Extracting variant forms of chemical .names .for information retrieval

Ari Pirkola
Department of Information Studies, University of Tampere, 33014 Tampere, Finland

Abstract

Introduction. Chemical substance .names are long, complex and prone to variation. . This study investigates the retrieval effects of the variation. .
Method. A+ .large set of acronyms and associated text parts was extracted from a subset of the Medline collection .and used to construct a full name - acronym index. A+ .longest common subsequence and statistics .based technique .(named FNV-Finder) was devised .to identify MeSH term .variants from the full name - acronym index for use as query terms in searching. The average number of variants for each MeSH term, . the performance .of the FNV-Finder technique .and retrieval performance .were evaluated.
Results. The average number of unique variants for each MeSH term .denoting a chemical .substance .is 2. .82. The FNV-Finder technique .achieved 95.0% recall and 97.1% precision. The retrieval experiments showed that the collection .contains a substantial number of documents .that contain only variant forms of the MeSH terms (and do not contain the MeSH terms or CAS registry numbers).
Conclusions. The selection .of variant forms for queries from a collection .would be very useful or even necessary in chemical .name searching. Variant forms can be selected readily from the full name - acronym index either manually or automatically using the FNV-Finder technique. .

Introduction

We investigate .the variation .of *chemical substance .names* and the retrieval effects of the variation. . Chemical names .pose a special challenge in information retrieval since they typically are long and complex expressions, . being thus prone to variation, . which in turn may cause .a decrease in retrieval performance .due to a mismatch between query terms and index terms.

A new technique .named *FNV-Finder* (where FNV stands for Full Name Variant) was developed to automatically identify the variant forms. Articles discussing a given chemical .substance .often use a canonical pattern where a full name is followed .by an acronym in parentheses, e.g., *N-methyl pyrrolidone (NMP)*. The FNV-Finder technique .is based on the fact that different variant forms of the same full name share the same acronym. The acronym is used as a pivot to find the variants of the same name. We extracted all the canonical patterns from a subset of the Medline collection, . i.e., TREC 2003 Genomics Track collection .containing some 525, 000 documents .(Hersh and Bhupatiraju 2004). We constructed a full name-acronym index which contains the extracted acronyms and associated text parts and where the acronyms are arranged in an alphabetical order. The index allows an efficient means of identifying the full names .of acronyms both manually and automatically. The FNV-Finder technique .uses a similarity measure (longest common subsequence, LCS) between an acronym and string sequences associated with it in the full name-acronym index and statistical .data contained .in the full name-acronym index to identify the full names .of acronyms and the variant forms of the same full name.

As test .data we used a set of chemical .acronyms and their MeSH (Medical Subject Headings) terms, for which variant forms were identified from the full name-acronym index manually and automatically using the FNV-Finder technique. . Using .these data we investigate .the following research problems:

1. What is the average number of variants for a MeSH term .denoting a chemical .substance?
2. How to effectively identify automatically different variant forms of a MeSH term? For this research problem we devised .the LCS and statistics .based FNV-Finder technique, . which identifies the variants from the full name-acronym index for use as query terms in chemical .name searching.
3. What are the recall and precision of the proposed FNV-Finder technique?
4. Does chemical .name searching benefit from using variant forms in queries?

The main .contributions .of this paper are: to present a novel .technique (FNV-Finder) to identify chemical .name variants from a collection, . and to present evaluation results for the approach; to demonstrate how to effectively organize the full name-acronym patterns contained .in the collection .(the construction of the full name-acronym index); and to report the effects of chemical .name variation .in information retrieval.

Methods and data

Test collection .and test .words

The test .collection .used in the study was the TREC 2003 Genomics Track collection, . which is a subset of the Medline collection. . The test .collection .contains some 525,000 article abstracts, which were indexed between January 2002 and January 2003. Medline's documents .are indexed with the National Library of Medicine's Medical Subject Headings (MeSH). Chemical names .are also indexed with Chemical Abstract Service registry numbers.

A Medline record consists of several fields of which the fields TI (title), AB (abstract), MH (MeSH terms), and RN (Chemical Abstract Service Registry Number) were indexed for the retrieval experiments conducted in this study.

The full names .of chemical .acronyms in the full name-acronym index are often terms that are contained .in the Medical Subject Headings vocabulary: they are either MeSH terms or so-called .substance .names, or the full names .in the full name-acronym index are their variants. In this study, variation .is considered from the viewpoint of the MeSH terms and the substance .names. A+ .MeSH term .or substance .name is regarded as a standard full name of an acronym while the full names .that denote the same chemical .substance .as the MeSH term .or substance .name and that are similar to it but written differently are regarded as variant forms. For example, *hydroxyethyl methacrylate* is a substance .name and *hydroxyethylmethacrylate* and *2-hydroxyethyl methacrylate* are its variant forms. As can be seen, the latter two names .are similar, but not identical, . to *hydroxyethyl methacrylate*. Below are more examples of variant forms. (For simplicity, in this paper *MeSH term* refers .both to the actual .MeSH terms and the substance .names contained .in the MeSH.)

From the viewpoint of the research problems examined in this study we differentiate between three types of names .that denote a chemical .substance: a MeSH term, . its variant forms, and an acronym that refers .to the MeSH term .and the variant forms. We do not study orthographic variation .(see below) but *lexical variation*. Here lexical .variation means that the MeSH term .and a variant form denote the same chemical .substance .and are similar, but there are differences in components, letters, or numbers. We can distinguish .several types of variant cases: the MeSH term .and its variant form have one or more common components but differ from each other in the number or the order of the components (e.g., MeSH term *.inosine monophosphate* and a variant form *inosine 5 .monophosphate*); the components are written together (i.e., as a compound word) vs. the components are written separately (i.e., as a phrase) (e.g., MeSH term *.carboxymethylcellulose* and a variant form *carboxymethyl cellulose*); the corresponding components in the MeSH term .and a variant form differ (typically) in one or two characters (e.g., MeSH term *.1 methyl 2 .pyrrolidinone* and a variant form *1 methyl 2 .pyrrolidone*); a component that appears in a variant may be an abbreviation of a component that appears in the MeSH term, . or the other way around (e.g., MeSH term *.methyl tert butyl ether* and a variant form *methyl t butyl ether*). All these cases affect information retrieval. For example, it is obvious that query terms *carboxymethylcellulose* (compound) and *carboxymethyl cellulose* (phrase) give different retrieval results.

Since we examine lexical .variation rather than orthographic variation .we performed orthographic normalisation in documents, . queries, and in the full name-acronym index: other characters than letters and numbers were replaced with spaces; case was normalised into lower case. As an example of orthographic normalisation, the substance .name *N-Formylmethionine Leucyl-Phenylalanine* was converted into a normalised form of *n formylmethionine leucyl phenylalanine*. In a retrieval phase, phrasal names .were searched for using the ordered window proximity operator of the InQuery retrieval system (Allan *et al* . 2000) which was used as a test .system in the retrieval experiments. The benefits of orthographic normalisation in information retrieval are obvious and it is commonly used in retrieval systems .even though it may create ambiguity. However, in this study orthographic normalisation did not affect performance .(the issue of wrong .identifications by FNV-Finder is discussed in the Findings section).

We used a training set of fifty acronyms (see Appendix .1) to devise the FNV-Finder technique .and to set the thresholds to get the best possible results. The test .acronyms were taken randomly from the full name-acronym index; every Nth acronym was selected iteratively until fifty substance .name acronyms for which there were both MeSH terms and Registry Numbers were obtained. From the original set of 55 acronyms five were removed since there were no MeSH terms or Registry Numbers for them. In the case of ambiguous chemical .acronyms having more than one MeSH term, . the first term .was selected for the tests. .

The number of the test .acronyms is relatively small because of the time-consuming manual identification of variants and document relevance assessments , which are necessary for reliable .results. It should be noted, however, that using this test .data we are able to report statistically significant results in retrieval experiments.

As an aid in manual name identification we used printed reference .books in chemistry .and the National Library of Medicine's ChemIDplus Lite Web service. In some (rare) cases it was not possible determine which string sequence .in the full name-acronym index was acronym's full name. In these cases we looked at the document from where the text part was extracted and used document's text to determine the correct full name.

Identification of full name variants

Constructing the full name-acronym index

The full name-acronym index was constructed by extracting from the Medline test .collection .all strings that consisted .of letters, numbers or both letters and numbers and that were located in parentheses. For each such string a text part of the length .of up to nine strings prior to it was also extracted. Strings inside parentheses containing .other characters than letters or numbers were not included in the full name-acronym index. The first version of the index was cleaned: strings within parentheses that only contained .numbers and associated text parts were removed.

To take an example of the extraction phase, consider .a Medline record containing .the following phrases: *end-stage renal disease (ESRD) were glomerulonephritis (26%)*, and, *with antithymocyte globulin (ATG) or orthoclone thymocyte 3 .(OKT3)*. The strings *ESRD*, *ATG*, and *OKT3* were included in the full name-acronym index while the string "26%", which does not contain letters and which contains a percentage sign was not included.

The index was arranged in an alphabetical order. The final index contains 479,882 lines. . Most of the strings inside parentheses are not acronyms but noise in the context of this study. However, such strings do not play any role in variant identification and thus do not have any effects on the FNV-Finder effectiveness or retrieval results.

An *index entry* contains all the extracted instances of an acronym *Ai* (with *Ai* denoting any acronym in the index) and the associated text parts. Let us take an example of an index entry. For the acronym MTBE part of the entry is as follows (here we only present seven .lines, . the full entry includes fifty-one lines):

entry: mtbe
etbe t butyl alcohol tba methyl t butyl ether mtbe
mtbe from soil samples .ab methyl tert butyl ether mtbe
9 .isopropyl alcohol ipa in methyl tert butyl ether mtbe
a procedure for determination of methyl tert butyl ether mtbe
an increase in levels of methyl tertiary butyl ether mtbe
and in bus exposures to methyl tertiary butyl ether mtbe
butyl ether in water .ab methyl tert butyl ether mtbe

The string positions .in a line *.l* are numbered from $s_1(A_i(l))$ (the first position) to $s_n(A_i(l))$ (the last position immediately left to the acronym).

For the fifty test .acronyms, the number of lines .in entries .ranged from 1 .to 214. Overall, the fifty entries .contained 1, .595 lines. . Manual identification of variants was done using all 1, .595 lines .(the first step of FNV-Finder also used all 1, .595 lines).

FNV-Finder algorithm

FNV-Finder-based identification of acronyms' full name variants has six steps. Next we describe the FNV-Finder algorithm .which is presented in Appendix 2. It is important to note that in steps one to five no difference is made between MeSH terms and variant forms. Only in the last step variant forms are separated from the MeSH terms.

In the *first step*, all lines .in an entry that do not contain any of the components of a phrasal MeSH term, . either as a separate .string or as a string embedded in a compound word, are filtered out. Single word and compound word MeSH terms (e.g., *resiniferatoxin*) are divided .into non-overlapping 6-grams (with the exception of the last n-gram which may contain more than 6 .characters), which are used as a filter as in the case of phrasal MeSH terms. The experiments showed that this step removed most of the lines .that did not contain a MeSH term .or a variant form.

The *second step* is based on the observation that chemical .substance .names only very rarely contain function words and other so-called .stop words. Therefore, the rightmost stop word in each line .and all strings to the left of the rightmost stop word are removed from an index entry. Also the acronyms are removed at this stage. The stop word list .used was that of the retrieval system. As single letters are found in chemical .names, . they were removed from the list. . The list .was supplemented with the collection .specific abbreviations *ti* and *ab* (*ti* refers .to title and *ab* to abstract).

In the *third step*, the *longest common subsequence (LCS)* technique .(see for example Pirkola *et al*. 2002) is used to identify probable full names .which we call *full name candidates*. The algorithm .scans strings from right to left in each line .to find a string that starts with the same letter .as the acronym. If such string is found, it is marked as a *temporary first component (TFC)* of a full name. Next LCS is computed for the acronym and the string sequence .starting with the TFC and ending with $s_n(A_i(l))$. If LCS = |the number of characters in the acronym| the string sequence .is marked as a full name candidate and is passed to the fifth step. The lines .where full name candidates are not found are passed to the fourth step.

Numbers 0-100 as well as the strings *alpha, beta, gamma, tert, nalpha, d, l, n, p, r, s*, are often found as left components in chemical .substance .names (e.g., *1 6 .diphenyl 1 .3 5 .hexatriene*), and if such string or a combination .of such strings appears immediately left to the TFC it is included in the full name candidate.

In the experiments, the third .step correctly identified most of the full names .(note, at this stage they are still candidates). Below is an example of the second and third .steps. The sample .entry shown above is presented. The hyphen shows the location from which the left part of the entry was removed in the second step. The remaining part (called as a post-second-step entry in the following text) was processed in the third .step. The asterisk indicates the TFC. As the example shows the LCS method identifies the full names *.methyl tert butyl ether* (MeSH term), *methyl t butyl ether* and *methyl tertiary butyl ether* (variant forms).

entry: mtbe
etbe t butyl alcohol tba *methyl t butyl ether [LCS=4]
mtbe from soil samples .ab- *methyl tert butyl ether [LCS=4]
9 .isopropyl alcohol ipa in- *methyl tert butyl ether [LCS=4]
a procedure for determination of- *methyl tert butyl ether [LCS=4]
an increase in levels of- *methyl tertiary butyl ether [LCS=4]

and in bus exposures to- *methyl tertiary butyl ether [LCS=4]
butyl ether in water .ab- *methyl tert butyl ether [LCS=4]

The *fourth step* handles the lines .which the third .step could not solve. . A+ .frequency-based *FNV indicator* value is computed for each string in a post-second-step entry using a *FNV-Finder computation .scheme*. The idea is that the string that appears frequently .in an entry and that is the leftmost string among the frequent strings in a line .is likely to be a start component of a full name.

The FNV-Finder computation .scheme computes a FNV indicator value for the string $s_k(A_i)$ as follows:

$$Fr(s_k(A_i)) / \ln(N(A_i))$$

$Fr(s_k(A_i))$ = frequency of the string s_k in the entry of an acronym A_i.
$N(A_i)$ = total number of strings in the entry of an acronym A_i.

The frequency of s_k is divided .by the total number of strings N in an entry to get figures .that are in proper proportion between different entries. . Similarly the natural logarithm (ln) in a denominator equalizes figures .between different entries. .

FNV indicator values .are computed for each string in a post-second-step entry. The leftmost string whose value is above a set threshold (the value of 0. .50 was used based on the training tests) is marked as a TFC. As in the third .step, the whole string sequence .from the TFC to the $s_n(A_i(l))$ is marked as a full name candidate. Similarly to the third .step, numbers 0-100 and the strings *alpha, beta, gamma, tert, nalpha, d, l, n, p, r, s*, and combinations of them are included in the full name candidates (also in cases where the indicator value of such string is =< 0. .50). The lines .which the fourth step does not solve .are rejected.

To take an example of full name identification based on the FNV-Finder computation .scheme, consider .the post-second-step entry of the acronym NMP:

entry: nmp
n methyl 2 .pyrrolidone [LCS=3]
1 .methyl 2 .pyrrolidinone
n methyl pyrrolidone [LCS=3]
1 .methyl 2 .pyrrolidone

The LCS method of the third .step found the names *.n methyl 2 .pyrrolidone* and *n methyl pyrrolidone* (which are variant forms) in lines .1 .and 3. . But it did not find the names *.1 methyl 2 .pyrrolidinone* (MeSH term) in the second line .nor *1 methyl 2 .pyrrolidone* (a variant form) in the fourth line, . whereas the FNV-Finder computation .scheme did so based on the indicator value of 0. .74 (> 0. .50) of the string *1*.

The *fifth step* checks if there is a contradiction between the endings of the full name candidate and the MeSH term. . In other words, this step checks whether they designate different types of chemical .substances. . The substance .types are identified by means of a suffix list .built in the study on the basis of a set of substance .names extracted from the full name-acronym index. The list .contains core word endings (n=21) typical .of different types of substance .names. The end parts of the MeSH term .and the candidate are matched against the suffix list. . If the end parts match different suffixes in the list .(e.g., one is *amide* and matches *-ide*, and the other is *amine* and matches *-ine*), the candidate is rejected. Otherwise, the candidate is passed to the sixth step.

In the *sixth step*, the candidates are mapped to the MeSH terms to see if they are MeSH terms or variants. A+ .full match indicates a MeSH term .while a mismatch indicates a variant form.

Evaluation

FNV-Finder effectiveness (research problem three) was evaluated using the measures of *recall* and *precision*. In both cases we consider *.unique* variants (as opposed .to variant occurrences). Recall is defined as the proportion of variants FNV-Finder identified correctly from the full name-acronym index to intellectually identified variants from the full name-acronym index. Recall:

|Correct variants identified by FNV-Finder| / |Intellectually identified (correct) variants|

The intellectually identified variants thus constitute a recall base for variants identified by FNV-Finder. A+ .full match between a FNV-Finder variant and an intellectually identified variant is considered a correct identification. A+ .partial .match and a mismatch are considered wrong .identifications.

Precision is defined as the proportion of correctly identified variants by FNV-Finder among all entities FNV-Finder identified as variants. Precision:

|Correct variants identified by FNV-Finder| / |All entities identified by FNV-Finder|

In this study, a *relevant document* is defined as a document that contains at least one of the following items that denote the chemical .substance .in question: the MeSH term, . the Chemical Abstracts Service registry number, a variant form. In retrieval experiments (research problem four) we ran the following three queries:

MeSH+Registry Number (hereafter, RN) . Baseline. These queries (n=50) contained .the MeSH terms and registry numbers, and they serve as baseline for the following test .queries. For each acronym, they retrieved documents .that contained .the MeSH term .or the registry number (or both of these). By definition, all the documents .retrieved by MeSH+RN queries are relevant documents. .

MeSH+RN+manual-variant . These queries (n=50) contained .the MeSH terms, registry numbers, and the manually

identified variants. For each acronym, they retrieved documents .that contained .at least one of the following items: the MeSH term, . the registry number, at least one of the manually identified variants of the MeSH term. . The substance .names selected manually from the full name-acronym index were all correct names .and these queries were expected .to retrieve only relevant documents .(see below for relevance assessments).

MeSH+RN+automatic-variant . These queries (n=50) contained .the MeSH terms, registry numbers, and the variant forms identified by the FNV-Finder technique. . For each acronym, they retrieved documents .that contained .at least one of the following items: the MeSH term, . the registry number, at least one of the automatically identified variants of the MeSH term. . Due to the automatically identified variant forms these queries retrieved also irrelevant documents. .

We computed recall for MeSH+RN, MeSH+RN+manual-variant and MeSH+RN+automatic-variant queries, and precision for MeSH+RN+automatic-variant queries. These measures provide an answer to the research question whether, and to what extent, substance .name searching benefits from using variant forms.

The relevance of the documents .containing only variant forms was assessed, firstly, to ensure that the manually identified variants retrieved only relevant documents, . so that reliable .recall values .are obtained for MeSH+RN+manual-variant queries, and secondly, to compute recall and precision for MeSH+RN+automatic-variant queries.

The documents .retrieved by the MeSH+RN+manual-variant queries constitute a recall base for the 50 queries. We are aware that the recall base is not complete, in particular due to the synonyms of MeSH terms. Therefore, we measure *relative recall*, in other words, recall relative to the set of known relevant documents, . i.e., documents .retrieved by MeSH+RN+manual-variant queries. (By the term .*synonym* we mean a term .that designates the same chemical .substance .as the MeSH term .but is dissimilar to it.)

A chemical .name may be embedded in another, longer name. Therefore prior to the runs we checked .the MeSH and ChemIDplus Lite databases to see which MeSH terms and variants in our data appeared .in longer names. . For such terms and variants we constructed a Boolean .negation-based (InQuery's #bandnot-operator) query which excluded wrong .names from the results. It should be noted that this type of query modification is not a feature .of FNV-Finder based retrieval, but to obtain .good results, modification is necessary in many cases. A+ .drawback of this approach that some relevant documents .may be lost. . However, we analysed a sample .set of such documents, . and the analysis results suggest .that only a fraction of relevant documents .are lost. .

Findings

Table 1 .presents the results of the first research problem, i.e., the frequency of variant forms. It can be seen that the average number of unique variants per a MeSH term .is 2. .82.

Table 2 .shows the distribution .of the variants between the 50 MeSH terms. The MeSH terms were divided .into groups on the basis of the number of variants for each term. . The left column shows the cardinality of a MeSH term .group, and the right column the number of variants in a group. Table 2 .shows that out of the 50 MeSH terms 39 have variants. 32 terms have 1-4 variants. One term .has 10, one has 13, and one (*n formylmethionine leucyl phenylalanine*) has even 18 variants. These three terms with many variants were long terms. Of the 11 terms that did not have variants 10 were single words, compound words, or phrases with only two components. Thus, the results suggest .a general trend: the longer a MeSH term .is the more variant forms it has.

Table 3 .reports the results of the FNV-Finder effectiveness experiments. It can be seen that FNV-Finder performed very well - it achieved .95.0% recall and 97.1% precision. These figures .indicate that there were only a few variant forms that FNV-Finder could not identify, and only a few wrong .identifications.

Table 4 .reports the results of the retrieval experiments. As shown, MeSH+RN queries retrieved 4, .110 documents. . MeSH+RN+manual-variant queries retrieved 4, .468 documents, . and MeSH+RN+automatic-variant queries retrieved 4, .456 *relevant* documents. . For MeSH+RN+manual-variant queries the difference (third column) is 358 documents. . This means that - for the 50 queries - the collection .contains 358 documents .that contain only variant forms (and do not contain the MeSH terms or RNs). For MeSH+RN+automatic-variant queries the difference is 346 documents. . For MeSH+RN+manual-variant queries we selected only correct variants, and all the retrieved documents .were relevant.

Both for MeSH+RN+manual-variant and MeSH+RN+automatic-variant queries recall was improved (with respect to MeSH+RN queries) for 29 queries out of the 50 queries. The statistical .significance was analysed by one-tailed non-parametric Wilcoxon signed rank test. . For both query types the results were statistically significant at the level of $p < 0.0001$.

MeSH+RN+automatic-variant queries yield a high precision, i.e., 99.8%. They retrieved 4, .467 documents, . of which 11 were irrelevant documents. . In the case of an ambiguous acronym GDP (*guanosine diphosphate* in our data) FNV-Finder gave a wrong .name *geranyl diphosphate*, and all the irrelevant documents .dealt with this substance. . Other wrong .identifications by FNV-Finder were too short and too long sequences, but they did not affect retrieval results. In particular short sequences may be useful query terms though sometimes their use may decrease precision.

.solid.rgb(153, 245, 251); font-size: smaller; font-style: normal; font-family: verdana,geneva,arial,helvetica,sans-serif; background-color: rgb(253, 255, 221);" align="center" border="1" . cellpadding="5" . cellspacing="0" . width="60%">

Condition	Number/Average Number
Number of MeSH terms	50
Number of unique variants	141
Number of unique variants per a MeSH term	2.82

Table 1. . The frequency of variants.

.solid.rgb(153, 245, 251); font-size: smaller; font-style: normal; font-family: verdana,geneva,arial,helvetica,sans-serif; background-color: rgb(253, 255, 221);" align="center" border="1" . cellpadding="5" . cellspacing="0" . width="60%">

Number of MeSH terms in a group	Number of variants
11	0
11	1
8	2
6	3
7	4
2	5
1	8
1	9
1	10
1	13
1	18

Table 2. . Distribution of variants between MeSH terms.

.solid.rgb(153, 245, 251); font-size: smaller; font-style: normal; font-family: verdana,geneva,arial,helvetica,sans-serif; background-color: rgb(253, 255, 221);" align="center" border="1" . cellpadding="10" cellspacing="0" . width="60%">

Recall/Number of instances	Recall%
134/141	95.0
Precision/Number of instances	**Precision%**
134/138	97.1

Table 3. . FNV-Finder effectiveness.

.solid.rgb(153, 245, 251); font-size: smaller; font-style: normal; font-family: verdana,geneva,arial,helvetica,sans-serif; background-color: rgb(253, 255, 221);" align="center" border="1" . cellpadding="10" cellspacing="0" . width="80%">

Query type n=50	Relevant Retrieved	2-1 / 3-1	Recall%	All Docs. Retrieved	Precision%
1.MeSH+RN, baseline	4110	-	92.0	4110	(100)
2.MeSH+RN+manual-variant	4468	358	100.0	4468	(100)
3.MeSH+RN+automatic-variant	4456	346	99.7	4467	99.8

Table 4. . Retrieval performance .of test .queries.

Related work

Larkey *et al* . (2000) extracted from the Web a large collection .of acronyms and their expansions (full names, . definitions). They developed and evaluated several methods to extract acronyms' expansions. The corpus of the acronyms and expansions built on the basis of the best method was comparable to a high-quality hand-crafted Web site providing acronyms and expansions. Other pattern matching methods for extracting the full names .of acronyms are presented in (Schwartz and Hearst 2003; Terada *et al* . 2004; and Yu *et al* . 2002).

Liu and Friedman (2003) developed a statistics-based method that associates .abbreviations with their expansions. For example, the acronym ER is linked to the expansion .*estrogen receptor* in a text containing .a parenthetical expression .*estrogen receptor (ER)*. Text parts containing .parenthetical expressions .were first extracted from documents. . Potential collocations were then generated and their frequencies were computed. The method uses the number of elements in a potential collocation and the ratios .of the frequencies of potential collocations to eliminate words co-occurring by chance. For example, for *include pneumonia* versus *pneumonia* the frequency ratio is low, and *include pneumonia* is eliminated. . The method achieved .96.3% precision and 88.5% recall.

The FNV-Finder technique .proposed in this paper differs .from the above studies in that it identifies variant forms of the same name. The above studies focused on the question of the demarcation of the full name (expansion) boundaries in text.

It seems that chemical .structure and substructure searching (e.g., Willett 2000) has been studied more than chemical .name searching which seems to be a relatively unexplored area. Research on chemical .name processing started several decades ago (see for example, Vander Stouw *et al*. 1976). However, there are few studies that have looked at how different techniques and factors affect retrieval performance. . Wren (2006) used a first-order Markov Model to evaluate its ability to recognize chemical .terms and reject non-chemical terms. The method achieved .93% recall and 99% precision. Based on the results of a small scale information retrieval experiment .the researcher concluded .that chemical .name variation .affects information retrieval. This is in agreement .with the results of this study.

Luque Ruiz *et al* . (1996)developed an approximate .string matching algorithm .for repairing lexical .errors in inorganic chemical .names. . The National Library of Medicine has several powerful tools .to process biological texts .(Aronson 1994; Wilbur *et al*. 1999). Lexical Variant Generator and MetaMap generate lexical .variants for words appearing in texts. . Wilbur *et al* . (1999) explored the problem of recognizing chemical .terms and tested three methods: a lexical .method where chemical .terms were analysed into their constituent chemical .morphemes, and two statistical .methods based on the Bayesian .classifier. The

evaluation results showed that one of the statistical .methods achieved .the best results, an overall classification .accuracy of 97%.

Discussion and conclusions

We found in this study that the average number of unique variants for a chemical .MeSH term .is 2. .82. The retrieval experiments showed that the collection .contains a substantial number of documents .that contain only MeSH term .variants (and do not contain the MeSH terms or RNs). The FNV-Finder technique .effectively identified variant forms, and it achieved .a high recall and precision. The findings .of the retrieval experiments are in agreement .with the FNV-Finder effectiveness experiments. The use of the variant forms identified by the FNV-Finder technique .improved the recall of queries with respect to the recall of queries that only contained .MeSH terms and RNs, and only slightly decreased precision.

MeSH and the Chemical Abstracts Service registry number system are the main .terminological .resources in chemical .name searching. However, in exhaustive retrieval where the aim is to retrieve a large set of documents .discussing a given chemical .substance, . as well as some other retrieval situations, the selection .of additional names .(variant forms) for queries either manually from a full name-acronym index constructed from a collection .or automatically using the FNV-Finder technique .would be very useful or even necessary.

It seems that the proposed FNV-Finder technique .could be used effectively for name variant identification in any domain .where names .are expressed using the full name - acronym pattern. Naturally, the domain-specific components of the technique .(e.g., suffix lists) need to be constructed according to the selected domain. . The proposed technique .could also be used in other areas than information retrieval (e.g., in information extraction), and for other languages .than English. . The next step in the FNV-Finder research will be the application of the technique .in the Web environment. . We have crawled large amount of data from the Web which will be used to construct Web based full name-acronym indexes. In addition to English, . we plan to construct Web based full name-acronym indexes for French, . German, . and Spanish. These will be used to identify the full names .of chemical .acronyms in Web pages, and also other domains will be investigated. . We would also like to investigate .whether the technique .could be applied in the area of converting chemical .names .into structures.

Acknowledgements

This research was funded by the Academy of Finland (project names .*NLP-based information retrieval systems .for the biological literature* and *Focused retrieval of Web documents* .).

The InQuery search .engine was provided by the Center for Intelligent Information Retrieval at the University of Massachusetts.

.navy; border-bottom: 1px solid .navy; padding: 0. .1ex 0. .5ex; color: white; background-color: #5E96FD;; font-size: medium; font-weight: bold;">References

- Allan, J++. ., Connell, M.E., Croft, W.B., Feng, F.-F, Fisher, D. & Li, X++. . (2000). InQuery and TREC-9. In *Proceedings of the Ninth Text REtrieval Conference (TREC-9)*. (pp. 551-600). Gaithersburg, MD: National Institute of Standards and Technology. (NIST Special Publication 500-249) Retrieved .13 August, 2008 from http://trec.nist.gov/pubs/trec9/papers/umass-trec9.pdf (Archived by WebCite® at http://www.webcitation.org/5a2pwybjy)
- Aronson, A+. .R. (1994). *Comparison of LVG and MetaMap functionality* . Bethesda, MD: Lister Hill National Center for Biomedical Communications. Retrieved .13 August, 2008 from http://skr.nlm.nih.gov/papers/references/LVG-MetaMap.comparison. .pdf (Archived by WebCite® at http://www.webcitation.org/5a2qR12a7)
- Hersh, W. & Bhupatiraju, R++. .T. (2004). TREC genomics track overview. In *Proceedings of the Twelfth Text Retrieval Conference (TREC-2003)*, (pp. 14-23). Gaithersburg, MD: National Institute of Standards and Technology. (NIST Special Publication 500-255) Retrieved .13 August, 2008 from http://trec.nist.gov/pubs/trec12/papers/GENOMICS.OVERVIEW3.pdf (Archived by WebCite® at http://www.webcitation.org/5a2qzVCvD)
- Larkey, L., Ogilvie, P., Price, A+. . & Tamilio, B. (2000). Acrophile: an automated acronym extractor and server. In *Proceedings of the ACM Fifth .International Conference on Digital Libraries, DL '00, Dallas TX.* (pp. 205-214). New York, NY: ACM Press.
- Liu, H. & Friedman, C. (2003). Mining terminological .knowledge in large biomedical corpora. In Russ B Altman, A+ .Keith Dunker, Lawrence Hunter, Tiffany A+ .Jung & Teri E Klein (Eds.). *Proceedings of the 8th Pacific Symposium on Biocomputing (PSB 2003), Lihue, Hawaii.* (pp. 415-426). Singapore: World Scientific Publishing Co.
- Luque Ruiz, I., Cruz Soto, J++. . L. & Gómez-Nieto, M. A+. . (1996). Error detection, recovery, and repair in the translation of inorganic nomenclatures. 2: a proposed strategy. *Journal of Chemical Information and Computer .Sciences*, **36**(1), 16-24.
- Pirkola, A+. ., Keskustalo, H., Leppänen, E. , Känsälä, A+. .-P. & Järvelin, K. (2002). Targeted s-gram matching: a novel .n-gram matching technique .for cross- and monolingual word form variants. *Information Research*, **7**(2), paper 126. Retrieved .13 August, 2008 from http://informationr.net/ir/7-2/paper126.html (Archived by WebCite® at http://www.webcitation.org/5a2sJHoaw)
- Schwartz, A+. . & Hearst, M. (2003). A simple algorithm .for identifying abbrevation definitions .in biomedical texts. In Russ B Altman, A+ .Keith Dunker, Lawrence Hunter, Tiffany A+ .Jung & Teri E Klein (Eds.). *Proceedings of the 8th Pacific Symposium on Biocomputing (PSB 2003), Lihue, Hawaii.* Singapore: World Scientific Publishing Co. Retrieved .13 August, 2008 from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1. .1.13.2481&rep=rep1&type=pdf (Archived by WebCite® at http://www.webcitation.org/5a2slmvxP)
- Terada, A+. ., Tokunaga, T. & Tanaka, H. (2004). Automatic expansion .of abbrevations by using context and character information. *Information Processing and Management* , **40**(1), 31-45.
- Vander Stouw, G.G., Gustafson, C., Rule, J++. .D. & Watson, . C.E. (1976). The Chemical Abstracts Service chemical .registry system. IV: use of the registry system to support the preparation of index nomenclature. *Journal of Chemical Information and Computer .Sciences*, **16**(4), 213-218.
- Wilbur, W.J++. ., Hazard, G.F., Divita, G., Mork, J++. .G., Aronson, A+. .R. & Browne, A+. .C. (1999). Analysis of

biomedical text for chemical .names: a comparison .of three methods. *Proceedings of the AMIA Annual Symposium* , (pp. 176-180). Bethesda, MD: American Medical Informatics Association. Retrieved .13 August, 2008 from http://skr.nlm.nih.gov/papers/references/chemicals.pdf (Archived by WebCite® at http://www.webcitation.org/5a2t0w0BE)
- Willett, P. (2000). Textual and chemical .information processing: different domains but similar algorithms . *Information Research*, **5**(2), paper 69. Retrieved .25 July, 2007 from http://informationr.net/ir/5-2/paper69.html (Archived by WebCite® at http://www.webcitation.org/5a2tL872j)
- Wren, J++. .D. (2006). A+ .scalable machine-learning approach to recognize chemical .names .within large text databases. *BMC Bioinformatics*, **7**(Suppl 2): S3.
- Yu, H., Hatzivassiloglou, V., Rzhetsky, A+. . & Wilbur, W.J++. . (2002). Automatically identifying gene/protein terms in MEDLINE abstracts. *Journal of Biomedical Informatics* , **35**, 322-330.

Appendix 1 .- Training data for FNV-Finder

. cellspacing="0" . cellpadding="3" . align="center" style="border-right: #99f5fb solid; border-top: #99f5fb solid; font-size: smaller; border-left: #99f5fb solid; border-bottom: #99f5fb solid; font-style: normal; font-family: verdana, geneva, arial, helvetica, sans-serif; background-color: #fdffdd">

Acronym	MeSH term/ Substance name	Variant forms in the FN-A index
adomet	s adenosylmethionine	adenosyl methionine; s adenosyl l methionine
adpr	adenosine diphosphate ribose	adenosine 5 .diphosphoribose; adp ribose
aibn	azobis isobutyronitrile	2 2 .azobis isobutyronitrile; 2 .2 azobisisobutyronitrile
alum	aluminum hydroxide	aluminium hydroxide
ampt	alpha methyltyrosine	alpha methyl p tyrosine; alpha methyl para tyrosine; alpha methylpara tyrosine
ap4a	diadenosine tetraphosphate	p 1 .p 4 .di adenosine 5 .tetraphosphate
azttp	3 azido 3 .deoxythymidine 5 .triphosphate	azt triphosphate
bbp	butylbenzyl phthalate	benzyl butyl phthalate; benzylbutyl phthalate; butyl benzyl phthalate; butylbenzylphthalate
bche	butyrylcholinesterase	butyrylcholine esterase
cccp	carbonyl cyanide m chlorophenyl hydrazone	carbonyl cyanide 3 .chlorophenylhydrazone; carbonyl cyanide m chlorophenylhydrazone; carbonyl cyanide m cholorophenyl hydrazone; carbonylcyanide m chlorophenyl hydrazone; carbonylcyanide m chlorophenylhydrazone
dcc	dicyclohexylcarbodiimide	dicyclohexyl carbodiimide; n n dicyclohexylcarbodiimide
dep	diethyl phthalate	diethylphthalate
dmn	dimethylnitrosamine	n nitrosodimethylamine

dmpo	5 5 .dimethyl 1 .pyrroline 1 .oxide	5 5 .dimethyl 1 .pyrroline n oxide; 5 .5 dimethyl pyrroline n oxide; 5 .5 dimethylpryrroline n oxide; dimethyl pyrroline n oxide
emate	estrone 3 .o sulfamate	oestrone 3 .sulphamate
galn	galactosamine	d galactosamine
gsno	s nitrosoglutathione	nitrosoglutathione; s nitroso glutathione; s nitroso l glutathione
hba	4 hydroxybenzoic acid	p hydroxybenzoate; p hydroxybenzoic acid
hdi	1 6 .hexamethylene diisocyanate	hexamethylene diisocyanate; hexane diisocyanate
iptg	isopropyl thiogalactoside	isopropyl beta d thiogalactopyranoside; isopropyl beta d thiogalactoside; isopropylthiogalactoside
kic	alpha ketoisocaproic acid	2 ketoisocaproic acid; alpha ketoisocaproate
kyna	kynurenic acid	kynurenate
m6g	morphine 6 .glucuronide	morphine 6 .beta glucuronide
mat	methionine adenosyltransferase	methionine adenosyl transferase
mcpa	2 methyl 4 .chlorophenoxyacetic acid	4 chloro 2 .methyl phenoxyacetic acid; 4 .chloro 2 .methylphenoxy acetic acid; 4 .chloro 2 .methylphenoxyacetic acid
mdp	acetylmuramyl alanyl isoglutamine	muramyl dipeptide: muramyldipeptide
mlsb	streptogramin b	streptograminb
mstfa	n methyl n trimethylsilyl trifluoroacetamide	n methyl n trimethylsilyltrifluoroacetamide
naag	n acetyl 1 .aspartylglutamic acid	n acetylaspartylglutamate
nata	n acetyltryptophanamide	n acetyl tryptophan amide; n acetyl l tryptophanamide
neuac	n acetylneuraminic acid	n acetyl neuraminic acid
nmm	omega n methylarginine	n g monomethyl l arginine; ng monomethyl l arginine
npa	alpha naphthylphthalamic acid	1 n naphthylphthalamic acid; n 1 .naphthylphthalamic acid; naphtylphtalamic acid
npe6	monoaspartyl chlorin e 6	mono l aspartyl chlorin e 6; n aspartyl chlorin e 6
nppb	5 nitro 2 .3 phenylpropylamino benzoic acid	5 nitro 2 .3 phenylpropyl amino benzoate; 5 .nitro 2 .3 phenylpropylamino benzoate
oag	1 oleoyl 2 .acetylglycerol	1 oleoyl 2 .acetyl sn glycerol; 1 .oleoyl acetyl sn glycerol; oleoyl acetyl glycerol
ocdd	octachlorodibenzo 4 .dioxin	octachlorinated dibenzo p dioxin; octachlorinated dibenzodioxin; octachlorodibenzo p dioxin; octachlorodibenzodioxin
otz	2 oxothiazolidine 4 .carboxylic acid	2 oxothiazolidine 4 .carboxylate
pam	peptidylglycine monooxygenase	peptidyl glycine alpha amidating monooxygenase; peptidyl glycine alpha amidating mono oxygenase; peptidylglycine alpha amidating monooxygenase
pcmb	p chloromercuribenzoic acid	para chloromercuribenzoate
pcmbs	4 chloromercuribenzenesulfonate	4 chloromercuri benzene sulfonic acid; p chloromercuribenzene sulfonate; p chloromercuribenzenesulphonic acid; p chloromercurybenzenesulfonate
prpp	phosphoribosyl pyrophosphate	phosphoribosylpyrophosphate
rubp	ribulose 1 .5 diphosphate	ribulose 1 .5 bisphosphate
sdma	n n dimethylarginine	symmetric dimethylarginine; symmetric n g n g dimethyl l arginine; symmetric ng ng dimethyl l arginine
so2	sulfur dioxide	sulphur dioxide
tbh	tert butylhydroperoxide	t butylhydroperoxide; tert butyl hydroperoxide; tertiary butyl hydroperoxide
tcne	tetracyanoethylene	tetracyano ethylene
tpmt	thiopurine methyltransferase	thio purine methyl transferase;

		thiopurine methyl transferase; thiopurine s methyltransferase
utp	uridine triphosphate	uridine 5 .triphosphate
vpa	valproic acid	valproate

. />

Appendix 2 - FNV-Finder algorithm

Step 1. Input: an index entry; components extracted from a phrasal MeSH term, . or 6-grams extracted from a single word or compound word MeSH term. .

Remove from an index entry all lines .that do not contain a MeSH term .component or a 6-gram. #Such lines .do not contain acronym's MeSH term .or its variant.

Step 2. Input: output file from step 1; a stop word list. .

For each line .in an entry, match strings in a line .against a stop word list. . If there is a match then print the end of the line .located between the rightmost stop word and the acronym. Else print the whole line, . excluding the acronym.

Step 3. Input: post-second-step-entry; a prefix string list .containing strings 0-100, alpha, beta, gamma, tert, nalpha, d, l, n, p, r, s; acronym.

For each line, . start from the right to find a string whose first letter .is identical .to the first letter .of the acronym. If such string is found then mark it as a temporary first component (TFC). Else go to step 4. .

Compute LCS for the acronym and the string sequence .right to and including the TFC. If LCS = |the number of characters in the acronym| mark the string sequence .as a full name candidate. Else go to step 4. .

Match the line .containing the full name candidate against a prefix string list. . If there is a prefix string that appears immediately left to the full name candidate, include it and all consecutive prefix strings to the left of it in the full name candidate, and go to step 5. . Else print the full name candidate and go to step 5. .

Step 4. Input: lines .that were not resolved in step 3; post-second-step-entry; a prefix string list .containing strings 0-100, alpha, beta, gamma, tert, nalpha, d, l, n, p, r, s.

Compute a FNV indicator value for each string in the post-second-step-entry. If a line .contains one string with FNV indicator > 0. .50, mark it as a TFC. If a line .contains more than one string with FNV indicator > 0. .50, mark the leftmost string as a TFC. Mark the string sequence .right to and including the TFC as a full name candidate. Else, if a line .does not contain a string with FNV indicator > 0. .50, reject the line. .

Match the line .containing the full name candidate against a prefix string list. . If there is a prefix string that appears immediately left to the full name candidate, include it and all consecutive prefix strings to the left of it in the full name candidate, and go to step 5. . Else print the full name candidate and go to step 5. .

Step 5. Input: output files from steps 3 .and 4; a suffix list; MeSH term. .

Match the end of the full name candidate against a suffix list. . Match the end of the MeSH term .against a suffix list. . If the full name candidate and the MeSH term .match different suffixes, reject the candidate. Else go to step 6. .

Step 6. Input: output file from step 5; MeSH term. .

Match the full name candidate against the MeSH term. . If there is a match, mark the full name candidate as a MeSH term. . Else mark the full name candidate as a variant form. Print the variant form.

. />

. style="color: #5E96FD;" />