Towards a Multilingual Medical Lexicon

Kornél Markó[1], Robert Baud[2], Pierre Zweigenbaum[3], Lars Borin[4], Magnus Merkel[5], Stefan Schulz[1]

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

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

Table 1: Fields of the Lexicon Interchange Format

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 2 shows an excerpt of different lexicons encoded in the Interchange Format. One obvious 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).

[1]Common Specifications and Notation for Lexicon Encoding and Preliminary Proposal for the Tagsets (http://nl.ijs.si/ME/V3/msd/related/msd-multext/)

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

Lng	Id	Typ	Lem	Mul	Frm	Mfr	Prt	Str
FR	UMLF:10081	B	doigt	Ncms				
EN	LIU:EN8427	T	finger nail	Nc-sn				nail
SV	LIU:SV6663	B	digital	Afp-sn				
SV	UGOT:3373	C	fingeravtryck	Nc-sn			finger–avtryck	avtryck
DE	UKLFR:39556	B	Fingerpanaritium	Ncnsn	Fingerpanaritien	Ncnpa		

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

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 a definite or indefinite object follows or 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 MULTEXT features
2. **Rel2:** A and B share the same POS, but at least one MULTEXT 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 *Mul* 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.

Field	Description	Definition
Src	Source Entry ID	The Id of the source entry to be linked to a target entry
Tar	Target Entry ID	The Id of the target entry linked from the source entry
Typ	Link Type	Type of relation

Table 3: Fields of the Linking Format

Src	Tar	Typ	Lng1	Lem1	Mul1	Lng2	Lem2	Mul2
LIU:EN147	LIU:SV151	REL1	EN	abdominal hernia	Nc-sn	SV	bukbrck	Nc-sn
LIU:EN143	UKLFR:34985	REL2	EN	abdominal aorta	Nc-sn	DE	Bauchaorten	Ncfpn
LIU:EN947	UMLF:1123	REL3	EN	alveolar	Afp–n	FR	alvole	Ncfs

Table 4: Sample Links between Lexical Items. Additional information and MULTEXT 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.

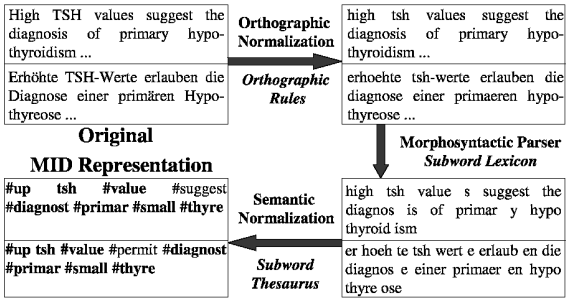

Figure 1: Morpho-Semantic Indexing Pipeline

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.[3] 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).

[3]Thus, multi-word entries and chemical compounds are not considered in the following discussion.

Lng	UMLS	Covered	Syns.	Addit.
EN	122,035	32,668	3,807	68,842
DE	21,162	2,832	1,269	23,379
FR	10,260	3,590	309	25,923
SV	12,012	8,520	994	17,579
$\sum$	165,469	189,712		

Table 5: Comparison of Lexical Entries: UMLS Metathesaurus and Multilingual Lexicon

Lng Pair	UMLS	Covered	Syns.	Addit.
EN-DE	15,979	1,259	8,801	21,484
EN-FR	12,589	1,783	6,974	15,611
EN-SV	9,554	3,403	10,124	20,503
DE-FR	9,859	850	773	8,835
DE-SV	10,063	810	1,699	9,596
FR-SV	6,793	1,109	1,911	5,292
$\sum$	64,837	120,817		

Table 6: Comparison of Cross-Lingual Mappings

Finally, the number of additional lexemes in the lexicon that are neither found in the Metathesaurus, nor constitute morpho-syntactic variants of existing UMLS entries, are depicted in Column five. All in all, the multilingual lexicon contains 189,712 different lemmas.

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 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

[1] Robert Baud, Mikaël Nyström, Lars Borin, Robert Evans, Stefan Schulz, and Pierre Zweigenbaum. Interchanging lexical information for a multilingual dictionary. In *Proc AMIA Symp 2005*, pages 31–35, 2005.

[2] Olivier Bodenreider, Anita Burgun, and Joyce A. Mitchell. Evaluation of WordNet as a source of lay knowledge for molecular biology and genetic diseases: A feasibility study. In *Proc MIE 2003*, pages 379–384, 2003.

[3] Anita Burgun and Olivier Bodenreider. Comparing terms, concepts and semantic classes in WordNet and the *Unified Medical Language System*. In *Proc NAACL 2001 Workshop 'WordNet and Other Lexical Resources: Applications, Extensions and Customizations'*, pages 77–82, 2001.

[4] Y.C. Chiao and P. Zweigenbaum. Looking for french-english translations in comparable medical corpora. In *Proc AMIA Symp 2002*, pages 150–154, 2002.

[5] Christiane Fellbaum, editor. WordNet*: An Electronic Lexical Database*. Cambridge, MA: MIT Press, 1998.

[6] Kornél Markó, Stefan Schulz, and Udo Hahn. MorphoSaurus: Design and evaluation of an interlingua-based, cross-language document retrieval engine for the medical domain. *Meth Inf Med*, 44(4):537–545, 2005.

[7] Fiammetta Namer and Robert Baud. Guessing lexical relations between biomedical terms: towards a multilingual morphosemantics-based system. In *Proc MIE 2005*. 2005.

[8] Umls. *Unified Medical Language System*. Bethesda: National Library of Medicine, 2005.

[9] Piek Vossen, editor. EuroWordNet*: A Multilingual Database with Lexical Semantic Networks*. Dordrecht: Kluwer Academic Publishers, 1998.

[10] Gesa Weske-Heck, Albrecht Zaiss, Stefan Schulz, Wolfgang Giere, Michael Schopen, and Rüdiger Klar. The German Specialist Lexicon. In *Proc AMIA Symp 2002*, pages 884–888, 2002.

[11] Pierre Zweigenbaum, Robert Baud, Anita Burgun, Fiammetta Namer, Eric Jarrousse, Natalia Grabar, Patrick Ruch, Franck Le Duff, Jean-François Forget, Magaly Douyère, and Stéfan Darmoni. UMLF – a unified medical lexicon for French. *Int J Med Inf*, 74(2/4):119–124, 2005.

[4] http://tagmatica.fr/doc/ISO24613cdRev7.pdf