Cross-Language Mining for Acronyms and Their Completions from the Web

Udo Hahn[1], Philipp Daumke[2], Stefan Schulz[2], and Kornél Markó[2]

[1] Jena University Language and Information Engineerings (JULIE) Lab, Germany
http://www.coling.uni-jena.de
[2] Department of Medical Informatics, Freiburg University Hospital, Germany
http://www.imbi.uni-freiburg.de/medinf

Abstract. We propose a method that aligns biomedical acronyms and their long-form definitions across different languages. We use a freely available search and extraction tool by which abbreviations, together with their fully expanded forms, are massively mined from the Web. In a subsequent step, language-specific variants, synonyms, and translations of the extracted acronym definitions are normalized by referring to a language-independent, shared semantic interlingua.

1 Introduction

The understanding of acronyms and abbreviations in biomedical texts is crucial for various NLP applications, such as text mining [1], information extraction [2], or information retrieval systems [3]. This is witnessed, in particular, for protein and gene expressions from biomedical texts [4] (as well as the relations between them). Those expressions frequently consist of acronyms, but their definitions in the text might differ from the ones found, e.g., in external databases, such as ARGH, ACROMED, or SARAD [5] (cf. also [6] for an overview).

Multiple expansions for the same acronym, or multiple acronyms for the same definition, will lead to difficulties when one tries to match natural language expressions with a standardized vocabulary such as the UMLS or MESH [7]. In an information retrieval scenario, unresolved acronyms will possibly lead to a loss of precision: Does ”AD” refer to ”Alzheimer’s Disease” or to ”allergic dermatitis”? The problem of ambiguity becomes even harder when multilingual documents are encountered. This is likely to happen to Web search engines. In this case, the acronym ”AD” may have a German expansion (”atopische Dermatitis”), a Spanish one (”aurícula derecha”), or a Portuguese one (”agua destilada”), and possibly many more. Even worse, the German acronym equivalent to ”Alzheimer’s Disease” is ”AK” (”Alzheimer Krankheit”) or ”MA” (”Morbus Alzheimer”), while for Spanish the equivalent short-cut is ”EA”(”enfermedad de Alzheimer”).

Many research efforts have been spent on the automatic extraction of short-form/long-form (SF/LF) pairs (abbreviations and acronyms mapped to their expansions/completions) within a single language [8, 9, 10, 11, 5, 12, 13, 14]. Different ways of how abbreviations are actually used in written (medical) language were also studied [15], while little attention has been paid to how acronyms behave across languages.

A. Hoffmann, H. Motoda, and T. Scheffer (Eds.): DS 2005, LNAI 3735, pp. 113–123, 2005.

This is a particular challenge for intelligent Web search engines and it is the focus of this paper.

2 Analysis of Terms into Subwords

We propose a method that automatically aligns acronyms and their definitions across different languages. It is based upon a dictionary the entries of which are *equivalence classes* of subwords, i.e., semantically minimal units [1]. From a linguistic perspective, subwords are often closer to formal Porter-style stems [2] rather than to lexicologically orthodox basic forms, e.g., of verbs or nouns or linguistically plausible stems. Hence, their merits have to be shown in experiments. These equivalence classes capture intralingual as well as interlingual synonymy. As equivalence classes abstract away from subtle particularities within and between languages and reference to them is realized via a language-independent concept system they form an *interlingua*.

Subwords are assembled in a multilingual lexicon and thesaurus, with the following considerations in mind:

- Subwords are listed, together with their attributes such as language (English, German, Portuguese, Spanish) or subword type (stem, prefix, suffix, invariant). Each subword is assigned one or more morpho-semantic class identifier(s), we call *MID*(s), representing the corresponding synonymy equivalence class.
- Intralingual synonyms and interlingual translation synonyms of subwords are assigned the same equivalence class (judged within the context of medicine only).
- Two types of meta relations can be asserted between synonymy classes:
  (i) a paradigmatic relation *has-meaning*, which relates one ambiguous class to its specific readings, as with:
  $\{head\} \Rightarrow \{kopf, zephal, caput, cephal, cabec, cefal\}$ *OR* $\{boss, leader, lider, chefe\}$.
  (ii) a syntagmatic relation *expands-to*, which consists of predefined segmentations in case of utterly short subwords, such as:
  $\{myalg\} \Rightarrow \{muscle, muskel, muscul\} \oplus \{pain, schmerz, dor\}$.[1]

We refrain from introducing additional hierarchical relations between MIDs because such links can be acquired from domain-specific vocabularies, e.g., the Medical Subject Headings [3] (cf. experimental evidence from Markó *et al.* [4]).

Figure 1 depicts how source documents (top-left) are converted into an interlingual representation by a three-step procedure. First, each input word is orthographically normalized in terms of lower case characters and according to language-specific rules for the transcription of diacritics (top-right). Next, words are segmented into sequences of subwords as found in the lexicon (bottom-right). Finally, each meaning-bearing subword is replaced by a language-independent semantic identifier, the corresponding MID, which unifies intralingual and interlingual (quasi-)synonyms, thus producing the interlingual output representation of the system (bottom-left). In Figure 1, bold-faced MIDs co-occur in both document fragments (after conversion into the interlingua format).

[1] '$\oplus$' denotes the string concatenation operator.

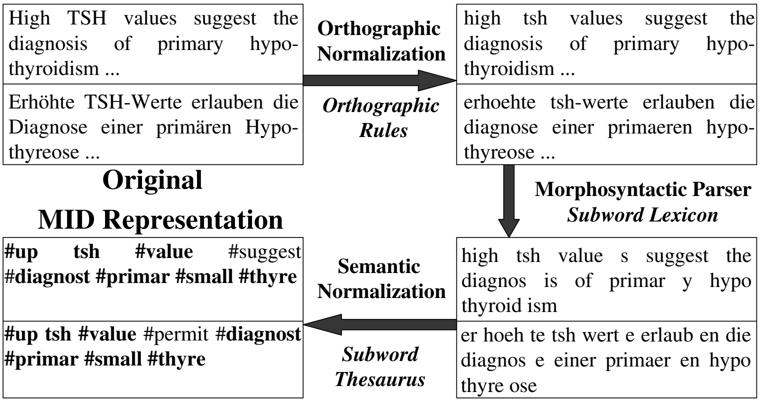
High TSH values suggest the diagnosis of primary hypothyroidism ...
Erhöhte TSH-Werte erlauben die Diagnose einer primären Hypothyreose ...
Original
Orthographic Normalization
Orthographic Rules
high tsh values suggest the diagnosis of primary hypothyroidism ...
erhoehte tsh-werte erlauben die diagnose einer primaeren hypothyreose ...
Morphosyntactic Parser
Subword Lexicon
high tsh value s suggest the diagnos is of primar y hypo thyroid ism
er hoeh te tsh wert e erlaub en die diagnos e einer primaer en hypo thyre ose
Semantic Normalization
Subword Thesaurus
MID Representation
#up tsh #value #suggest #diagnost #primar #small #thyre
#up tsh #value #permit #diagnost #primar #small #thyre

Fig. 1. Morpho-Semantic Indexing (MSI)

In the meantime, the entire subword lexicon (as of July 2005) contains 72,513 entries, with 22,067 for English,[2] 22,497 for German, 14,888 for Portuguese, and 13,061 for Spanish. All of these entries are related in the thesaurus by 20,990 equivalence classes. We also found a well-known logarithmic growth behavior as far as the increase of the number of subwords are concerned [1]. Under this observation, at least the English and German subword lexicons have already reached their saturation points.

Our project started from a bilingual German-English lexicon, while the Portuguese part was added in a later project phase (hence, its size still lags somewhat behind). All three lexicons and the common thesaurus structure were manually constructed, which took us about five person-years. While we simultaneously experimented with various subword granularities as well as weaker and stronger notions of synonymy, this manual approach was even heuristically justified. With a much more stable set of criteria for determining subwords emerging from these experiments, we recently switched from a manual to an automatic mode for lexicon acquisition. The Spanish sublexicon, unlike all other previously built sublexicons, was the first one generated solely by an automatic learning procedure which is specifically targeted at large-scale lexical acquisition. It makes initial use of cognate relations (roughly, string similarities) that can be observed for typologically related languages [5] and has recently been embedded into a bootstrapping methodology which induces new subwords that cannot be found by considering merely cognate-style string similarities. This extended acquisition mode makes heavy use of contextual co-occurrence patterns in comparable corpora [6].

In earlier experiments on cross-language information retrieval [1] and multilingual document classification [7], we showed the usefulness of representing medical documents on an interlingual layer. However, we were not able to properly account for acronyms, since they were completely missing in our lexicons. Therefore, we here

[2] Just for comparison, the size of WordNet assembling the lexemes of general English in the 2.0 version is on the order of 152,000 entries (`http://wordnet.princeton.edu/man/wnstats.7WN`, last visited on May 13, 2005). Linguistically speaking, the entries are basic forms of verbs, nouns, adjectives and adverbs.

adapt previous work on automatic acronym detection to the needs of our interlingual representation approach.

3 Extracting Biomedical Acronyms and Completions

Our work reuses, without any modification, a simple and fast algorithm for the extraction of abbreviations and their completions from biomedical documents, which has been developed by Schwartz and Hearst [8].[3] The algorithm achieves 96% precision and 82% recall on a standardized test collection and, thus, performs at least as good as other existing approaches [9, 10, 11, 12, 13]. It deals with the extraction of acronyms and abbreviations together with their full forms (completions) in a two-step process. First, a list of candidate short-form/long-form (SF-LF) pairs is determined, which are then validated by taking additional selection criteria into account. In the following, we briefly describe the principles underlying both steps.

Extraction of possible SF-LF terms. Basically, SF-LF pairs are identified by their adjacency to parentheses. Two basic patterns, *LF (SF)* and *SF (LF)*, have to be distinguished. According to Schwartz and Hearst, a *short* form has the following characteristics: it contains between 2 and 10 characters, has a maximum of two words, at least one character is a letter, and its first character is alphanumeric. The *long* form must immediately appear before or after the corresponding short form and the maximum number of words is constrained by $min(|A| + 5, |A| * 2)$.[4] In practice, the *LF (SF)* pattern occurs more frequently. Therefore, only if a criterion for an *LF (SF)* pattern is not fulfilled (e.g., more than two words inside the parentheses), the complemenatry pattern, *SF (LF)*, is tried.

Selection of the correct SF-LF term. Next, rules are applied to identify the correct SF-LF pair from the list of candidates which were extracted in the first step. Most importantly, each character in the short form must match a character in the long form and characters of the short form must appear in the same linear order as in the long form. Furthermore, the first character of the SF has to be the same in the LF. Finally, all LFs are removed which are shorter than the corresponding SF, or which include the corresponding SF within one of their single words.

4 Experiments

The WWW is here taken as the authoritative textual resource where the largest and most up-to-date variety of acronyms and their associated completions can be found. Hence, for our experiments, we generated very large corpora directly from different, heterogeneous WWW sources, including MEDLINE. With more than 250m text tokens, the derived English corpus was much larger than those for the other languages involved (37m tokens for German, 14m for Portuguese, and 11m for Spanish, cf. Table 1). The

[3] The source code (in Java) is made available on the Web; see http://biotext.berkeley.edu/software.html.

[4] |A| is the number of characters in the corresponding SF.

Table 1. Corpus and Acronym Extraction Statistics

Language	Corpus Tokens	Proportion of Acronyms
English	250,258,039	1,253,311 (0.50%)
MSI-Covered		1,033,929 (82.5%)
German	37,105,363	29,967 (0.08%)
MSI-Covered		26,770 (89.3%)
Portuguese	13,904,790	8,532 (0.06%)
MSI-Covered		7,065 (82.8%)
Spanish	11,103,066	7,714 (0.07%)
MSI-Covered		4,723 (61,2%)

contribution of this paper lies in the cross-language linking of these data items by applying the MSI procedure outlined in Section 2.

Using the algorithm described above, we collected over 1.2m abbreviations together with their long forms for English, while we extracted some 30K pairs for German, 9K pairs for Portuguese and 8K pairs for Spanish (for exact numbers, cf. Table 1). In contradistinction to the other languages, the English corpus included a large number of expert-level MEDLINE abstracts. As a consequence, every 200th token in the collection was classified as an acronym. For the other languages (for which the corpora included a larger amount of consumer information), this ratio is much smaller (0.06 to 0.08 percent of the text tokens in the corpora).

After the acquisition of SF-LF pairs, the long forms were processed by the MSI procedure as described in Section 2. Upon prior manual inspection of document samples we observed that English long forms also tended to frequently occur in German,

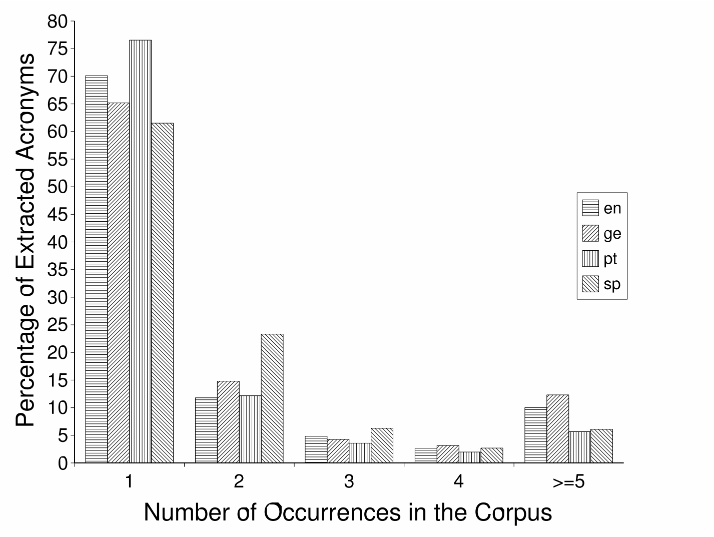

Fig. 2. Distribution of SF-LF Occurrences per Corpus

Table 2. Effects of Morpho-semantic Normalization in Terms of Unique SF-LF Pairs and Tokens per Type

Language	Surface		MSI	
	Unique	Ratio	Unique	Ratio
English	212,470	4.87	189,639	5.45
German	4,276	6.26	3,653	7.33
Portuguese	3,934	1.20	3,633	1.95
Spanish	2,037	2.32	1,911	2.47

Portuguese, and Spanish texts. Therefore, a decision had to be taken which lexicon to use for the MSI process. Our approach was to segment the long forms using *every* lexicon available (so no a priori decision was taken). Those language hypotheses were kept for which the underlying lexicon yielded *complete* lexical coverage with regard to the specific long form. If there were more than one remaining language hypothesis, the document language (if not English) was preferred over English.

This procedure led to over one million SF-LF pairs completely covered by the MSI procedure for English (83%), and approximately 27K pairs (89%) for German, 7K pairs (83%) for Portuguese, and 5K pairs (61%) for Spanish (cf. Table 1 for detailed numbers). In the following, we will only focus on this subset of extracted abbreviations. Figure 2 yields an impression of how frequent unique SF-LF pairs occur in the corpora considered, for each language condition. 61% to 76% of all acronyms extracted occur only once, 12% to 23% appear two times, whilst five or more occurrences are found for 6% to 12% of all SF-LF pairs.

As depicted in Table 2 (Column 2), 212,470 unique SF-LF pairs were generated for English, 4,276 for German, 3,934 for Portuguese, and 2,037 for Spanish. Column 3 of the table shows the average number of corpus occurrence for each unique SF-LP pair. After the MSI normalization of long forms, the number of unique SF-LF pairs decreases to 189,639 for English (3,653 for German, 3,633 for Portuguese and 1,911 for Spanish). Accordingly, the number of tokens per type increases, as depicted in the fifth column of Table 2. As an example, morpho-syntactic variants in long forms such as in *"CTC"*-*"computed tomographic colonography"* and *"CTC"*-*"computed tomography colonography"* are unified, an immediate effect of term normalization based on the interlingua (composed of equivalence classes of subwords).

4.1 Intra-Lingual Phenomena

Two basic ambiguity phenomena have to be considered when we discuss the results for a given language: First, one short form can have multiple long forms (SF ambiguity), and, second, one long form can have multiple short forms (LF ambiguity). An example for an SF ambiguity is given with *"ABM"* mapped to *"acute bacterial meningitis"* and to *"adult bone marrow"*. Table 3 shows the average numbers of different long forms for each short form, both for the baseline condition (lower-case surface form) and the MSI condition. For English, 82,501 unique short forms were extracted. The average number of long forms associated to unique SFs decreases from 2.56 to 2.30 for

Table 3. SF Ambiguity

Language	SFs	Average LF	
		Surface	MSI
English	82,501	2.56	2.30
German	2,954	1.45	1.24
Portuguese	2,517	1.56	1.44
Spanish	1,450	1.41	1.32

Table 4. LF Ambiguity

Language	Surface		MSI	
	LFs	Average SF	LFs	Average SF
English	184,639	1.15	154,693	1.23
German	4,187	1.02	3,515	1.04
Portuguese	3,798	1.04	3,395	1.07
Spanish	1,979	1.03	1,825	1.05

MSI, as expected. A similar tendency can also be observed for the other languages we considered.

The second phenomenon, one long form which comes with multiple different short forms, can also be observed in all languages involved in our experiments. For example, the noun phrase *"acid phosphatase"* has nine different abbreviations in the English corpus we processed (case insensitive): *"AcP"*, *"acPAse"* *"ACP-ase"*, *"Acph"*, *"ACPT"*, *"AP"*, *"APase"*, *"AphA"*, and *"APs"*. Table 4 depicts the numbers describing this phenomenon. For English, a total of 184,639 different long forms were extracted, arising from 212,470 different SF-LF pairs (cf. Table 2). Thus, each LF is associated with 1.15 SFs, on the average. For the MSI condition, fewer different long forms are encountered. Hence, the ratio slightly increases, for all languages.

4.2 Inter-Lingual Phenomena

4.2.1 Identical SF-LF Pairs

The first observation we made is that quite often SF-LF pairs are appear in other languages, such as *"WHO"* and its expansion *"World Health Organization"*, *"PCR"* and its completion *"polymerase chain reaction"*, or *"IL"* associated with *"interleukin"*. Summarizing (cf. Table 5, Column 2), we found 584 identical SF-LF for English-German, 181 for English-Portuguese, 192 for English-Spanish, 35 for German-Portuguese, 40 for German-Spanish, and 106 for Portuguese-Spanish (the latter sets also may contain some English SF-LF pairs).

4.2.2 Identical SF, Different LF

One way of identifying possible translations of long forms is to collect those long forms which are connected to a unique short form at the surface level. For example, if an English document contains *"WHO"*-*"World Health Organization"* and a German document contains *"WHO"*-*"Weltgesundheitsorganisation"*, the long forms can be regarded as possible translations of each other. For English-German, 100,915 of these pairs can be extracted, for English-Portuguese 151,037, for English-Spanish 109,568, for German-Portuguese 2,468, for German-Spanish 1,709, and for Portuguese-Spanish we counted 3,454 of these hypothesized translations (Table 5, Column 3). Of course, these sets also contain syntactic variants and a large number of false positives, since short forms are used differently across languages. Therefore, we switched our perspective to the interlingual layer of long form representations.

4.2.3 Identical SF, Translation of LF

In this condition, we examined those cases, in which short forms were identical and long forms were different at the surface level, but identical at the interlingual layer, by comparing SF-LF pairs extracted from the different source corpora. As a result, we obtained lists of bilingually aligned terms, such as English *"acute lymphatic leukemia"* linked to the German *"akute lymphatische Leukämie"* via the common short term *"ALL"*. As an example, 2,479 translations were generated for English-German using this heuristics (cf. Table 5 for additional data covering the remaining language pairs, as well).

Table 5. Statistics on Cross-Lingual Acronym Extraction: Results for Identical (I), Different (D) and Translations (T) of Short Forms (SF) and Long Forms (LF)

	Surface		MSI	
Language Pair	I(SF) I(LF)	I(SF) D(LF)	I(SF) T(LF)	D(SF) T(LF)
EN-GE	584	100,915	2,479	3,212
EN-PT	181	151,037	665	3,982
EN-SP	192	109,568	573	2,136
GE-PT	35	2,468	81	328
GE-SP	40	1,709	110	290
PT-SP	106	3,454	250	207
Total	1,138	369,151	4,158	10,155

4.2.4 Different SF, Translation of LF

In this scenario, we examined those cases, for which the long forms were identical or translations of each other (i.e., identical at the interlingua layer), but with different short forms. This captures interesting constellations such as English *"AIDS"* (*"acquired immune deficiency syndrome"*) aligned to Spanish or Portuguese *"SIDA"* (*"síndrome de inmunodeficiencia adquirida"*). We collected 207 of these translations for Portuguese-Spanish, and up to 3,212 for English-German (cf. Table 5, Column 5, for additional data covering the remaining language pairs, as well).

5 Lexicon Integration

In order to enhance the existing lexicons with acronyms automatically, the quality of the derived associations of short forms to long forms had to be ensured. To the best of our knowledge, we know of no multilingual acronym repository in the biomedical field which might serve as a suitable gold standard. With 96% precision, as measured by Schwartz and Hearst [8] on a standardized test set, we expect, however, about 8,500 false positives in the set of unique SF-LF pairs, only considering English (cf. Table 2). Furthermore, since our work focuses on cross-language information retrieval [1] and multilingual text classification [7], we are interested in the cross-lingual mapping of lexical entries. Both challenges are met by a simple heuristics, based upon the idea that "two languages are more informative than one" [14]. Hence, we incorporated those extracted SF-LF pairs in our subword lexicons, for which the long form is a translation of another, at least one, long form in a different language (after mapping on the interlingua layer). Thus, we collected those pairs for which the number of occurrences are depicted in Column 4 and 5 in Table 5. As a result, we obtained an intersection of 4,931 English SF-LF forms, and, correspondingly, 1,149 for German, 1,077 for Portuguese, and 647 for Spanish (a total of 7,804). For the monolingual mapping of short forms to long forms, we decided to additionally collect those language-specific SF-LF pairs, which occur at least 2 times on the layer of the interlingua (cf. Table 2, right). As Table 6 reveals, the lexicon size for the specific languages increased from initially 72,513 entries to 138,343 lexical items (61,081 new entries for English, 2,055 for German, 1,585 for Portuguese, and 1,109 for Spanish). Hence, our approach can truly be considered as a cross-language mining methodology for boosting lexicon growth through the incorporation of acronyms and abbreviations, as well as their associated completions.

Table 6. Enhancement of the Size of the Subword Lexicon

Language	Initial Size	New Acronyms
English	22,067	61,081
German	22,497	2,055
Portuguese	14,888	1,585
Spanish	13,061	1,109
Sum	72,513	65,830
Total	138,343	

6 Related Work

Several different techniques for the automatic extraction of abbreviations and their definitions from biomedical text (particularly from MEDLINE abstracts) have been developed up until now. Schwartz and Hearst [8] offer a simple and fast algorithm for the extraction of abbreviations and their completions from biomedical documents, to which we completely adhere in our approach. The algorithm achieves 96% precision and 82% recall on a standardized test collection and, thus, performs at least as good as other existing approaches [9, 10, 11, 12, 13].

Comprehensive databases with millions of entries are provided by different research groups [15, 9, 11, 12, 13]. They adopt similar sorts of heuristics such as identifying and processing parenthetical phrases within texts. Some of them rely on pattern matching only [16], some use stemming [9, 13], and/or apply term normalization routines to abbreviations and full forms [9, 11, 13] or employ statistical metrics [17]. In addition, Pustejovsky *et al.* [9] even incorporate a shallow parsing approach. A general overview of four large databases and their algorithms can be found in [18].

Our approach for the multilingual alignment of acronyms and their definitions is tied up to the research from these precursors. Unlike most previous research, however, we heavily exploit the WWW for gathering evidence for the linkage between abbreviations and their expanded forms. Furthermore, by mapping extracted long forms onto an interlingual representation layer, an approach which has not been considered so far, acronyms and their definitions are made comparable across different languages with a high coverage. The interlingua layer also serves as a conceptual filter to eliminate false friends (incorrectly linking short and long forms), which are likely to occur in a multilingual Web environment.

7 Conclusions

We introduced a method for aligning biomedical short forms (acronyms, abbreviations) and their associated long forms (completions) across four different languages. A total of 65,830 new lexicon entries were added to an already existing multilingual subword lexicon, boosting its original size by more than 90% of new lexical material.

Acknowledgements. This work was partly supported by Deutsche Forschungsgemeinschaft (DFG), grant KL 640/5-1, and the European Network of Excellence ”Semantic Mining” (NoE 507505).

References

[1] Schulz, S., Hahn, U.: Morpheme-based, cross-lingual indexing for medical document retrieval. International Journal of Medical Informatics **59** (2000) 87–99

[2] Porter, M.F.: An algorithm for suffix stripping. Program **14** (1980) 130–137

[3] MESH: Medical Subject Headings. Bethesda, MD: National Library of Medicine (2004)

[4] Markó, K., Hahn, U., Schulz, S., Daumke, P., Nohama, P.: Interlingual indexing across different languages. In: RIAO 2004 – Conference Proceedings: Coupling Approaches, Coupling Media and Coupling Languages for Information Retrieval, Avignon, France, 26-28 April 2004. Paris: Centre de Hautes Etudes Internationales d'Informatique Documentaire (CID) (2004) 82–99

[5] Schulz, S., Markó, K., Sbrissia, E., Nohama, P., Hahn, U.: Cognate mapping: A heuristic strategy for the semi-supervised acquisition of a Spanish lexicon from a Portuguese seed lexicon. In: COLING Geneva 2004 – Proceedings of the 20th International Conference on Computational Linguistics. Volume 2., Geneva, Switzerland, August 23-27, 2004. Association for Computational Linguistics (2004) 813–819

[6] Markó, K., Schulz, S., Medelyan, A., Hahn, U.: Bootstrapping dictionaries for cross-language information retrieval. In: SIGIR 2005 – Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Salvador, Brazil, August 15-19, 2005. New York, NY: ACM (2005)

[7] Markó, K., Daumke, P., Schulz, S., Hahn, U.: Cross-language MESH indexing using morpho-semantic normalization. In Musen, M.A., ed.: AMIA'03 – Proceedings of the 2003 Annual Symposium of the American Medical Informatics Association. Biomedical and Health Informatics: From Foundations to Applications, Washington, D.C., November 8-12, 2003. Philadelphia, PA: Hanley & Belfus (2003) 425–429

[8] Schwartz, A.S., Hearst, M.A.: A simple algorithm for identifying abbreviation definitions in biomedical text. In Altman, R.B., Dunker, A.K., Hunter, L., Jung, T.A., Klein, T.E., eds.: PSB 2003 – Proceedings of the Pacific Symposium on Biocomputing 2003, Kauai, Hawaii, USA, January 3-7, 2003. Singapore: World Scientific Publishing (2003) 451–462

[9] Pustejovsky, J., Castaño, J., Cochran, B., Kotecki, M., Morrell, M.: Automatic extraction of acronym-meaning pairs from MEDLINE databases. In Patel, V.L., Rogers, R., Haux, R., eds.: MEDINFO 2001 – Proceedings of the 10th World Congress on Medical Informatics. Vol. 1. Number 84 in Studies in Health Technology and Informatics, London, U. K., September 2001. Amsterdam: IOS Press (2001) 371–375

[10] Yu, H., Hripcsak, G., Friedman, C.: Mapping abbreviations to full forms in biomedical articles. Journal of the American Medical Informatics Association **9** (2002) 262–272

[11] Chang, J.T., Schütze, H., Altman, R.B.: Creating an online dictionary of abbreviations from MEDLINE. Journal of the American Medical Informatics Association **9** (2002) 612–620

[12] Wren, J.D., Garner, H.R.: Heuristics for identification of acronym-definition patterns within text: Towards an automated construction of comprehensive acronym-definition dictionaries. Methods of Information in Medicine **41** (2002) 426–434

[13] Adar, E.: SARAD: A simple and robust abbreviation dictionary. Bioinformatics **20** (2004) 527–533

[14] Dagan, I., Itai, A., Schwall, U.: Two languages are more informative than one. In: Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics, Berkeley, CA, USA, 18-21 June 1991. Association for Computational Linguistics (1991) 130–137

[15] Rimer, M., O'Connell, M.: BIOABACUS: A database of abbreviations and acronyms in biotechnology and computer science. Bioinformatics **14** (1998) 888–889

[16] Taghva, K., Gilbreth, J.: Recognizing acronyms and their definitions. International Journal on Document Analysis and Recognition **1** (1999) 191–198

[17] Nenadić, G., Spasic, I., Ananiadou, S.: Automatic acronym acquisition and term variation management within domain-specific texts. In Rodriguez, M., Paz Suarez Araujo, C., eds.: LREC 2002 – Proceedings of the 3rd International Conference on Language Resources and Evaluation. Vol. 6, Las Palmas de Gran Canaria, Spain, 29-31 May, 2002. Paris: European Language Resources Association (ELRA) (2002) 2155–2162

[18] Wren, J.D., Chang, J.T., Pustejovsky, J., Adar, E., Garner, H.R., Altman, R.B.: Biomedical term mapping databases. Nucleic Acids Research **33** (2005) D289–293