Amol J. Bapat. A system to extract abbreviation-expansion pairs from biomedical literature. A Master's Paper for the M.S. in I.S degree. April, 2009. 38 pages. Advisor: Catherine Blake.

We present a system to identify abbreviation expansion pairs from scientific articles. We work with the Genomics track of the TREC collection. Authors report abbreviations in two places - an abbreviations section and within the body of a scientific article. Articles with an abbreviations section had fewer abbreviations than those that did not have an abbreviations section (an average of 7.1 versus 13.2 abbreviations per article). For articles that do have an abbreviations section, authors report 98.2% of the abbreviations present in the document in that section. Inspired by Schwartz & Hearst's earlier work our program identified 2.1 million abbreviations from 162,259 documents. A manual inspection of a randomly selected set of articles revealed that our system achieved 86.7% precision and 81.9% recall.

Headings:

Abbreviations Expansion Pairs

Abbreviations

Text Mining

A SYSTEM TO EXTRACT ABBREVIATION-EXPANSION PAIRS FROM THE BIOMEDICAL LITERATURE

by
Amol J. Bapat

A Master's paper submitted to the faculty
of the School of Information and Library Science
of the University of North Carolina at Chapel Hill
in partial fulfillment of the requirements
for the degree of Master of Science in
Information Science.

Chapel Hill, North Carolina

April 2009

Approved by

_____

Catherine Blake

Table of Contents

1. Introduction 6

2 Related Work 9

3. Material and Methods 13

3.1 Genomics TREC Collection 13

3.2 Operational Definitions 14

3.2.1 Abbreviations 14

3.2.2 Expansions 14

3.3 Approach and Algorithm 14

3.3.1 Identifying abbreviations from the Abbreviations section 15

3.3.2 Identifying abbreviations in the body of the article 18

3.3.2.1 Extracting Abbreviations 18

3.3.2.2 Extracting Expansions 19

3.4 Manual Identification 22

4 Results and Discussion 23

4.1 Pilot Study 23

4.1.1 Abbreviations extracted from the ‘Abbreviations’ section 23

4.1.2 Abbreviations extracted from the body of the article 24

1 Abbreviations missed due to punctuations 25

2 Abbreviations missed due to numbers 25

3 Abbreviations missed due to other symbols 25

4 Abbreviations missed due to multiple parentheses 26

5 Abbreviations missed due to presence of type / classification information 26

4.2 Detailed Comparison 27

4.2.1 Total number and distribution of abbreviations 27

4.2.2 Number of expansions per abbreviation 29

4.3 Further Discussion 30

5 Conclusion 31

6 Acknowledgements 33

7 References 34

Table of Figures

Figure 1: An example demonstrating abbreviations and expansions in scientific articles PMID: 9298989 (Kallunki et al., 1997)……8

Figure 2: An example of abbreviations and expansion from a scientific article. PMID: 15994197 (Taylor & Fei, 2005)…….14

Figure 3: Implementation model giving overall picture of the procedure employed for identifying abbreviations and expansions given in the Abbreviations section……...…15

Figure 4: An example of abbreviations and expansion from a scientific article PMID: 16192347 (Chen et al., 2005)…….…….....16

Figure 5: Pseudo code to extract abbreviation-expansion pairs from abbreviations section……...17

Figure 6: Implementation model giving overall picture of the procedure employed for identifying and extracting abbreviations and expansions defined in the body of the scientific article…….……18

Figure 7: An example representation of abbreviations and expansions PMID:14977639 (Sharp & Little, 2004)……..……...……19

Figure 8: Pseudo code to extract abbreviation expansion pairs from the body of the article……..……21

Figure 9: Subroutine to identify best expansion for an abbreviation……...22

Figure 10: Chart indicating numbers of abbreviations and expansions present in different parts of the scientific article……27

Figure 11: Graph showing a power law curve for number of abbreviations versus the number of expansions……29

Table of Tables

Table 1: Results obtained from abbreviation-expansion pairs extracted from a separate section..…….23

Table 2: Results obtained by extracting abbreviation-expansion pairs from the body of the article…..….24

1. Introduction

Abbreviations play an important role in scientific literature. Correctly expanding abbreviations is critical if a reader is to understand the contents of a scientific article accurately. To alleviate the challenge of correctly identifying abbreviations and expansions, researchers like Jablonski (1993) and Sola (1992) have developed abbreviation dictionaries and web based acronym and abbreviation finders. (Jablonski, 1993; Sola, 1992). Several researchers have proposed methods to extract abbreviations from scientific literature. Approaches such as Chang and colleagues worked on building an automatically generated and maintained lexicon of abbreviations (Chang, Schütze, & Altman, 2002) while Yu and colleagues developed a software program called AbbRe (Yu, Hripcsak, & Friedman, 2002). Although the dictionary approach is a powerful approach to identify existing abbreviations, researchers continue to create new abbreviations and acronyms that require the dictionary be updated. Moreover, the same abbreviation can have different expanded versions, so a universal dictionary is unsustainable.

Correct abbreviation expansions play an important role in text mining systems that extract meaning from text. The aim of the research is to extract machine understandable structured text from a human understandable unstructured collection of words presented as scientific articles. This in-turn can help develop multi-document summarization and recognizing textual entailment. At the heart of systems that recognize textual entailment or summarize multiple documents lies a relevant information extractor module. Abbreviations can provide key links to this information extractor module about relevant information. Expanding abbreviations incorrectly or missing abbreviations might cause the systems to miss an important relevant information source. Text mining approaches also have many other potential applications and advantages such as identifying implicit relationships between literatures. Accurate

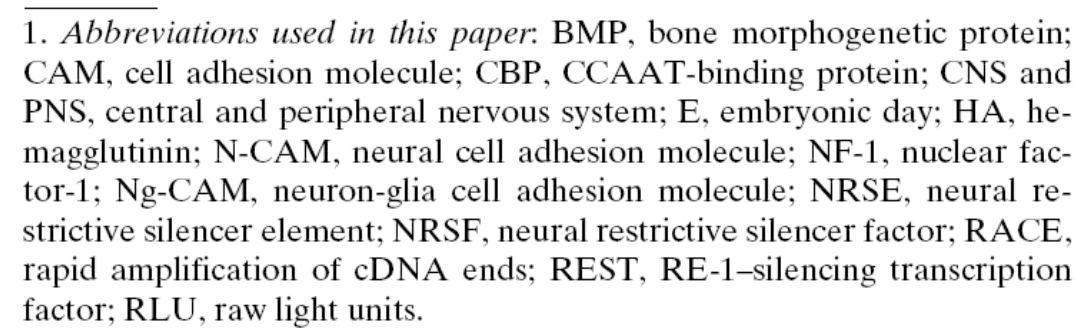
abbreviation expansion plays a major role in bridging the gap between what an author writes in an article and what the expanded version of the abbreviation actually mean. Thus, correctly identifying abbreviations and their expansions forms an important aspect of text mining studies. Automated methods of abbreviation extraction cannot achieve 100% accuracy (Yoshida, Fukuda, & Takagi, 2000). However Ao and Takagi's system ALICE achieved 97.5% precision and 98% recall (Ao & Takagi, 2003), Schwartz & Hearst's system achieved 96% precision and 82% recall (Schwartz & Hearst, 2003), Yoshida and colleagues built a system that achieved 98.98% precision and 95.56% recall (Yoshida et al., 2000) while Yu and associates' system achieved 95% precision and 70% recall for abbreviations defined in scientific articles (Yu et al., 2002).

Baeza-Yates and Ribeiro-Neto define the term precision as "The fraction of retrieved documents that are relevant" (Baeza-Yates & Ribeiro-Neto, 1999). We reuse this definition in the context of abbreviation-expansion pairs, as "Precision is the fraction of retrieved abbreviation-expansion pairs that are relevant." Baeza-Yates and Ribeiro-Neto define recall as "The fraction of relevant documents that have been retrieved" (Baeza-Yates & Ribeiro-Neto, 1999). We reuse this definition in the context of abbreviation-expansion pairs, as "Recall is the fraction relevant documents that have been retrieved."

In addition to scientific literature, ambiguous and unsanctioned abbreviations are pervasive in clinical reports (Pakhomov, Pedersen, & Chute, 2005). Many studies have been conducted that disambiguate clinical text have been suggested, such as (Pakhomov et al., 2005), (HaCohen-Kerner, Kass, & Peretz, 2008a), (HaCohen-Kerner, Kass, & Peretz, 2008b) . This again emphasizes the need to have a system to extract abbreviations and expansions from the articles.

Schwartz & Hearst built a system that achieved precision values of 96% and recall values of 82%, required no training data, and did not employ complex mathematical operations. We present a system based on the approach taken by Schwartz & Hearst that uses rules based pattern matching to extract abbreviations and relevant expansions. Our goal is to identify abbreviations and their corresponding

expansions from scientific articles automatically. To evaluate the results, we compare manually extracted abbreviation expansions to those extracted from the body of the article using the Schwartz & Hearst's approach (Schwartz & Hearst, 2003). We also extract abbreviation expansions listed in an 'abbreviations' section in the scientific article and use it to evaluate the results. For example:



1. *Abbreviations used in this paper*: BMP, bone morphogenetic protein; CAM, cell adhesion molecule; CBP, CCAAT-binding protein; CNS and PNS, central and peripheral nervous system; E, embryonic day; HA, hemagglutinin; N-CAM, neural cell adhesion molecule; NF-1, nuclear factor-1; Ng-CAM, neuron-glia cell adhesion molecule; NRSE, neural restrictive silencer element; NRSF, neural restrictive silencer factor; RACE, rapid amplification of cDNA ends; REST, RE-1–silencing transcription factor; RLU, raw light units.

Figure 1: An example demonstrating abbreviations and expansions in scientific articles PMID: 9298989 (Kallunki et al., 1997)

Our work involves identifying abbreviations and their respective expansions for scientific articles from collection of documents from the TREC collection. To verify our approach the following three sets of annotations were collected against one another.

1. Abbreviations defined in an abbreviations section
2. Abbreviations defined in the body of the scientific article
3. Manually extracted abbreviations from the body of the scientific article

2 Related Work

Schwartz and Hearst provided an algorithm that identifies abbreviations from biomedical text (Schwartz & Hearst, 2003). Similar to other work done on abbreviation extraction, their approach employed pattern matching, based on the abbreviation length and on number of words in the candidate-expanded form. They found that abbreviations occurred in one either of two places, following the expanded form or just before it, in parentheses. e.g.:

abbreviation (expansion)

expansion (abbreviation)

For the Schwartz and Hearst's approach to work, the abbreviation and expanded form needed to occur in one of the two patterns above. Unlike learning approaches, the pattern matching approach does not require training data. Their algorithm achieved 96% precision and 82% recall on a collection of 1000 randomly selected abstracts from MEDLINE containing the word "yeast", and 95% precision and 82% recall on another larger test collection (Schwartz & Hearst, 2003).

Yoshida and colleagues built their own hybrid system composed of PROPER and PNAD systems (Yoshida et al., 2000). PNAD is acronym for Protein Name Abbreviation Dictionary. The PNAD System could extract the pairs from parenthetical-paraphrases involved in protein names and the PROPER System identified these pairs. Yoshida and colleagues argued that machine-readable natural language resources appear much more promising and yield more abbreviation expansion pairs in quicker time than humans do. Their system was web based and used PERL programming language. Their eight-step system achieved 98.85% precision, 95.56% recall and 97.58 % complete precision. They identified six different

types of abbreviations reported in various abstracts in biomedical literature. The web based system consisted on six different subsystems: web server, PNAD-CSS server, dictionary management server, text database management server, PROPER system and an abbreviation extraction system (Fukuda, Tamura, Tsunoda, & Takagi, 1998).

Yu and associates also analyzed biomedical documents (Yu et al., 2002). They developed a system called AbbRE (Abbreviation Recognition and Extraction) to map defined abbreviations to their full forms. Similar to approaches described above, AbbRE bases itself on pattern matching principles too. They used opinions of domain experts as a reference standard and evaluated the performance of their system, by noting the recall and precision values of AbbRE for defined abbreviations in ten biomedical articles randomly selected from the ten most frequently cited medical and biological articles. They also measured the percentage of undefined abbreviations in the same set of articles, and they investigated whether they could map undefined abbreviations to any of four public abbreviation databases (GenBank LocusLink, SWISSPROT, LRABR of the UMLS Specialist Lexicon, and BioABACUS). AbbRE had an average 0.70 recall and 0.95 precision for the defined abbreviations. The authors found that average 25% of the abbreviations defined in biomedical articles and a randomly selected subset of undefined abbreviations, 68% can map to any of four abbreviation databases. They also found that many abbreviations are ambiguous (i.e., they map to more than one full form in abbreviation databases).

Pakhomov and colleagues worked on abbreviation and acronym disambiguation by selecting a sample of about 1.7 million notes from the Mayo Clinic (Pakhomov et al., 2005). They built a sense disambiguation system, one phase of which was extracting correct abbreviations and acronyms. Each of the eight acronyms they identified had over an average of 540 occurrences. They too used regular expressions to identify the abbreviations and later provided the results to experts for annotating. Their study indicated that established sources of acronyms and abbreviations might not be entirely suitable as sources of sense inventories for acronyms in clinical notes.

Liu and associates described a method to extract abbreviations from the UMLS. (Liu, Lussier, & Friedman, 2001) They evaluated the method and studied the ambiguous nature of the abbreviations. They extracted 163,666 unique (abbreviation, expansion) pairs from the UMLS with a precision of 97.5%, and a recall of 96%. They showed that the UMLS abbreviations were highly ambiguous. Abbreviations with six characters or less had multiple meanings were 33.1%. They used a combination of manual and automated extraction methods to come to a gold set of abbreviations that they further used in their disambiguation study.

Liu and colleagues did a study on three-letter abbreviations that were defined using parenthetical expressions (Liu, Aronson, & Friedman, 2002). This study also concentrated on abbreviation disambiguation techniques. They developed a program called PW3 to extract abbreviations. PW3 used a matching method for three-letter abbreviations and was designed to search for a possible expansion from candidate text strings. PW3 searched within a window size of six words that were to the left of a parenthetical expression containing the abbreviation.

To identify abbreviations and their expansions Ao and Takagi built a mining system called ALICE (Abbreviation LIfter using Condition based Extraction) (Ao & Takagi, 2003). Their system also accepts Acronyms and their definitions successfully excluding synonyms, hypernyms and citations. The system worked in three phases: information retrieval, information extraction, confirmation judgment. Similar to (Schwartz & Hearst, 2003), their system also works on the hypothesis that parentheses are used for abbreviations. One important distinguishing point between our approach based on Schwartz & Hearst and Ao & Takagi's approach is that Ao & Takagi's rules do not restrict the first word of a expansion to begin with its initial letter of the abbreviation (Ao & Takagi, 2003). The information extraction phase in their approach classifies candidate abbreviations in nine different types according to how they are composed. Their system achieved 96 % precision and 98 % recall over 1000 abstracts randomly selected from MEDLINE.

Chang and associates used machine-learning approach to find and score abbreviations. (Chang, Schütze, Altman 2002). Their system consisted of four phases: to scan text, to find possible abbreviations, align them with their prefix strings, and then collect a feature vector based on eight characteristics of the abbreviation and alignment. In the final step, they applied binary logistic regression to generate a score from the feature vector. Their system generated a precision of 80% and a maximum recall of 83% on Medstract gold standard. They searched for abbreviations from the China Medical Tribune in MEDLINE and generated a recall of 88%. They confirmed that there has been a growth in the number of abstracts and abbreviations added to MEDLINE from a period of 1975 to 2000. Their system gave a comparable performance to the Acromed system developed by Pustejovksy and colleagues (Pustejovsky, Castaño, Cochran, Kotecki, & Morrell, 2001). The system was computationally intensive requiring 70 hours of CPU time using five processors on a Sun Enterprise E3500 running Solaris 2.6

3. Material and Methods

In this section, we describe the materials and methods used in this study. We conducted this study on documents from the Genomics Track of the Text REtrieval Conference (Hersh, Cohen, Roberts, & Rekapalli, 2006)

3.1 Genomics TREC Collection

The Text REtrieval Conference (TREC) is an on-going series of workshops focusing on a list of different information retrieval (IR) research areas, or tracks. The National Institute of Standards and Technology (NIST) co-sponsors the conference with the Disruptive Technology Office of the U.S. Department of Defense. The first conference was held in 1992 as part of the TIPSTER Text program with a goal to support and encourage research within the information retrieval community by providing the infrastructure necessary for large-scale evaluation of text retrieval methodologies and to increase the speed of lab-to-product transfer of technology.

Each track in TREC has a challenge wherein NIST provides participating groups with data sets and test problems. Depending on track, test problems might be questions, topics, or target extractable features. NIST also provides uniform evaluation scoring systems. After evaluation of the results, a workshop provides a place for participants to collect together thoughts and ideas and present current and future research work (http://trec.nist.gov/overview.html). The publications are at the website: http://trec.nist.gov/pubs.html.

TREC test collections are large enough so that they realistically model operational settings. Most of today's commercial search engines include technology first developed in TREC.

3.2 Operational Definitions

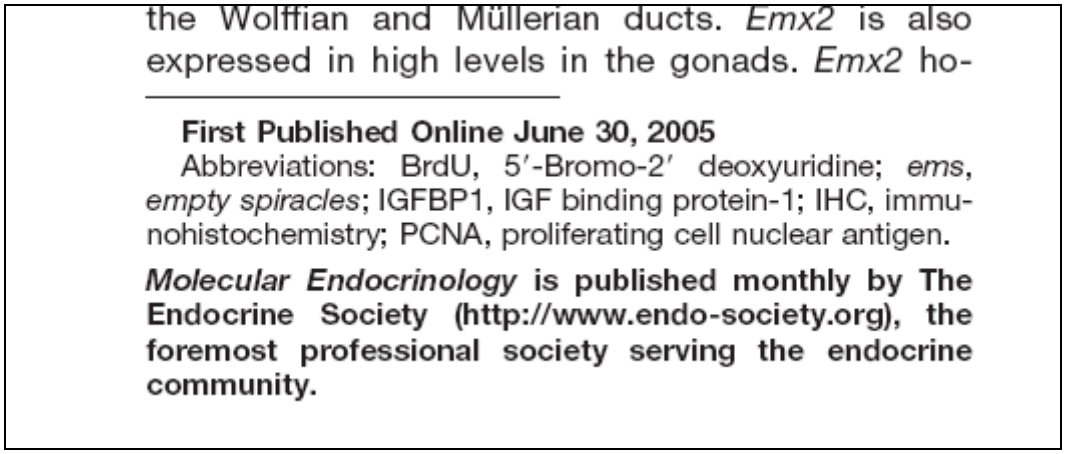
This section gives us operational definitions of the two most important terms used in this paper.

3.2.1 Abbreviations

Abbreviation - An abbreviation (from Latin brevis "short") is a shortened form of a word or phrase. Usually, but not always, it consists of a letter or group of letters taken from the word or phrase. For example, we can represent the word "abbreviation" as "abbr." or "abbrev." (*Abbreviation*). Figure 2 has the following five abbreviations: BrdU, ems, IGFBP1, IHC and PCNA

3.2.2 Expansions

Expansion – An expansion here refers to the definition of the abbreviation mentioned in the scientific article in question. For example, Figure 2 has the following expansions: 5'-Bromo-2' deoxyuridine; empty spiracles, IGF binding protein 1, immunohistochemistry, and proliferating cell nuclear antigen.


the Wolffian and Müllerian ducts. *Emx2* is also expressed in high levels in the gonads. *Emx2* ho-

First Published Online June 30, 2005

Abbreviations: BrdU, 5′-Bromo-2′ deoxyuridine; *ems*, *empty spiracles*; IGFBP1, IGF binding protein-1; IHC, immunohistochemistry; PCNA, proliferating cell nuclear antigen.

Figure 2: An example of abbreviations and expansion from a scientific article. PMID: 15994197 (Taylor & Fei, 2005)

3.3 Approach and Algorithm

This section identifies the three approaches taken by the authors to identify various abbreviations and their respective expansions used in scientific articles.

3.3.1 Identifying abbreviations from the Abbreviations section

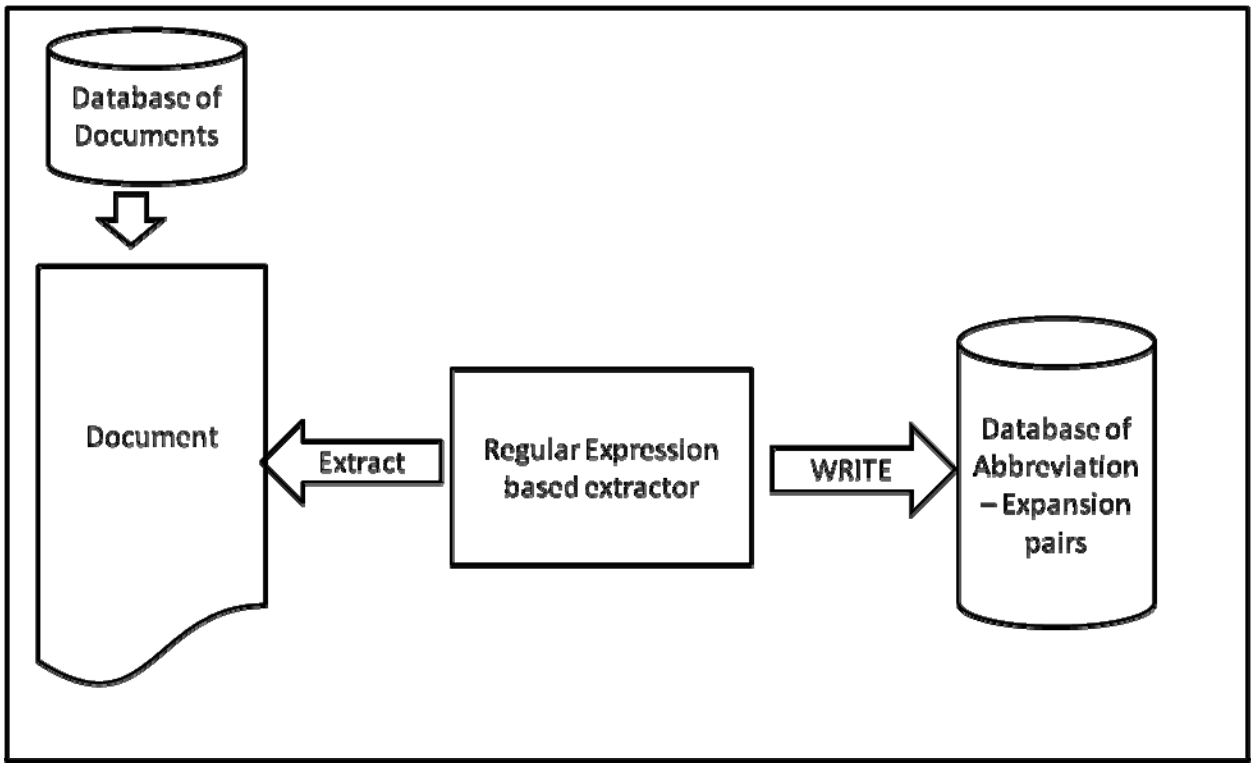
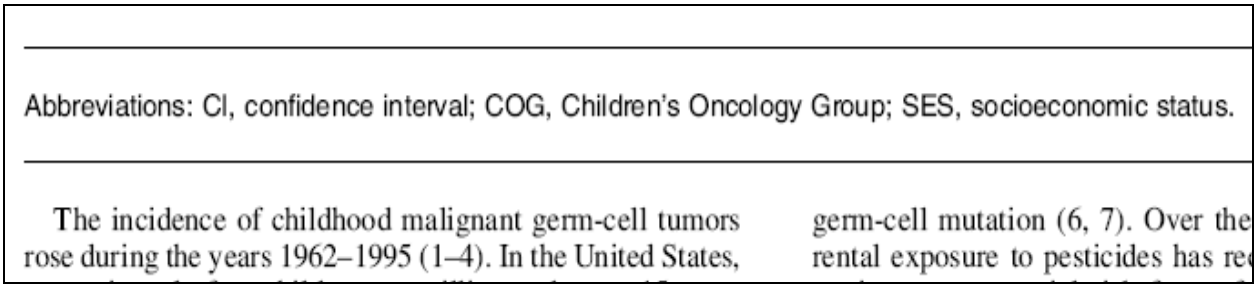


Figure 3: Implementation model giving overall picture of the procedure employed for identifying abbreviations and expansions given in the abbreviations section.

Documents in the TREC collection that contain abbreviations reported as separate entities were either present as a separate abbreviations section in the scientific article or followed the word “Abbreviations:” Hence, we developed a program to address each of these situations and location where the abbreviations separately.

We observed that that where authors reported abbreviation–expansion pairs separately; they followed the pattern similar one shown in the figure below:



Abbreviations: CI, confidence interval; COG, Children's Oncology Group; SES, socioeconomic status.

The incidence of childhood malignant germ-cell tumors rose during the years 1962–1995 (1–4). In the United States, germ-cell mutation (6, 7). Over the rental exposure to pesticides has re

Figure 4: An example of abbreviations and expansion from a scientific article PMID:16192347 (Chen et al., 2005)

As shown in Figure 4, authors use well-defined delimiters when describing abbreviations. The program that identified the position of these abbreviations involved regular expression based logic to extract the correct abbreviations and their respective expansions. The program wrote the following information to a database:

1. The identifying information for the document as a whole - PUBMED Identifier,
2. A unique sentence identifier giving the exact location of the abbreviation in the scientific article
3. The abbreviation extracted
4. The expansion extracted

The pseudo code for this process is as shown in Figure 5.

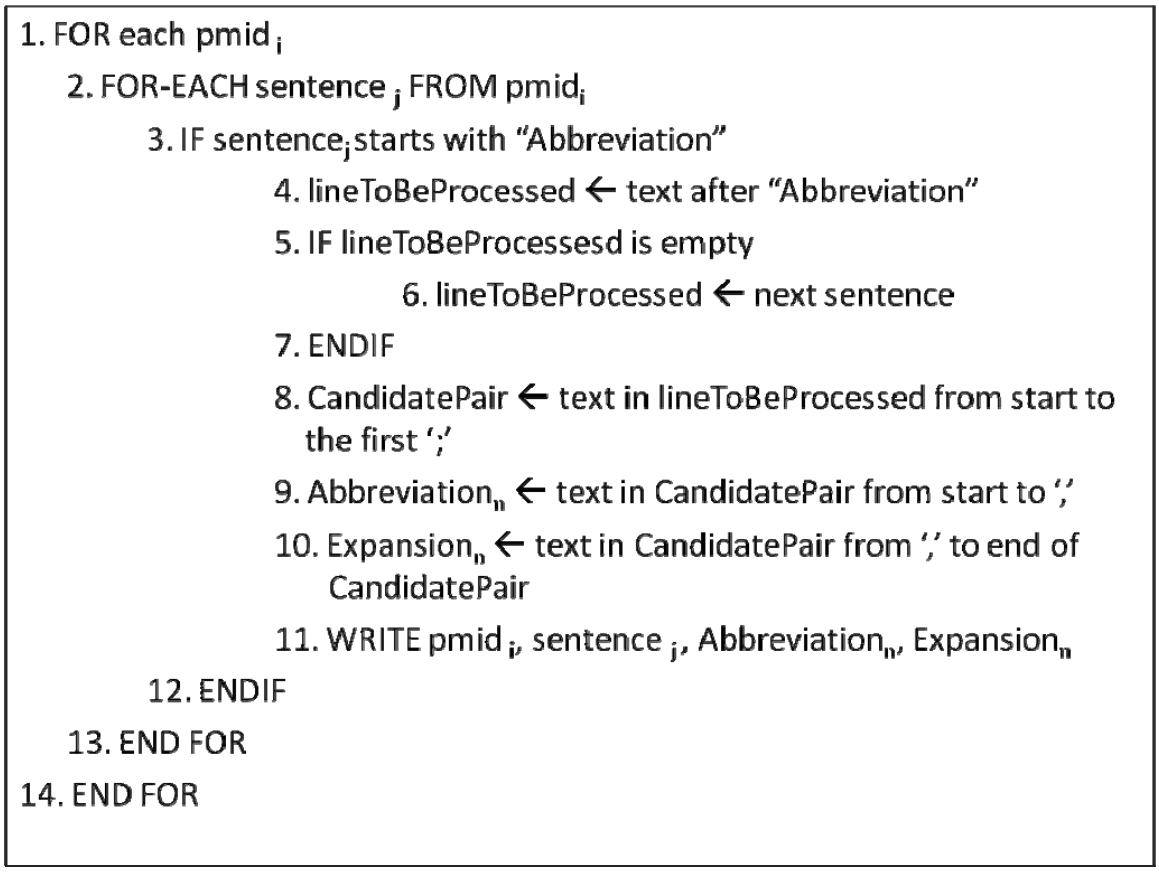


Figure 5: Pseudo code to extract abbreviation-expansion pairs from abbreviations section

3.3.2 Identifying abbreviations in the body of the article

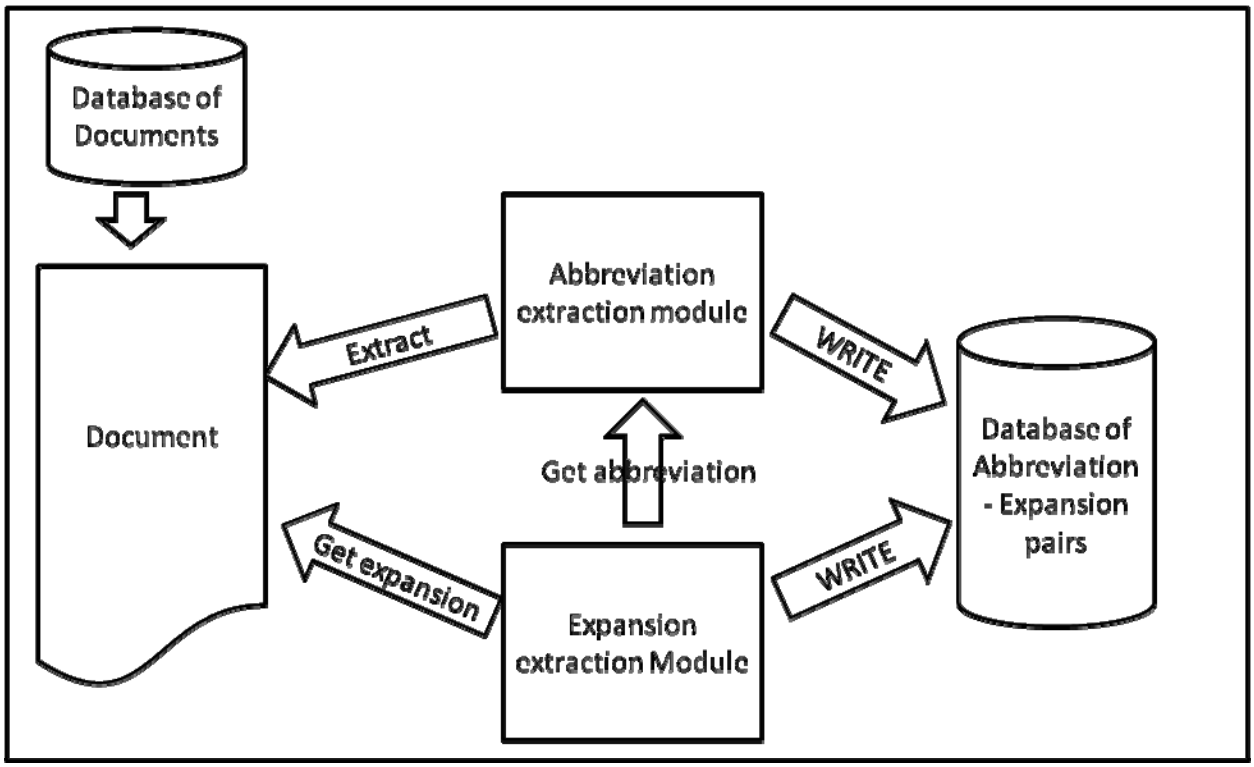


Figure 6: Implementation model giving overall picture of the procedure employed for identifying and extracting abbreviations and expansions defined in the body of the scientific article.

We base our approach primarily on the work done by Schwartz & Hearst (Schwartz & Hearst, 2003). The algorithm identifies abbreviations and their respective expansions from an input text, in our case scientific articles from the Genomics TREC collection.

3.3.2.1 Extracting Abbreviations

Our working hypothesis is that abbreviations used in a scientific article in the Genomics TREC collection typically follow a uniform pattern, where the first letter of each word in the expansion corresponds to one letter in the abbreviation. This simple fact enables the program to identify abbreviations easily. The algorithm concentrates on identifying abbreviation candidates by looking near parentheses. The system uses the following constraints:

1. Abbreviations consist of at most two words

2. The length of the abbreviation is between two to ten characters
3. The abbreviation has at least one letter
4. The first character of the abbreviation is alphanumeric

For example, consider the phrase Heat Shock Transcription Factor (HSF). In this case, HSF the system considers HSF a valid short form candidate as it satisfies all the system constraints.

3.3.2.2 Extracting Expansions

Based on manual inspection our study revealed that in the Genomics TREC collection, authors report abbreviation-expansion pairs in type-two format as defined by Schwartz and Hearst's classification as shown below. (Schwartz & Hearst, 2003)

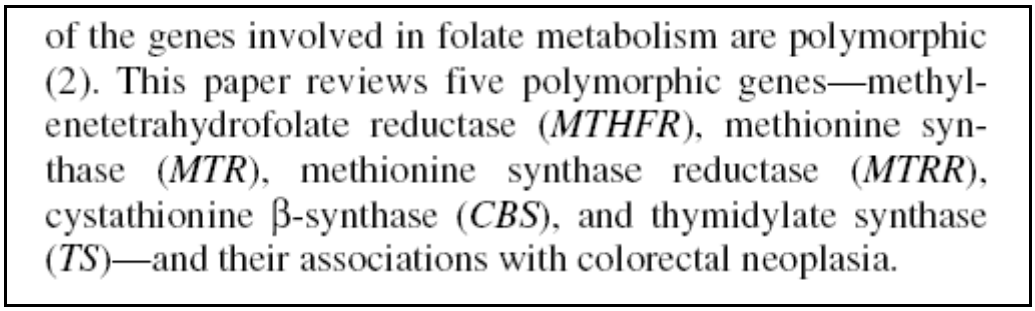

of the genes involved in folate metabolism are polymorphic (2). This paper reviews five polymorphic genes—methylenetetrahydrofolate reductase (*MTHFR*), methionine synthase (*MTR*), methionine synthase reductase (*MTRR*), cystathionine β-synthase (*CBS*), and thymidylate synthase (*TS*)—and their associations with colorectal neoplasia.

Figure 7: An example representation of abbreviations and expansions. PMID:14977639 (Sharp & Little, 2004)

The abbreviation and relevant expansion occur in the order given in the Figure 7 above and follow the pattern of parentheses above. Hence, while reading the document, once the program finds an abbreviation, it passes the part of the sentence before the reported abbreviation to the expansion extractor. Let us call this section of the sentence as Phrase A. The program then trims this Phrase A based on the following constraints on expansion candidates:

1. The expansion candidate must appear in the same sentence as the abbreviation
2. As per Park and Byrd (Park & Byrd, 2001), the expansion should have no more than minimum from |A| + 5 or |A| * 2 words, where |A| is the number of characters in the abbreviation.

For example, if an abbreviation contains 3 characters, then |A| becomes 3. So the algorithm will look for the number of words to look for in abbreviation is minimum of either (3+5=8) or (3*2=6). As six is less than eight, the algorithm will look for expansion of the abbreviation in six words to the left of the abbreviation.

The algorithm now proceeds by looking letter by letter, for character-matches starting from the right in both the abbreviation and expansion. The algorithm works backward until it finally reaches the first letter of abbreviation.

One downside of this approach is that it finds the correct definition, but not necessarily the correct alignment. Let us take an example of the abbreviation and expansion: Heat shock tranScription Factor (HSF). The algorithm parses the phrases backward. It starts from ‘r’ and compares with the last letter of abbreviation ‘F’. These do not match, so the algorithm compares the second last letter ‘o’ with ‘F’. Again, there is no match, so the algorithm checks the previous character. In this way, the algorithm keeps checking until it finds a match. In this case, the algorithm finds a match for ‘F’ in the first character of the word ‘Factor’. This is what we call, a correct alignment between the letters in the and the expansion. The algorithm now moves to the previous letter in the abbreviation, ‘S’. Working backward again, it searches for S in the word transcription. This however, results in an incorrect alignment. The abbreviation contained the letter ‘S’ corresponding to the word Shock, but the algorithm determined ‘S’ from abbreviation to be the ‘s’ in transcription.

Once the program identifies the information, it writes to a database the following information:

1. The identifying information for the document as a whole – PMID
2. A unique sentence identifier giving the exact location of the abbreviation in the scientific article
3. The abbreviation extracted
4. The expansion extracted

The pseudo code for this process is as shown in Figure 8:

```
1. FOR EACH
 2. FOR-EACH sentence_j in PMID_i
  3. IF sentence_j contains a '(' character and a ')' character
    4. AbLength ← number of characters in parentheses
     5. IF AbLength > 2 AND AbLength < 10
        6. CandAbbrev ← characters from the parentheses
        7. IF first character of CandAbbrev is not a digit
           8. PotExp ← words in sentence_j preceding the CandAbbrev
           9.  trimPotExp<-- minimum(AbLength + 5, AbLength * 2 words)
           10. CandExp <-- GetBestExpression(CandAbbrev, sentence_j, TrimPotExp)
           11. IF CandExp is not empty
              12. Write PMID_i, sentence_j, CandAbbrev, CandExpansion
           13. ENDIF
        14. ENDIF
     15. ENDIF
    16. END IF
  17. END FOR
18. END FOR
```

Figure 8. Pseudo code to extract abbreviation expansion pairs from the body of the article

Figure 9 shows the pseudo code for the program that searches for the best possible expansion for a recognized abbreviation. (Schwartz & Hearst, 2003) is the base of this code

```
1. S ← length of CandAbbrev
2. N ← length of PotExp
3. WHILE S > 0
    4. WHILE N > 0
        5. IF sth character in CandAbbrev matches nth character in PotExp
                6. S = S – 1
                7. break out of while in step 4
        8. ELSE
                9. N  = N - 1
        6. END IF
    7. END WIHLE
    8. IF N = 0
        9. RETURN empty
    10. ENDIF
11. END WIHLE
12. CandExp ← substring of PotExp starting from index n upto end of string
13. RETURN CandExp
```

Figure 9. Subroutine to identify best expansion for an abbreviation.

3.4 Manual Identification

We browsed through eight different papers in the TREC collection manually to identify all the abbreviations and their respective expansions. We then manually inserted the identified information into a database.

4 Results and Discussion

In this section, we discuss the results obtained from the study.

4.1 Pilot Study

We performed a pilot study to gain an idea about the number of abbreviation-expansion pairs defined in the 'abbreviations' section. This also helps us get an idea about how the abbreviations section organized abbreviation-expansion pairs. We observed that the abbreviations section contained a comma or a semi-colon as delimiters to separate the abbreviations from the expansions.

4.1.1 Abbreviations extracted from the 'Abbreviations' section

Table 1 compares the abbreviation-expansion pairs picked up manually with those identified by the regular expression based abbreviation-expansion extractor.

Table 1. Results obtained from abbreviation-expansion pairs extracted from an abbreviations section

PubMED Identifier	Manual Extraction	Extractor Program – Correctly Identified	Extractor Program
15280570	0	0	0
15994197	5	4	4
16192347	3	2	3
14769635	6	5	6
15870514	7	4	4
14977639	9	8	9
9298989	15	13	15
14977639	9	8	9
Total	54	44	50

As shown in the Table 1, we looked at eight documents and compared results from manual and automatic extraction of abbreviation-long form pairs reported in an 'Abbreviations' section. The program extracted 50 abbreviation-expansion pairs from the eight documents. The precision - the correctly identified pairs out of all pairs identified (44/50) was 88.9%. As shown Table 1, manual extraction revealed 54 pairs while program extracted only 44 pairs making the recall – the correctly identified pairs out of total pairs in the document 81.8%.

There was only one characteristic identified with the missed abbreviations. As shown in Figure 2 in the paper, some abbreviations contained a single quote or a comma causing the program to fail from adding the abbreviation in the database. The authors did not fix the problem due to time constraints.

4.1.2 Abbreviations extracted from the body of the article

Table 2 compares abbreviation-expansions pairs picked up manually to the ones identified by the algorithm based on Schwartz and Hearst (Schwartz & Hearst, 2003).

Table 2. Results obtained by extracting abbreviation-expansion pairs from the body of the article

PubMED Identifier	Manual Extraction	Extractor Program – Correctly Identified	Extractor Program
15280570	17	13	15
15994197	7	5	5
16192347	4	4	4
14769635	6	6	6
15870514	5	3	5
14977639	6	4	6
9298989	16	15	15
14977639	12	9	12
Total	73	59	68

As shown in the table above, we looked at eight documents and compared results from manual and automatic extraction of abbreviation-long form pairs defined in the body of the article. The program extracted 59 pairs correctly from a total of 68 pairs in the eight documents, making the precision, the

correctly identified pairs out of all pairs identified (59/68) = 86.7%. The program identified the 68-59 = 9 abbreviation expansion pairs incorrectly because of alignment problem that we have discussed incorrect in section 3.3.2.2. To summarize the problem, the program stops searching for a letter in abbreviation as soon as it encounters that letter in the expansion. However, this letter identified in the expansion may not be correct letter. As shown in the table, manual extraction revealed 73 pairs while program extracted only 59 pairs making the recall (correctly identified pairs out of total pairs) of = 81.9%.

Our analysis revealed the following reasons why the system 14 missed abbreviations:

1 Abbreviations missed due to punctuations

Abbreviations where incorrectly missed due to punctuation marks. For example, Standard Deviation is abbreviated as (S.D.) in the article by Hallert & others (Hallert et. al., 2004). In this case, the program failed to recognize the two dots present in the brackets and missed the abbreviation expansion pair completely. We observed this type of error in five out of the 14 missed abbreviation and expansion pairs.

2 Abbreviations missed due to numbers

There were cases where the sentence looked like: “Disease activity was also assessed by calculating the 28-joint count disease activity score (DAS-28) as described.”(Hallert et al., 2004) In this case, the program looked for characters in the expansion in the order in which they appear in the abbreviations, starting from D and ending at eight. However, in this nonconventional abbreviation, the number 28 appeared before the expansion for DAS, causing the algorithm to miss the abbreviation - expansion pair. We observed this type of error in two out of the 14 missed abbreviations and expansion pairs.

3 Abbreviations missed due to other symbols

Consider the following case: Microsatellite Instability (MSI+) (Sharp & Little, 2004). In this case, our program missed the abbreviation because of the ‘+’ symbol. In such cases along with the normal abbreviation, the abbreviated word also contains a definition of the class. In such cases, the program fails

to pick up the abbreviation information due to the presence of type or class information. We observed this type of error in one out of the 14 missed abbreviation and expansion pairs.

4 Abbreviations missed due to multiple parentheses

Some abbreviation and expansion pairs looked like: 2,3-dioleyloxy-N-2[(sperminecaboxamido)ethyl]-N-N-dimetheyl- 1- propanaminnium trifluoroacetate (DOPSA) (Taylor & Fei, 2005). Complex compounds named in chemistry literature sometimes follow the IUPAC nomenclature that can looks like the example mentioned above. In this case, the program looked at the word "sperminecaboxamido" as an abbreviation, skipped it because it contained more than ten characters. This is correct as the entire compound name occurs at the end of the string. However, after each potential abbreviation and expansion is considered, the program trims it out from the string under consideration in order to avoid duplication. So, the program compared DOPSA only against the phrase after the word sperminecaboxamido, causing the program to miss the abbreviation-expansion pair.

Another instance where multiple parentheses are problematic occurs in the format: "…(Relative Risk (RR), …)" (Hallert et al., 2004). In this case, "relative risk" is abbreviated as RR, however, the phrase "relative risk" occurs in a set of parentheses that the program missed. We observed this type of error in three out of the 14 missed abbreviation and expansion pairs

5 Abbreviations missed due to presence of type / classification information

Other nonconventional abbreviation techniques include such examples as Breast cancer 1 and 2 (BRAC 1/2) (Mondugno, 2004). In this case, the program missed the pair because of the punctuation '/' in the abbreviation, but there is also an important point to note here that 2 expansions Breast Cancer Gene 1 and Brest Cancer Gene 2 where abbreviated using just one abbreviation BRAC1/2. Such phrases that come together need some other treatment so that the programs recognize them.

Another example where the program missed the pair is "methlyenetetrahydrofolate reduactase (MTHFR C677T and A1298C)"(Sharp & Little, 2004). In this case, MTHFR is the abbreviation for the preceding

phrase, but the parentheses also contain C677T and A1298C. Having such examples in the parentheses caused the program to miss the abbreviations. We observed this type of error in three out of the 14 missed abbreviation and expansion pairs

4.2 Detailed Comparison

After the initial pilot study, we processed all the articles and obtained the following results:

4.2.1 Total number and distribution of abbreviations

As we have discussed earlier, we observed that abbreviations and their expansions occur in different places in the scientific article. Out the total 162,259 documents that we analyzed, we observed that 24,990 documents contained an ‘abbreviations’ section. The rest of the document contained no such section, but had abbreviations and relevant expansions in the body of the scientific article.

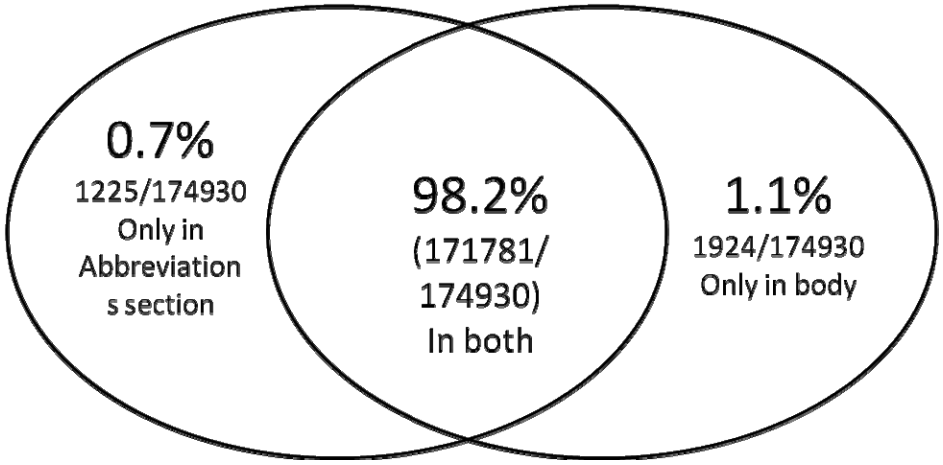


Figure 10: Chart indicating numbers of abbreviations and expansions present in different parts of the scientific article

For those articles having an abbreviations section, we observed the distribution as given in the Figure 10 above. As shown above, out of the total 24990 documents that contained an abbreviations section, 0.7% (1225 / 174930) abbreviations and expansions where given only in the abbreviations section and nowhere else, while 1.1% (1924 / 174930) of the total abbreviations and expansions occurred only in the body of

the article and not the abbreviations section. Rest of the 98.2% (171781 / 174930) abbreviations and expansions were given, both in the abbreviations section as well as in the body of the scientific article.

To verify these results, we randomly selected 20 articles and analyzed them manually. The 20 documents had 136 abbreviations and expansions in abbreviations section and 142 abbreviations in the body of the article. We found that for this small case study, the results match our observations for the full collection. We observed 1.1% abbreviations and expansions occurred only in the abbreviations section. For example, COMT is defined as catechol-O-methyltransferase in the abbreviations section, but is not defined in the body of the article.(Masson et al., 2005; Masson, Sharp, Cotton, & Little, 2005). There were 1.3% abbreviations and expansions absent from the abbreviations section but present only in the body of the article. For example, OR is defined as Odds Ratio, but is missing from the abbreviations section in the article (Chen et al., 2005). The rest of the 97.6% abbreviations and expansions occurred in both, the body of the article and the abbreviations section. Our program extracted 2,189,596 abbreviations from the collection of 162,259 documents.

4.2.2 Number of expansions per abbreviation

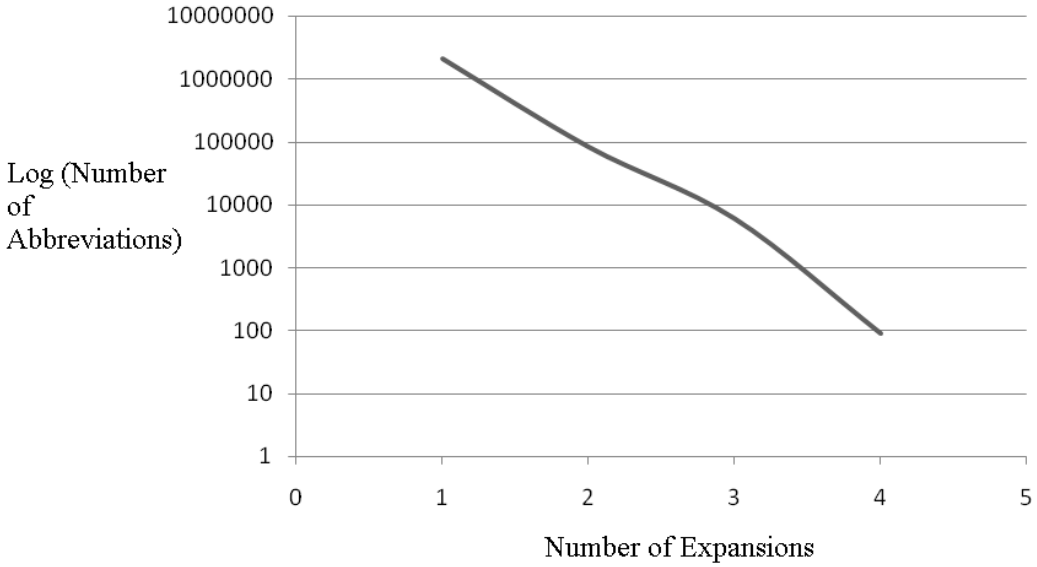


Figure 11: Graph showing a power law curve for number of abbreviations versus the number of expansions

We observed that many abbreviations had more than one expansion. Figure 11 shows the distribution of number of abbreviations to number of expansions. Our program extracted 2,189,596 abbreviations from the collection of 162,259 documents. Out of these 2,100,247 abbreviations had a single expansion, while 83246 abbreviations had two expansions, 6013 abbreviations had three expansions and 90 abbreviations had four expansions. There have been a number of studies concentrating on disambiguation of these abbreviations. The studies include (HaCohen-Kerner, Kass, & Peretz, 2008a); (HaCohen-Kerner, Kass, & Peretz, 2008b; Pakhomov et al., 2005). We believe that our study will help further research in this field due to ready availability of algorithms to extract abbreviation-expansion pairs.

4.3 Further Discussion

As mentioned in the previous sections, our approach of extracting abbreviation-expansion pairs consisted of looking at two places – the abbreviations section and in the body of the article. While looking at abbreviations defined in an abbreviations section, the regular expression worked well, having just one characteristic where it failed as explained in section 4.1.1 above. However, the extraction of abbreviation-expansion pairs from the body of the article was more problematic.

We found that the abbreviations missed due to punctuations, numbers and symbols formed the highest percentage – (8/14) 57 %. To solve this problem, we propose running the current system, while stripping off symbols, numbers and punctuation marks. The authors expect that the number of false positives may increase because of this, but the algorithm will pick up missed abbreviations too, causing improved recall number.

Abbreviations with multiple parentheses are relatively simple to resolve. The current matching algorithm strips off the part of the sentence extracted as abbreviation – expansion pair as soon as the program judges it correct or incorrect. Keeping this phrase as a part of the sentence until the program fully parses the sentence will help solve the problem. This might however reduce the processing speed of the algorithm.

5 Conclusion

We have presented a system that extracts abbreviation-expansion pairs from a scientific article. One advantage of our system is that is does not require training data. The only input needed is the actual body of the text and the system extracts abbreviation-expansion pairs into a separate database table.

Our method involved looking for abbreviation-expansion pairs at two separate locations with the scientific article. Abbreviation-Expansion pairs were extracted from the abbreviations section where they were defined in a delimited fashion. The second method involved extracting abbreviation-expansion pairs form the body of the article. For the Genomics TREC Collection, abbreviations and expansions where organized in a pattern. The expansion followed the abbreviation and the abbreviation was written in a set of parentheses. We manually inspected eight full text documents and the algorithm resulted in 86.7% precision and 81.9% recall.

We conducted an error analysis that revealed the following reasons why an abbreviation was missed:

- punctuations marks
- numbers
- symbols
- multiple parentheses

We recommend that future abbreviation systems remove punctuation and symbols and consider numbers as our error analysis suggests that these changes would improve system performance. This study revealed that the authors of biomedical articles report abbreviations in the abbreviations section, or in a sentence

that starts with the word Abbreviations. This finding has two important implications. For articles containing an abbreviations section, the difficult task of extracting abbreviations from the full text of a document can be avoided in biomedical journals because the majority of abbreviations are captured in the abbreviations section (171781 / 174930 = 98.2%), which is easier to parse automatically. Provided that abbreviations in biomedicine reflect abbreviations in other disciplines, the second implication of this research is that the abbreviations stated in the abbreviations section can be used to evaluate methods that identify abbreviations from the full text.

Our method can contribute to a larger studies involving abbreviation disambiguation providing assistance in the abbreviation extraction phase of the studies. To better understand our method's contribution, we propose that designers integrate our method into such a disambiguation system along with other components and that they evaluate the contributions of each component.

Extracting abbreviations is an important area for research in text mining applications. Our work adds to the available knowledge on abbreviation definition features by providing a means to identify the abbreviations and their expansions correctly.

6 Acknowledgements

Thanks to Dr. Catherine Blake for motivating me to conduct this study, teaching me a systematic way of conducting such a study, providing material assistance by way of preprocessed TREC collection as a part of RENCI Faculty Fellowship Program, database space, processing machines, and helping me produce a well-designed report.

Thanks for the faculty and staff of UNC-School of Information and Library Science for helping me in every way possible and encouraging me to work hard.

7 References

Abbreviation .http://en.wikipedia.org/wiki/Abbreviation

Ao, & Takagi. (2003). An algorithm to identify abbreviations from MEDLINE. *Genome Informatics, 14*, 697-698.

Baeza-Yates, & Ribeiro-Neto. (1999). Model information retrieval, *Addison Wesley*.

Chang, Schütze, & Altman. (2002). Creating an online dictionary of abbreviations from MEDLINE. *Journal of the American Medical Informatics Association, 9*(6), 612-620.

Chen, Stewert, Davies, Giller, Krailo, Davis, et al. (2005). Parental occupational exposure to resticides and childhood germ-cell tumors. *American Journal of Epidemiology, 162*(9), 858-867.

Fukuda, Tamura, Tsunoda, & Takagi. (1998). Toward information extraction identifying protein names from biological papers, *Pacific Symposium on Biocomputing*, 707-718.

HaCohen-Kerner, Kass, & Peretz. (2008a). Abbreviation disambiguation: Experiments with various variants of the one sense per discourse Hypothesis. *Natural language and information systems* (pp. 27-39) Springer Berlin / Heidelberg.

HaCohen-Kerner, Kass, & Peretz. (2008b). Combined one sense disambiguation of abbreviations. , *HLT, Short Papers (Companion Volume)* 61-64.

Hallert, Husberg, Jonsson, & Skogh. (2004). Rheumatoid arthritis is already expensive during the first year of the disease (the sweedish TIRA project). *Rheumatology, 43*(11), 1374-1382.

Hersh, Cohen, Roberts, & Rekapalli. (2006). TREC 2006 genomics track overview. *In TREC Notebook, NIST*

Jablonski. (1993). Dictionary of medical acronyms and abbreviations. 2nd ed. *Hanley & Belfus*.

Kallunki, Edelman, & Jones (1997). Tissue-specific Expression of the L1 Cell Adhesion Molecule Is Modulated by the Neural Restrictive Silencer Element. *The Journal of Cell Biology,*138:1343 - 1354.

Liu, Aronson, & Friedman. (2002). A study of abbreviations in MedLINE abstracts, *Proceedings of the AMIA Symposium,* 464–468

Liu, Lussier, & Friedman. (2001). A study of abbreviations in the UMLS. *Proceedings of AMIA Symposium,* 393–397

Masson, Sharp, Cotton, & Little. (2005). Cytochrome P-450 1A1 gene polymorphisms and risk of breast cancer: A HuGE review. *American Journal of Epidemiology, 161*(10), 901-915.

Mondugno. (2004). Ovarian cancer and polymorphisms in the androgen and progesterone receptor genes: A HuGE review. *American Journal of Epidemiology, 159*(4), 319-335.

Pakhomov, Pedersen, & Chute. (2005). Abbreviation and acronym disambiguation in clinical discourse. *AMIA Symposium*, 589-593.

Park, & Byrd. (2001). Hybrid text mining for finding abbreviations and their definitions. *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing,*126-133.

Pustejovsky, Castaño, Cochran, Kotecki, & Morrell. (2001). Automatic extraction of acronym-meaning pairs from medline databases. *Medinfo, 10(Pt 1)*, 371-375.

Schwartz, & Hearst. (2003). A simple algorithm for identifying abbreviation definitions in biomedical text. *Pacific Symposium on Biocomputing*, 451-462.

Sharp, & Little. (2004). Polymorphisms in genes involved in folate metabolism and colorectal neoplasia: A HuGE review. *American Journal of Epidemiology, 159*(5), 423-443.

Sola. (1992). Abbreviations dictionary. 8th ed. *CRC Press.*

Taylor, & Fei. (2005). Emx2 regulates mammalian reproduction by altering endometrial cell proliferation. *Molecular Endocrinology, 19*(11), 2839-2846.

Yoshida, Fukuda, & Takagi. (2000). Pnad-css: A workbench for constructing a protein name abbreviation dictionary. *Bioinformatics (Oxford, England), 16*(2), 169-175.

Yu, Hripcsak, & Friedman. (2002). Mapping abbreviations to full forms in biomedical articles. *Journal of the American Medical Informatics Association, 3*(262), 272.