Towards a Multilingual Medical Lexicon

Kornél Markó1, Robert Baud2, Pierre Zweigenbaum3, Lars Borin4, Magnus Merkel5, Stefan Schulz1

1Freiburg University Hospital, Department of Medical Informatics, Freiburg, Germany
2University Hospitals of Geneva, Service of Medical Informatics, Geneva, Switzerland
3Inserm, U729; Assistance Publique – Paris Hospitals, STIM; Inalco, CRIM, Paris, France
4Göteborg University, NLP Section, Department of Swedish, Göteborg, Sweden
5Linköping University, Department of Computer and Information Science, Linköping, Sweden

Abstract

We present results of the collaboration of a multinational team of researchers from (computational) linguistics, medicine, and medical informatics with the goal of building a multilingual medical lexicon with high coverage and complete morpho-syntactic information. Monolingual lexical resources were collected and subsequently mapped between languages using a morpho-semantic term normalization engine, which captures intra- as well as interlingual synonymy relationships on the level of subwords.

Introduction

Lexicons, especially designed for natural language processing purposes, can generally be characterized along several dimensions. Firstly, lexicons can provide different amounts of lexical information, such as part-of-speech, number, gender and case. Secondly, the coverage of a lexicon, which often captures the terminology of a specialized domain, indicates the proportion of words of a (domain specific) text collection, for which lexical information is available. For translation dictionaries, finally, a special attention is drawn on the multilingual dimension.

There is currently no large electronic dictionary in the medical domain which is characterized by a true multilingual dimension, relevant coverage, and substantial lexical information at the same time. Of course, the UMLS Metathesaurus [8] constitutes a widely used multilingual resource with high coverage in the medical domain. However, detailed lexical information is restricted to the English language only.

For non-specialized domains, remarkable effort of developing mono- and multilingual dictionaries has been made. For example, WordNet [5] provides a good coverage for general English. It may be useful for covering lay terminology of medicine [3] or biology [2], for example within a consumer-oriented health information system. The European counterpart, EUROWORDNET [9] tends toward a multilingual system, but with considerably diverse levels of lexical coverage.

Whenever medical terminology has been addressed in the construction of an expressive multilingual dictionary, it lacks convenient coverage or has been developed as a demonstrative prototype [4].

Within the European Network of Excellence “Semantic Interoperability and Data Mining in Biomedicine”, a multinational team of researchers from (computational) linguistics, medicine, and medical informatics, including the authors, organized a series of meetings with the goal of building a multilingual medical lexicon with high coverage and complete lexical information. That lexicon should account for several languages, with at least 300,000 entries.

Multilinguality means at least that the corresponding entries in different languages are connected. Therefore, syntactical as well as semantic criteria have to be developed, or, at least, a consensus of different lexical input providers has to be found.

Of course, monolingual resources exist for different languages, so the first step to merge them is to create a common framework for the integration of lexical entities from different languages, with respect to their intrinsic peculiarities.

Interchanging Lexical Information

The Interchange Format is a specification for exchanging linguistic information entering in the building process of a medical multilingual lexicon [1]. The basic idea is that the exchange of information is performed through the Interchange Format only, and each contributor of lexical resources is converts his or her data into that representation.
Table 1 lists the fields of the Interchange Format. The most important ones are the following:

- **Lng**: The language field determines to which language this particular entry belongs.
- **Id**: The unique identifier of the multilingual lexicon entry is composed by the concatenation of the name of the input provider and a consecutive number.
- **Typ**: The basic entry (B) encodes single words. The subword entry (S) is a marker for parts of words entering in the composition of a compound entry (C). Finally, a term entry (T) describes a sequence of words.
- **Lem**: The lemma is the representation of the entry in its basic form (singular, nominative for nouns; infinitive for verbs).
- **Mul**: The code for encoding morphological and syntactic information is defined as in the open standard MULTEXT.\(^1\)
- **Frm**: Inflected form that is linked to an entry for its lemma through the Ref field.
- **Mfr**: The morpho-syntactic features of the inflected form using MULTEXT exactly as for the Mul field.
- **Ref**: If the entry consists of an inflected form, a unique ID of its lemma entry is given.

<table>
<thead>
<tr>
<th>Field</th>
<th>Description</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lng</td>
<td>Language</td>
<td>the language to which pertains the present entry</td>
</tr>
<tr>
<td>Id</td>
<td>Multilingual Identifier</td>
<td>the unique identifier of this entry</td>
</tr>
<tr>
<td>Typ</td>
<td>Entry Type</td>
<td>one of the 4 allowed types of entry (B,C,S,T)</td>
</tr>
<tr>
<td>Err</td>
<td>Correctness</td>
<td>flag for correctness of this entry</td>
</tr>
<tr>
<td>Lem</td>
<td>Lemma</td>
<td>the entry in its basic form</td>
</tr>
<tr>
<td>Mul</td>
<td>Morpho-syntactic Features</td>
<td>the MULTEXT morpho-syntactic tag of the lemma</td>
</tr>
<tr>
<td>Frm</td>
<td>Inflected Form</td>
<td>any inflected form</td>
</tr>
<tr>
<td>Mfr</td>
<td>Features of Inflected Form</td>
<td>the MULTEXT morpho-syntactic tag of the inflected form</td>
</tr>
<tr>
<td>Inf</td>
<td>Inflection Model</td>
<td>language specific information</td>
</tr>
<tr>
<td>Mis</td>
<td>Language Specific Argument</td>
<td>to be used freely by provider of entries</td>
</tr>
<tr>
<td>Prt</td>
<td>Decomposition</td>
<td>the decomposition of a compound entry into its parts</td>
</tr>
<tr>
<td>Str</td>
<td>Head</td>
<td>the head word of the term</td>
</tr>
<tr>
<td>Ref</td>
<td>Reference Lemma</td>
<td>ID of its lemma’s entry (if inflected form)</td>
</tr>
<tr>
<td>Exa</td>
<td>Typical Usage</td>
<td>a sentence presenting a typical usage of this entry</td>
</tr>
<tr>
<td>Com</td>
<td>Comment</td>
<td>any comment or warning about this entry</td>
</tr>
</tbody>
</table>

Table 1: Fields of the Lexicon Interchange Format

shortcoming is that different lexical resources provide information of different granularity. For example, the German word *Fingerpanaritien* is a compound, though the decomposition is missing and the type of the entry is marked as a base form (cf., on the other hand, the Swedish compound *fingeravtryck* and its segmentation).

**Monolingual Resources**

After agreeing upon the Interchange Format, partners from five different institutions collected their monolingual lexical resources. These are:

- the French UMLF lexicon from different French health-related organizations and the University Hospitals of Geneva, Switzerland (33,718 entries) [11]
- an English medical lexicon from Linköping University, Sweden (22,686 entries)
- a Swedish medical lexicon from Linköping University (23,223 entries)
- a Swedish medical lexicon from Göteborg University, Sweden (6,786 entries)
- the German Specialist Lexicon from Freiburg University Hospital, Germany (41,316 entries) [10]

In addition,

- the English Specialist Lexicon, which is part of the UMLS (96,621 entries, avoiding acronyms and chemical names) [8],

has also been converted into the Interchange Format. So far, 224,351 lexical entries for the biomedical domain, fully encoded with morpho-syntactic
features, were collected covering four languages (cf. Table 2 for a sample\(^2\)). The number of different lemmas (ignoring multiple lexical information) is 105,317 for English, 29,822 for French, 27,480 for German, and 27,093 for Swedish (a total of 189,712, therefore, 1.2 morpho-syntactic variants are given per lexical entry, in average).

### Linking Format Definition

The cross-lingual grouping of corresponding entries is the essence of a multilingual dictionary. Unfortunately, this is not a straightforward process and a couple of cross-lingual phenomena are problematic to capture, especially regarding the different characteristics of case, gender and number in different languages, as well as multiple derivations, e.g. for adjectives, dependent on whether their use is attributive or predicative. Consider the German words *Schere* and *Hose* (both noun, singular) and the English equivalents, *scissors* and *trousers* (both noun, plural). Singular forms of the latter examples do not exist, whilst for other examples, of course, singular forms can be translated to a corresponding singular form in the other language. Such information should be kept in a multilingual lexicon, e.g. for the use in machine translation applications.

Different languages also make different use of grammatical gender or noun classes. Whilst in German, Greek or Latin, three grammatical genders are distinguished (*masculine*, *feminine* and *neuter*), French and Italian only use two (*masculine*, *feminine*). Swedish and Danish discriminate the classes *common* and *neuter*. Finally, English does not account for any of these features at all. In a first version, in order to find an agreement on the question, in which cases two lexical items from different languages, \(A\) and \(B\), can be regarded as translations of each other, we defined the following "levels" of bi-directional relationships:

1. **Rel1**: \(A\) and \(B\) share the same part of speech (POS) and all MULTEX T features

2. **Rel2**: \(A\) and \(B\) share the same POS, but at least one MULTEX T feature differs

3. **Rel3**: \(A\) and \(B\) do not share the same POS

Having these types of relations in mind, we created a simple Linking Format, which is depicted in Table 3.

So far, the meaning of words and their possible translations have not been discussed. In the following section, we show how lexical entities can be aligned on the semantic level.

### Cross-Lingual Alignment

For the medical domain, methods for the automatic search for translation candidates have already been explored. One promising idea is to use already existing translations at a subword level in order to support the acquisition of translations at a term level [7]. For the linkage of lexemes on the semantic level, we make use of the MorphoSaurus system [6], a text normalization engine using subword lexicons for different languages, as well as a multilingual thesaurus.

### Morpho-Semantic Indexing

The MorphoSaurus system is based on the assumption that neither fully inflected nor automatically stemmed words constitute the appropriate granularity level for lexicalized content description. Especially in scientific sublanguages, we observe a high frequency of complex word forms such as in ‘pseudo\(\_\)hypo\(\_\)para\(\_\)thyroid\(\_\)ism’. To properly account for particularities of ‘medical’ morphology, the notion of subwords was introduced as self-contained, semantically minimal units. Subwords are assembled in a multilingual dictionary and thesaurus, which contain their entries, special attributes and semantic relations between them. Entries are listed together with their attributes such as language and subword type (stem, prefix, suffix, invariant). Each lexicon entry is

---

\(^2\)The first character of the Mut field encodes the part-of-speech: \(N\) (noun), \(A\) (adjective). In case of nouns, \(c\) denotes common nouns, \(m\) masculine, \(s\) singular, \(n\) neuter or nominative, depending on the position. For adjectives, \(f\) stands for qualitative, \(p\) positive. The character ‘-’ indicates that a particular feature does not fit into the language given (e.g. gender in English) or is unspecified for this entry.

### Table 2: Sample of Compiled Lexical Resources (some fields omitted)

<table>
<thead>
<tr>
<th>Lng</th>
<th>Id</th>
<th>Typ</th>
<th>Lem</th>
<th>Mul</th>
<th>Frm</th>
<th>Mfr</th>
<th>Pri</th>
<th>Str</th>
</tr>
</thead>
<tbody>
<tr>
<td>FR</td>
<td>UMLF:10081</td>
<td>B</td>
<td>doigt</td>
<td>Nems</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EN</td>
<td>LIU:EN8427</td>
<td>T</td>
<td>finger nail</td>
<td>Ne-sn</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SV</td>
<td>LIU:SV6663</td>
<td>B</td>
<td>digital</td>
<td>Alp-sn</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SV</td>
<td>UGOT:33733</td>
<td>C</td>
<td>fingeravtryck</td>
<td>Ne-sn</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DE</td>
<td>UKLFR:39556</td>
<td>B</td>
<td>Fingerpanaritium</td>
<td>Ne-nm</td>
<td></td>
<td>Fingerpanariten</td>
<td>Nenpa</td>
<td></td>
</tr>
</tbody>
</table>

For each language, the Linking Format Definition (Linking Format Definition) includes a simple Linking Format, which is depicted in Table 3.
<table>
<thead>
<tr>
<th>Field</th>
<th>Description</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Src</td>
<td>Source Entry ID</td>
<td>The Id of the source entry to be linked to a target entry</td>
</tr>
<tr>
<td>Tar</td>
<td>Target Entry ID</td>
<td>The Id of the target entry linked from the source entry</td>
</tr>
<tr>
<td>Typ</td>
<td>Link Type</td>
<td>Type of relation</td>
</tr>
</tbody>
</table>

Table 3: Fields of the Linking Format

Table 4: Sample Links between Lexical Items. Additional information and MULTTEXT values of the corresponding items are depicted in Column four to nine (Cf. Footnote 2 for the explanation of Mul values).

assigned to one or more morpho-semantic identifier(s) representing the corresponding synonymy class(es) (MIDs). Intra- and interlingual semantic equivalence are judged within the context of medicine only.

Figure 1 depicts how source documents (top-left) are converted into an interlingual representation by a three-step morpho-semantic indexing procedure. First, each input word is orthographically normalized (top-right). Next, words are segmented into sequences of subwords or left unaffected when no subwords can be decomposed (bottom-right). Finally, each meaning-bearing subword is replaced by a language-independent semantic identifier, its MID, thus producing the interlingual output representation of the system (bottom-left). MIDs which co-occur in both document fragments appear in bold face.

Figure 1: Morpho-Semantic Indexing Pipeline

Linking Algorithm

In a first step, all lexical entries are processed with the MORPHOSAURUS system. Afterwards, a quite simple algorithm was used to perform the mappings between all entries: Every lexeme \( i \) and its attributes is compared to any other lexeme \( j \) in the list. If their representations in the interlingua format are identical, they are considered as potential translations or synonyms and linked. Then the relation type (REL1, REL2 or REL3, cf. previous section) is determined, by comparing the lexical attributes.

Results

Using the algorithm introduced, we obtained 651,542 bi-directional relations between lexemes, a sample of which is depicted in Table 4. For English-German, 126,504 translations have been generated (31,544 when only different lemmas are taken into account, thus ignoring ambiguous lexical informations), for English-French 70,680 (24,368, respectively) and for English-Swedish 86,655 (34,030). Furthermore, 21,604 (8,312) relations have been extracted for French-Swedish, 32,659 (10,458) for French-German and finally, 41,469 (12,105) for German-Swedish. All other relations (271,971) cover intralingual synonymy. The distribution of different types of relations is 66,641 occurrences for REL1 (10%), 286,880 for REL2 (44%) and 298,021 for REL3 (46%).

Coverage

The UMLS Metathesaurus is the most comprehensive resource for medical terminology. Therefore, it is particularly interesting how many terms of the UMLS are covered by the multilingual lexicon. Table 5 gives the numbers for those items in the UMLS, which are marked as an preferred entry and only contain alphabetic characters. Column two depicts the number of UMLS terms for the corresponding languages, Column three gives the number of those UMLS entries, which are covered by the multilingual lexicon. Values range between 13% for German up to 71% for Swedish. The numbers in Column four show how many synonyms and morpho-syntactic variants of UMLS terms are listed in the lexicon which are not part of the Metathesaurus, and, therefore, could be added. This consideration only takes those variants into account, which share at least the same part of speech with the corresponding UMLS entry (only REL1 and REL2).

3Thus, multi-word entries and chemical compounds are not considered in the following discussion.
system, we showed that a substantial number of translations can be generated. First examinations of the data proved many alignments to be valid. Of course, an extensive evaluation of the multilingual medical lexicon is still due. Further work will also examine relations with the Lexical Markup Framework of ISO/TC 37/SC 4.4

Acknowledgments: This work was supported by the European Network of Excellence “Semantic Mining” (NoE 507505).

References


Cross-Lingual Mappings

For the language pairs considered, the UMLS Metathesaurus already contains between 6,700 and 16,000 translations (cf. Table 6, Column two). Within a range of 8% (EN-DE and DE-SV) to 36% (EN-SV), these mappings are also included in the multilingual lexicon (Column three). A total of 30,282 synonymous entries (Column four) could be added to 64,837 existing UMLS translations. Finally, those cross-lingual mappings which are captured in the multilingual lexicon but not in the UMLS Metathesaurus, sum up to 81,321 alignments (again, only considering REL1 and REL2). While there are 64,837 word-to-word translations in the UMLS for the languages considered, the multilingual lexicon contains 120,817 different translations.

Conclusion

We introduced a common framework for the integration of heterogeneous lexical resources covering different languages. The second issue of this contribution concerns the Linkage Format, in which lexical relations can be coded. We endorse a simple architecture that is easy to apply for different language pairs. Finally, using morpho-semantic normalization in terms of the MorphoSaurus...