# Evaluation of Index Term Discovery in Medical Reference Text

**Dennis Wollersheim, Wenny Rahayu and James Reeve**

*Abstract* - **Currently, and in the foreseeable future, the amount of text in electronic form is multiplying. While retrieval of this text is facilitated by increased indexing, there is a danger of overwhelming the user with too much information. As well, the current rise in available computing power means that we can consider computationally intense solutions. One such solution is dynamic taxonomy. Dynamic taxonomy uses an ontology based index, giving it the ontological advantages of structure, and increased number of interconnections. Subsequent complexity is managed by dynamic pruning.**

**Dynamic taxonomy needs a text base that has been deeply indexed with concepts from a taxonomy. In this paper, we evaluate various algorithms for generating such an index. For the text, we use a subset of an electronic medical reference database, and the indexing terms are extracted from an established medical ontology, the Unified Medical Language System (UMLS).**

**This paper compares an index created by various algorithms against one devised by medical domain experts. The algorithms extract candidate phrases from the text using techniques such as NLP based phrase extraction, combinations of N consecutive words, and combinations of preexisting book index terms. These candidate phrases are normalised, and then matched against the UMLS database.**

*Index Terms* - **dynamic taxonomy, indexing, information retrieval, natural languages, ontology**

## I. INTRODUCTION

Since the advent of the book, the computer, and especially the internet, the amount and availability of text has increased dramatically. To make effective use of this resource, we need ways to access and retrieve this text. Indexing, the assigning of keywords to chunks of text, is one such way.

An index functions as a way to access text, group similar text, and summarise text content. Traditionally, indexing has been an activity of human communication. A index term is chosen because someone thinks it will be a useful way to access a particular piece of text. Indexing distils the meaning of text.

While paper based indexing has conventionally been at 'page' granularity, computer based indexing has the potential to index at deeper and finer levels. A example of this in the medical field is differentiation between child and adult dosages. One problem with these levels of indexing is that they are labour intensive. This paper investigates the automated construction of such fine grained indexes.

These indexes could help the user by increasing recall, providing more targeted retrieval and eliminating extraneous text, but there is a danger of overwhelming the user with too many index terms. Because of this, we look at using ontologies to add structure to the index.

An ontology is a formal specification of concepts and objects, and the relationships that hold among them. The amount of ontological material is also recently increasing (for example, see [1]). By connecting an ontology to an index, the index is changed from an alphabetised list of words into a structure that has intrinsic meaning. This also enriches the underlying text, by expanding its connections within itself, and by connecting it to a view of the world.

The use of deep indexing and ontological connection as an access methods is complex, but the current rise in available computing power means that we can consider computationally intensive solutions. One such solution is dynamic taxonomy (DT).

## II. BACKGROUND: DYNAMIC TAXONOMY

DT, proposed by Sacco [2], uses a taxonomic subset of ontology, consisting only of concepts and the IS-A type relationships that join them. It reduces the complexity of an index in two ways. First, it provides a standard taxonomic ordering of an index term set, allowing complexity to be hidden under higher level terms. Second, it provides a zoom operator, which dynamically reduces the term set.

Zoom functions in the following manner. In response to a request to zoom in on a index phrase, the system prunes the taxonomy, retaining only the taxonomic elements which categorise the set of text atoms which are also categorised by the zoom phrase. More importantly, any terms in the taxonomy that do not either directly or indirectly classify the remaining text atoms are also pruned. For example, fig. 1 shows a sample taxonomy derived from the medical field. Fig. 2 shows the pruned taxonomy after a zoom on the concept amoxycillin. If a text atom is not classified by amoxycillin, it is not included in the pruned taxonomy.

A similar medically based dynamic hierarchical categorisation search system has been shown to be an effective tool for users. *Dynacat* [3, 4] users found significantly ($P<0.05$) more answers in a fixed amount of time, and were significantly ($P<0.05$) more satisfied with this tool over cluster or ranking systems.

To function, a dynamic taxonomy needs a text base that has been multiply classified by concepts from a taxonomy. The taxonomy will be derived from an established medical ontology, the Unified Medical Language System [5] (UMLS, 2001 version) metathesaurus. In this work, we investigate three index generation methodologies:

This paper describes research being carried out within projects funded by the Australian Research Council, Australian National Prescribing Service, and Therapeutic Guidelines Limited.

Dennis Wollersheim is a Phd student at La Trobe University, Bundoora Victoria, Australia (email dewoller@cs.latrobe.edu.au, telephone: +61 3 9479 1280 ).

Dr. Wenny Rahayu is a senior lecturer at La Trobe University, Bundoora Victoria, Australia (email: wenny@cs.latrobe.edu.au, telephone: +61 3 9479 1282).

James Reeve is a researcher at Therapeutic Guidelines Ltd, Level 2, 55 Flemington Rd, North Melbourne, Australia (email: jreeve@tg.com.au, telephone: +61 3 9326 9959)

ISBN: 1-86467-114-9

- extraction from source text,
- extraction from book index terms, and
- manual domain expert indexing.

The procedure for building the actual taxonomy from the set of found base terms is detailed in our earlier work [6].

For the text base, we use a subset of an electronic medical reference database, provided by Therapeutic Guidelines Limited (TGL), an Australian medical guideline publisher. The text was originally in handbook format [7], and has since been broken into HTML based fragments, which are the source for this work.

This text was designed for quick reference, and so has dense, wide ranging information content. This makes it appropriate for fine grained classification and small chunk retrieval, as the result will be useful for busy medical practitioners, who are often looking for single points of information. Initially, we plan to classify (and retrieve) the text at paragraph granularity.

UMLS, a project of the USA National Institute of Health, is an amalgamation of different sets of medical terminology. A key constituent of the metathesaurus is a *concept*, which serves as nexus of terms across the different terminology sets. This means that it is strong in the area of synonymy. For example, table I shows UMLS terms that are synonyms for the concept 'skin' (concept ID C0037267). The value of UMLS concepts are enhanced by being interconnected through a set of relationships.

## III. DETAILED METHOD FOR GENERATION OF INDEX PHRASES

This section details three methods for generation of potential index phrases from medical reference text. These methods are described briefly below, and their advantages and disadvantages set out in table II.

In the area of source text phrase extraction, we use two methods: *phrase extraction*, and *moving window word combinations*. Phrase extraction looks at phrases found by existing part of speech (POS) taggers. Recent research in this area is encouraging, finding 80% true positives when using phrase extraction tools to automate shallow, document

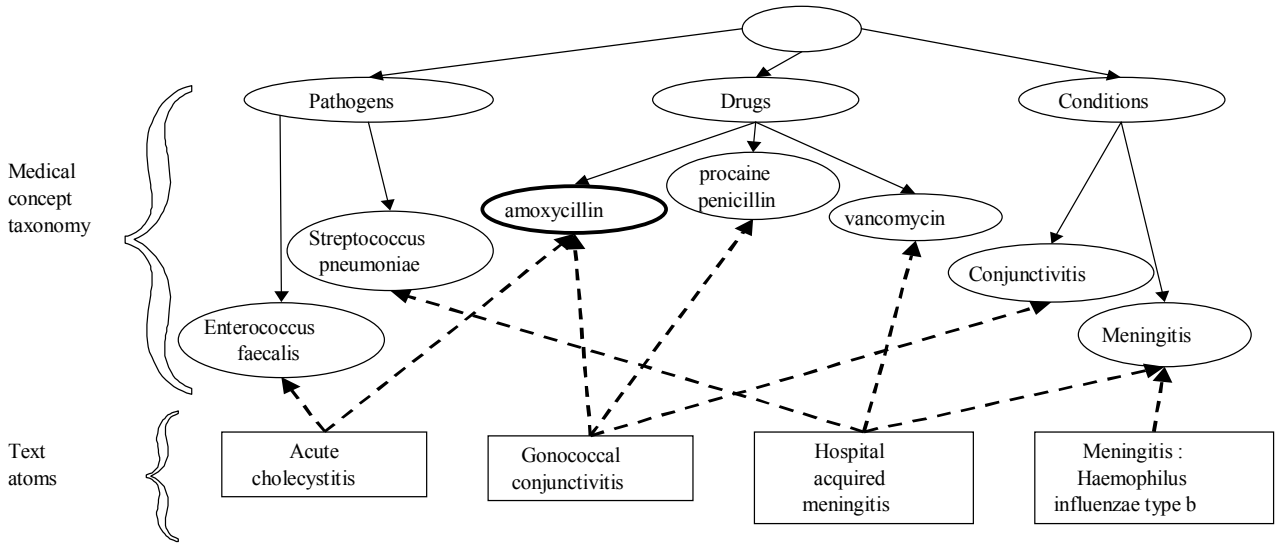


Fig. 1. Example taxonomy, and subsequently classified text atoms. Solid lines shows taxonomic IS-A links, while dotted lines denoted classification by a taxonomic term

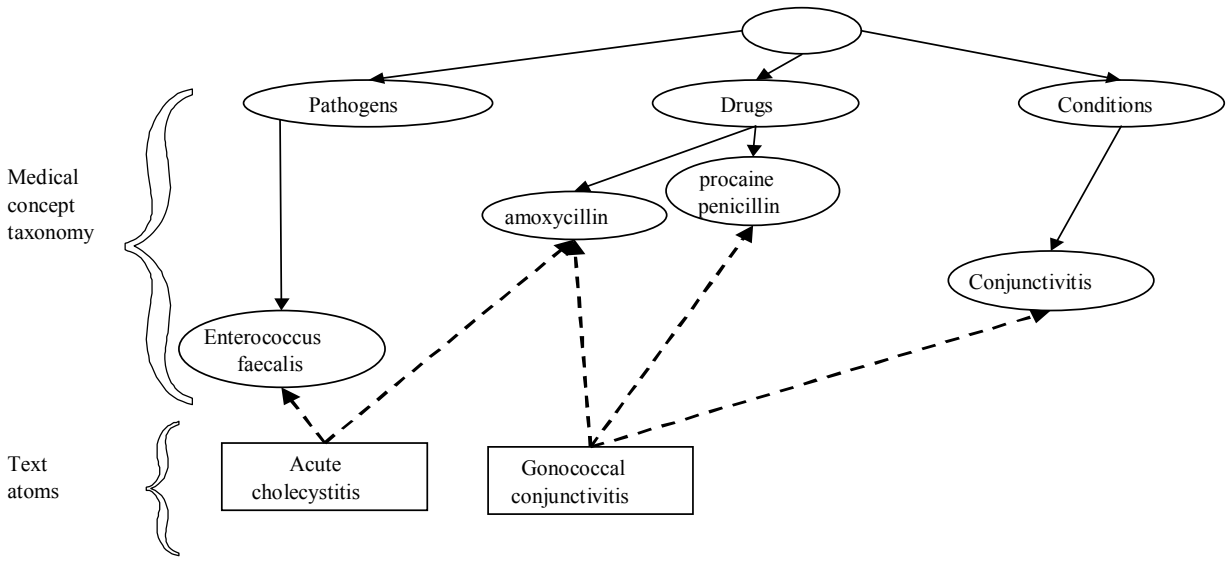


Fig. 2 Remaining taxonomy after a zoom on the concept amoxycillin. Atoms that are not classified by amoxicillin are pruned; also pruned are the sections of the taxonomic tree that would be unused.

| | |
|---|---|
| cutaneous tissue | skin |
| integument | skin (anatomy) |
| integumentary system | skin <1> |
| integumentary system, nos | skin nos |
| integumentum commune | skin system |
| section 0 integumentary system | skin, nos |
| | 010-012 skin |

Table I: UMLS terms that are synonyms for 'skin', concept ID C0037267

level indexing of medical text by UMLS [8]. We compare three web available part of speech taggers: two statistical taggers, namely *Brill Tagger [9],* and *Treetagger [10],* and a more recent rule-based constraint grammar tagger, the *EngCG Tagger* [11].

A second method of index term generation uses the existing *book index* as a base. It gets candidate terms from the list of key words that was assigned to the text when it was in book form. The generated phrases from each of the above methods are then put into a canonical form, and filtered for membership in UMLS.

In contrast to the above two algorithmic methods, the *expert index* was manually produced. Each paragraph from

| Index type | Positive | Negative |
|---|---|---|
| Indices generated from original book index ($T_b$) | • Deliberately chosen – *the* most important terms<br>• Pruned | • Limited number of terms<br>• Vocabulary not controlled<br>• Page level granularity<br>• Text must be pre-indexed |
| Expert index ($T_e$) | • Derived from UMLS<br>• Human expertise used<br>• Paragraph granularity | • Huge time commitment<br>• Impractical to do large scale text volume<br>• Not full coverage |
| Indices generated from source text ($T_s$) | • From UMLS<br>• No human intervention necessary<br>• Paragraph granularity | • Unknown validity |

Table II: Positives and negatives of the various indexing methods.

A more detailed description of each of the methods follows, and the entire process is detailed in fig. 3.

*A. Expert Indexing*

To construct the expert index, a domain expert was instructed to assign a maximum number of applicable

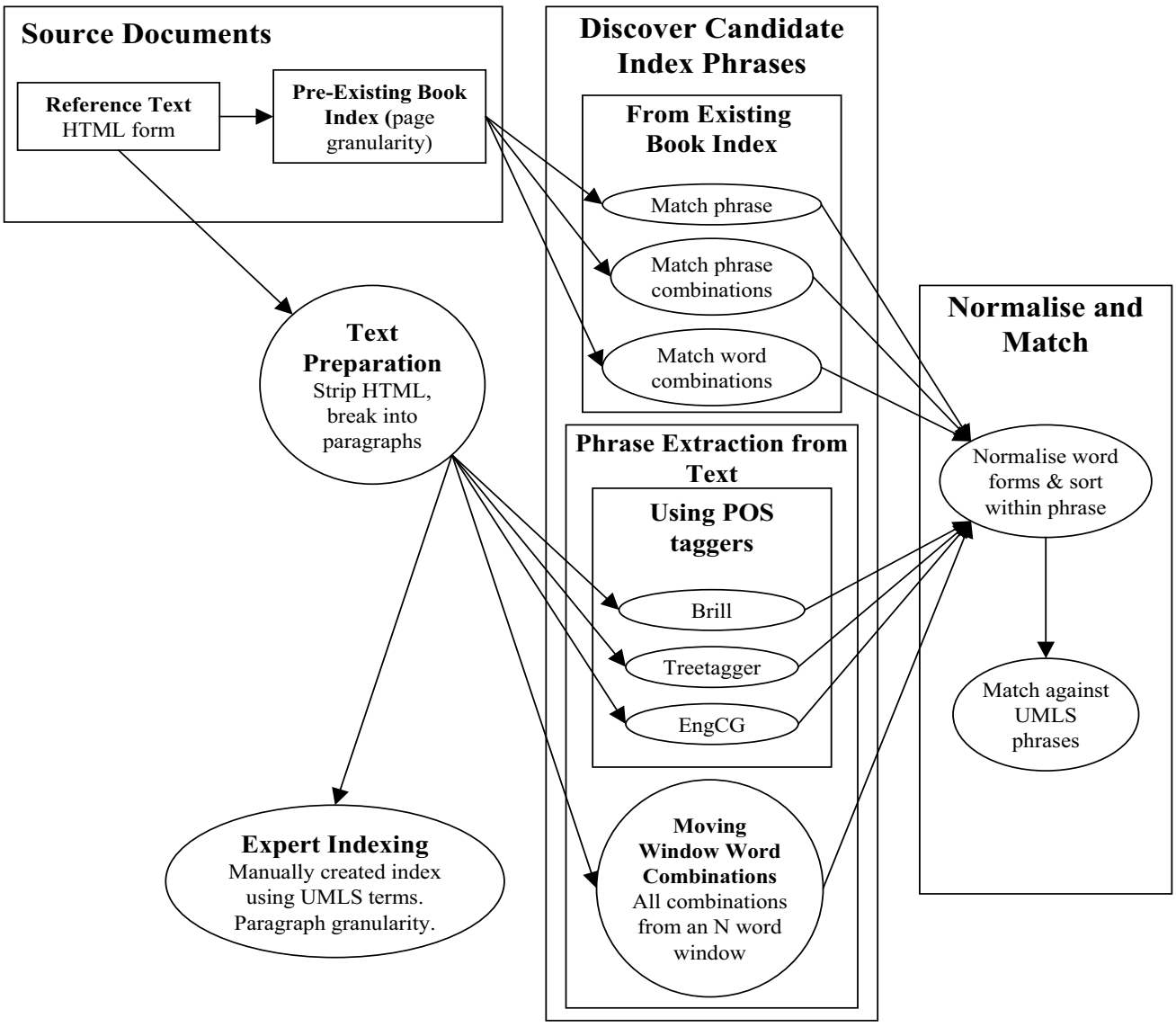


Fig. 3: Workflow outlining methods used to generate index phrases. The input is medical reference text, and the output is various sets of UMLS derived index words.

concepts to each paragraph. This is due to the requirements of DT, which needs multiple indexing of concepts, and that of our application, where we want to retrieve minimum useful length chunks of text.

Because the UMLS has a large number of concepts (almost 800,000), it would be labour intensive to construct the desired index from scratch. On the other hand, we did not want to provide a list of all the concepts that our algorithms had found in the text, theorising that this would bias the index creation process, and discourage the search for other valid concepts, thus limiting the index. We compromise by offering a targeted subset of the UMLS, generated as follows.

For each word in the text, a list of all the UMLS concepts that contain this word are extracted. If the number of unique concepts is less than 50, the preferred form of these concepts are offered to the expert to choose from. Similarly, if there are greater than 50 concepts containing this word, but one of them is a concept consisting of exactly that word, it too is offered to the expert. The cutoff point of 50 terms was chosen because we believe that 1) it is infeasible for the expert to process more than that many terms, and 2) words that were this common would be unlikely to lead to index terms, for example, the terms 'metallic' and 'numb'.

*B. Automatic Extraction of Potential Index Phrases*

Potential index phrases are produced by two sets of methods. The first set extracts phrases from original text, variously selective (phrase identification), or indiscriminate (moving window word combination). The second set of methods uses the existing book index.

*i. Text Preparation*

Phrase extraction begins with text preparation. The source text is in HTML form, so we first strip out the HTML tags, as the following extraction algorithms function best with plain text. Simple format semantics are captured through the preservation of text breaks. The result is that the extracted phrases come from only within the sentence. This is accomplished by inserting end of sentence markers in place of HTML tags (such as <BR> and </TD> that would have marked such breaks. Additionally, the resulting paragraphs are numbered, so that results can be consistently compared.

*ii. Candidate Phrase Extraction from Source Text*

Both the phrase extraction (PE) and moving window (MW) techniques start with the raw text prepared above. The phrase extraction process begins with tagging of the raw text by a part of speech tagger. We use the Brill tagger, Treetagger, and the EngCG taggers for this purpose, to provide a comparison of effectiveness. From the output of the taggers, consecutive words tagged as nouns or adjectives, uninterrupted by punctuation, are chosen as candidate phrases for the next step. Table III shows the phrases extracted by the Treetagger algorithm from one sentence of raw text.

As an alternative to PE, we also devised the MW algorithm. It is more exhaustive, using a combinatorial approach. It works as follows. For each sentence, each possible combination of N consecutive words (with word length > 3) becomes a candidate phrase.

| **Input Sentence** | **Treetagger extracted phrases** |
|---|---|
| Dermatitis is a non-specific inflammatory response of the skin to a combination of exogenous and endogenous factors. | – Dermatitis<br>– non-specific inflammatory response<br>– skin<br>– combination<br>– exogenous<br>– endogenous factors |

Table III: Phrases derived from an input sentence by Treetagger algorithm.

*iii. Candidate Phrase Extraction from Book Index*

Because the existing book index is set out in a 'level 1, level 2, … level N' format, there are multiple related phrases that make up a single index entry,. An example of a three level index entry is "exogenous dermatitis, retinoids, phototoxicity". From this, we generate candidate phrases in three ways, variously maximising precision, maximising recall, or compromising these two goals. The first algorithm simply uses the individual index phrases as candidate phrases, the second tries all combinations of entire index phrases, and the third tries all combinations of single index words. Table IV shows an example.

*C. Phrase Normalisation and Matching*

After a candidate phrase is generated, it is then checked against the master list of all English language UMLS phrases (the UMLS file MRNXS.ENG) to see if it exists there. The phrases in this list are stored in a normalised form, so we too normalise our phrases before checking them. UMLS provides a text manipulation tool (called LVG) which does various transformations that result in a normalised word. In our case, we use the LVG command line switches '-fltgpB -fC', which normalises the phrase using 2 different algorithms. After normalisation, the candidate phrases are sorted in alphabetic order, and checked against the UMLS master list. Each exact match of any synonym of a UMLS concepts is deemed to be an index concept.

| **Technique** | **Candidate phrases generated** |
|---|---|
| Individual index phrases | – Exogenous dermatitis<br>– Retinoids<br>– Phototoxicity |
| Combinations of phrases | *As above, and*<br>– exogenous dermatitis retinoids<br>– retinoids phototoxicity<br>– exogenous dermatitis phototoxicity retinoids |
| Combinations of words | *As above, and*<br>– exogenous retinoids<br>– exogenous phototoxicity<br>– exogenous retinoids phototoxicity<br>– dermatitis retinoids<br>– dermatitis phototoxicity<br>– dermatitis retinoids phototoxicity |

Table IV: Candidate phrases generated by various methods from three level index entry "exogenous dermatitis, retinoids, phototoxicity".

## IV. IMPLEMENTATION AND RESULTS

The input medical reference text consisted of 6 pages of text, containing a total of 2982 words. The expert index process found 615 UMLS concepts in the text, the Brill tagger found 982, and the six word moving window found 2894.

For a given text, the theoretical set of index terms $T_p$ index it perfectly. Tp includes all the possible avenues that a user would use to access the text. In the following results, we compare the three indexing methods used against $T_p$.

The set of index terms $T_b$ was derived from the original book index, $B_{original}$. $B_{original}$ is a concise subset of chosen index terms; it was carefully produced, and methodically verified over a long production cycle. Because of this, $T_b$ will have the high precision. We assume that almost all of the terms in $T_b$ would be in $T_p$. Unfortunately, $T_b$ has low recall because of the limited number of original index terms, and ultimately, it not useful as a basis for $T_p$, because it has only page level granularity. It becomes valuable as a way to evaluate the other methods.

An example of this is when we compare $T_b$ to the set of expert index derived terms, $T_e$. While the terms that made up $T_e$ were not subjected to the same scrutiny that $B_{original}$ received, they were vetted by an expert. As such, $T_e$ should have high precision. But $T_e$ found only 35.5% of $T_b$ (the Tb derived with the most restrictive parameters, compared on a per-page basis). So, while $T_e$ has high precision, it must have low recall, as it misses 64.5% of terms that must be in $T_p$.

On the other hand, the set of index terms derived from the source text ($T_s$) has much higher recall of $T_b$ terms; the moving window technique (with window length=2 words) retrieved 73.6% of $T_b$, and TreeTagger retrieved 56.2%. While this is not surprising (it merely shows that medical phrases that are in the index are also likely to be in the source text), it does point to the specific deficiencies of $T_e$. The problem is not that $T_e$ was small; it is of the same magnitude of $T_s$. It is merely a different set of terms. An error analysis follows.

$T_e$ was produced by choosing concepts from a UMLS subset. The subset's size was greater than 10,000 terms; this reduced the expert's motivation to index further. This problem could have been alleviated by the production of a better interface into the UMLS database.

An analysis of the missing book index terms (the set $T_b - T_e$) show that concepts were left out because they 1) were unrepresented in the text, 2) were not included in the list offered because they were represented in the text by a word that was too common in UMLS (e.g. *water*, and *arm*), or 3) they were overlooked in the list.

Even though there are deficiencies in $T_e$, it is the best fine grained approximation of $T_p$ that exists. Because of this, we now look at precision and recall of the different $T_s$ algorithms in comparison to $T_e$.

Among $T_s$ derivation algorithms, there was little surprise. The phrase extraction algorithms (PE) all had higher precision and lower recall compared to the moving window algorithms (MW); this makes sense, as PE is generally a subset of MW (where window length > 2 words). Full results are in table V.

$T_s$ did not perform as well as $T_e$ in the retrieval of multi-word phrases. This makes intuitive sense; a human indexer is more likely to be able to distil and match meaning from multi word phrases than a automated tool. Many of the multi-word phrases in $T_e$ are specific examples of a more general term found in $T_s$. For example, $T_s$ holds the term *contact dermatitis*, while $T_e$, in addition to holding the parent term, also holds child terms *contact dermatitis due to solvents*, *contact dermatitis due to jewellery, contact dermatitis due to detergent, contact dermatitis due to poison ivy, contact dermatitis due to poison oak, contact dermatitis due to poison sumac,* and *contact dermatitis due to poison vine.*

| **Part of Speech Taggers** | # of terms | Precision | Recall |
|---|---|---|---|
| Brill | 948 | 26.3% | 17.1% |
| Conexor | 877 | 27.8% | 19.5% |
| Treetagger | 820 | 26.0% | 19.5% |
| **Moving Window** (N: Window width) | | | |
| Combined Words (N=1) | 2521 | 47.0% | 11.5% |
| Combined Words (N=2) | 2719 | 49.9% | 11.3% |
| Combined Words (N=3) | 2803 | 50.7% | 11.1% |
| Combined Words (N=4) | 2834 | 50.9% | 11.0% |
| Combined Words (N=5) | 2863 | 51.1% | 11.0% |
| Combined Words (N=6) | 2894 | 51.2% | 10.9% |

Table V: Precision and recall of source text based index term generation algorithms ($T_s$) compared to expert indexing ($T_e$).

This phenomenon can be explained as follows. Because the expert maximised index size, groups of specific, highly described terms were included. The string representations of these concepts are wordy, and so, unlikely to be specifically matched by any of our algorithms. On the other hand, the broader parent concept, which often has a more concise string representation is more likely to be included. The DT algorithm itself will ameliorate this phenomena, due to its hierarchical nature.

Ts includes a set of false positive type errors that rise out of the UMLS itself. UMLS is a conglomeration; it was not designed systematically. As such, it includes words that are specific to the medical domain and make high quality index terms (e.g. *dermatitis*); it also includes common words, that would not be such good index terms, such as 'week', 'year', or 'response'. It is possible that there is a way to filter UMLS so that these low quality words are excluded, possibly by looking at word frequency.

Another error is that of scope. While we are trying to index at a paragraph level, the meaning of a paragraph does not neatly stop at the end of the paragraph. Any valid indexing of a document would need to bring the meaning from a one paragraph to bear on the next. This explains the rise in precision and recall for all algorithms when comparing paragraph level term set to document level term set.

## V. FUTURE WORK

The first place to work is to implement the dynamic taxonomy itself, and so provide a place to test the indexes created. This will give us some feedback as to the validity of each of the different algorithms.

There is a case for better support of the expert indexer, and providing expanded tools for this job. This includes a smoother interface to UMLS, and a facility to allow vetting of index phrases found by the other algorithms. While the latter would lessen the validity of using the expert index as a comparison, it would increase the quality. Additionally, it would be useful to be able to assign terms a both book and paragraph level.

There is room for improvement in our algorithm. Nadkarni et al. [8] obtain substantially better recall and precision using an algorithm specialised for the medical domain. While our work is more generic, (by plugging in of a new word list, applicable to any domain), we can consider their techniques. They use a commercial phrase extraction tool, and match against a chosen subset of UMLS. The extracted phrases are stripped of stop words, and matched against the UMLS preferred term only, (which is possibly not appropriate in the Australian context; the UMLS preferred form has cultural specificity). The checking for matches are done using descending length combinations of words chosen from the extracted phrases, and when a partial phrase match is found, the remaining words are checked using the same technique.

A more detailed error analysis would help us tune the algorithm. There are several variables that possibly relate to inclusion in expert index. These include:

- intrinsic factors from UMLS; for example, some of the UMLS source term sets would be likely more applicable to the material being indexed than others,
- syntactical and semantic meaning from the text itself; this is exemplified by the detail returned by the phrase extraction tool EngCG (we currently only use the POS information), or
- the semantic features within the format of the document, e.g.: title.

The information found would then be used to bias the phrase generation routines.

## VI. CONCLUSION

This work establishes a comprehensive framework for the construction of base index term sets that will be used by dynamic taxonomies. It allows us to check and compare both algorithms for index generation, and tools for manual indexing. The use of automated index construction methods for use in dynamic taxonomy has not been previously done. Dynamic taxonomies themselves are relatively unexplored. Outstanding questions include the appropriate level of detail to an index, best methods for construction of manual indexing and verification tools, etc.

This work is the first step in creation of a detailed connection between a taxonomy and a medical reference text. Dynamic taxonomy is a way to make use of a detailed index to improve retrieval. Here, we explore automated ways of making the connections between the taxonomy and the text base.

## *ACKNOWLEDGMENT*

This project could not have been possible without the support and contributions of Dr. Bryn Lewis, Dr. Ken Harvey, Dr. Jonathan Dartnell, Ms. Mary Hemming, Assoc. Prof. Teng Liaw, and Ms. Elizabeth Deveny.

## *REFERENCES*

[1] Lenat, D.B., *Cyc: A Large-Scale Investment in Knowledge Infrastructure.* Communications of the ACM, 1995. **38**(11).

[2] Sacco, G., *Dynamic taxonomies: a model for large information bases.* IEEE Transactions on Knowledge & Data Engineering, 2000. **12**(3): p. 468-79.

[3] Pratt, W., *Dynamic Categorization: A Method for Decreasing Information Overload*, in *Medical Information Sciences*. 1999, Stanford University. p. 171.

[4] Pratt, W. and L. Fagan, *The Usefulness of Dynamically Categorizing Search Results.* Journal of American Medical Informatics Association, 2000. **7**(6): p. 605-617.

[5] Lindberg, D.A.B., B.L. Humphreys, and A.T. McCray, *The Unified Medical Language.* System. Meth. Inform. Med., 1993. **32**: p. 281-291.

[6] Wollersheim, D. and W. Rahayu. *Methodology For Creating a Sample Subset of Dynamic Taxonomy to Use in Navigating Medical Text Databases*. in *IDEAS 2002*. 2002 (Accepted for publication). Edmonton, Alberta, Canada.

[7] Stevens, D.M., ed. *eTG: Dermatology*. 1 ed. 1999, Therapeutic Guidelines Ltd: Melbourne.

[8] Nadkarni, P., R. Chen, and C. Brandt, *UMLS Concept Indexing for Production Databases: A Feasibility Study.* Journal of American Medical Informatics Association, 2001. **8**: p. 80-91.

[9] Brill, E., *Some Advances In Rule-Based Part of Speech Tagging.* AAAI, 1994.

[10] Schmid, H. *Probabilistic Part-of-Speech tagging Using Decision Trees*. in *Conference on New Methods in Language Processing*. 1994. Manchester, UK.

[11] Voutilainen, A., *EngCG tagger, Version 2*, in *Sprog og Multimedier*, T. Brondsted and I. Lytje, Editors. 1997, Aalborg Universitetsforlag: Aalborg.