Mining a Lexicon of Technical Terms and Lay Equivalents

Noemie Elhadad and **Komal Sutaria**
Computer Science Department
The City College of New York
New York, NY 10031
noemie@cs.ccny.cuny.edu, kdsutaria@gmail.com

Abstract

We present a corpus-driven method for building a lexicon of semantically equivalent pairs of technical and lay medical terms. Using a parallel corpus of abstracts of clinical studies and corresponding news stories written for a lay audience, we identify terms which are good semantic equivalents of technical terms for a lay audience. Our method relies on measures of association. Results show that, despite the small size of our corpus, a promising number of pairs are identified.

1 Introduction

The field of health literacy has garnered much attention recently. Studies show that most documents targeted at health consumers are ill-fitted to the intended audience and its level of health literacy (Rudd et al., 1999; McCray, 2005). While there are many components involved in health literacy that are specific to the reader (e.g., reading level and cultural background), we investigate what can be done from the standpoint of the text to adapt it to the literacy level of a given reader. As such, we set ourselves in the context of a text-to-text generation system, where a technical text is edited to be more comprehensible to a lay reader. An essential resource for such an editing tool is a lexicon of paraphrases, or semantically equivalent terms. In this paper, we investigate a corpus-driven method for building such a lexicon. We focus on terms that are recognized by the UMLS (UMLS, 1995), both for technical and lay candidate terms for equivalence.

Because we have lay audiences in mind, our definition of semantic equivalence must be broader than a notion of strict medical equivalence utilized by medical experts. Thus, while a medical dictionary like UMLS assigns different concept unique identifiers (CUIs) to two particular terms, such as *percutaneous transluminal coronary angioplasty* and *angioplasty*, these terms should be considered semantically equivalent for the purposes of lay readers.

Besides enabling a text tailoring system to adapt technical texts for a lay audience, a lexicon of semantically equivalent technical/lay terms would benefit other tools as well. For instance, the Consumer Health Vocabulary initiative[1] is a comprehensive list of UMLS terms familiar to lay readers. Our lexicon could help augment the terms with equivalence links to technical terms. While much research of late has been devoted to identifying terms incomprehensible to lay readers, such research has not established links between technical terms and equivalent lay terms beyond their CUI information (Zeng et al., 2005; Elhadad, 2006).

The key points of our approach are: (1) the use of combined measures of association to identify pairs of semantically equivalent terms, and (2) a knowledge-based heuristic which acts as a powerful filter for identifying semantically equivalent pairs. Our method does not rely on human labeling of semantically equivalent term pairs. As such, it is unsupervised, and achieves results that are promising considering the small size of the corpus from which the results are derived.

This paper is organized as follows. The next section describes our parallel corpus of paired technical/lay documents. The Methods section describes the different measures of association we experimented with, how we combine them to leverage their complimentary strengths, and our semantic filter. The Results section reports the evaluation against our gold standard and a discussion of our results.

[1] http://www.consumerhealthvocab.org

BioNLP 2007: Biological, translational, and clinical language processing, pages 49–56,
Prague, June 2007.

2 Data Description

Because our ultimate goal is to learn, in a data-driven fashion, semantic equivalents of terms that are too technical for lay readers, we can benefit from having instances of texts which relay similar information but are conveyed in different styles. We collect a corpus similar in structure to those used in the field of statistical machine translation. But, instead of having two collections in different languages, we collect texts written for two different audiences: medically trained readers (technical collection) and health consumers (lay collection).

The lay collection is composed of news stories from the ReutersHealth E-line newsfeed[2] summarizing research in the medical field. Reuters journalists take technical publications and report the main findings and methods and, on occasion, include interviews with the authors of the scientific publication. The stories are targeted at a lay audience with a 12th-grade reading level. Furthermore, every story in our collection contains a reference to the original scientific publication. Thus, it is possible to gather the original texts, which convey similar information but were written for a technical audience. The stories draw upon studies from reputable medical journals, such as Annals of Internal Medicine, New England Journal of Medicine and Lancet.

The technical collection in our corpus is composed of the original scientific articles corresponding to each news story in the lay collection. Accordingly, the lay and technical collections contain the same number of documents and are parallel at the document level. That is, each technical document has a lay equivalent and vice-versa. Because a lay document is a summary of a technical article and is, hence, much shorter than the original scientific article, we decided to include only the abstract of the technical document in our collection. This way, the technical and lay documents are comparable in content and length. It should be noted, however, that the content in a technical/lay document pair is not parallel, but comparable (McEnery and Xiao, 2007): there is no natural sentence-to-sentence correspondence between the two texts. This is to be expected: technical abstracts contain many technical details, while lay stories, to provide background, introduce information entirely absent from abstracts. In addition, the lay stories drastically rearrange the order in which information is typically conveyed in technical abstracts. For these reasons, our corpus is not parallel at the sentence level and, thus, differs from other bilingual parallel corpora used in machine translation.

[2]http://www.reutershealth.com

	Words			Sentences		
	Min	Max	Avg	Min	Max	Avg
Technical	137	565	317	5	18	10
Lay	187	1262	444	6	42	15

Table 1: Statistics for the Technical and Lay collections. Each contains 367 documents.

To ensure that some significant number of terms appears with sufficient frequency in our corpus in order to induce equivalent pairs automatically, we focused on articles and stories in a single domain: cardiology. We identified the original scientific article manually, as the lay document only contains a reference, not an actual link. For this reason, only a relatively small amount of data could be collected: 367 pairs of documents (see Table 1 for statistics).

3 Methods

3.1 Data Processing

We focus in this paper on finding term equivalents when both terms are recognized by the UMLS. Thus, our first step in processing our collections is to identify terms as defined by the UMLS. Both collections are processed by our tool TermFinder (Teufel and Elhadad, 2002). Sentences are identified and the texts are tokenized and tagged with part-of-speech information. Noun phrases are identified with a shallow parser. Next, terms are identified by looking up the noun phrases in the meta-lexicon of UMLS for an exact match. Terms are tagged with their concept unique identifier (CUI) and a semantic type, both provided by UMLS. For our purposes, we only consider a subset of all the terms listed in UMLS, based on their semantic type. This is due to the fact that certain UMLS semantic types are unlikely to yield technical terms in need of simplification. As such, terms belonging to semantic types such as “Activity,” “Family Group” or “Behavior” were left untagged. Terms with semantic types such as “Disease or Syndrome” or “Therapeutic or Preven-

	Corresponding lay doc. contains *lay_term*	Corresponding lay doc. does not contain *lay_term*
Technical doc. contains *tech_term*	a	b
Technical doc. does not contain *tech_term*	c	d

Table 2: Contingency table for (*tech_term*, *lay_term*).

tive Procedure," on the other hand, were considered terms. For instance, both the terms *PTCA* and *percutaneous transluminal coronary angioplasty* have the same CUI C0002997, as they are considered synonyms by UMLS. The term *balloon angioplasty* has the CUI C0002996. Both C0002997 and C0002996 have the semantic type "Therapeutic or Preventive Procedure."

3.2 Contingency Table

We call (*tech_term*, *lay_term*) a term pair, where *tech_term* is a term occurring in one or more technical documents and *lay_term* is a term present in at least one of the corresponding lay documents.[3] For any such pair, we can compute a contingency table based on co-occurrence. Our definition of co-occurrence is slightly unusual: *tech_term* and *lay_term* co-occur in one document pair if *tech_term* appears at least once in the technical document and *lay_term* appears at least once in the corresponding lay document. Our unit of content is document frequency for a CUI, *i.e.*, the number of documents in which a given CUI appears. For instance, in our data, the contingency table for the term pair *(MI, heart attack)* shows the following counts: the document frequency of the CUI corresponding to *MI* in the technical collection is 98; the document frequency of the CUI corresponding to *heart attack* in the lay collection is 161. Among these documents, there are 84 technical/lay document pairs (out of the total of 367 paired documents) in which the CUI for *MI* occurs on the technical side and the CUI for *heart attack* occurs on the lay side. Hence, the contingency table for this term pair is, following the notations of Table 2: a = 84, b = 98-84 = 14, c = 161-84 = 77, and d = 367-98-161+84 = 192.

At this stage of processing, lexical terms are abstracted by their CUIs. We do this to maximize the possible evidence that two terms co-occur. For instance, the document frequency for *MI* in our technical collection is 20, while the document frequency for its corresponding CUI is 98. Section 3.7 describes how we proceed from identifying equivalent terms at the CUI level to finding lexical equivalents.

3.3 Gold Standard

To evaluate the validity of our approach, we collected all possible term pairs at the CUI level in our corpus (that is, all the term pairs for which a contingency table is computed). We then whittled this set down to those pairs where each CUI occurs in at least two documents. This resulted in 2,454 pairs of CUIs. We asked our medical expert, an internist in practice who interacts with patients on a daily basis, to indicate for each pair whether the terms were equivalent from a medical standpoint *in the context of communicating with a patient*.[4] An operational test for testing the equivalence of two terms is whether he would use one term for the other when talking to a patient. We indicated to our expert that the terms should be equivalent *out of context*. So, for instance, while the pair (myocardial infarction, complication) could be deemed equivalent in certain specific contexts, these terms are not generally considered equivalent. Table 3 shows examples of pairs annotated as semantic equivalents for lay readers.[5] The list of terms contained only the actual lexical terms and no information from the UMLS to avoid biasing our expert.

Out of the 2,454 CUI pairs provided to our medical expert, 152 pairs were labeled as equivalent. Out of the 152 pairs, only 8 (5.3%) had different semantic types. Interestingly, 84 pairs (55.3%) had different CUIs. This confirms our intuition that the notion of semantic equivalence for lay readers is looser than for medically knowledgeable readers.

[3] This means that if *tech_term* and *lay_term* have no technical/lay document in common, *lay_term* is not considered a possible candidate for semantic equivalence for *tech_term*.

[4] While it is in some ways counterintuitive to rely on a technical expert to identify lay semantic equivalents, this expertise helps us validate equivalences from a medical standpoint.

[5] In the table, DIGN stands for "Diagnostic Procedure," DISS for "Disease or Symptom," FIND for "Finding," and PATH for "Pathological Finding."

Technical term	Lay term
myocardial infarction | C0027051 | DISS	heart attack | C0027051 | DISS
SBP | C0428880 | DIGN	systolic blood pressure | C0428880 | DIGN
atrial fibrillation | C0004238 | PATH	arrhythmia | C0003811 | PATH
hypercholesterolemia | C0020443 | DISS	high cholesterol | C0848569 | FIND
mental stress | C0038443 | DISS	stress | C0038435 | PATH

Table 3: Examples from the gold standard of term pairs considered equivalent.

3.4 Measures of Association

Given a term pair (*tech_term*, *lay_term*) and its corresponding contingency table, we want to determine whether *lay_term* is a valid semantic equivalent of *tech_term* from the standpoint of a lay reader. We rely on three alternative measures of association introduced in the Statistics literature: the χ^2 statistic, the λ measure, and odds ratio. All of these measures are computed as a function of the contingency table, and do not rely on any human labeling for equivalence. Measures of association have been used traditionally to identify collocations (Manning and Schütze, 1999). Here we investigate their use for building a lexicon.

3.4.1 The χ^2 Statistic

The standard chi-square statistic (χ^2) is used to determine whether the deviation of observed data from an expected event occurs solely by chance (Goodman and Kruskal, 1979). Our null hypothesis for this task is that the presence of *lay_term* in a lay document is independent of the presence of *tech_term* in its correspondent technical document. Thus, any pair of terms for which the χ^2 is above the critical value at a given level of significance are considered semantic equivalents. One important constraint for the measures to be valid is that the observed data be large enough (more than five observations per cell in the contingency table).

The χ^2 statistic for our 2x2 contingency table, and with N being the total number of document pairs, is calculated as follows:

$$\chi^2 = \frac{N(ad-bc)^2}{(a+b)(a+c)(c+d)(b+d)}$$

Since χ^2 is a true statistic, we can rely on critical values to filter out pairs with low associative power. In our case, we set the significance level at .001 (with a critical value for χ^2 of 10.83).

	C0011847	¬ C0011847	Sum
C0011849	a = 13	b = 8	21
¬ C0011849	c = 40	d = 306	346
Sum	53	314	367

Table 4: Contingency table for (C0011849, C0011847).

3.4.2 The λ and λ* Measures

The lambda measure (λ) assesses the extent to which we can predict the presence of *lay_term* in a lay document by knowing whether the original technical document contained *tech_term* (Goodman and Kruskal, 1979). λ is an asymmetrical measure of association. Since a lay document is always written based on an original technical document, it is a plausible assumption that the presence of a specific term in the technical document influenced the lexical choices of the author of the lay document. Thus, we consider the presence of *tech_term* in a technical document the antecedent to the presence of *lay_term* in the corresponding lay document, and, accordingly, operate in the setting of predicting the presence of *lay_term*.

We present the intuition behind λ in the context of the following example. Consider the contingency table for the technical CUI C0011849 (*diabetes mellitus*) and C0011847 (*diabetes*) in Table 4. The task is, given a random lay document, to predict which of two available categories it belongs to: either it contains the lay CUI (in our example, CUI C0011847 for *diabetes*) or it does not. There are two possible cases: either (1) we do not have any knowledge about the original technical document, or (2) we know the original technical document and, therefore, we know whether it contains the antecedent (in our example, CUI C0011849 for *diabetes mellitus*).

Without any prior knowledge (case (1)), the safest prediction we can make about the lay document is the category with the highest probabil-

ity. The probability of error in case (1) is $P_{err1} = \frac{N-Max(a+c,b+d)}{N}$.
In our example, the safest bet is $\neg$ C0011847, with a raw count of 314 documents, and a probability of error of $P_{err1} = 0.1444$.

If we have prior knowledge about the original technical document (case (2)), then our safest prediction differs. If we know that the technical document contains the CUI C0011849 (*diabetes mellitus*), then our safest prediction is the category with the highest probability: C0011847, with a raw count of 13 documents. If, on the other hand, we know that the technical document does not contain the CUI C0011849, our safest prediction is the category $\neg$ C0011847, with a raw count of 306 documents. Thus, overall the probability of error in case (2) is $P_{err2} = \frac{N-(Max(a,b)+Max(c,d))}{N}$.
In our example, knowledge about the original technical document lowers the probability of error to $P_{err2} = 0.1308$.

The λ measure is defined as the relative decrease in probability of error in guessing the presence of *lay_term* in a lay document $\lambda = \frac{P_{err1}-P_{err2}}{P_{err1}}$ which, using our notation for contingency tables, can be expressed as

$$\lambda = \frac{Max(a,b) + Max(c,d) - Max(a+c,b+d)}{N - Max(a+c,b+d)}$$

In our example, $\lambda = 0.094$. λ ranges between 0 and 1. A value of 1 means that knowledge about the presence of *tech_term* in the original technical document completely specifies the presence of *lay_term* in its corresponding lay document. A value of 0 means that knowledge about the presence of *tech_term* in the original technical document does not help in predicting whether *lay_term* is present in its corresponding lay document.

The λ measure is not a test of significance like χ^2. For instance, while two independent variables necessarily have a λ of 0, the opposite is not necessarily true: it is possible for two dependent variables to have a λ of 0. In our setting in particular, any contingency table where a=b will provide a λ of 0.

Since λ is computed as a function of maxima of rows and columns, λ can easily be biased toward the original proportions in the antecedent. In our example, for instance, a very large proportion of technical documents has no occurrence of C0011849, *diabetes mellitus* (94.3% of the technical documents). But for our purposes, such contingencies should not affect our measure of association, as the proportion of technical documents happening not to contain a particular term is just an artificial consequence of corpus collection. λ* is a variant of λ also proposed by Goodman and Kruskal (1979) and is able to take this fact into account. It is computed using the same formula as λ, but the elements of the contingency table are modified so that each category of the antecedent is equally likely. In our case, this means: N*=1, a*=0.5a/N(a+b), b*=0.5b/N(a+b), c*=0.5c/N(c+d), and d*=0.5d/N(c+d). Going back to our example of *diabetes mellitus* and *diabetes*, we now find $\lambda* = 0.324$, which is much higher than the original λ of 0.094, and which indicates a strong association.

We focus on λ* as a measure of association for semantic equivalence of term pairs. Since λ and λ* are not true statistics, there is no significance level we can rely on to set a threshold for them. Instead, we estimate an optimal threshold from the performance of λ* on a development set. The development set was obtained in the same manner as the gold standard and contains 50 term pairs. This is a small number of pairs, but the term pairs in the development set were carefully chosen to contain mostly semantically equivalent pairs. In our experiments, the optimal value for λ* was 0.3. Thus, λ* is used as a binary test for our purposes: *tech_term* and *lay_term* are considered semantically equivalent if their λ* is above 0.3.

3.4.3 Odds Ratio

Odds ratio is a measure of association that focuses on the extent to which one category in the contingency table affects another (Fleiss et al., 2003). For our contingency table, the odds ratio is expressed as follows:

$$OR = \frac{ad}{bc}$$

For instance, given the contingency table of Table 4, the odds ratio for the pair *(diabetes mellitus, diabetes)* is 12.43, which means that a lay document is 12.43 times more likely to contain the CUI C0011847, for *diabetes*, if its original technical document contains the term C0011849, for *diabetes mellitus*.

Like λ*, odds ratio is not a true statistic and, therefore, does not have any critical value for statistical significance. We estimated the optimal value of a threshold for OR based on the same development set described above. The threshold for OR is set to 6. Thus, OR is used as a binary test for our purposes: *tech_term* and *lay_term* are considered semantically equivalent if their OR is above 6.

3.5 Combining the Measures of Association

Each of the measures of association described above leverages different characteristics of the contingency tables, and similarly, each has its limitations. For instance, χ^2 cannot be computed when there are not sufficient observations, and λ* can equal 0, even when there is a strong association between the two terms. We combine measures of association in the following fashion: two terms are considered equivalent if at least one of the measures determined so.

3.6 Semantic Filtering

The measures of association described above and their combination provide information solely based on corpus-derived data. Since all our counts are based on co-occurrence, a measure of association by itself can encompass many types of semantic relations. For instance, the pair for (*stroke, brain*) tests positive with our three measures of association. Indeed, there is a strong semantic association between the two terms: strokes occur in the brain. These terms, however, do not fit our definition of semantic equivalence.

We rely on knowledge provided by the UMLS, namely semantic types, to help us filter equivalent types of associations among the candidate term pairs. One can assume that sharing semantic types is a necessary condition for semantic equivalence. Our semantic filter consists of testing whether *tech_term* and *lay_term* share the same semantic types, as identified by our tool TermFinder.

3.7 Lexical Choice

So far, term pairs are at the CUI level. The measures of association and the semantic filter provide a way to identify candidates for semantic equivalence. We still have to figure out which particular lexical items among the different lexical terms of a given CUI are appropriate for a lay reader. For instance, the pair (C0027051, C0027051) is considered a good candidate for semantic equivalence. In the technical collection, the lexical terms contributing to the CUI are *AMI*, *AMIs*, *MI*, *myocardial infarction*, *myocardial infarct* and *myocardial necrosis*. In the lay collection, however, the lexical terms contributing to the same CUI are *heart attack*, *heart attacks*, and *myocardial infarction*. Clearly, not all lexical items for a given CUI are appropriate for a lay reader.

To select an appropriate lay lexical term, we rely on the term frequency of each lexical item in the lay collection (Elhadad, 2006). In our example, the lexical term “heart attack” has the highest term frequency in the lay collection among all the variants with the same CUI. Thus, we chose it as a semantic equivalent of any lexical term of the CUI C0027051 in the technical collection.

If a technical term has several candidate semantic equivalents at the CUI level, the lexical lay term is chosen among all the lay terms. For instance, *(adverse effect, side effect)* and *(adverse effect, complications)* are two valid equivalents, but *side effects* has a term frequency of 16 in our lay collection, and *complications* has a lay term frequency of 35. Thus, *complication* is selected as the lay equivalent for *adverse effect*.

4 Results

We report on the two steps of our system: (1) finding semantic equivalents at the CUI level, and (2) finding an appropriate lay lexical equivalent.

Finding Semantic Equivalents at the CUI Level Table 5 shows the precision, recall and F-measure (computed as the harmonic mean between precision and recall) against our gold standard for the three alternative measures of association, including different combinations of these, and also adding the semantic filter. In addition, we report results for a competitive baseline based solely on CUI information, where *tech_term* and *lay_term* are considered equivalent if they have the same CUI.

The baseline is fairly competitive only because of its perfect precision (CUI in Table 5). Its recall, however (44.7), indicates that building a lexicon of technical and lay equivalents based solely on CUI information would miss too many pairs within the UMLS.

Method	P	R	F	Method	P	R	F	Method	P	R	F
lam	40.8	20.4	27.2	chi,odds	20.6	78.3	32.6	CUI	100	44.7	61.8
chi	38.7	23.7	29.4	chi,lam,odds	20.6	80.3	32.8	sem,odds	57.8	71.1	63.7
sem,lam	76.3	19.1	30.5	sem,chi	81.8	23.7	36.7	sem,lam,odds	57.4	73.7	64.6
odds	20.4	74.3	32	chi,lam	38.2	39.5	38.8	sem,chi,odds	58.5	75	65.7
lam,odds	20.5	77	32.3	sem,chi,lam	79.5	38.2	51.6	**sem,chi,lam,odds**	**57.9**	**77**	**66.1**

Table 5: Precision, Recall and F measures for different variants of the system.

Relying on only one measure of association without any semantic filtering to determine semantic equivalents is not a good strategy: λ* (lam in Table 5), χ^2(chi) and OR (odds), by themselves, yield the worst F measures. Interestingly, the measures of association identify different equivalent pairs in the pool of candidate pairs. Thus, combining them increases the coverage (or recall) of the system. For instance, λ* by itself has a low recall of 20.4 (lam). When combined with OR, it improves the recall from 74.3 (odds) to 77 (lam,odds); when combined with χ^2, it improves the recall from 23.7 (chi) to 39.5 (chi,lam). Combining the three measures of association (chi,lam,odds) yields the best recall (80.3), confirming our hypothesis that the measures are complementary and identify pairs with different characteristics in our corpus.

While combining measures of association improves recall, the semantic filter is very effective in filtering inaccurate pairs and, therefore, improving precision: λ*, for instance, improves from a precision of 40.8 (lam) to 76.3 (sem,lam) when the filter is added, with very little change in recall. The best variant of our system in terms of F measure is, not surprisingly, combining the three measures of association and adding the semantic filter (sem,chi,lam,odds in Table 5).

The results of these experiments are surprisingly good, considering that the contingency tables are built from a corpus of only 367 document pairs and rely on document frequency (not term frequency). These quantities are much smaller than those used in machine translation, for instance.

Finding Lay Lexical Equivalents We evaluate our strategy for finding an appropriate lay lexical item on the list of 152 term pairs identified by our medical expert as semantic equivalents. Our strategy achieves an accuracy of 86.7%.

5 Related Work

Our work belongs to the field of paraphrase identification. Much work has been done to build lexicons of semantically equivalent phrases. In generation systems, a lexicon is built manually (Robin, 1994) or by relying on an electronic thesaurus like WordNet (Langkilde and Knight, 1998) and setting constraints on the type of accepted paraphrases (for instance, accepting only synonyms as paraphrases, and not hypernyms). Building paraphrase lexicons from a corpus has also been investigated. Jacquemin and colleagues (1997) identify morphological and syntactic variants of technical terms. Barzilay and McKeown (2001) identify multi-word paraphrases from a sentence-aligned corpus of monolingual parallel texts. One interesting finding of this work is that the mined paraphrases were distributed across different semantic links in WordNet: some paraphrases had a hypernym relation, while others were synonyms, and others had no semantic links at all. The composition of our gold standard confirms this finding, since half of the semantically equivalent terms had different CUIs (see Table 3 for examples of such pairs).

If we consider technical and lay writing styles as two sublanguages, it is easy to see an analogy between our task and that of machine translation. Identifying translations for words or phrases has been deeply investigated in the field of statistical machine translation. The IBM models of word alignments are the basis for most algorithms to date. All of these are instances of the EM algorithm (Expectation Maximization) and rely on large corpora aligned at the sentence level. We cannot apply an EM-based model to our task since we have a very small corpus of paired technical/lay documents, and EM requires large amounts of data to achieve accurate results. Moreover, the technical and lay documents are not parallel, and thus, we do not have access to a sentence alignment. Of course, our task is easier than the one of machine translation, since we focus on "translating" only technical terms, rather than every single word in a technical document.

Gale and Church (1991) do not follow the EM model, but rather find French translations of English words using a χ^2-like measure of association. Their corpus is the parallel, sentence-aligned Hansard corpus. Our method differs from theirs, as we do build the contingency table based on document frequencies. Gale and Church employ sentence-level frequencies. Our corpus is much smaller, and the sentences are not aligned (for comparison, we have 367 document-pairs, while they have nearly 900,000 sentence pairs). Another difference between our approach and theirs is our use of the semantic filter based on UMLS. We can afford to have such a filter because we focus on finding semantic equivalents of UMLS terms only.

6 Conclusions and Future Work

We presented an unsupervised method for identifying pairs of semantically equivalent technical/lay terms. Such a lexicon would benefit research in health literacy. In particular, it would benefit a system which automatically adapts a medical technical text to different levels of medical expertise.

We collected a corpus of pairs of technical/lay documents, where both documents convey similar information, but each is written for a different audience. Based on this corpus, we designed a method based on three alternative measures of association and a semantic filter derived from the UMLS. Our experiments show that combining data-driven statistics and a knowledge-based filter provides the best results.

Our method is concerned specifically with pairs of terms, as recognized from UMLS. While UMLS provides high coverage for technical terms, that is not the case for lay terms. In the future, we would like to extend our investigation to pairs consisting of a technical term and any noun phrase which is sufficiently frequent in our lay collection. Finding such pairs would have the side effect of augmenting UMLS, a primarily technical resource, with mined lay terms. One probable step towards this goal will be to increase the size of our corpus of paired technical and lay documents.

References

R. Barzilay and K. McKeown. 2001. Extracting paraphrases from a parallel corpus. In *Proc. ACL'01*, pages 50–57.

N. Elhadad. 2006. Comprehending technical texts: Predicting and defining unfamiliar terms. In *Proc. AMIA'06*, pages 239–243.

J. Fleiss, B. Levin, and M.C. Paik. 2003. *Statistical Methods for Rates and Proportions*. Wiley.

W. Gale and K. Church. 1991. Identifying word correspondences in parallel texts. In *Proc. Speech and Natural Language Workshop*, pages 152–157.

L. Goodman and W. Kruskal. 1979. *Measures of Association for Cross Classifications*. Springler Verlag.

C. Jacquemin, J. Klavans, and E. Tzoukermann. 1997. Expansion of multi-word terms for indexing and retrieval using morphology and syntax. In *Proc. ACL'97*, pages 24–31.

I. Langkilde and K. Knight. 1998. Generation that exploits corpus-based statistical knowledge. In *Proc. COLING-ACL'98*, pages 704–710.

C. Manning and H. Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press.

A. McCray. 2005. Promoting health literacy. *JAMA*, 12(2):152–163.

A. McEnery and Z. Xiao. 2007. Parallel and comparable corpora: What is happening? In *Incorporating Corpora. The Linguist and the Translator*. Clevedon.

National Library of Medicine, Bethesda, Maryland, 1995. *Unified Medical Language System (UMLS) Knowledge Sources*. http://www.nlm.nih.gov/research/umls/.

J. Robin. 1994. *Revision-Based Generation of Natural Language Summaries Providing Historical Background*. Ph.D. thesis, Columbia University.

R. Rudd, B. Moeykens, and T. Colton. 1999. *Annual Review of Adult Learning and Literacy*, chapter 5. Health and literacy: a review of medical and public health literature. Jossey Bass.

S. Teufel and N. Elhadad. 2002. Collection and Linguistic Processing of a Large-scale Corpus of Medical Articles. In *Proc. LREC'02*, pages 1214–1218.

Q. Zeng, E. Kim, J. Crowell, and T. Tse. 2005. A text corpora-based estimation of the familiarity of health terminology. In *Proc. ISBMDA'05*, pages 184–192.