Enha

***ANALYZING LEXICAL TOOLS’ FRUITFUL VARIANTS for concept mapping IN THE SYNONYM MAPPING TOOL***

***RoseMary Hedberg, MLIS Associate Fellow, 2012-2013 Project Leaders: Allen Browne and Chris Lu, PhD***

**Final Report**

***Spring 2013***

***TABLE OF CONTENTS***

Acknowledgements 3

Abstract 4

Introduction 5

Methodology 7

Results 11

Discussion 13

Recommendations 16

References 17

***ACKNOWLEDGEMENTS***

I would like to thank my project sponsors Allen Browne and Chris Lu for their support and advice.

Thank you to Kathel Dunn, Coordinator of the Associate Fellowship Program, for making allowances and letting me change my mind so many times.

I would also like to thank Kin Wah Fung and Julia Xu for their additional contributions to the project and for volunteering their time and expertise.

***ABSTRACT***

***OBJECTIVE:***

The Synonym Mapping Tool (SMT), developed by the Lexical Systems Group, maps terms to concepts in the UMLS Metathesaurus by using synonym substitution for terms found in a synonym corpus. SMT may be improved through the inclusion of lexical tool fruitful variants, which include spelling variants, inflectional variants, acronyms, abbreviations and their expansions, and derivational variants.

Individually, each of the variants contributes to the rate of recall and precision in synonymous concept mapping; however, the exact extent to which they each contribute is unknown. The goal of this project is to determine the individual weight of each variant.

***METHODS:***

Individual lexical variant synonym test scenarios were created by applying variants to the SMT process of subterm substitution. SMT was run for each variant test scenario using the UMLS-CORE Subset as an input term set. The UMLS-CORE Subset is known to be expertly mapped in the Metathesaurus and is seen as a gold standard. The results of running SMT with each variant synonym test scenario will be compared to the gold standard to determine the differences in mapping performance, thereby showing exactly how well each scenario (and thusly, each variant) performed individually in terms of precision and recall.

***RESULTS:***

The differences in variant test scenario mapping results will be calculated in order to determine individual variant type weight and provide reference for precision and recall.

***CONCLUSIONS:***

Finding a balance point between recall and precision will assist users who want to choose the combination of variant types in their searches, as discovering the actual cost to performance of each variant will allow users to accurately choose the best set of variants to use for their natural language processing research.

***INTRODUCTION***

The Lexical Systems Group developed the Sub-Term Mapping Tools (STMT), which is a generic tool set that provides sub-term related features for query expansion and other natural language processing applications. The Synonym Mapping Tool (SMT) is one of the most commonly used tools in the STMT package and is designed to find concepts in the Unified Medical Language System (UMLS)-Metathesaurus using synonym substitutions. Synonyms for sub-terms of an input term are found by loading the input term into a corpus of normalized synonyms. Terms with the same, similar, or related meanings are considered synonymous within the SMT, and may be used as substitute sub-terms to improve coverage without sacrificing accuracy. The synonyms substitute sub-terms in various patterns to form new terms for concept mapping. For example, the term “decubitus ulcer of sacral area” does not map to a corresponding Concept Unique Identifier (CUI) in the UMLS-Metathesaurus. However, if the synonyms “pressure ulcer” and “region” are substituted for the sub-terms “decubitus ulcer” and “area” respectively, the resulting term “pressure ulcer of sacral region” maps to CUI C2888342 within the Metathesaurus. By applying this sub-term synonym mapping query expansion technique in an earlier UMLS-CORE project, the SMT was able to increase the coverage rate of CUI mapping by 10%, while still maintaining accuracy1.

The performance (precision and recall) of the SMT in finding mapped concepts depends mainly on the comprehensiveness of the synonym corpus, which is itself dependent on the effectiveness of sub-term substitution through the application of lexical variant types. These lexical, or “fruitful”, variant types are found within another tool set, the Lexical Tools package, and include spelling variants, inflectional variants, acronyms and abbreviations, expansions, and derivational variants.

Spelling variants relate to orthography and include minute differences such as align – aline, anesthetize – anesthetize, and foetus – fetus. Inflectional and derivational variants refer to word morphology. Inflectional variants of terms include the singular and plural for nouns, verb tenses, and various changes to adjectives and adverbs. Nucleus – nuclei, cauterize – cauterizes, and red – redder – reddest are all examples of inflectional variants. Derivational variants derived new words from existing words by adding or removing a prefix or suffix, such as laryngeal – larynx and transport – transportation. Acronym and abbreviation variants are included in the Lexical Tools package, and they form the initial or shortened components of a phrase or word. Expansion variants react inversely, and map acronyms or abbreviations to their long-form versions, such as NLM – National Library of Medicine.

Each of these variant types was previously assigned a rough distance score to denote the cost of query expansion. Each variant type increases recall, and ideally, the lower the distance score, the less damage to precision. The greater the distance score, the higher a likelihood of meaning drift or error in the form of false positives and true negatives in mapping.

| Operation | Notation | Distance Score |
| --- | --- | --- |
| Spelling Variant | s | 0 |
| Inflectional Variant | i | 1 |
| Synonym Variant | y | 2 |
| Acronym/Abbreviation | A | 2 |
| Expansion | a | 2 |
| Derivational Variant | d | 3 |

Individually, each of the variant types, when applied, contributes to the coverage rate of synonymous concept mapping; however, the exact extent to which they each contribute is unknown. The performance of SMT in concept mapping could be improved by identifying the weight of each lexical variant type and determining how they individually contribute to the coverage rate of finding mapped concepts. These individual weights could then be translated into non-arbitrary distance scores.

***METHODOLOGY***

CREATING THE INDIVIDUAL LEXICAL VARIANT SYNONYM TESTING SCENARIOS

The first step in testing performance involved creating individual lexical variant synonym testing scenarios so the mapping performance of each variant type could be compared to that of the baseline Specialist Lexicon synonym corpus. These variant synonym testing scenarios were created by running each variant type through Java programs and shell scripts that run both SMT and the Lexical Variants Generation (LVG) program, which is a suite of utilities that can generate, transform, and filter lexical variant types from a given input. The first script used was GetBaseVars, which creates and normalizes a file of all synonyms found by applying a variant type to the baseline synonym corpora. The GetBaseVars script also transforms the output from a standard LVG format of:

| Field 1 | Field 2 | Field 3 | Field 4 | Field 5 | Field 6 | Field 7+ |
| --- | --- | --- | --- | --- | --- | --- |
| Input | Output | Categories | Inflections | Flow History | Flow Number | Add’l Information |

to a much more manageable format, selecting only base forms and removing duplicate words. The resulting files were then input into GetSynonyms, a SMT script that adds the variant synonyms generated by GetBaseVars to the baseline synonym corpus, thus creating new normalized variant synonym testing scenarios (Baseline, Baseline + Spelling Variants, Baseline + Inflectional Variants, etc.).

RUNNING THE INITIAL TEST

The comparative mapping performances of the newly created individual variant synonym testing scenarios and the existing baseline synonym corpus were tested through another SMT script, FVTest. FVTest is the script that actually runs a given set of input terms through SMT using the new variant synonym testing scenarios, while also accounting for normalization, resulting in a list of input terms mapped to CUIs through sub-term substitution. Note in the following workflow that running SMT is not a one-step progression. Normalization, performed using STMT’s Normalization Tool, is applied at two distinct instances in the mapping process. Synonym Norm, or SynNorm, maps input terms to synonyms by abstracting away from genitives (possessive case), parenthetical plural forms, punctuation, and cases, while also removing duplicated results. Lexical Tools Norm, or LvgNorm, is used within the Metathesaurus to normalize terms for CUI mapping by abstracting away from genitives, parenthetical plural forms, punctuation, and cases, as well as symbols, stop words, Unicode, and word order.

The FVTest script also generates a log file, which logs the performance of the particular synonym testing scenario in mapping a set of input terms by calculating the number of total input terms, the number of terms mapped to CUIs with simple normalization, the number of terms mapped to CUIs with 1 sub-term synonym substitution, the number of terms mapped to CUIs with 2 sub-term synonym substitutions, the number of terms that were not mapped to CUIs using that particular synonym testing scenario, and the number of errors, if any exist.

A simple difference operation creates an output file listing the difference in mapping performance between the baseline synonym corpus and any of the variant synonym testing scenarios, based on the set of input terms.

VERIFYING RESULTS BY TESTING AGAINST GOLD STANDARD

The initial findings came about by using a list of 1000 randomly selected terms. While this was sufficient for a preliminary test of the model, the test set was too small to produce significant, representative results. In order to verify the results, SMT was run again with the individual lexical variant synonym testing scenarios, just as in the initial test. However, for the second test the UMLS-CORE Subset, consisting of over 15,000 terms, was used as an input term set. The UMLS-CORE (Clinical Observations Recording and Encoding) Subset was developed by Dr. Kin Wah Fung as a result of his research into defining a UMLS subset useful for the documentation and encoding of clinical information. The Subset is based on datasets submitted by eight healthcare institutions: Beth Israel Deaconess Medical Center, Intermountain Healthcare, Kaiser Permanente, Mayo Clinic, Nebraska University Medical Center, Regenstrief Institute, Hong Kong Hospital Authority, and the Veterans Administration1. These terms are not found within the UMLS-Metathesaurus; however, they are mapped to UMLS concepts through previous lexical matching supplemented by manual expert review.

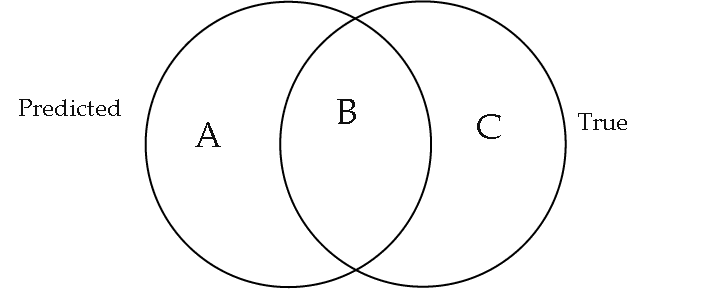
RETESTING WITH NEW SCENARIOS

We ran into a problem in the results of our initial test. We noticed that the variant testing scenarios we developed were inaccurate, because the baseline upon which each was built was not pure. It was in fact corrupted because the baseline was built already including inflectional and spelling variants. This was due to the Normalization process, in which it takes all variations and collapses them. An impure baseline negated the validity of the individual lexical variant testing scenarios we created, so they inevitably had to be discarded. This means that now our experiment would only run variants on subterms of input strings, to match inputs to CUIs through direct mapping and subterm substitution, bypassing the need for the previously created testing scenarios.

CALCULATING PRECISION AND RECALL

We decided to determine the weight of each individual lexical variant by comparing their relative precision and recall, as well as their corresponding F1-measures. This would assist us in assigning an accurate distance score. Precision and recall are the basic measures that are used in evaluating a search strategy, answering whether all relevant materials have been found or if some correct materials have been left out. These measures assume that there is a set of records in any given database which is retrieved in regards to a search topic. These records are either predicted, true, or both predicted and true.

In the following diagram, the left circle (A+B) represents our system’s predictions. The right circle (B+C) represents what is true.



Section A represents the records our system predicted that were wrong, called false positives. Section B represents the records our system predicted correctly, called true positives. Section C represents the true records our system failed to predict, called false negatives.

Precision is the ratio of correctly predicted records to the total number of predicted records. It asks what percentage of our predictions was right. It is expressed as the formula:

Recall is the ratio of correctly predicted records to the total number of true records. It asks what percentage of the true records we got right. It is expressed as the formula:

The F1-measure is used as a combination of precision and recall, and it is known as a balanced F-score because precision and recall are evenly weighted. It is calculated:

The F1-measure will be used to objectively determine both the performance (weight) of each variant type as well as their individual contributions to the coverage rate of finding mapped concepts.

***RESULTS***

The initial FVTest of UMLS-CORE terms yielded interesting results. Out of the 1000 total input terms, 640 of the terms mapped to CUIs just through the application of Norm.

|  | Test ID | No Sub. | 1 Sub. | 2 Sub. | No CUI Found | Total CUIs Found |
| --- | --- | --- | --- | --- | --- | --- |
| Spelling | **s** | 640 | 64 | 10 | 286 | 714 |
| Inflectional | **i** | 640 | 61 | 10 | 289 | 711 |
| Acronym | **A** | 640 | 89 | 21 | 250 | 750 |
| Expansion | **a** | 640 | 91 | 23 | 246 | 754 |
| Derivational | **d** | 640 | 64 | 12 | 284 | 716 |
| All Variants | **Ge** | 640 | 97 | 28 | 235 | 765 |

This initial set of terms served well as a preliminary testing set for our experimental model, but we knew that we needed to use a larger set of input terms to really find significant results. Therefore, we ran the test again, starting with the entire UMLS-CORE Subset as a list of input terms. The UMLS-CORE Subset included 15,447 terms. We removed multiple instances of terms, as well as terms that mapped to more than one CUI. This left us with a testing set of n=13,077.

|  | No Sub. | 1 Sub. | 2 Sub. | Precision | Recall | F1-Measure |
| --- | --- | --- | --- | --- | --- | --- |
| SMT syn | 78.23% (10,230) | 4.74% (620) | 0.31% (40) | 63.48% | 77.52% | 0.6980 |
| SMT syn + s | 78.23% (10,230) | 5.00% (654) | 0.34% (45) | 63.44% | 77.76% | 0.6987 |
| SMT syn + i | 78.23% (10,230) | 4.76% (623) | 0.31% (40) | 63.47% | 77.53% | 0.6980 |
| SMT syn + y | 78.23% (10,230) | 4.74% (620) | 0.31% (40) | 63.48% | 77.52% | 0.6980 |
| SMT syn + A | 78.23% (10,230) | 6.35% (830) | 0.63% (82) | 63.52% | 79.64% | 0.7067 |
| SMT syn + a | 78.23% (10,230) | 6.42% (840) | 0.68% (89) | 63.50% | 79.71% | 0.7069 |
| SMT syn + d | 78.23% (10,230) | 5.21% (681) | 0.42% (55) | 63.37% | 77.99% | 0.6992 |
| SMT syn + Ge | 78.23% (10,230) | 7.33% (959) | 0.89% (117) | 62.63% | 80.65% | 0.7050 |

After running this test, we realized that our variant testing scenarios were incorrect, because the baseline was impure. We changed our experimental model to run individual variants on the subterms of input strings, which would give us more accurate results.

|  | No Sub. | 1 Sub. | 2 Sub. | Precision | Recall | F1-Measure |
| --- | --- | --- | --- | --- | --- | --- |
| baseline | 78.23% (10,230) | 0.00% (0) | 0.00% (0) | 64.49% | 72.02% | 0.6805 |
| spelling | 78.23% (10,230) | 0.30% (39) | 0.01% (1) | 64.44% | 72.24% | 0.6812 |
| inflectional | 78.23% (10,230) | 0.02% (3) | 0.00% (0) | 64.48% | 72.03% | 0.6805 |
| synonyms | 78.23% (10,230) | 0.18% (23) | 0.00% (0) | 64.43% | 72.10% | 0.6805 |
| acronyms | 78.23% (10,230) | 3.04% (398) | 0.13% (17) | 64.18% | 75.41% | 0.6934 |
| expansion | 78.23% (10,230) | 3.12% (408) | 0.17% (22) | 64.17% | 75.48% | 0.6936 |
| derivational | 78.23% (10,230) | 0.83% (109) | 0.04% (5) | 64.40% | 72.72% | 0.6831 |
| all fruitful var | 78.23% (10,230) | 4.60% (602) | 0.39% (51) | 63.20% | 76.75% | 0.6932 |

Out of the 13,077 input terms run through SMT and calculated for precision and recall, 10,230 of the terms mapped through normalization alone. It was determined that the numbers should be calculated again, excluding the terms mapped through normalization, in order to gain a better perspective on the true individual variant impact on precision and recall. Now, rather than n=13,077, n=2,847.

|  | 1 Sub. | 2 Sub. | Precision | Recall | F1-Measure |
| --- | --- | --- | --- | --- | --- |
| baseline | 0.00% (0) | 0.00% (0) | N/A | 0.00% | N/A |
| spelling | 1.37% (39) | 0.04% (1) | 47.92% | 0.81% | 0.0159 |
| inflectional | 0.11% (3) | 0.00% (0) | 25.00% | 0.04% | 0.0007 |
| synonyms | 0.81% (23) | 0.00% (0) | 27.59% | 0.28% | 0.0056 |
| acronyms | 13.98% (398) | 0.60% (17) | 55.91% | 10.96% | 0.1833 |
| expansion | 14.33% (408) | 0.77% (22) | 56.67% | 11.35% | 0.1891 |
| derivational | 3.83% (109) | 0.18% (5) | 52.63% | 2.46% | 0.0470 |
| all fruitful var | 21.15% (602) | 1.79% (51) | 45.58% | 15.38% | 0.2300 |

***DISCUSSION***

From our results of the test running individual variants on the subterms of input strings,

|  | No Sub. | 1 Sub. | 2 Sub. | Precision | Recall | F1-Measure |
| --- | --- | --- | --- | --- | --- | --- |
| baseline | 78.23% (10,230) | 0.00% (0) | 0.00% (0) | 64.49% | 72.02% | 0.6805 |
| spelling | 78.23% (10,230) | 0.30% (39) | 0.01% (1) | 64.44% | 72.24% | 0.6812 |
| inflectional | 78.23% (10,230) | 0.02% (3) | 0.00% (0) | 64.48% | 72.03% | 0.6805 |
| synonyms | 78.23% (10,230) | 0.18% (23) | 0.00% (0) | 64.43% | 72.10% | 0.6805 |
| acronyms | 78.23% (10,230) | 3.04% (398) | 0.13% (17) | 64.18% | 75.41% | 0.6934 |
| expansion | 78.23% (10,230) | 3.12% (408) | 0.17% (22) | 64.17% | 75.48% | 0.6936 |
| derivational | 78.23% (10,230) | 0.83% (109) | 0.04% (5) | 64.40% | 72.72% | 0.6831 |
| all fruitful var | 78.23% (10,230) | 4.60% (602) | 0.39% (51) | 63.20% | 76.75% | 0.6932 |

We can then determine the precision, recall, and F-1 measure of each individual variant type.

|  | Precision | Recall | F1-Measure |
| --- | --- | --- | --- |
| spelling | -0.05% | 0.22% | 0.0007 |
| inflectional | -0.01% | 0.01% | 0 |
| synonyms | -0.06% | 0.08% | 0 |
| acronyms | -0.31% | 3.39% | 0.0129 |
| expansion | -0.32% | 3.46% | 0.0131 |
| derivational | -0.09% | 0.70% | 0.0026 |
| all fruitful var | -1.29% | 4.73% | 0.0127 |

Reordering the variant types by F1-measure gives a more accurate idea of the relative strengths and weaknesses of each variant type.

|  | Precision | Recall | F1-Measure |
| --- | --- | --- | --- |
| inflectional | -0.01% | 0.01% | 0 |
| synonyms | -0.06% | 0.08% | 0 |
| spelling | -0.05% | 0.22% | 0.0007 |
| derivational | -0.09% | 0.70% | 0.0026 |
| all fruitful var | -1.29% | 4.73% | 0.0127 |
| acronyms | -0.31% | 3.39% | 0.0129 |
| expansion | -0.32% | 3.46% | 0.0131 |

Generally, these results show the relationship between precision and recall in relation to using the variant types. If we choose to broaden our chances of getting results by adding variant types, recall increases. However, the more right answers we get, the greater possibility of getting wrong answers as well, which is a cost to precision. This chart shows by how much each variant type increases recall and decreases precision, relative to the baseline. For example, the inflectional variant flow increased recall by 0.01% and decreased precision by only 0.01%. This means that it is better than the expansion flow which increased recall by 3.46% but decreased precision by 0.32%.

Inflectional and spelling variants have a very low effect. This is because they are already accounted for in the normalization process, as they are embedded in LvgNorm (which is the baseline). The synonyms variants also have a low effect, but that was to be expected. The synonym list used is very small, and the variant was never intended to be used alone.

RE-EXAMINING THE GOLD STANDARD

After examining the precision and recall results from our test against the gold standard, we noticed that precision was much lower than anticipated. There was a problem somewhere that may have negatively affected the scores for precision and recall. As such, we decided to conduct an error analysis to take a closer look at our mappings and the gold standard mappings. The precision was so low that we hypothesized that some of our findings were incorrectly marked as false positives. They were instead true positives – we were counting ourselves as wrong when we were actually right. However, this would mean that they would disagree with the gold standard, and therefore the gold standard would be wrong.

In order to find an answer to this issue, we focused initially on the results for mapping with derivational variants. We extracted all the results from derivational mapping that had been marked as a false positive, meaning that while they did map to CUIs, they mapped to CUIs other than the ones found by the gold standard. The script written for this extraction created an output file with the format:

| Field 1 | Field 2 | Field 3 | Field 4 | Field 5 | Field 6 |
| --- | --- | --- | --- | --- | --- |
| Input | Our CUI | Our Term | FLAG\_FP | Gold CUI | Gold Term |

This created a file of 67 terms determined by the system to be false positives.

| Input Term | Our CUI | Our Term | Gold CUI | Gold Term |
| --- | --- | --- | --- | --- |
| Ca Urethra | C0700101 | Urethral Carcinoma | C0153620 | Malignant neoplasm of urethra |
| Conjunctiva Red | C2673789 | Redness of the conjunctivae | C0235267 | Redness of eye |
| Coronary arterial disease | C0010068 | Coronary heart disease | C0010054 | Coronary Arteriosclerosis |

From this we attempted to determine which of our CUIs were correct and which of the gold standard CUIs were correct. After discussion of the differences with one of Dr. Kin Wah Fung’s team members, Julia Xu, we determined that our mapping, rather than the gold standard mapping, should be the correct mapping in 38 out of 67 terms. This change results in a significant shift in precision.

Initially, it seemed that derivational variants had a precision of 52.63% and a recall of 2.46%, with an F1-measure of 0.0470. This was calculated by using the formulas:

However, after the re-examination of the gold standard, and the resultant changing in mappings, the new precision, recall, and F-1 measure of derivational variants are:

***RECOMMENDATIONS***

Based on these findings, the Lexical Systems group might be able to optimize the flow for concept mapping, once all the variant types have been tested again by the gold standard and adjusted for mapping accuracy. After the gold standard has been adjusted, the flow can be re-examined in order to find the optimal combination that delivers the highest performance in both precision and recall for concept mapping.

It is not necessarily true that a test using a combination of every lexical variant scenario will yield the highest results. The variants are transitive, which may actually worsen mapping performance by linking two terms that are not actually conceptually related. No matter what, users will then be able to make their own choices for finding a variant flow most conducive to their searching needs.

***REFERENCES***

1. Fung KW, McDonald C, and Srinivasan S. The UMLS-CORE project: a study of the problem list terminologies used in large healthcare institutions. J Am Med Inform Assoc. 2010; 17:675-680.