French or most other languages. We propose a simple and powerful method to help acquire automatically such knowledge. This method takes advantage of the synonym terms present in medical terminologies. In a bootstrapping step, it detects morphologically related words from which it learns "derivation rules". In an expansion step, it then applies these rules to the whole vocabulary available. Our goal is to acquire data for French and other languages for which they are not available. However, to evaluate the efficiency of the method, we tested it on English in a setting which is close to that prevailing for French, and we confronted its results to those obtained with the Specialist lexical variant generation tool.

INTRODUCTION

Medical words exhibit a rich and productive morphology. In French or English, as well as in many other European languages, they are often formed using Greek or Latin roots and affixes. The decomposition of a word into its component *morphemes* is useful to get at its elementary meaning units. This is a key to more relevant and more principled semantic processing of medical utterances. In an even simpler way, this allows finer grained indexing of medical texts and terms, and potentially better accuracy for information retrieval and coding assistants[1].

Three types of morphological variations are classically distinguished: (i) inflection (*e.g.*, plural form creation) creates variant forms of the same word; (ii) derivation adds affixes (prefixes or suffixes) around one root (*e.g.*, to obtain the adjectival form of a noun); and (iii) composition combines several roots (and possibly affixes).

Medical morphology has been studied by many researchers for several different languages. Wingert[1] presents the classical models of word morphology and applies these models to an indexing task. Pierre Dujols[2], following a line of work opened by Pacak and colleagues[3], has focussed on a precise model of words ending with the suffix *"-osis"*. Alexa McCray[4] describes the morphological knowledge and procedures available for English in the UMLS Specialist lexicon[5]. Peter Spyns[6] has designed a "guesser" which proposes syntactic information for unknown Dutch words based on their affixes. Christian Lovis[7] has implemented the derivation and composition of French medical roots and affixes as a finite state transducer. The morphological knowledge involved in the above works was, to our knowledge, collected manually: most of this work relies on labor-intensive coding of morphological knowledge, although some tools have been used in some instances to help collect data or organize knowledge[4]. In a more general setting, computational linguistics models such as "two-level morphology"[8] have formed the basis for general descriptions of natural languages, also mainly acquired manually. Nevertheless, whereas rich morphological resources are available for the English medical language with the UMLS Specialist lexicon, no such resources are publicly available for French or most other languages. We therefore propose a simple and powerful method to help acquire automatically such knowledge.

Some researchers have tried to obtain morphological knowledge from the observation of language data. Christian Jacquemin[9] compares the (two-word) terms present in a reference thesaurus with word collocations in a corpus: his hypothesis is that when two contiguous words of the corpus are similar to (share some common prefix with) the words of a reference term, they may be morphologically and semantically related. This method is able to find, for instance, the relation between *"gene expression"* and *"genic expression"*, where *"gene"* and *"genic"* are really related. He applied this method to the Medic corpus and the Pascal thesaurus (from the INIST Institute, France). Jinxi Xu and Bruce Croft[10] worked on a large corpus with no a priori thesaurus: they rely on mutual information statistics to select those morphologically similar words

[0] To appear in *Journal of the American Medical Informatics Association*, 6(suppl), 1999, Proceedings of AMIA Annual Fall Symposium 99, Washington, DC.

that may be morphological variants.

Our method[11] acquires morphological information from a thesaurus with no a priori linguistic knowledge. It takes advantage of the synonym terms present in medical terminologies. In a bootstrapping step, it detects morphologically related words from which it learns "derivation rules". In an expansion step, it then applies these rules to the whole vocabulary available. Noise is kept at a very low level by enforcing constraints at two crucial places in the process: (i) the bootstrapping morphological rules are learnt on words found in synonymous terms and (ii) the acquired pairs of morphologically related words only include word forms that are attested in the available language data.

Our goal is to acquire data for French and other languages for which they are not available. However, to evaluate the method, we tested it on English in a setting close to that prevailing for French, and we confronted its results to those obtained with the UMLS Specialist lexical variant generation (`lvg`) tool.

We first describe the English and French language data, drawn from SNOMED International and ICD-10. We detail the specifics of the acquisition method, and illustrate it on this data. We present the results in both languages. We discuss the advantages and limitations of the method, and draw perspectives for further work.

MATERIAL

SNOMED International, as many other medical terminologies, not only assigns a preferred term to each of its concepts, but also lists synonymous terms for some concepts. These synonyms are flagged with a '02' or '05' class code, while the preferred term is flagged with a '01' class code. Table 1 shows example terms. The English terms were extracted from the English SNOMED International, which includes 128,855 terms, among which 26,312 have both a preferred and synonym terms. Our base material for bootstrapping the acquisition process consists of the 26,312 series of such synonyms. The same method was applied to the French Microglossary for Pathology[12] (the whole SNOMED is not yet available in French), which includes 12,555 terms, among which 2344 have synonym terms.

Table 1: Preferred and synonym SNOMED terms.

Code	Class	French term	English term
F-00470	01	symbiose	symbiosis
F-00470	02	commensalisme	commensalism
F-00470	05	symbiotique	symbiotic
F-00470	05	commensal	commensal
T-51110	01	palais dur	hard palate
T-51110	02	voûte palatine	–

Our second sample of language data is a reference list of word forms. In a general setting, the largest possible coverage of medical words should probably be aimed at. Since the ICD-10 classification is widely used in France, we were interested in enhancing our SNOMED word list with ICD-10 words. The resulting English word list contains 49,627 distinct word forms, and the French word list contains 7,490 word forms.

The validation of the output morphological data was performed with the UMLS 1999 Specialist Lexicon `lvg V1.83` tool[5]. `lvg` generates morphologically related English words for an input word. We ran it with options (i) `lvg -m -fi` to produce inflections (*e.g.*, input *"thesaurus"* yields output *"thesauri"*) (ii) `lvg -m -fRf` for derivations (*e.g.*, *"abdomen"* yields *"abdominal"*).

METHODS

Overview

We consider as *morphologically related* a pair of words that are derived from a common root and, therefore, share a more or less large part of their meanings. A pair of such words quasi-universally share a common set of characters. In many Indo-European languages this set of characters is most often a string of contiguous characters, *e.g.*, **symbio** in "**symbio***sis*" / "**symbio***tic*". Very often too, this string is found at the beginning of each word, as in the above example. A simple way to find a clue that two words are potentially morphologically related is to examine their longest common prefix. If this prefix is "large enough", they might belong to the same morphological paradigm.

Such a simple approach might however lead to much noise. For instance the pair of words "**admi***nistrative*" / "**admi***ssion*", found in our word list, share a four-character prefix but they are not obtained by morphological rules operating from a common root. Even with a higher threshold on the minimal length of the longest common prefix, one still finds word pairs such as "**antidia***betic*" / "**antidia***rrhea*" (common prefix length = 7) which are not derived from a common root (although here they do share a common *prefix*).

Bootstrapping step

The idea put forth in this method is to apply this simple approach *in a very specific, favorable context.* This context must be such that it focuses the comparison of word pairs on words that have, by their context of occurrence, a high chance of sharing some meaning. We found such a favorable situation in pairs of synonyms, as are included in medical vocabularies such as SNOMED. Examples such as those in table 1 show that in the restricted context of pairs of synonymous terms, morphologically related word pairs can be found: *e.g.*, “**symbio***sis*” / “**symbio***tic*”, “**commensal**” / “**commensal***ism*”. Such words must be aligned by comparing their longest common prefix. We set experimentally the threshold on minimal common prefix length to 3, *i.e.*, two words are considered related in this context iff they share at least their first three letters.

The advantage of working with pairs of synonymous terms is that given this alignment procedure, there is very little risk of finding morphologically similar but semantically unrelated word pairs. Therefore, the set of word pairs found by this procedure is fairly accurate. In contrast, an algorithm that would consider morphologically related any two words that share a three-character prefix would produce a high level of noise: for instance, the 17 forms in the French corpus that start with *“tro-”* actually come from 11 different roots.

Word pairs organized around the same prefix are then joined into morphological families, for instance: “**aborti***cide*”/“**aborti***ent*”/“**aborti***facient*” “**cardi***aque*” / “**cardi***o*” / “**cardi***omégalie*” / “**cardi***opathie*” / “**cardi***te*”.

The word pairs identified as morphologically related are instances of potentially *more general derivation rules* akin to those of `lvg`[4]. We therefore hypothesize the existence of these rules, and register each of them. Since we do not work here on syntactic categories, we adopt an even simpler notation than `lvg` rules. A morphological rule consists of a pair of suffixes, such as “*sis/tic*”. It allows to transform a word ending in *“-sis”* into another word where the suffix *“-sis”* has been removed and replaced with the suffix *“-tic”* (*e.g.*, from *“stenosis”* to *“stenotic”*). These rules are symmetric, so that the same rule allows to go from a *“-tic”* word to a *“-sis”* word. Since these rules have been induced from mostly accurate word pairs, they are attested at least in one instance. The next step of the procedure attempts to apply these rules to an additional language sample to identify more related words.

Expansion step

When applying these rules to new words, there is a high risk that many of the resulting word forms simply do not exist in the language studied. This is all the more expectable as some of the rules have one empty suffix. For instance, the rule *“/al”*, which relates inter alia *“ombilic”* to *“ombilical”*, can add the suffix *“-al”* to any word form. Therefore, it is important here again to constrain the application of this simple method. Our rationale is to *relate attested word forms* rather than to try to hypothesize new word forms. We try to apply each rule to each word form in our reference list, and accept the derivation when (i) the word ends with one of the suffixes of the rule, (ii) the derived word is found in the reference list, and (iii) the common prefix is at least 3 characters long. For instance, the rule *“/al”* only succeeded on three word pairs for French: *“médiastin”* / *“médiastinal”*, *“vagin”* / *“vaginal”*, and *“ombilic”* / *“ombilical”*. The new word pairs discovered in this second step extend initial families collected in the first step. For instance, the initial family “**abnormal**” / “**abnormal***ities*” / “**abnormal***ity*” was extended into “**abnorm***al*” / “**abnorm***alis*” / “**abnorm***alities*” / “**abnorm***ality*” / “**abnorm***is*”. Families that share a common word form are joined.

Preparing data for validation

Our method produced pairs of morphologically related words as well as morphological families. To evaluate this output, we focused on the word pairs. We used `lvg` to automatically evaluate the recall of English word pairs for inflection and derivation. To make things comparable, we only considered the resulting word pairs where the two word forms are different and exist in the reference list. The rest of the evaluation was performed by human review.

RESULTS

We implemented our method using `perl` programs and Unix scripts using mainly `sed`, `awk` and `sort`. Table 2 summarizes the results obtained on the English and French data.

Table 2: Morphological knowledge acquired.

Knowledge elements	English	French
Morphological rules	3188	566
Suffixes	2610	453
Initial families	3207	755
Words per family	3.94	3.39
Families after expansion	6424	1304
Words per family	4.84	3.67

Manual review: types of rules

We performed a manual review of the acquired morphological rules (3188 for English and 566 for

French). The above-mentioned three types of morphological variation can be found. It may be useful to make a difference between the rules that preserve meaning, possibly allowing only slight variations and the rules that modify meaning, generally by adding some information.

The inflection rules and a part of the derivation rules are essentially morpho-syntactic variations on the base form: they are formed with grammatical suffixes. They count 257 (12% of the total 2610) suffixes for English and 149 (33% of the total 453) for French. Examples are the “*/s*” plural inflection rule and the “*e/ien*” derivation rule that relates the noun “*oesophag***e**” to the adjective “*oesophag***ien**”.

The compounding rules and the rest of the derivation rules use lexical suffixes (*e.g.*, “*-itis*”) and other roots (*e.g.*,“*-centesis*”), which add specific information. An example derivation is “*bronch***ial**” / “*bronch***itis**”, and an example compounding rule is “*id/blastoma*” as in “*andro***id**” / “*andro***blastoma**”. 2024 (88%) English and 299 (66%) French suffixes belong to this type.

Recall and precision analysis

We used `lvg` data as the reference for the automated computation of **recall** (sensitivity) on the English data. For inflections, 235 `lvg`-generated pairs (8%) were not found by our method, so that its recall is 92%. For instance the pair “*ag***ed**” / “*ag***ing**” (rule “*ed/ing*”) was not found. We acquired this rule, but the common prefix in this case is shorter that 3 characters. Another example “*ben***ds**” / “*ben***t**” (rule “*ds/t*”) involves verbal paradigms, and SNOMED includes very few verbs. For derivational variants, 632 pairs (21%) were not found and recall is 79%. The pair “*adjust*” / “*adjust***able**” is absent in our result, because rule “*/able*” was not acquired, although “*-able*” forms rules with other suffixes.

We also examined the **precision** (specificity) obtained by looking at pairs of words obtained only by our program. `lvg` inflection and derivation produced 5,670 word pairs; our method found a total of 25,740 pairs, among which 4,817 (*i.e.*, 19%) have also been produced by `lvg`. This difference is mainly due to the fact that our method does not separate compounding from inflection and derivation, whereas `lvg` does not currently treat compounding. As we found out, a very large subset of the suffixes (88% for English and 66% for French) involve compounding operations.

A manual analysis on a sample of 206 word pairs produced by our program (one every 250) revealed 33 errors, corresponding to a precision of $84\% \pm 5\%$ ($\alpha = 0.05$). This shows the overall relevance of the morphologically related word pairs obtained. Examples of pairs not found by `lvg` include: (i) composition: “*abdomen***ocentesis**” / “*abdomen*”; (ii) derivation: “*abdomin***alis**” / “*abdomin***al**”; (iii) inflection: “*abdomin***o**” / “*abdomin***al**”. As for the following pair “*ablephar***ia**” / “*ablephar***on**”, the rule is absent from `lvg` files. We should stress the fact however that in many instances, there may be very good reasons for `lvg` not to incorporate specific cases. For instance, rule “*ia/on*” seems to be valid only in the above case.

We performed a manual analysis of the **precision** on English and French **morphological families**. As the examples below show, two types of errors were encountered. (i) The words obtained actually share a *prefix*, that is, they are based on different roots in front of which the same prefix has been added. Nothing in the current method allows to make a difference between such a prefix and an actual root form. (ii) At least one pair of words in the family have no morphological relation at all. The analysis of English data was realised on a sample of 257 families (one every 25). Among these, 35 families (14%) are erroneous (precision = $86\% \pm 4\%, \alpha = 0.05$). A prefix-type error is “**ante***sternal*” / “**ante***thoracic*”, while “**clo***al*” / “**clo***sed*” / “**clo***thes*” correspond to erroneous derivations. Among the 1304 French families, 30 families (5%, this corresponds to 1% of erroneous pairs of words) were considered erroneous (precision = 95% of families, 99% of pairs): for instance, prefix “*auto-*” in “**auto***greffe*” / “**auto***logue*” / “**auto***plastique*”, and erroneous derivation “**chro***me*” / “**chro***nique*”.

DISCUSSION AND PERSPECTIVES

This method relies on very little a priori linguistic knowledge. The only hypotheses made are that (i) a segmentation of a word into base + suffix is relevant; (ii) setting a minimum length for a base sufficiently reduces noise while not bringing much silence (the choice of a 3-character threshold is a parameter of the method, which can be changed if desired). The simple segmentation of words into base and suffix, with a minimal length constraint on the base, limits the kinds of morphological rules that can be identified. We expect other romance and germanic languages to satisfy these hypotheses. In constrast, semitic languages do not verify hypothesis (i); and morphology is simply not relevant for isolating languages such as Vietnamese. We have started to work with Russian, where results are encouraging. It would be interesting to see how this method would fare with agglutinative languages such as Turkish or Finnish.

By comparing the results of our method for English with the `lvg` output, we showed that it obtains a good recall: 92% for inflections and 79% for derivations. Actually, if the transitive closure of rules were applied, much more new pairs could be found. The same level of recall should be attainable for French by starting from a synonym list of comparable size. The precision estimated on a sample of English morphological families is $86 \pm 4\%$, whereas the precision for word pairs is $84 \pm 5\%$. As for French, 95% of the morphological families and 99% of the derived pairs are correct.

A next step will consist in the automatic decomposition of words into minimal, normalized morphemes (roots, and grammatical and lexical affixes) and their division into inflectional, derivational and compounding morphemes. This would be a useful asset for assessing semantic proximity. Distinguishing part of the semantic suffixes might be possible, *e.g.*, by trying to identify known words used as suffixes[7]. Adding to each rule syntactic categories and a list of exceptions, as in `lvg`, would introduce complementary constraints. Besides, this could help to suppress some of the errors related to prefixes in the current method. The next step would be to cope with more complex morphological alternations with deletion, insertion or modification of characters inside morphemes such as *"détruire"* / *"destruction"* or (*e.g.*, *"strangulation"* / *"étranglement"*). More elaborate morphological models[8] would probably prove useful here, as well as methods to acquire morphological descriptions from morphologically related word pairs[13].

Finally, other sources of constraining contexts might be found elsewhere to bootstrap the method, for instance in corpora instead of thesauri [9,10].

CONCLUSION

As this method is entirely automated, we believe it is an economical way to obtain immediately a large volume of mostly reliable morphological data, which can then be refined by human processing. Besides, the expansion step currently uses a rather crude method, and more sophisticated computational linguistics models could bring even better results. Finally, since it does not rely on a priori knowledge of the language, it can be applicable as well to other languages.

ACKNOWLEDGEMENTS

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 making the UMLS available.

References

1. Wingert F, Rothwell D, and Côté RA. Automated indexing into SNOMED and ICD. In: Scherrer JR, Côté RA, and Mandil SH, eds, *Computerised Natural Medical Language Processing for Knowledge Engineering*. North-Holland, Amsterdam, 1989:201–39.
2. Dujols P, Aubas P, Baylon C, and Grémy F. Morphosemantic analysis and translation of medical compound terms. *Methods Inf Med* 1991;30:30–5.
3. Pacak MG, Norton LM, and Dunham GS. Morphosemantic analysis of -ITIS forms in medical language. *Methods Inf Med* 1980;19:99–105.
4. 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.
5. National Library of Medicine. UMLS Knowledge Sources Manual, 1999.
6. Spyns P. A robust category guesser for Dutch medical language. In: Proceedings of ANLP 94 (ACL), 1994:150–5.
7. 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.
8. Koskenniemi K. *Two-level morphology: a general computational model for word-form recognition and production*. PhD thesis, University of Helsinki Department of General Linguistics, Helsinki, 1983.
9. Jacquemin C. Guessing morphology from terms and corpora. In: Actes, 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'97), Philadelphia, PA. 1997:156–67.
10. Xu J and Croft BW. Corpus-based stemming using co-occurrence of word variants. *ACM Transactions on Information Systems* 1998;16(1):61–81.
11. Zweigenbaum P and Grabar N. Automatic acquisition of morphological knowledge for medical language processing. In: Horn W, Shahar Y, Lindberg G, Andreassen S, and Wyatt J, eds, *Artificial Intelligence in Medicine*, LNAI. Springer-Verlag, 1999:416–20. also available online.
12. Côté RA. Répertoire d'anatomopathologie de la SNOMED internationale, v3.4. Université de Sherbrooke, Sherbrooke, Québec, 1996.
13. Theron P and Cloete I. Automatic acquisition of two-level morphological rules. In: ANLP97, Washington, DC. 1997:103–10.