2nd Workshop on Building and Evaluating Resources for Biomedical Text Mining

Tuesday, 18th March 2010

Valletta, Malta

Organisers:

Sophia Ananiadou
Kevin Cohen
Dina Demner-Fushman

Workshop Programme

9:15 – 9:30 Welcome

9:30 – 10:30 **Invited Talk** (chair: Sophia Ananiadou)
Pierre Zweigenbaum, Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur (LIMSI-CNRS), France

10:30 – 11:00 Coffee break

11:00 – 12:30 **Session 1** (chair: Kevin Cohen)

11:00 Spelling Correction in Clinical Notes with Emphasis on First Suggestion Accuracy
Jon Patrick, Mojtaba Sabbagh, Suvir Jain and Haifeng Zheng

11:25 Automatically Building a Repository to Support Evidence Based Practice
Dina Demner-Fushman, Joanna Karpinski and George Thoma

11:50 An Empirical Evaluation of Resources for the Identification of Diseases and Adverse Effects in Biomedical Literature
Harsha Gurulingappa, Roman Klinger, Martin Hofmann-Apitius and Juliane Fluck

12:15 A Task-Oriented Extension of Chinese MeSH Concepts Hierarchy
Xinkai Wang and Sophia Ananiadou

12:40 – 14:10 Lunch break

14:10 – 15:10 **Invited talk** (chair: Sophia Ananiadou)
Simonetta Montemagni, Istituto di Linguistica Computazionale (ILC-CNR), Italy

15:10 – 16:55 **Session 2** (chair: Dina Demner-Fushman)

15:10 Structuring of Status Descriptions in Hospital Patient Records
Svetla Boytcheva, Ivelina Nikolova, Elena Paskaleva, Galia Angelova, Dimitar Tcharaktchiev and Nadya Dimitrova

15:35 Annotation of All Coreference in Biomedical Text: Guideline Selection and Adaptation
K. Bretonnel Cohen, Arrick Lanfranchi, William Corvey, William A. Baumgartner Jr., Christophe Roeder, Philip V. Ogren, Martha Palmer and Lawrence Hunter

16:00 – 16:30 Coffee break

16:30 Contribution of Syntactic Structures to Single Word Term Extraction
Xing Zhang and Alex Chengyu Fang

16:55 – 17:10 **Concluding remarks** (chair: Sophia Ananiadou)

Workshop Organisers

- Sophia Ananiadou, NaCTeM, University of Manchester, UK
- Kevin Cohen, Center for Computational Pharmacology and The MITRE Corporation, USA
- Dina Demner-Fushman, National Library of Medicine, USA

Workshop Programme Committee

- Olivier Bodenreider, National Library of Medicine, USA
- Wendy Chapman, University of Pittsburgh, USA
- Aaron Cohen, Oregon Health and Science University
- Liu Hong Fang, Georgetown University Medical Center, USA
- Martin Krallinger, National Biotechnology Center, Spain
- John McNaught, NaCTeM, University of Manchester, UK
- John Pestian, Computational Medicine Center, University of Cincinnati, USA
- Andrey Rzhetsky, University of Chicago, USA
- Jian Su, Institute for Infocomm Research, Singapore
- Junichi Tsujii, University of Tokyo, Japan and NaCTeM, University of Manchester, UK
- Yoshimasa Tsuruoka, JAIST, Japan
- Karin Verspoor, Center for Computational Pharmacology, University of Colorado, USA
- Xinglong Wang, University of Manchester, UK
- Bonnie Webber, University of Edinburgh, UK
- John Wilbur, NCBI, NLM, NIH, USA
- Pierre Zweigenbaum, LIMSI-CNRS, France

Table of Contents

Author Index

FOREWORD

This volume contains the papers accepted at the 2nd workshop on Building and Evaluating Resources for Biomedical Text Mining held at LREC 2010, Malta. Biomedical text mining over the last decade has become one of the driving application areas for the NLP community, resulting in a series of very successful yearly specialist workshops at ACL since 2002, BioNLP, as well as the launch of the BioMed special interest group in 2008.

In the past, most of the work has focused on solving specific problems, often using task-tailored and private data sets. This data was rarely reused, in particular outside the efforts of the providers. This has changed during the last years, as many research groups have made available resources that have been built either purposely or as by-products of research or evaluation efforts. A number of projects, initiatives and organisations have been dedicated to building and providing biomedical text mining resources (e.g., the GENIA suite of corpora, PennBioIE, TREC Genomics track, BioCreative, Yapex, LLL05, BOOTStrep, JNLPBA, KDD data, Medstract, BioText, etc.). There is an increasing need to provide community-wide discussions on the design, availability and interoperability of resources for bio-text mining, following on from specific applications such as gene name identification and protein-protein interactions in BioCreative I/II, the BioNLP'09 shared task on event extraction etc., and recognition of clinically relevant entities and relations in the i2b2 challenges to the use of common resources and tools in real life applications.

The papers in this volume reflect the current transitional state of the biomedical text mining field and range from named entity recognition in MEDLINE abstracts (Gurulingappa et al., Zhang and Fang) to reuse of annotation tools and guidelines to fully annotate 97 journal articles (Cohen et al.). The corpus of 97 fully annotated articles will be made publicly available upon completion of syntactic, semantic, and discourse structure annotation with the emphasis on coreference annotation.

The biomedical text mining field is expanding in two directions: enrichment and use of cross-lingual resources and work in resource-poor languages on the one hand, and significant inroads in processing clinical narrative on the other hand. The enrichment and evaluation of cross-lingual resources are exemplifies in the work on extension of the Chinese Medical Subject Headings vocabulary (Wang and Ananiadou). Spanning a resource-poor language and clinical domain, Boytcheva et al. focus on the mining of Bulgarian clinical notes for descriptions of patients' status and complications. Patrick et al. focus on spelling errors in clinical notes and present a detailed description of a spelling suggestion generator. Clinical notes are linked to MEDLINE citations in the collection under construction at NIH, as described by Demner-Fushman et al.

We wish to thank the authors for submitting papers for consideration, and the members of the program committee for their time and effort during the review process. We would also like to thank our invited speakers, Simonetta Montemagni and Pierre Zweigenbaum for their contribution.

Sophia Ananiadou, Kevin Cohen and Dina Demner-Fushman

Spelling correction in Clinical Notes with Emphasis on First Suggestion Accuracy

Jon Patrick, Mojtaba Sabbagh, Suvir Jain, Haifeng Zheng

Health Information Technology Research Laboratory
School of IT, University of Sydney
jonpat@it.usyd.edu.au, ssab8677@uni.sydney.edu.au, suvir@cs.usyd.edu.au, hzhe6571@uni.sydney.edu.au

Abstract

Spelling correction in a large collection of clinical notes is a relatively lesser explored research area and serious problem as typically 30% of tokens are unknown. In this paper we explore techniques to devise a spelling suggestion generator for clinical data with an emphasis on the accuracy of the first suggestion. We use a combination of a rule-based suggestion generation system and a context-sensitive ranking algorithm to obtain the best results. The ranking is achieved using a language model created by bootstrapping from the corpus. The described method is embedded in a practical work-flow that we use for Knowledge Discovery and Reuse in clinical records.

1. Introduction

Our work specializes in processing corpora of medical texts ((Wang and Patrick, 2008), (Ryan et al., 2008), (Zhang and Patrick, 2007) and (Asgari and Patrick, 2008)). These corpora usually contain many years of patient records from hospital clinical departments. Discovered knowledge about the texts needs to be reused immediately to automate certain language extraction tasks that support the workflow of staff in the hospital.

However, to be able to extract the required knowledge, first various lexical verification steps need to be undertaken, such as expansion of abbreviations and acronyms, recognition of named entities, drug dosage details etc. An initial analysis of a corpus is to identify unknown words and non-words (tokens containing any non-alphabetic characters) which typically constitute 30% of a corpus. A large majority of unknown words are misspellings but a significant number (10%-15%) are meaningful and need to be correctly interpreted.

The nature of the requisite processing techniques is more complex than simple spelling correction because the medical texts also contain various abbreviations, acronyms, names of people, places and even a number of neologisms. In such a text, every word that does not appear in a dictionary need not be an invalid word. Therefore, during the process of knowledge discovery, extreme care has to be taken that valid tokens are not misclassified as spelling errors. For abbreviations, acronyms and named entities, we apply separate workflows for their detection and resolution.

Our study is based on those words in the corpus that have been designated by medical experts as genuine spelling mistakes. These words include both medical and non-medical words. The diverse nature of these words renders traditional spell checkers ineffective and requires a novel spell checking system for their resolution.

Currently the nature of spelling correction of this system is ‘offline’, i.e., these suggestions are not being generated during entry of data in the hospitals. Instead, these methods are applied to the corpus as a whole. We look to implement this system as an ‘online’ process in the near future.

Our knowledge base to begin the knowledge discovery includes the following resources:

1. Systematised Nomenclature of Medicine-Clinical Terms, i.e., SNOMED-CT: 99860 words
2. The Moby Lexicon: 354992 words
3. The Unified Medical Language System, i.e., UMLS: 427578 words
4. Lists of abbreviations and acronyms discovered in previously processed clinical corpora: 1483 words
5. Gazetteers built of named entities discovered in previously processed clinical corpora: 19264 words
6. List of the gold standard corrections of misspellings in previously processed texts: 78889 words

Combined, this knowledge base provides us with 758,481 unique words at our disposal. The fact that the test corpus of 20 million tokens which contained 57,523 words and initially had 24,735 unknown tokens, i.e., tokens which did not exist in our knowledge base, only shows the infinite diversity in clinical notes and the magnitude and complexity of the problem.

2. Related Work

Various works have explored the problem of spelling correction in the past but literature on spelling correction in clinical notes, which is a distinct problem, is quite sparse. A previous work (Crowell et al., 2004) used word frequency based sorting to improve the ranking of suggestions generated by programs like GSpell and Aspell. Their method does not actually detect any misspellings nor generate suggestions. Work (Ruch, 2002) has also been done to study contextual spelling correction to improve the effectiveness of an IR system.

Another group (Tolentino et al., 2007) created a prototype spell checker using UMLS and Wordnet as their sources of knowledge. See (Mykowiecka and Marciniak, 2006) for a program for automatic spelling correction in mammography reports. They utilized edit distances and bigram probabilities and their method is applied to a very specific domain. None of these methods scale up satisfactorily to the size and diversity of our problem.

Due to the lack of previous work on spelling correction in our very large medical corpus (60 million tokens), we were motivated to devise our own method.

The remainder of this paper is organized in the following manner. Section 3 contains details about the suggestion generation, context sensitive ranking algorithm and the heuristics employed by our method. Section 4 gives details about the data sets that we used in our experiments. Section 5 presents the experiments with the results. Sections 6, 7 and 8 include the interpretations, possible improvements in future work and conclusion.

3. Method

Our method for spelling correction uses a combination of rule-based suggestion generation and a context-sensitive ranking algorithm.

3.1. Suggestion Generation

The suggestion generator that we used follows a series of rules to choose the suggestions that would be ranked in subsequent processing. The following are various rules that are applied to the unknown word to choose valid suggestions. It should be noted that if the algorithm finds that a rule matches atleast one valid suggestion, it does not process the subsequent rules. It simply passes on the computed set of suggestions to the ranking algorithm. For example, if the 'Edit Distace One' algorithm finds atleast one suggestion that matches a word in our knowledge base, the subsequent rules are not processed. Otherwise, the next rule would have been processed and so on.

Table 1 provides the average individual contribution of each of the following rules if they are used in isolation from the rest.

1. *Edit Distance 1:* Edit distance (also known as Damerau-Levenshtein distance (Damerau, 1964), (Levenshtein, 1966)) is a "distance" (string metric) between two strings, i.e., finite sequence of symbols, given by counting the minimum number of operations needed to transform one string into the other, where an operation is defined as an insertion, deletion, or substitution of a single character, or a transposition of two characters. The suggestions are generated by simulating **deletion**, **transposition**, **alteration** and **insertion** in one character position of the unknown word and, all those strings that are verified as legal from our knowledge base are put into the bag of suggestions.

   Misspellings like *blokage*: (blockage), *patholigical*: (pathological) and *maliase*: (malaise) are corrected.

   This method is based on Damerau's(Damerau, 1964) findings, according to which a majority of all spelling mistakes are represented by these actions. This list covers common typographic errors such as those caused due to pressing adjacent keys. It also covers those misspellings where the typist forgot to press a key or exchanged adjacent letters.

   The suggestions are sorted according to their frequency in the corpus or the knowledge base, the differences of which are evaluated later.

   It must be noted that these suggestions are not passed to the subsequent context-sensitive ranking algorithm. This is a heuristic that we have employed. The sorting according to the word frequency greatly increased the accuracy of the suggestions generated by this process(see Table 2). In the results we have shown the accuracy achieved by sorting both with corpus frequencies (CFSORT) and knowledge base frequencies (KBSORT) and their combination (CFKBSORT).

2. *Missing white space between 2 words:* In this step, most of the tokens which consist of two valid words concatenated due to missing white-space, are identified. This is achieved by inserting a space between every adjacent letter pair and then checking if the two substrings exist in the knowledge base. The set of suggestions generated by this function is passed on to the ranking algorithm.

   Misspellings like *patientrefused*: (patient refused) are corrected.

3. *Edit Distance 2:* In this step we compute all the strings within edit distance 2 of the misspelling. This is achieved by finding all strings within edit distance 1 of the unknown word and repeating the process for all generated words. This method too has a good coverage and covers misspellings which might have a single error at 2 distinct positions in the word. It also covers those words which have multiple errors at a single site. Of these, all those strings that exist in our database are passed to the ranking algorithm.

   Misspellings like *statarating*: (saturating) are corrected.

4. *Phonetic*: This process generates all the valid strings which are phonetically similar to the misspelling. The set of strings generated is verified against the knowledge base and all those strings which are valid are passed on to the ranking algorithm.

   The algorithm used here is the "Double Metaphone" algorithm.The Double Metaphone search algorithm is a phonetic algorithm and is the second generation of the Metaphone algorithm (Phillips, 2000). It is called "Double" because it can return both a primary and a secondary code for a string; this accounts for some ambiguous cases as well as for multiple variants of surnames with common ancestry. For example, encoding the name "Smith" yields a primary code of SM0 and a secondary code of XMT, while the name "Schmidt" yields a primary code of XMT and a secondary code of SMT–both have XMT in common.

   Misspellings like *simptamotolgi*: (symptomatology) are corrected.

5. *Phonetic edit distance 1*: In our medical lexicon, many words are long and complex. Typists often tend to miss one syllable or do a contraction. With this function, such words will be coded by phonetics and then we generate all words with edit distance one for this phonetic code. The suggestion list is populated by all words which have the same phonetic code as we generated here. It will largely extend our suggestion list.

   Misspellings like *amns*: (amounts) and *ascultion*: (auscultation) are corrected.

6. *Edit distance 1 and concatenated word:* The combination of rules 1 and 2 is used. There are often words in the text where a valid word is joined with an invalid one, which is usually within 1-edit distance of its correction. Such cases are dealt with at this stage. As before, the set of suggestions is passed on to the ranking algorithm.

   Misspellings like *affectinback*: (affecting back) are corrected.

7. *Multiple concatenated words:* Sometimes, 3 or more words have been concatenated due to missing white space. Such words are filtered here and the suggestions are passed on to the ranking algorithm.

   Misspelling like *explorationforbleedingand*: (exploration for bleeding and) is corrected.

3.2. Context-Sensitive Ranking Algorithm

Previous studies (see (Elmi and Evens, 1998), (Golding and Schabes, 1996) and (Kukich, 1992)) have suggested that spelling correction is much more effective when the method takes into account the context in which the word occurs. We take into account the context considering the misspelling, one word before it and one word after it. Unlike some previous studies, we did not experiment with context defined as adjacent letters. Since our corpus is large, we could get reliable estimates if we considered the surrounding words as the context.

For this purpose, we used the CMU-Cambridge Statistical Language Modelling Toolkit (Clarkson and Rosenfeld, 1997). It is a suite of UNIX software tools to facilitate the construction and testing of statistical language models. It includes tools to process general textual data to word frequency list and vocabularies, n-gram counts and backoff models.

The toolkit was used to build a vocabulary using all the words in the corpus which occurred more than once. Using these words, a language model was constructed. This tool was used to calculate all unigram, bigram and trigram probabilities along with their corresponding backoff values. The probability values were smoothed to increase the reliability. One of the salient features of our method is that the corpus was bootstrapped to correct its own mistakes. This was done by ranking the suggestions using the language model and the frequency of the misspellings in the corpus itself. This method helped promote the more relevant suggestions in the suggestion set to the top positions. This method greatly increased the accuracy of the first suggestion.

3.3. Complex Frequency Ranking Algorithms

The basic ranking algorithm for candidate corrections is a frequency-based technique. (Crowell et al., 2004) posed their frequency-based model to improve the spelling suggestion rank in medical queries. Our trained dictionaries have two kinds of frequency values. The first is a knowledge based word frequency which computes the number of times the same word is repeated when we are training the dictionary set. The second is a corpus word frequency which uses the word frequencies in the corpus. As shown in previous works, frequency-based methods can improve the first suggestion accuracy of spelling correction, however, the dictionary configuration is a real problem, and is fairly hard to optimize. At this stage, we train our frequency values based on the following three methods: the first is based on corpus word frequency; the second is based on knowledge base word frequency; the third is based on the combined frequency in the knowledge base and corpus. Considering there are many misspellings from missing white space(s), it is not fair to compute the frequency of the misspelling as a whole or just compute the frequency of the first separated word. This paper introduces two new frequency ranking algorithms:

1. First two words ranking: This algorithm computes the frequency of the first two words when a suggestion contains more than one word.

2. First and last word ranking: This algorithm computes the combined frequency of the first and last word of a suggestion.

3.4. Heuristics Employed

A set of the interesting heuristics that was evaluated to improve the first suggestion's accuracy is also provided.

1. Suggestions generated by Edit Distance 1 were not ranked by the context-sensitive ranking algorithm. They were instead sorted by their word frequencies in the knowledge base. The rationale behind this strategy was that suggestions that occur more frequently in the knowledge base are more likely to occur in the corpus. The accuracy was found to drop sharply when we tried to rank the suggestions by this method.

2. The misspellings of adverbs ending with 'ly' were not being corrected in a satisfactory manner. For this purpose, we devised a simple heuristic. If a misspelling ended with 'ly', then the first suggestion in the suggestion-set ending with 'ly' was arbitrarily promoted to the first position in the suggestion set. The application of this heuristic enabled us to correct almost all such misspelt adverbs.

   For example, if the misspelling was *accidentaly*, the first suggestion might be *accidental* whereas the correct suggestion *accidentally* might be third or fourth in the list.

3. The orthographic case for the misspelling was not considered. All words were converted to lower case. It was noted that although words beginning with upper

case do give an idea if the word is a named entity, many such named entities did not begin with upper case. Therefore, to simplify processing the case was ignored for all words.

4. The Edit Distance methods used Inverse Edit Distance instead of direct edit distance due to the large size of our dictionaries. The use of Direct Edit Distance would have led to unnecessarily long processing times. Instead Inverse Edit Distance allowed us to calculate all the possible strings and then check the dictionaries for them.

5. In some cases, one or more suggestions have the same probability when we try to rank them by language model. This was due to sparseness in corpus. In such cases, we summed frequency of suggestions in the corpus and frequency of the suggestion in knowledge base and included them in computation of the language model. This helped us get more reliable ranking for suggestion.

4. Data Set

1. The **training data** consisted of misspellings in a corpus of clinical records of the Emergency Department at the Concord Hospital, Sydney. The total numbers of unique words in the corpus was 57,523. This corpus contained 7442 misspellings for which we already had gold standard corrections verified by expert analysis. We found that this set served our purpose quite well because it contained a good mixture of medical as well as non-medical words. It also contained many rare words and some very convoluted misspellings which could not be corrected by usual spelling correction techniques.

   Interestingly, due to this reason the methods show a higher accuracy on the test data than the training data.

2. The **test data** contained 2 separate data sets as described below.

   - The first data set consisted of 12,438 words in the Concord Emergency Department Corpus which has been described above. These words were hitherto unknown words which did not exist in our knowledge base. These words were manually corrected and along with their gold standard they serve as one of the test data sets of our method.
   - The second data set consists of 65,429 misspellings from the Intensive Care Unit patient records at the Royal Prince Alfred Hospital, Sydney, the RPAH-ICU corpus. This corpus contained a total of 164,302 unique words.

5. Experiments and Results

5.1. Experiments

Initial experiments started with a simple spell checker to establish a baseline. At this stage the spellchecker only checked for a known word, word with 1-edit distance or 2-edit distance, as described in section 3.1. The first suggestion was matched with the gold standard and the accuracy was calculated. This method gave an accuracy of 60.52% on our test data and 63.78% on the training data. We then incrementally added the different heuristics to further improve the accuracy. The percentage increase with the addition of each step is presented in Table 2. Finally, the accuracy reached 62.31% on the training data and 75.11% on the test data.

At this point sorting of the suggestions based on their frequency in the knowledge base was incorporated. Words which are more frequent in the knowledge base should logically be ranked higher in the suggestion list.

This option was only available because of the large knowledge base. Therefore, the suggestions based on their word frequencies in the corpus was also used. In both cases, the accuracy improved by a significant margin. But the sorting based on corpus frequencies provided slightly higher accuracy than that of knowledge base frequencies.

Looking at an example to illustrate why **sorting based on knowledge base frequencies** is useful. The word *haemodynamically* has been misspelt in 381 different ways in previously processed texts. In the ICU corpus alone, it occurs 30,162 times in its correct form. This makes it highly probable that if 'haemodynamically' is present anywhere in the suggestion list, it deserves to be among the top suggestions.

Similarly, **sorting based on corpus word frequencies** is also helpful and is a method which can be be used for any large corpus. Also, the texts belong to a single department and a great deal of medical terms that are used in one department would not have been used at all in another. Therefore this method helps improve the ranking especially for words that are typical of the specific corpus. Phone1 algorithm works better in the Concord-ED collection because it contains more complex errors than 'Train' and the RPAH-ICU data set.

Sorting based on combined frequencies in corpus and knowledge base was slightly more effective than these methods individually(See Table 2).

After trying a number of different methods to generate suggestions, the next improvement was generated from experimenting with ranking suggestions. When the number of suggestions is large ranking is difficult because there are a lot of high-frequency small words which are within 'Edit Distance 1' or 'Edit Distance 2' of many misspellings. So, some dictionaries were progressively removed from our knowledge base. As shown in Table 3, removing each dictionary improves the result slightly. While having rich resources may increase the coverage (recall) however it injects noise into the suggestions and leads to a decrease in accuracy. We need to find a tradeoff between the coverage and the number of suggestions for each misspelling. By reducing the number of dictionaries, number of suggestions is reduced. The last row of this table shows the effect of a language model on the overall accuracy. For this, a language model was created from all the lexically verified words that occurred more than once in the specific medical text that is being processed. The language model stores all the word unigrams, bigrams and trigrams with their corresponding backoff values and probability values. The language model probabilities were added to the corpus and knowledge base frequencies as explained in Heuristic 5. The reason RPAH-

Rule	Percentage	Description
Edit distance 1	69.80	Edit distance one
Two words	6.73	Missing white space
Edit distance 2	85.52	Contains Edit distance one and two
Phonetic	42.52	Pure Phonetic
Two word edit distance 1	7.09	Contains Two words and Two edit distance one
Multi word	6.86	Contains two or more concatenated words
Phonetic edit distance 1	84.78	Pure Phonetic and Phonetic with Edit distance one

Table 1: Mean Individual Contribution of each Spelling Correction Algorithm to the Concord-ED Corpus

Rule	Training data	Concord-ED	RPAH-ICU
(Baseline)Edit Distance 1 and 2	60.62	63.78	62.28
Adding ‘Two words’	62.31	74.99	66.42
Adding ‘Phonetic’	62.31	75.11	66.57
Adding ‘Two word edits 1’	62.31	75.11	66.58
Adding ‘Multi word’	62.31	75.11	66.61
Adding ‘Sorting based on corpus word frequency’	87.19	93.17	81.24
Adding ‘Sorting based on knowledge base word frequency’	83.41	93.07	79.99
Adding ‘Sorting based on combined frequency in knowledge base and corpus’	87.25	**93.54**	81.45
Replace ‘Phonetic’ with ‘Phonetic edit distance 1’	87.24	93.43	81.69
Adding ‘First two words ranking’	87.25	92.89	81.75
Adding ‘First and last word ranking’	**87.26**	93.05	**81.83**

Table 2: Percentage Accuracy After Addition of each Process Described in Section 3.1

ICU data collection was used here is, it contains a complete set of all errors and so is a good test set for this method. The language model was applied to the suggestions computed after application of each of the sorting strategies. The result drops a little bit because the frequency ranking is very good and adding each ranking method may worsen the accuracy.

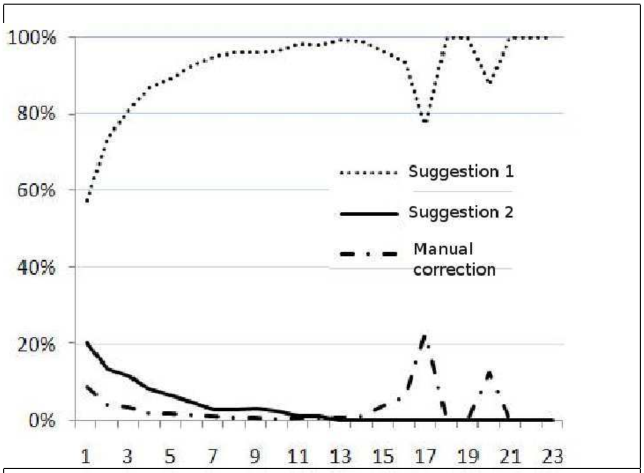


Fig 1. Variation of Accuracy of Suggestions with Word length

5.2. Observations

1. Each of the three data sets that has been used in this experiment is unique in itself. This is the reason that the accuracies are very different for each data set. But analysis of an individual data set’s results shows that the increase with the addition of sorting and/or language model was equally pronounced. The training set has some from very complex misspellings which could not be corrected by a general heuristic. The Concord Test Corpus had more non-medical words and fewer medical terms. The RPAH-ICU corpus had more of a mix of medical terms and non-medical words.

2. As shown in Table 2, both ‘First two words ranking’ and ‘First and last word ranking algorithms’ improved the accuracy, but obviously the latter is better. According to the analysis of over 200 multiple concatenated word cases, the small word coverage of ‘First two words ranking’ is 89 against 81 of ‘First and last word ranking algorithms’. The small words are defined as prepositions, conjunctions, pronouns and some common adjectives, adverbs, verbs and nouns, which have fairly short word length and fairly high frequency value. The short length makes them easier to be considered as suggestions, and the high frequency value can adversely affect the suggestion ranking. Thus, FLW ranking is more suitable as it skips the short words and only considers the first and last word.

3. An analysis of the results showed that the accuracy of our method was found to increase as the word length increased. This is because the bag of suggestions for

Rule	Correct Num (out of 65429)	RPAH-ICU corpus
Remove UMLS	54038	82.59
Remove SNOMED-CT	53769	82.18
Remove Both	54857	83.84
Remove Gazetteers	53614	81.94
Remove All	55197	84.36
Reomve All + Adding Language Model	54985	84.04

Table 3: Percentage Accuracy After Removing Dictionaries and Adding Language Model

longer words contains less number of suggestions and we can easily rank them.

Figure 1 shows the gradual increase of the first suggestion's accuracy with increasing word length. It also shows the decrease in accuracy of second suggestion and manual corrections with increasing word length. Spikes at word length 17 and 20 are anomalies due to low frequency of very long tokens in the corpus.

4. The results show that an exhaustive knowledge base is a boon for a spell checker.

5. With a large corpus, both corpus word frequencies and context based probabilities are quite effective. A combination of these two techniques yields improved results.

6. Possible Improvements

The authors believe that the method described in this paper can be improved by one or more of the following:

1. Both language model based sorting and simple word frequency sorting to improve the ranking of suggestions, have given us comparable results. This does not mean that the use of language model is futile. We believe that there is room for improvement in the language model. We have used only trigram probabilities but context is often beyond trigrams. In the future, we plan to implement a more sophisticated language model to augment the spelling correction system.

2. Some words could have been properly corrected with a good US-UK spelling disambiguation system. It was seen that sometimes the first suggestions was *anesthetize* but the corresponding gold standard was *anaesthetize*. In our case, the obvious work-around to this problem was adding both to the gold standard. In some cases, we looked at the Wordnet synonym list of the first suggestion. For non-medical terms, this method often worked fine. But far better than all such methods, would have been the use of a robust US-UK spelling disambiguation system with good coverage.

3. Sometimes when the misspelling was that of a verb, the spelling suggestions preserved the root but not the morphology. A system which incorporates such functionality would be another possible improvement.

4. A small percentage of words in the gold standard are erroneous and this slightly lowered the accuracy.

5. A limitation of the context sensitive ranking algorithm was that the suggestion generated did not always exist in the corpus itself. We tried countering this error by adding the Knowledge Base word frequencies to the trigram probabilities. A semantic approach to such ambiguous cases would be more effective and, at the same time, elegant.

7. Conclusions

We have presented a method for spelling correction based on combining heuristic-based suggestion generation and ranking algorithms based on word frequencies and trigram probabilities. We have achieved high accuracies on both the test data sets with both the methods. We believe the results are satisfactory and an improvement upon previous work. We look to implement more sophisticated methods in the future to utilize the context in a more effective manner. This spelling suggester has been incorporated in the group's workflow and is used in the manual process of spelling correction to great effect.

8. References

P. Asgari and J. Patrick. [2008]. The new frontier - analysing clinical notes for translation research - back to the future. In Violaine Prince and Mathieu Roche, editors, *Information Retrieval in Biomedicine: Natural Language Processing for Knowledge Integration.*, pp 357–378. IGI Global.

P.R. Clarkson and R. Rosenfeld. [1997]. Statistical language modeling using the cmu-cambridge toolkit. In *Proceedings ESCA Eurospeech*, pp 2701–2710, September.

Jonathan Crowell, Qing Zeng, Long Ngo, and Eve-Marie Lacroix. [2004]. A frequency-based technique to improve the spelling suggestion rank in medical queries. *J Am Med Inform Assoc.*, 11:179–185.

F.J. Damerau. [1964]. A technique for computer detection and correction of spelling errors. *Communications of the ACM*, 3(7):171–176, March.

Mohammed Ali Elmi and Marthe Evens. [1998]. Spelling correction using context. In *Proceedings of the 17th international conference on Computational linguistics*, volume 98, pp 360–364.

A. Golding and Y. Schabes. [1996]. Combining trigram-based and feature-based methods for context-sensitive spelling correction. In *Proc. 34th ACL*, pp 71–78.

K. Kukich. [1992]. Techniques for automatically correcting words in text. *ACM Computing Surveys*, 24(4):377–349.

V.I. Levenshtein. [1966]. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10:707.

Agnieszka Mykowiecka and Magorzata Marciniak. [2006]. Domaindriven automatic spelling correction for mammography reports. *Intelligent Information Processing and Web Mining*, 5:521–530.

Lawrence Phillips. [2000]. The double metaphone search algorithm. Technical report, C/C++ Users Journal.

Patrick Ruch. [2002]. Using contextual spelling correction to improve retrieval effectiveness in degraded text collections. In *Proceedings of the 19th International conference on Computational linguistics.*, volume 1, pp 1–7. Association for Computational Linguistics.

A. Ryan, J. Patrick, and R. Herkes. [2008]. Introduction of enhancement technologies into the intensive care service, royal prince alfred hospital, sydney. *Health Information Management Journal*, 37(1):39–44.

Herman D Tolentino, Michael D Matters, Wikke Walop, Barbara Law, Wesley Tong, Fang Liu, Paul Fontelo, Katrin Kohl, and Daniel C Payne. [2007]. A umls-based spell checker for natural language processing in vaccine safety. *BMC Med Inform Decis Mak.*, 7(3), February.

F. Wang and J. Patrick. [2008]. Mapping clinical notes to medical terminologies at point of care. In D. Demner-Fushman, S. Ananiadou, K. Bretonnel Cohen, J. Pestian, J. Tsujii, and B. Webber, editors, *In Proc of BioNLP2008*, pp 102–103. Association for Computational Linguistics (ACL).

Y. Zhang and J. Patrick. [2007]. Extracting semantics in a clinical scenario. In *Health Knowledge and Data Mining Workshop, Ballarat, Research and Practice in Information Technology*, volume 249, pp 217–224. Health Knowledge and Data Mining Workshop, Ballarat, Research and Practice in Information Technology.

Automatically Building a Repository to Support Evidence Based Practice

Dina Demner-Fushman, Joanna L. Karpinski, George R. Thoma
Lister Hill National Center for Biomedical Communications (LHNCBC)
U.S. National Library of Medicine, National Institutes of Health, Bethesda, MD 20894
{ddemner|karpinskij|gthoma}@mail.nih.gov

Abstract

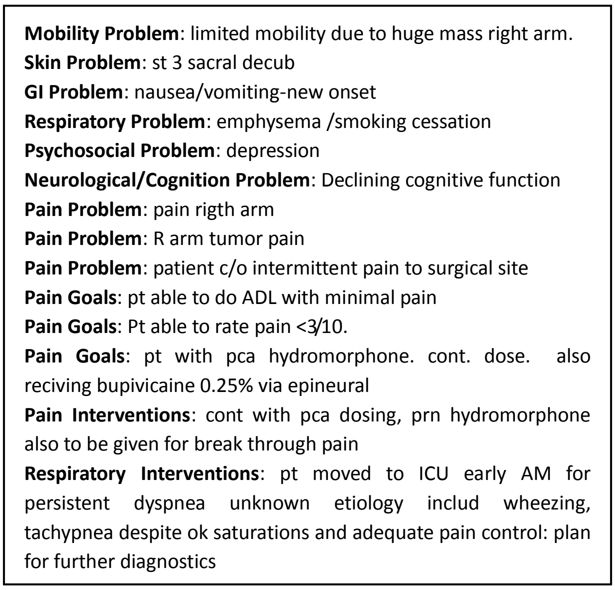
Our long-term goal is to find, store, update, and provide access to key facts needed to support clinical decision making. Presently, the facts are extracted automatically from clinical narrative and biomedical literature sources, primarily MEDLINE, and stored in the Repository for Informed Decision Making. We envision expert community validation of the extracted facts and peer-reviewed direct deposit of key facts in the future. The key facts reported in publications do not change and can be extracted in advance and retrieved as needed. We chose an alternative approach to building the repository: extraction of key facts for specific clinical tasks and clinical scenarios. Combined with providing extracted facts (linked to the original publications) as service at the point of care, this approach allows capturing clinicians' relevance judgments and leads to gradual, weakly-supervised construction of a collection of facts and documents pertaining to specific clinical situations and expert judgments of relevance and quality of the documents. In this paper, we demonstrate our approach to corpus construction using the clinical task of patient care plan development. We provide an overview of the process and focus on the automatic construction of PubMed queries – an essential step in finding documents containing key facts.

1. Introduction

Collections containing biomedical documents, key facts extracted from the documents, and judgments on relevance and value of the documents to specific clinical tasks and questions are essential for many reasons, including clinical decision support and further development of biomedical natural language processing methods. Our research into methods for building collections of key clinical facts relatively fast and at low cost is motivated by the well-known desire of clinicians to have research evidence provided in the form of bottom-line advice (Ely et al., 2005) on the one hand, and significant manual efforts presently needed to find and summarize key facts in the biomedical domain, on the other hand. For example, creation of the 2007 Text REtrieval Conference (TREC) Genomics track collection involved extensive interviewing of biologists to obtain real-life questions of interest to biological domain, as well as recruiting judges with significant domain knowledge, typically in the form of a PhD in a life science (Roberts et al., 2009).

We propose inferring clinical questions using formal representation of a patient's case, rather than actively soliciting information needs. The second step of our repository building process is fairly typical for building collections and consists of literature retrieval. Our retrieval process is complicated by the fact that rather than having a short description of information need provided by a user or a relevant document (which a search engine could use to find similar documents) we are presented with the description of the patient's status and need to find documents relevant both to the patient's status and the clinical task to be performed. We describe our approach to automatic construction of expert queries in Section 3 and the evaluation of the method in Section 4. The rest of the paper is organized as follows: in Section 2 we introduce the clinical task of patient care plan development and the framework for formal representation of a patient's case. Section 5 presents the structure of our repository and the mechanism for obtaining relevance judgments as part of the healthcare workflow. We conclude with a discussion of the preliminary results of our approach to building the repository.

2. Patient Care Plan Development



Mobility Problem: limited mobility due to huge mass right arm.
Skin Problem: st 3 sacral decub
GI Problem: nausea/vomiting-new onset
Respiratory Problem: emphysema /smoking cessation
Psychosocial Problem: depression
Neurological/Cognition Problem: Declining cognitive function
Pain Problem: pain rigth arm
Pain Problem: R arm tumor pain
Pain Problem: patient c/o intermittent pain to surgical site
Pain Goals: pt able to do ADL with minimal pain
Pain Goals: Pt able to rate pain <3/10.
Pain Goals: pt with pca hydromorphone. cont. dose. also reciving bupivicaine 0.25% via epineural
Pain Interventions: cont with pca dosing, prn hydromorphone also to be given for break through pain
Respiratory Interventions: pt moved to ICU early AM for persistent dyspnea unknown etiology includ wheezing, tachypnea despite ok saturations and adequate pain control: plan for further diagnostics

Figure 1: Semi-structured interdisciplinary team note. The format combines problem types restricted to controlled vocabulary (shown in bold) and free text description of the problems entered by the team members

Care plan development starts with assessment of a patient's status[1]. The assessment results are documented

[1] The assessment part of the note entered into a patient's chart is sometimes preceded by patient's description of the patient's current condition (mostly in narrative form) and registration of objective conditions, such as vital signs, patient's status observed by the clinician during examination, results of laboratory tests, and other observations.

in the care plan as descriptions of the patient's problems. The form of the description ranges from a list of problems selected from a controlled vocabulary to a narrative summary of the problems. Once the problems are established, the clinician reviews each problem and establishes goals to be achieved while addressing the problems, and plans interventions to achieve the goals. An example of a de-identified note derived from the interdisciplinary notes used to build our collection is shown in Figure 1. Ideally, a clinician would seek evidence support for all three steps of care plan development.

The elements of the care plan are in essence the same elements that are used in the evidence based practice (EBP) framework for finding information to ensure the best possible care in a given clinical situation. The elements of the framework, called PICO, are: the description of the patient and problem, intended interventions and comparisons, and desirable outcomes (Richardson et al., 1995). Clearly, these are the problems, interventions, and goals sections of the care plan and we, therefore, can use the framework developed within EBP for construction of well-formed clinical questions to formally represent the clinical situation. We use an existing EBP-based question-answering system, CQA, (Demner-Fushman and Lin, 2007) to find and extract key facts from publications relevant to the patient's case. The CQA system showed good performance answering clinical questions when queries and clinical scenario frames were developed manually. Manual formulation of the query is not ideally suited for use in clinical setting because it interrupts the workflow and is often perceived as less useful than spending the time with the patient (Bond, 2007).

To replace manual initiation of the search for key facts, we developed an automatic process for extracting information from a patient's record and properly formulating a query to identify appropriate evidence using the National Library of Medicine (NLM) resource, PubMed®. Our solution to automatic construction of clinical questions and initiation of the search process is described next.

3. Query Formulation Algorithm

To identify search strategies that would yield relevant results, we manually developed a set of reference PubMed search strings to analyze for text elements, query forms, and search processes that are most likely to yield successful search results. The set was developed by a medical librarian (the second author) using 254 records of patient encounters from 52 patients selected from a dataset of more than 4500 patient encounters. The 254 records were selected because the formal representations of the encounters using the PICO framework and simple searches (that combined all identified PICO elements) retrieved at least one MEDLINE® citation or other evidence (for example, a MedlinePlus® article). The reference queries retrieved the greatest proportion of relevant results, presented the most relevant results at the top of the results display, and retrieved a total number of hits that could be easily perused by a busy clinician in two minutes or less. These strings were constructed using the patient's primary diagnosis (*Chief Complaint*) and problems found in the interdisciplinary notes (*IDP Problem*).

3.1 PICO Representation of Patient Records

The *Problem* and *Intervention* extraction modules of the CQA system were used to represent patients' cases. Given a clinical note, the system automatically generates a question frame using one of the Named Entity Recognition (NER) tools (MetaMap (Aronson, 2001), NER modules of the NLM experimental search engine Essie (Ide et al., 2007) or CQA internal dictionary-based NER module) and a set of rules for extraction of the elements of a clinical scenario.

IDP Pain Problem: patient c/o intermittent pain to surgical site
IDP Pain Interventions: cont with pca dosing, prn hydromorphone also to be given for break through pain
NER (MetaMap Mappings): Intermittent pain [Sign or Symptom]; Surgical Site (Operative site) [Spatial Concept]; Hydromorphone [Organic Chemical,Pharmacologic Substance]; …; Pain [Sign or Symptom]
CQA Problem(s): Intermittent pain
CQA Patient:
CQA Intervention(s): Hydromorphone

Figure 2: Clinical question frame is used to formally represent a patient's note using NER

To generate question frames (see Figure 2 for an example), the CQA system extracts from the NER output concepts that belong to the following semantic groups: *Problems/findings* (meant to represent a patient's problem list), *Interventions*, and *Anatomy* (which provides details about the patient). The semantic groups are based on the Unified Medical Language System® (UMLS®) (Lindberg et al., 1993) Metathesaurus semantic types. The *Problems/findings* semantic group is based on the UMLS semantic group *Disorders* (McCray et al., 2001). The *Interventions* group includes therapeutic and diagnostic procedures, drugs, and drug delivery devices. The *Anatomy* group includes semantic types in the anatomy and physiology groups excluding those on the cell and molecular level (for example, *Cell* or *Molecular Function*).

3.2 Analysis of Reference Queries

We evaluated the manually constructed search string using SAS®[2] hypothesis testing (we used the SAS 9.0 SURVEYLOGISTIC Procedure with the Cumulative Logit logistic regression model and Fisher's Scoring optimization) and SPSS[3] linear and cubic regression analysis. The set of citations retrieved by each search

[2] http://www.sas.com/
[3] http://www.spss.com/

string was evaluated on the following criteria:

- Overall success of the query (evaluated by the second author on a scale of zero to four, 0=no hits, 4=ideal results set) as a function of the number of relevant results in the top 10, number of hits retrieved, and the positions of all relevant results
- Total number of citations retrieved
- Number of relevant citations in top 10
- Position of the first relevant citation
- Number of queries executed prior to success or termination
- Number of review article citations retrieved
- Total number of relevant review article citations in top 10
- Position of first relevant review article citation.

Five variables of query construction were evaluated for their effects on quality of search results:

1. Increased use of Medical subject headings (MeSH® terms) -- controlled vocabulary terms assigned to MEDLINE citations during NLM manual indexing process.
2. Varied use of the Chief Complaint
3. Increased use of advanced search strategies (use of subheadings when appropriate; identification of terms to search as major topic headings)
4. Use of complex query forms (use of Boolean AND/OR/NOT; addition of nested search strings)
5. Application of search limits to retrieve only review articles to reduce total number of results retrieved when the retrieved set is too large.

SAS hypothesis testing identified variable 1, increased use of MeSH terms, variable 3, increased use of advanced search strategies, and variable 4, use of complex query forms, as statistically significant ($p < 0.001$) to a successful search outcome. SPSS regression analysis was then performed on the number of MeSH terms in the query, revealing that searches that used between two and five MeSH terms were far more likely to be successful than searching using fewer than two or more than five MeSH terms. These factors were used to guide the development of the query formulation process.

3.3 Query Formulation Rules

Based on the experience gained during creation of the reference queries and the SAS and SPSS analysis results, the second author derived the following rules for the automatic query formulation:

1. Identify MeSH terms
2. Construct Chief Complaint String: If multiple MeSH terms are identified, combine terms with Boolean “AND”.
3. Prior to constructing IDP Problem string
    3.1. Extract any subheading terms (Identify terms of an identical semantic type to PubMed subheadings and apply to Chief Complaint string (i.e. A drug name in the IDP Problem field would give the subheading “drug therapy” to the Chief Complaint string)
    3.2. Identify terms in the IDP Problem field that are explicitly subheadings. Apply those to the Chief Complaint string. (eg. Text phrase “surgery on Tuesday” would give the subheading “surgery” to the Chief Complaint terms.)
4. Construct IDP Problem string:
    4.1. Use dependency parser to identify relationships between MeSH terms
    4.2. Remove second-child terms
    4.3. Combine parent terms with Boolean “AND”
    4.4. Combine first-child terms with Boolean “OR” and nest this string.
5. Combine Chief Complaint and IDP Problem strings with Boolean “AND.”

Chief Complaint: Diffused B Cell Lymphoma
IDP Problem Text: Maintain perfusion of right hand and fingers, preserve neural function right hand and wrist, track hematoma for status
Baseline Query: Diffused B Cell Lymphoma AND perfusion AND right hand AND fingers AND neural function AND wrist AND hematoma
Advanced Query: (Diffused B Cell Lymphoma) AND (Perfusion AND (Hand OR Fingers) AND (Nerves OR Wrist) AND hematoma)

Figure 3: Automatic query formulation strategies

6. Run iterative searches as necessary:
    6.1. If, when Chief Complaint and IDP Problem strings are combined and hits retrieved are less than or equal to 5, then combine any multiple Chief Complaint terms with Boolean “OR.” Re-execute search.
    6.2. If retrieved set is less than or equal to three, then re-execute the search using only the IDP Problem string.
    6.3. If the results set retrieved from step (a) is between 20 and 100 hits, then search for all Chief Complaint terms as major topic headings.
    6.4. If retrieved set is still greater than or equal to 75, limit results display to review articles.

We evaluated the developed algorithm on 30 additional randomly-selected patient encounter records using a naïve ANDing of all identified PICO elements as the baseline. Figure 3 demonstrates the differences in the advanced and baseline query formulation.

4. Experimental Evaluation of the Query Formulation Algorithm

We evaluated citation sets retrieved by the baseline and advanced search strategies using the evaluation criteria for the reference searches (scale of zero to four; zero being the lowest, with no results retrieved, and four being the highest, with the greatest overall relevancy and usability of the results). Table 1 presents the results of this evaluation.

Relevance Ranking	Baseline	Advanced search
0 (no results retrieved)	4	0
1,2 (not relevant)	26	20
3,4 (relevant)	0	10

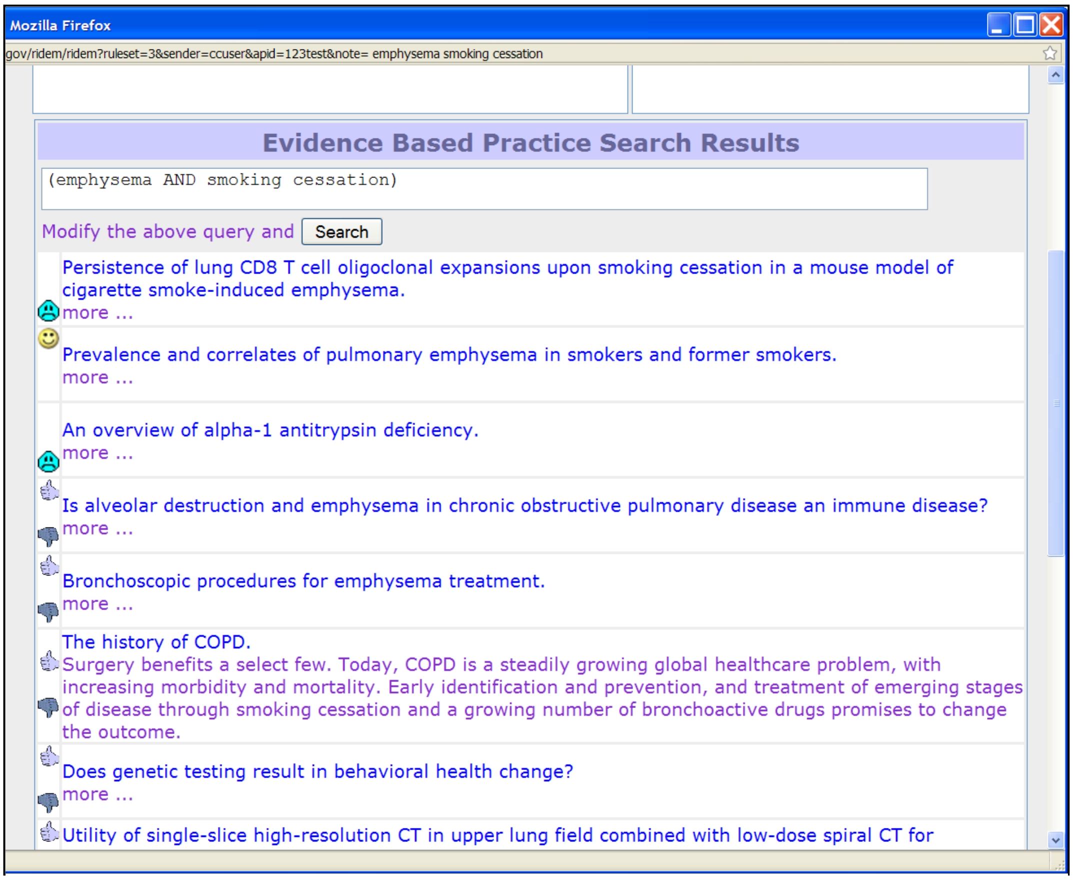
Table 1: Comparison of results retrieved using the baseline and advanced search strategies.

Assuming that a search query that retrieved results receiving a 0, 1, or 2 relevance score were not likely to be useful to a clinician, and that only those receiving a 3 or 4 should be considered successful retrievals that would be useful for patient care plan development, none of the results retrieved by the baseline strategy would thus be considered useful for care plan development, compared with 33.3% (10) of the results retrieved with the updated algorithm.

Result sets retrieved for all 30 queries were also evaluated for the placement of the first relevant result. Of the ten sets derived with the baseline algorithm that contained relevant citations in the top ten results, the mean location was 5.5; median location was 6. Of the 30 sets derived with the advanced search algorithm, mean location of the first relevant result was 2.4, with a median of 2. We observed that placement of the first relevant result was elevated by a mean of four citations when retrieved using the advanced search algorithm.

The overall number of relevant results in the top ten citations retrieved increased from a mean of 0.4 relevant citations in the top ten using the baseline algorithm to a mean of 2.3 relevant citations in the top ten using the advanced search algorithm.

We attribute better performance of the advanced search to reduction of the number of search terms and establishing better relations between the terms due to dependency parsing and rules (as opposed to ANDing all terms found in the note.)


Mozilla Firefox
gov/ridem/ridem?ruleset=3&sender=ccuser&apid=123test&note= emphysema smoking cessation
Search

Evidence Based Practice Search Results

(emphysema AND smoking cessation)

Modify the above query and Search

Persistence of lung CD8 T cell oligoclonal expansions upon smoking cessation in a mouse model of cigarette smoke-induced emphysema.
more ...

Prevalence and correlates of pulmonary emphysema in smokers and former smokers.
more ...

An overview of alpha-1 antitrypsin deficiency.
more ...

Is alveolar destruction and emphysema in chronic obstructive pulmonary disease an immune disease?
more ...

Bronchoscopic procedures for emphysema treatment.
more ...

The history of COPD.
Surgery benefits a select few. Today, COPD is a steadily growing global healthcare problem, with increasing morbidity and mortality. Early identification and prevention, and treatment of emerging stages of disease through smoking cessation and a growing number of bronchoactive drugs promises to change the outcome.

Does genetic testing result in behavioral health change?
more ...

Utility of single-slice high-resolution CT in upper lung field combined with low-dose spiral CT for

Figure 4: Obtaining expert judgments at the point of care

5. Obtaining Relevance Judgments at the Point of Care

Obtaining relevance judgment by the intended consumers of evidence at the point of care is a non-trivial task. Ideally, it has to involve minimal effort and be perceived as part of the workflow. We hope to achieve this goal by providing a service that delivers key facts extracted by the above described tools directly to an electronic patient record (EHR).

The next section provides an overview of the system that delivers evidence to an EHR and, at the same time, provides information to the system that automatically builds the repository for informed decision making.

5.1 System for Evidence Based Practice Support

Delivery of evidence to the point of care starts when a clinician requests evidence[4]. The EHR generates a request to our service. The request sent to our service contains the Chief Complaint and the de-identified patient's note (similar to the one shown in Figure 1). Our system then extracts the PICO elements and MeSH terms found in the note using the CQA modules described in Section 3.1; constructs the query following rules described in Section 3; searches MEDLINE; and responds with an overview of the retrieved evidence to be displayed in the EHR. Figure 4 shows the part of the information dashboard delivered to the EHR that contains the overview of evidence. An overview of the evidence delivery system is provided in (Demner-Fushman et al., 2008).

The evidence provided to clinicians in the information dashboard consists of the titles of MEDLINE citations (linked to the citation and full text paper, if available) and the summary of key facts extracted from the citation using the CQA system.

Logging of the clinicians' navigation of the dashboard and their judgments is described next.

5.2 Capturing Expert Judgments

The key facts are displayed under the title of a MEDLINE citation on demand, The "thumbs up" and "thumbs down" icons to the left of each article allow for a quick one-click judgment. When a clinician clicks one of the thumbs icons, the smiley- or sad-face icons are displayed to illustrate judgment results. At the same time, judgments linked to the citation PubMed unique identifier and the unique identifiers of the clinical scenario are stored in the repository, maintained as MySQL database. The system also registers if the judgment is based on viewing the title and the summary alone, or after following the link to the full citation.

5.3 Preliminary Results

The system is under evaluation through delivering evidence to an EHR at a major clinical center since August 2009. Although viewing evidence is the third popular activity (after accessing information about drugs and the specifics of the patient's case), the absolute number of viewed citations is small (the 350 followed links constitute less than 0.02% of all interactions with the information dashboard). At the same time, we obtained expert relevance judgments for 267 citations (116 positive and 151 negative). The relatively large proportion of papers with judgments compared to the total number of viewed papers is not surprising. The primary goal of judging is to improve other interdisciplinary team members experience when looking for evidence support: papers judged negatively for a given scenario are lowered in rank (based on the cumulative judgment scores), whereas papers judged positively are promoted in rank at the subsequent evidence deliveries.

Recently, clinicians requested that the system allows modifying the automatically constructed query and repeating the search. The search box shown in the top part of Figure 4 is about to be deployed to the EHR. We are looking forward to augmenting our collection with manually corrected searches.

6. Related Work

To the best of our knowledge, the proposed approach to building a collection is new. The mechanism for capturing experts' judgments is related to collaborative filtering widely used in commercial systems to predict a user's interest by collecting information about many users' taste in music, books, etc. (Goldberg et al., 1992), and adaptive information retrieval (Jose et al., 2008).

Our capturing of relevance judgments provided to improve colleagues experience is related to secondary use of biomedical literature, such as using inclusion or citation of a paper by the American College of Physicians (ACP) Journal Club as an indication of the paper's high quality and relevance to a specific clinical task (Aphinyanaphongs et al., 2005); use of MeSH heading indexing as the reference standard in information extraction task (Aronson et al., 2008); and use of journal descriptors for word sense disambiguation (Humphrey et al., 2005).

We also believe to have developed a new approach to automatic query creation. Several approaches have been previously taken to the process of automating query construction using contextual information while filtering for the best quality information.

Of these systems, many retrieve relevant information by relying on the categorization of information within the EHR to automatically identify terms and use those to populate the search fields of an appropriate evidence resource. KnowlegeLink system of drug information retrieval provides users with a search button within the EHR record where medication names appear (Maviglia et al., 2006). The system identifies drug names using text parsing and automatically populates the URL of one of two pre-specified drug databases (MicroMedEx or SkolarMD) with the name of the drug. Within the individual drug databases, users are able to select the type of information about the drug they wish to search for (therapy, adverse effects, etc.) Cimino's (2007) InfoButtons/InfoButton Manager system links the EHR to a repository of clinical questions and answers based on the user's selection of a limited number of different categories of information from the electronic record. A system designed by Rosenbloom et al. (2005) for use with Vanderbilt University's CPOE system similarly provided users with a method of automatically retrieving information relevant to patient care. Using terms derived from the active diagnosis and medication categories in the CPOE system, basic keyword searches could be constructed in PubMed at the point and moment of care.

[4] The approach to initiating the process largely depends on the EHR. In our current setting, the request is issued when a clinician clicks on the EBP tab of the EHR.

7. Conclusions

This paper presents an approach and an ongoing effort towards building a collection of clinical scenarios (that serve as topics of interest), combined with documents automatically retrieved to augment the scenarios with evidence, and expert judgments on the relevance of retrieved documents to the clinical scenario. The benefits of the proposed approach are in the relatively low cost and minimal supervision in construction of the collection, as well as obtaining expert judgments at the point of care.

Some of the drawbacks of the approach are in slow rate of obtaining judgments, sparseness of judgments and lack of reasons for judgments. The last two issues have been extensively studied in the context of TREC (Voorhees and Harman, 2005), but might need re-evaluation in the context of clinical applications. We plan to speed up the collection process through using the system in EBP educational sessions at the clinical center. During the sessions we hope not only to obtain more relevance judgments faster, but also to capture reasoning behind the judgments.

We are less concerned about the quality of obtained judgments – those are provided by members of tightly-knit teams for other members with the purpose of improving provided care, therefore we expect high-quality judgments.

8. Acknowledgements

This work was supported by the Intramural Research Program of the NIH, National Library of Medicine.

9. References

Aphinyanaphongs Y, Tsamardinos I, Statnikov A, Hardin D, Aliferis CF. (2005). Text Categorization Models for High-Quality Article Retrieval in Internal Medicine. *Journal of the American Medical Informatics Association*, 12(2), pp. 207--216.

Aronson AR. (2001) Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proc AMIA Symp*, pp. 17-21.

Aronson AR, Mork JG, Neveol A, Shooshan SE, Demner-Fushman D. (2008). Methodology for Creating UMLS Content Views Appropriate for Biomedical Natural Language Processing. In *Proceedings of the Annual Symposium of the American Medical Informatics Association (AMIA).* Washington, DC: AMIA, pp. 21--25.

Bond CS. (2007). Nurses and computers. An international perspective on nurses' requirements. *Medinfo*, 12(1), pp. 228—232.

Cimino JJ. (2007). An integrated approach to computer-based decision support at the point of care. *Transactions of the American Clinical and Climatological Association*, 118, pp. 273--288.

Demner-Fushman D, Lin J. (2007). Answering Clinical Questions with Knowledge-Based and Statistical Techniques. *Computational Linguistics*, 33(1) pp. 63--103.

Demner-Fushman D, Seckman C, Fisher C, Hauser SE, Clayton J, Thoma GR. A Prototype System to Support Evidence-based Practice. AMIA Annu Symp Proc. 2008 Nov 6:151-5.

Ely JW, Osheroff JA, Chambliss ML, Ebell MH, Rosenbaum ME. (2005) Answering physicians' clinical questions: obstacles and potential solutions. *Journal of the American Medical Informatics Association*, 12(2), pp. 217--224.

Goldberg D, Nichols D, Oki BM, Terry D. (1992) Using collaborative filtering to weave an information tapestry, *Communications of the ACM*, 35 (12), pp. 61—70

Humphrey SM, Rogers WJ, Kilicoglu H, Demner-Fushman D, Rindflesch TC. (2005). Word Sense Disambiguation by Selecting the Best Semantic Type Based on Journal Descriptor Indexing: Preliminary Experiment. *Journal of the American Society for Information Science and Technology*, 57(1), pp. 96--113.

Ide NC, Loane RF, Demner-Fushman D. Essie: a concept-based search engine for structured biomedical text. *J Am Med Inform Assoc*. 2007 May-Jun;14(3):253-63.

Jose JM, Joho H, van Rijsbergen CJ. (2008). Adaptive Information Retrieval, *Information Processing & Management*, 44(6), pp. 1819—1821

Lindberg DA, Humphreys BL, McCray AT. (1993) The Unified Medical Language System. *Methods of information in medicine*, 32(4), pp. 281--291.

Maviglia SM, Yoon CS, Bates DW, Kuperman G. (2006). KnowledgeLink: impact of context-sensitive information retrieval on clinicians' information needs. *Journal of the American Medical Informatics Association*, 13(1), pp. 67--73.

McCray AT, Burgun A, Bodenreider O. Aggregating UMLS semantic types for reducing conceptual complexity. *Stud Health Technol Inform*. 2001;84(Pt 1): 216–20.

Richardson W, Wilson MC, Nishikawa J, Hayward RSA. (1995). The well-built clinical question: a key to evidence-based decisions. *ACP Journal Club*, 123, pp. A-12.

Roberts PM, Cohen AM, Hersh WR. (2009). Tasks, topics and relevance judging for the TREC Genomics Track: five years of experience evaluating biomedical text information retrieval systems. *Information Retrieval*, 12(1), pp. 81 -- 97.

Rosenbloom ST, Geissbuhler AJ, Dupont WD, Giuse DA, Talbert DA, Tierney WM, Plummer WD, Stead WW, Miller RA. (2005). Effect of CPOE user interface design on user-initiated access to educational and patient information during clinical care. *Journal of the American Medical Informatics Association*, 12(4), PP. 458--473.

Voorhees EM, Harman DK. (2005). *TREC: Experiment and Evaluation in Information Retrieval.* Cambridge, MA: The MIT Press

An Empirical Evaluation of Resources for the Identification of Diseases and Adverse Effects in Biomedical Literature

Harsha Gurulingappa[*†], **Roman Klinger**[*], **Martin Hofmann-Apitius**[*†], **and Juliane Fluck**[*]

[*]Fraunhofer Institute for Algorithms and Scientific Computing
Schloss Birlinghoven, 53754 Sankt Augustin, Germany
[†]Bonn-Aachen International Center for Information Technology
Dahlmannstraße 2, 53113 Bonn, Germany
harsha.gurulingappa@scai-extern.fraunhofer.de,
{roman.klinger, martin.hofmann-apitius, and juliane.fluck}@scai.fraunhofer.de

Abstract

The mentions of human health perturbations such as the diseases and adverse effects denote a special entity class in the biomedical literature. They help in understanding the underlying risk factors and develop a preventive rationale. The recognition of these named entities in texts through dictionary-based approaches relies on the availability of appropriate terminological resources. Although few resources are publicly available, not all are suitable for the text mining needs. Therefore, this work provides an overview of the well known resources with respect to human diseases and adverse effects such as the MeSH, MedDRA, ICD-10, SNOMED CT, and UMLS. Individual dictionaries are generated from these resources and their performance in recognizing the named entities is evaluated over a manually annotated corpus. In addition, the steps for curating the dictionaries, rule-based acronym disambiguation and their impact on the dictionary performance is discussed. The results show that the MedDRA and UMLS achieve the best recall. Besides this, MedDRA provides an additional benefit of achieving a higher precision. The combination of search results of all the dictionaries achieve a considerably high recall. The corpus is available on `http://www.scai.fraunhofer.de/disease-ae-corpus.html`

1. Introduction

In the field of biomedical sciences, a huge amount of unstructured textual data is generated every year in the form of research articles, patient health records, clinical reports, medical narratives and patents (Karsten and Suominen, 2009; Cohen and Hersh, 2005). Enormous efforts have been invested in parallel to extract potentially useful information from these textual records (Wang et al., 2009; Chen et al., 2008). Therefore, automatic processing of literature data has gained popularity since over a decade, for example named entity recognition or key concept identification (Smith et al., 2008).

Named entity recognition serves as a basis for biomedical text mining in order to have key entities tagged before they can be subjected to relationship mining or semantic text interpretation. It deals with the identification of boundaries of terms in the text that represent biologically meaningful objects of interest such as genes, proteins, or diseases. Quite a lot of work has been done for the recognition of gene and protein names. For example, the BioCreAtIvE competitions address the challenges associated with the gene name recognition and normalization (Krallinger et al., 2008). Nevertheless, some groups have proposed different solutions for the identification of other interesting classes of biomedical entities such as drug names (Segura-Bedmar et al., 2008; Hettne et al., 2009) or disease names (Jimeno et al., 2008). However, in comparison to the gene and protein name recognition, only a little work has been invested for the recognition of disease names and particularly adverse effects in the free texts. This is partly due to a fact that the availability of annotated corpora is limited and they are of high cost for generation.

A disease in the context of human health is an abnormal condition that impairs the bodily functions and is associated with physiological discomfort or dysfunction. Similarly, an adverse effect is a health impairment that occurs as a result of intervention of a drug, treatment or therapy (Ahmad, 2003). The severity of adverse effects can range from mild signs or symptoms such as *nausea* and *abdominal discomfort* to irreversible damage such as *perinatal death*. Therefore, the mentions of both diseases and adverse effects in free texts denote special entity classes for the medical experts, clinical professionals as well as health care companies (Hauben and Bate, 2009; Forster et al., 2005). This not only helps in understanding the underlying hypothetical causes but also provide rationale means to prevent or diagnose such abnormal medical conditions. Specially in the clinical scenario, recognizing the adverse effects in medical literature can support the clinical decision making (Stricker and Psaty, 2004).

Some research work has been done in the past for the identification of diseases and adverse effects. Jimeno et al. (2008) proposed a statistical solution for the identification of diseases in a corpus of annotated sentences. They reused the corpus that was provided by Ray and Craven (2001) but the corpus has a limitation of being restricted to OMIM[1] diseases only that mostly include genetic disorders. Neveol et al. (2009) utilized the same corpus as well as PubMed[2] user queries for the detection of disease names. They adapted a statistical model and a natural language processing algorithm within their framework. Leaman et al. (2009) proposed a machine learning based technique for the identification of diseases in a corpus containing over

[1] Online Mendelian Inheritance in Man (OMIM): http://www.ncbi.nlm.nih.gov/omim/

[2] http://www.ncbi.nlm.nih.gov/pubmed/

2,500 sentences from PubMed. This corpus is made publicly available as the Arizona Disease Corpus (AZDC)[3] but the annotations are restricted to the diseases only and do not contain information about adverse effects. Curino et al. (2005) proposed a machine learning based solution for mining adverse effects of specific drugs from the web pages. They generated an adverse effect dictionary from the resources provided by the FDA[4]. However, the corpus utilized by Curino et al. (2005) is not openly available. McCray et al. (2001) proposed a statistical solution for mapping the terms in the corpus to the UMLS concepts. They determined the likelihood of a given UMLS string being found or not found in the corpus. A classical example of a tool for mapping the text to biomedical concepts in UMLS[5] meta-thesaurus is the MetaMap program (Aronson, 2001).

Several terminological resources are available that provide information about diseases and adverse effects. Few well known examples include the MeSH[6] thesaurus, the UMLS[7] meta-thesaurus, the ICD-10[8], and the NCI[9] thesaurus. These resources serve as a good basis for the dictionary-based named entity recognition in text but not all of them essentially suit the text mining needs. Although some of these resources have been utilized individually in the past for the detection of disease names (Jimeno et al., 2008; Chun et al., 2006), there is no common platform where most of these resources have been collectively evaluated.

The aim of this work is to provide an overview of the different data sources and evaluate the general usability of the contained disease and adverse effect terminology for named entity recognition. Although, a small set of corpus is available that contain sentences annotated with disease names, there is no freely available corpus containing the PubMed abstracts that are annotated with diseases as well as adverse effects. Therefore, a newly annotated corpora is made publicly available.

2. Terminological Resources

Dictionary-based named entity recognition approaches rely on comprehensive terminologies containing frequently used synonyms and spelling variants. Such resources include databases, ontologies, controlled vocabularies and thesauri. This section gives an overview of the available data sources for diseases and adverse effects. Examples of synonyms and term variants associated with the MeSH disease concepts are provided in Table 1.

Different resources have been designed to meet the needs of different user groups whereas some of them include certain disease specific information. For example, the NCI thesaurus serves as a reference terminology and an ontology providing a broad coverage of cancer domain including cancer related diseases, findings, abnormalities, gene products, drugs, and chemicals. Similarly, there are databases that include very specific organ or disease class related information such as the autoimmune disease database (Karopka et al., 2006) and the DSM-IV Codes[10] which is specific to mental disorders. On the other hand, sources such as the ICD-10, the UMLS and the MedDRA[11] provide a wider coverage of diseases, signs, symptoms, and abnormal findings irrespective of any kind of disease or any affected organ system. All these resources have their own advantages and areas of applicability. Therefore, the survey made here includes only those resources that encompass information about medical abnormalities that are associated with the entire human physiology.

From all the resources introduced here, individual dictionaries were generated and evaluated over a manually annotated corpus. Although, the MeSH, ICD-10, MedDRA, and SNOMED CT are already included as source vocabularies within the UMLS, these resources were separately downloaded from their respective official websites. The main reason is because when the terms from the source vocabularies are imported into the UMLS, they undergo a series of term modification steps [12]. This generates an impression that the terms present in the UMLS may not be identical to the terms present in the source vocabularies. Therefore, in order to validate the hypothesis of suitability of the individual resources for text mining, they were treated as independent terminologies.

Medical Subject Headings (MeSH) is a controlled vocabulary thesaurus from the NLM[13]. It is used by NLM for indexing articles from the PubMed database as well as books, documents, and audiovisuals acquired by the library (Coletti and Bleich, 2001). In MeSH, the terms are arranged in a hierarchical order that are associated with synonyms and term variants. A subset of MeSH that corresponds to the category *Diseases* (tree concepts with node identifiers starting with 'C') was extracted to generate a dictionary covering diseases and adverse effects. The MeSH dictionary contains over 4,500 entries.

Medical Dictionary for Regulatory Activities (MedDRA) is a standardized medical terminology that was developed to share regulatory information internationally about medical products used by human (Merrill, 2008). It provides a hierarchical structure of terms that include signs, symptoms, diseases, diagnosis, therapeutic indications, medical procedures, and familial histories. The MedDRA dictionary contains over 20,000 entries associated with synonyms and term variants.

International Classification of Diseases (ICD-10) is

[3] http://diego.asu.edu/downloads/AZDC/

[4] Food and Drug Administration (FDA): http://www.fda.gov/

[5] http://www.nlm.nih.gov/research/umls/

[6] Medical Subject Headings (MeSH): http://www.nlm.nih.gov/mesh/

[7] Unified Medical Language System (UMLS): http://www.nlm.nih.gov/research/umls/

[8] International Classification of Diseases Edition-10 (ICD-10): http://apps.who.int/classifications/ apps/icd/icd10online/

[9] National Cancer Institute (NCI): http://nciterms.nci.nih.gov/

[10] Diagnostic and Statistical Manual of Mental Disorders (DSM) 4th Edition: http://www.psych.org/mainmenu/research/dsmiv/dsmivtr.aspx

[11] Medical Dictionary for Regulatory Activities (MedDRA): http://www.meddramsso.com/

[12] http://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/source_faq.html#what_involved

[13] National Library of Medicine (NLM): http://www.nlm.nih.gov/

maintained by WHO[14] and it is used to classify diseases and heath problems recorded in many types of health and vital reports including death certificates and health records. The ICD-10 provides terms that are hierarchically ordered according to the organ system that is being affected. Unlike other resources, the ICD provides a flat list of terms and does not include synonyms or term variants. The complete ICD-10 was used for generating the dictionary and it contains over 70,000 entries altogether.

Systematized Nomenclature of Medicine–Clinical Terms (SNOMED CT) [15] is a comprehensive clinical terminology that is maintained and distributed by IHTSDO[16] (Cornet, 2009). It covers most areas of clinical information such as diseases, findings, procedures, microorganisms, pharmaceuticals etc. The SNOMED CT concepts are organized into hierarchies and the sub-hierarchy that corresponds to *Disorder* was used to generate a dictionary. The SNOMED CT dictionary contains over 90,000 concepts associated with synonyms and term variants.

Unified Medical Language System (UMLS) is a very large, multipurpose, and multilingual meta-thesaurus that contains information about biomedical and health related concepts (Browne et al., 2003). Overall, the UMLS has more than 2 million concepts that are associated with synonyms and relationships between them. The concepts in the UMLS are categorized into semantic groups. The semantic group *Disorders* contains semantic subgroups such as *Acquired Abnormality*, *Disease or Syndrome*, *Mental or Behavioral Dysfunction*, *Sign or Symptom*, etc. Although, the downloadable subset of the UMLS enclose large subsets of concepts from sub-thesauri such as the ICD-9, ICD-10, SNOMED CT, and MeSH, the level of ambiguity it contains has been well demonstrated (Aronson, 2000; Rindflesch and Aronson, 1994). Therefore, we presumed to test the UMLS separately in addition to its constituent sources. All concepts in the *Disorders* semantic group of the UMLS were used to generate a dictionary. This dictionary contains over 120,000 entries altogether.

3. Dictionary Characteristics

The dictionaries generated for the recognition of diseases and adverse effects were analyzed with regard to the following properties:

- Total number of entries,
- Number of synonyms provided, and
- Availability of mappings to other data sources

Table 2 provides a quantitative estimate of the entities present in the raw dictionaries. The UMLS has the largest collection of disease and adverse effect data followed by the SNOMED CT. Figure 1 shows the distribution of synonyms for all the analyzed dictionaries. Since the ICD-10 does not provide synonyms and term variants, it is visible only as a point in Figure 1. A large part of all the dictionaries contain less than 20 synonyms. Few entries in the UMLS, MeSH, and MedDRA[17] are associated with as much as more than 60 synonyms. Resources with high number of synonyms are of great value for dictionary-based named entity recognition approaches. They help to overcome a high false negative rate but may pose a risk of high number of false positives requiring a dedicated curation.

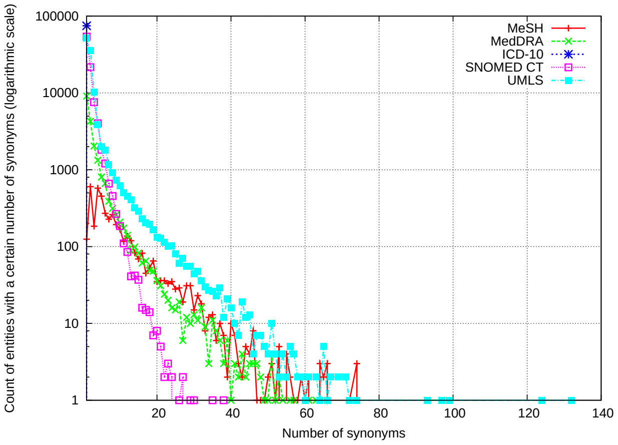


Figure 1: Plot of the synonym count distribution for all the analyzed dictionaries

Since UMLS is the largest resource, a survey was conducted to check the percentage of synonyms that overlap with synonyms in rest of the resources. The synonym comparison between the different resources was performed using a simple case-insensitive string match (i. e. only complete string matches were accepted). About 96 % of the MeSH and 23 % of the MedDRA synonyms are present in UMLS. Only 4 % of the ICD-10 and 13 % of the SNOMED CT synonyms are covered by UMLS. Hence, the outcome of this survey showed that integrating the smaller resources with UMLS would account for an enhanced terminology coverage.

Although, there is an enormous variation in size of the dictionaries used, their adaptability for finding terms in the text is questionable. A manual survey was performed concerning the quality of information contained in each of these dictionaries. The UMLS and SNOMED CT contained over 20,000 terms each that had special characters such as '@', '#&', '[X]', etc. enclosed within the terms. Examples of such ambiguous terms found in the UMLS are *5-@FLUOROURACIL TOXICITY* and *Congestive heart failure #&124*. A large subset of terms were too long and descriptive composed of more than 10 words. Such synonyms are seldom found in the text. An example of such descriptive term found in ICD-10 is *Nondisplaced fracture of lateral condyle of right femur, initial encounter for closed fracture*. ICD-10 has nearly 35,000 long descriptive terms

[14] World Health Organization (WHO): http://www.who.int/en/

[15] http://www.nlm.nih.gov/research/umls/Snomed

[16] International Health Terminology Standards Development Organisation (IHTSDO): http://www.ihtsdo.org/

[17] MedDRA, the Medical Dictionary for Regulatory Activities terminology is the international medical terminology developed under the auspices of the International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH). MedDRA is a registered trademark of the International Federation of Pharmaceutical Manufacturers and Associations (IFPMA)

ID	Concept	Synonyms
D000292	Pelvic Inflammatory Disease	Adnexitis, Inflammatory Disease; Pelvic, Inflammatory Pelvic Disease; Pelvic Disease, Inflammatory
D002534	Brain Hypoxia	Anoxia, Brain; Anoxic Brain Damage; Brain Anoxia; Brain Hypoxia; Cerebral Hypoxia; Encephalopathy, Hypoxic; Hypoxic Brain Damage; Hypoxic Encephalopathy

Table 1: Examples of synonyms and term variants associated with the concepts in the MeSH database.

	MeSH	MedDRA	ICD-10	SNOMED CT	UMLS
No. of entries	4,350	20,515	74,830	92,376	112,341
No. of synonyms (incl. concepts)	42,631	69,121	74,830	170,561	295,773
Percentage of synonyms covered by UMLS	96 %	23 %	4 %	13 %	100 %
Mappings	no	yes	no	yes	yes

Table 2: A quantitative analysis of the dictionaries generated for the disease and side effect named entity recognition. Total number of entries, number of synonyms, percentage of synonyms covered by UMLS, and the availability of inter data source mappings for individual dictionaries are reported. For the UMLS coverage, all synonyms of all the entries were compared.

which constitutes nearly 50 % of the entire dictionary. According to the experience of curators, MeSH and MedDRA were regarded as the specialized resources with considerably low level of ambiguity. Nevertheless, few vague entries such as *Acting out*, *Alcohol Consumption*, and *Childhood* were encountered in these dictionaries.

4. Corpus Characteristics and Annotation

For evaluating the performance of named entity recognition systems, an annotated corpus is necessary. Since, there is no freely available corpus that contains annotations of disease and adverse effect entities, a corpus containing 400 randomly selected MEDLINE abstracts was generated using 'Disease OR Adverse effect' as a PubMed query. This evaluation corpus was annotated by two individuals who hold a Master's degree in life sciences. All the abstracts were annotated with two entity classes, i. e., *disease* and *adverse effect*. In order to obtain a good estimate of the level of agreement between the annotators, they were insisted to carry out the task independently. First, one annotator participated in the development of a guideline for annotation. The corpus was iteratively annotated by this person along with the standardization of the annotation rules. Later, the second person annotated the whole corpus based on the annotation guideline generated by the first annotator. This procedure formed an evaluation corpus of 400 abstracts containing 1428 disease and 813 adverse effect annotations. Recognizing the boundaries without considering the different classes in the evaluation corpus, the inter-annotator agreement F_1 score and kappa (κ) between the two annotators are 84 % and 89 % respectively which indicates a substantial agreement.

The annotation of disease and adverse effect entities were performed very sensitively taking the context into account. Several instances occurred where the disease names and adverse effect names were the same. For example, in the sentence *Hypersensitivity reactions including fever, rash and (more seriously) agranulocytosis are associated with procainamide, and a frequent adverse effect requiring cessation of therapy is the development of systemic lupus erythematosus. (PMID: 2285495)*, the term *systemic lupus erythematosus* occurs as an adverse effect associated with procainamide treatment. In contrary, the sentence *IL-17 expression was found to be associated with many inflammatory diseases in humans, such as rheumatoid arthritis, asthma, systemic lupus erythematosus and allograft rejection and many in vitro studies have indicated a proinflammatory function for IL-17. (PMID: 20338742)* contains *systemic lupus erythematosus* as a disease associated with certain gene function. In such cases, the annotators were strictly insisted to use the contextual information for annotating the entities. Entities that overlap with semantic classes *disease* and *adverse effect* are difficult to be recognized unless a context-based disambiguation is performed. Altogether, there were 178 annotated entities had an overlap with the classes *disease* and *adverse effect*.

5. Results of Dictionary Performance

For the identification of named entities in text, the ProMiner (Hanisch et al., 2005) system was used along with different dictionaries. The text searching with ProMiner was performed using the raw or unprocessed dictionaries as well as with the processed dictionaries. The search was performed using case-insensitive, word order-sensitive and the longest string match as constraints.

The performance of the ProMiner runs with different dictionaries was evaluated using the Precision and Recall. The evaluations were performed for the complete match as well as partial match between the annotated entities and the dictionary terms. A partial match is a situation where either the left boundary or the right boundary of the annotated entity and the ProMiner search result are matched.

The results with raw dictionaries and such a simple search strategy gives a rough estimate of the coverage of different

	MeSH	MedDRA	ICD-10	SNOMED CT	UMLS
No. of entries	4,335	18,273	37,263	84,292	100,871
No. of synonyms (incl. concepts)	42,531	57,017	37,263	146,545	243,602

Table 3: A quantitative analysis of the curated dictionaries applied for the disease and side effect named entity recognition. Total number of entries and number of synonyms present within the individual dictionaries are reported.

		Raw			Curated			Disambiguation		
Dictionary	Match type	All	DIS	AE	All	DIS	AE	All	DIS	AE
MeSH	*Complete*	0.54/0.43	0.46	0.40	0.61/0.43	0.46	0.40	0.61/0.43	0.46	0.40
	Partial	0.73/0.58	0.64	0.51	0.80/0.57	0.62	0.51	0.80/0.57	0.62	0.51
MedDRA	*Complete*	0.48/0.62	0.64	0.59	0.57/0.61	0.63	0.59	0.60/0.61	0.62	0.59
	Partial	0.55/0.72	0.76	0.68	0.67/0.72	0.75	0.68	0.69/0.71	0.74	0.68
ICD-10	*Complete*	0.46/0.10	0.10	0.10	0.57/0.15	0.10	0.19	0.57/0.15	0.10	0.19
	Partial	0.59/0.15	0.15	0.14	0.66/0.19	0.14	0.23	0.57/0.19	0.14	0.23
SNOMED CT	*Complete*	0.38/0.18	0.18	0.18	0.40/0.20	0.22	0.18	0.43/0.18	0.20	0.15
	Partial	0.66/0.28	0.33	0.23	0.69/0.34	0.39	0.28	0.71/0.34	0.39	0.28
UMLS	*Complete*	0.18/0.58	0.60	0.55	0.33/0.57	0.60	0.54	0.36/0.57	0.60	0.54
	Partial	0.25/0.73	0.74	0.71	0.43/0.72	0.73	0.71	0.46/0.72	0.73	0.71
Combined	*Complete*	0.12/0.75	0.80	0.70	0.18/0.76	0.81	0.71	0.19/0.76	0.80	0.71
	Partial	0.14/0.92	0.92	0.91	0.21/0.91	0.92	0.89	0.22/0.91	0.92	0.89

Table 4: Comparison of the performance of different dictionaries tested over the evaluation corpus. The results are reported for the *complete matches* and *partial matches* of annotated classes disease (DIS), adverse effect (AE) and a combination of both the classes (All). For a combination of both the classes, i. e. *All*, the precision and recall values are reported. For the classes DIS and AE, only the recall values are reported. 'Combined' indicates the performance achieved by combining the results of all the dictionaries.

dictionaries and the effort that has to be invested to curate them. Table 4 shows the search results obtained with every individual dictionary when complete matches and partial matches were considered. The highest recall for complete matches were achieved by the MedDRA dictionary (62 %) and the UMLS dictionary (58 %). The recall of ICD-10 was the lowest of all dictionaries covering only 10 % of the entities annotated in the corpus. Unlike the other dictionaries, ICD-10 lacks information about the synonyms and term variants which hinders it from covering different types of variants mentioned in the text. The combination of results of all the dictionaries lead to a promising recall of 75 %.

Another important observation is the low recall (18 %) attained by the SNOMED CT dictionary. Although, this dictionary contains over 90,000 entries with 170,561 different terms, its usability for finding entities in the text seems extremely limited. One reason is because of the descriptive nature of most of the terms present in the SNOMED CT vocabulary such as *Spastic paraplegia associated with T-cell lymphotropic virus - 1 infection*. Although such long descriptive terms provide substantial information about the medical condition, they are not quite often used in the literature. Additional reasons are the perception of named entities in annotator's mind as well as the style adopted by the annotation guideline. Perhaps, our principle annotators would annotate such a textual description with *Spastic paraplegia* and *T-cell lymphotropic virus - 1 infection* as two distinct entities rather than annotating the entire phrase as a single entity.

Comparison of the results of complete matches and partial matches in Table 4 shows the granularity of information covered by different data sources and the textual explications. The UMLS and MedDRA achieved an overall recall of 73 % and 72 % respectively for the partial matches whereas the combined results of all the dictionaries achieved a highest recall of 92 %. This provides an indication that the terms contained in these dictionaries cover the head nouns associated with the disease and adverse effect entities but does not include different enumerations used in the literature. For example, in the case of *progressive neurodegenerative disorder*, only *neurodegenerative disorder* was identified whereas the adjective *progressive* was not covered. Based on the experience of the curators and the results from Table 4, nearly 10 % of the mismatches are caused by the medical adjectives such as *chronic*, *acute*, and *idiopathic* that are frequently used in texts but not provided by the resources. Another source of mismatch is the anatomical information often attached to the disease entity in texts. For example, in the case of *vaginal squamous cell carcinoma*, only the *squamous cell carcinoma* was recognized whereas the remaining anatomical substring remained unidentified.

The highest precision rates for the complete matches were achieved by the MeSH dictionary (0.54) and the MedDRA dictionary (0.48) hence validating the curator's opinion about the quality of these resources. The lowest precision of 18 % was achieved by the UMLS dictionary. The precision after combining the results of different dictionaries was considerably low due to the overlapping false positives generated by different dictionaries. The low precision is due to the presence of noisy terms such as *disease* or *response* within the dictionaries. The amount of such noisy terms considerably varies among the different resources with UMLS having the highest. Therefore, the curation of dictionaries is necessary in order to achieve better performance. Experiences from the previously reported dictionary-based named entity approaches let us assume that the precision could be greatly improved by the dictionary curation.

Since the MedDRA dictionary achieved the highest recall, the true positive matches obtained with this dictionary were mapped to the MedDRA level-2 superclasses in order to analyze the distribution of disease and adverse effect terminology over the complete MedDRA hierarchy. The analysis of distribution of annotated entities over the MedDRA subhierarchies is shown in Table 5 and Table 6. From the MedDRA tree distribution of disease or adverse effect matches, it is difficult to understand whether the entity is of kind disease or an adverse event. Here an additional context will be necessary to classify the matches into their respective classes.

MedDRA Superclass	No. of annotated entities
Infections and infestations	110
Psychiatric disorders	83
Neoplasms benign, malignant and unspecified	83
Nervous system disorders	47
Blood and lymphatic system disorders	38

Table 5: Analysis of the top five most frequently occurring disease entities distributed over different MedDRA level-2 superclasses.

MedDRA Superclass	No. of annotated entities
Cardiac disorders	96
Infections and Infestations	93
Injury, poisoning and procedural complications	29
Vascular disorders	23
Gastrointestinal disorders	19

Table 6: Analysis of the top five most frequently occurring adverse effect entities distributed over different MedDRA level-2 superclasses.

5.1. Dictionary Curation

The dictionaries were processed and filtered based on a subset of pre-defined rules in order to reduce the level of ambiguity associated with them. Most of the rules were adapted from Hanisch et al. (2005) and Aronson (1999). The rules that were applied for processing the dictionaries are listed below. All the rules were used in common to all the analyzed dictionaries.

Remove very short tokens: Single character alphanumericals that appear as individual synonyms were removed. For example, '5' was mentioned as a synonym of the concept *Death Related to Adverse Event* in the UMLS.

Remove terms containing special characters: Remove all the terms that contain unusual special characters such as '@', ':' and '&#'. An examples of such term in SNOMED CT is *Heart anomalies: [bulbus/septum] [patent foramen ovale]* .

Remove underspecifications: Substrings such as *NOS*, *NES* and *not elsewhere classified* were removed away from the terms. Such strings were often encountered at endings of the dictionary terms. An example of such a term from MedDRA is *Congenital limb malformation, NOS*

Remove very long terms: Very long and descriptive terms that contains more than 10 words were removed. An example of such a term found in SNOMED CT is *Pancreas multiple or unspecified site injury without mention of open wound into cavity*. Although such long terms do not appear in the text, filtering them from the dictionary gradually reduces the run time of the process.

Remove unusual brackets: Unusual substrings that often appear within the brackets were removed from the terms. Examples of such terms found in SNOMED CT include *[X]Papulosquamous disorders* and *[D]Trismus*.

Remove noisy terms: The ProMiner with different dictionaries was run over an independent corpus of 100,000 abstracts that were randomly selected from MEDLINE. The 500 most frequently occurring terms matched with the individual dictionaries were manually investigated to remove the most frequently occurring false positives. This process will improve the precision of entity recognition during the subsequent runs.

In addition to dictionary curation, the configuration of the ProMiner system was readjusted to match the possessive terms (e. g. *Alzheimer's disease*) that contain '*'s*' substring at the word endings. After the end of the dictionary processing and filtering, the number of entries and synonyms that remained in the individual dictionaries can be found in Table 3. The MeSH dictionary sustained minimum changes with only 15 entries being removed whereas ICD-10 underwent a large noticeable change. The size of the ICD-10 dictionary was reduced to nearly half of the previously used raw dictionary. The search results obtained with every individual curated dictionary can be found in Table 4.

As the result of dictionary curation, the performance of all the dictionaries improved remarkably well. For the complete matches, the precision of UMLS dictionary raised by 15 % with a drop in recall by just 1 %. Other dictionaries that benefited well from the curation process are ICD-10 and MedDRA with raise in their precision by 11 % and 9 % respectively. SNOMED CT showed only 2 % increase in the precision. The recall of all the dictionaries changed marginally except for ICD-10. Processing the synonyms of ICD-10 increased its recall on adverse effect entities by 9 % with an overall raise in the recall by 5 % for both the annotated classes.

5.2. Acronym Disambiguation

In spite of processing the dictionaries by removing the noisy terms as well as lexical modification of the synonyms, the acronyms present in the dictionaries turned out to be another source of frequent false positives. For example, *ALL* which is an acronym for *Acute Lymphoid Leukemia* generated a considerable noise. Therefore, acronyms present in all the dictionaries that have two to four characters were collected in a separate acronym list. Whenever there is a match between the term in the acronym list and the text tokens, a rule was defined in order to accept or neglect the match. This disambiguation facility is available within the ProMiner system. The acronym disambiguation rule accepts the match based on two criteria and they are:

- The match should be case sensitive.
- The acronym as well as any one of its synonym in the respective dictionary should co-occur anywhere within in the same abstract.

For example, the term *ALL* is associated with 17 synonyms in the MedDRA dictionary. Any case sensitive match between the *ALL* and tokens in the text would be accepted if any one synonym of the *ALL* occurs within the same abstract. The search results obtained with the individual curated dictionaries in addition to the acronym disambiguation can be found in Table 4. Considering the complete matches, the acronym disambiguation raised the precision of MedDRA, SNOMED CT and UMLS dictionaries by 3 % each. The performance of MeSH and ICD-10 remain unaffected indicating the presence of less acronyms within them. There was a marginal decline (less than 2 %) in the recall of the dictionaries after applying the disambiguation rule.

In summary, the experiments demonstrated that the performance of a simple search strategy using individual dictionaries for the identification of diseases or adverse effects is low. However, the precision of the dictionary look-up can be improved with the help of curation as well as rule-based filtering (e. g. the one adopted here for disambiguating the acronyms). When the performance of different dictionaries was compared, the MeSH and the MedDRA showed the highest quality with comparably low false positive rate and low ambiguity. The UMLS and SNOMED CT having the size five times as greater than MedDRA or MeSH reported low precision although there was an improvement after the subsequent curation. Depending on the user-specific needs, the UMLS and MedDRA cover large parts of the elementary disease names but does not include sufficient medical adjectives and anatomical specifications within the terms. Although, a sufficient effort has been invested to curate the SNOMED CT and UMLS, the amount of noise they contain overweighs their performance. The MedDRA and UMLS dictionaries demonstrated a competitive recall but the MedDRA being substantially smaller than UMLS reported comparatively low false positive rate. Finally, a combination of all the dictionaries reported the highest recall indicating the diversity of terms provided by different resources.

6. Conclusions

A survey of the performance of different resources for the identification of diseases and adverse effects in texts was performed. An outcome of the survey upheld the MedDRA as a compatible resource for the text mining needs having its recall competitive to the UMLS meta-thesaurus with considerably fair precision upon processing. The UMLS being the largest resource does not include all the names that are covered by the smaller resources. Hence, the combination of the search results from all the terminologies lead to a high increase in recall. This indicates a need for intelligent ways to integrate and merge the information spread across different resources. The amount of work that needs to be invested to curate very large resources such as the SNOMED CT and UMLS is also shown.

In addition to the performance comparison, the effect of dictionary curation and a limited manual investigation of the noisy terms shows to be effective. A rule-based processing coupled with the dictionary curation can substantially improve the performance of the named entity recognition.

In future, we will investigate more enhanced dictionary curation methods for improving the performance of dictionaries. Nevertheless, the performance of rule-based and machine learning-based approaches for identifying the disease and adverse effect named entities needs to be tested.

7. References

S. R. Ahmad. (2003). Adverse drug event monitoring at the Food and Drug Administration. *Journal of General Internal Medicine*, 18(1), pp. 57–60.

A. R. Aronson. (1999). Filtering the UMLS Metathesaurus for MetaMap. Technical report, National Library of Medicine, MD, USA. Available at http://skr.nlm.nih.gov/papers/references/filtering99.pdf.

A. R. Aronson. (2000). Ambiguity in the UMLS Metathesaurus. Technical report, National Library of Medicine, MD, USA. Available at http://skr.nlm.nih.gov/papers/references/ambiguity00.pdf.

A. R. Aronson. (2001). Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proceedings of the AMIA Symposium*, pp. 17–21.

A. C. Browne, G. Divita, A. R. Aronson, and A. T. McCray. (2003). UMLS language and vocabulary tools. *Proceedings of the AMIA Symposium*, p. 798.

E. S. Chen, G. Hripcsak, H. Xu, M. Markatou, and C. Friedman. (2008). Automated acquisition of disease drug knowledge from biomedical and clinical documents: an initial study. *Journal of the American Medical Informatics Association*, 15(1), pp. 87–98.

H. Chun, Y. Tsuruoka, J. Kim, R. Shiba, N. Nagata, T. Hishiki, and J. Tsujii. (2006). Extraction of gene-disease relations from Medline using domain dictionaries and machine learning. *Pacific Symposium on Biocomputing*, pp. 4–15.

A. M. Cohen and W. H. Hersh. (2005). A survey of current work in biomedical text mining. *Briefings in Bioinformatics*, 6(1), pp. 57–71.

M. H. Coletti and H. L. Bleich. (2001). Medical subject headings used to search the biomedical literature. *Journal of the American Medical Informatics Association*, 8(4), pp. 317–323.

R. Cornet. (2009). Definitions and qualifiers in SNOMED CT. *Methods of Information in Medicine*, 48(2), pp. 178–183.

C. A Curino, Y. Jia, B. Lambert, P. M. West, and C. Yu. (2005). Mining officially unrecognized side effects of drugs by combining web search and machine learning. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, pp. 365–372.

A. J. Forster, J. Andrade, and C. van Walraven. (2005). Validation of a discharge summary term search method to detect adverse events. *Journal of the American Medical Informatics Association*, 12(2), pp. 200–206.

D. Hanisch, K. Fundel, H. Mevissen, R. Zimmer, and J. Fluck. (2005). Prominer: rule-based protein and gene entity recognition. *BMC Bioinformatics*, 6 Suppl 1, pp. S14.

M. Hauben and A. Bate. (2009). Decision support methods for the detection of adverse events in post-marketing data. *Drug Discovery Today*, 14(7-8), pp. 343–357.

K. M. Hettne, R. H. Stierum, M. J. Schuemie, P. J. Hendriksen, B. J. Schijvenaars, E. M. van Mulligen, J. Kleinjans, and J. A. Kors. (2009). A dictionary to identify small molecules and drugs in free text. *Bioinformatics*, 25(22), pp. 2983–2991.

A. Jimeno, E. Jimenez-Ruiz, V. Lee, S. Gaudan, R. Berlanga, and D. Rebholz-Schuhmann. (2008). Assessment of disease named entity recognition on a corpus of annotated sentences. *BMC Bioinformatics*, 9 Suppl 3, pp. S3.

T. Karopka, J. Fluck, H. Mevissen, and A. Glass. (2006). The Autoimmune Disease Database: a dynamically compiled literature-derived database. *BMC Bioinformatics*, 7, pp. 325.

H. Karsten and H. Suominen. (2009). Mining of clinical and biomedical text and data: editorial of the special issue. *International Journal of Medical Informatics*, 78(12), pp. 786–787.

M. Krallinger, A. Morgan, L. Smith, F. Leitner, L. Tanabe, J. Wilbur, L. Hirschman, and A. Valencia. (2008). Evaluation of text-mining systems for biology: overview of the Second BioCreative community challenge. *Genome Biology*, 9 Suppl 2, pp. S1.

R. Leaman, C. Miller, and G. Gonzalez. (2009). Enabling Recognition of Diseases in Biomedical Text with Machine Learning: Corpus and Benchmark. In *Handbook of the 3rd International Symposium on Languages in Biology and Medicine*.

A. T. McCray, O. Bodenreider, J. D. Malley, and A. C. Browne. (2001). Evaluating umls strings for natural language processing. *AMIA Annual Symposium Proceedings*, pp. 448–452.

G. H. Merrill. (2008). The MedDRA paradox. *AMIA Annual Symposium Proceedings*, pp. 470–474.

A. Neveol, W. Kim, J. W. Wilbur, and Z. Lu. (2009). Exploring two biomedical text genres for disease recognition. In *BioNLP '09: Proceedings of the Workshop on BioNLP*, pp. 144–152.

S. Ray and M. Craven. (2001). Representing sentence structure in hidden markov models for information extraction. In *IJCAI'01: Proceedings of the 17th international joint conference on Artificial intelligence*, pp. 1273–1279, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

T. C. Rindflesch and A. R. Aronson. (1994). Ambiguity resolution while mapping free text to the UMLS Metathesaurus. *Proceedings of the Annual Symposium on Computer Applications in Medical Care*, pp. 240–244.

I. Segura-Bedmar, P. Martinez, and M. Segura-Bedmar. (2008). Drug name recognition and classification in biomedical texts. A case study outlining approaches underpinning automated systems. *Drug Discovery Today*, 13(17-18), pp. 816–823.

L. Smith, L. K. Tanabe, R. J. Ando, C. J. Kuo, I. F. Chung, C.N. Hsu, Y. S. Lin, R. Klinger, C. M. Friedrich, K. Ganchev, M. Torii, H. Liu, B. Haddow, C. A. Struble, R. J. Povinelli, A. Vlachos, W. A. Baumgartner, L. Hunter, B. Carpenter, R. T. Tsai, H. J. Dai, F. Liu, Y. Chen, C. Sun, S. Katrenko, P. Adriaans, C. Blaschke, R. Torres, M. Neves, P. Nakov, A. Divoli, M. Mana-Lopez, J. Mata, and W. J. Wilbur. (2008). Overview of BioCreative II gene mention recognition. *Genome Biology*, 9 Suppl 2, pp. S2.

B. H. Stricker and B. M. Psaty. (2004). Detection, verification, and quantification of adverse drug reactions. *British Medical Journals*, 329(7456), pp. 44–47.

X. Wang, G. Hripcsak, M. Markatou, and C. Friedman. (2009). Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: a feasibility study. *Journal of the American Medical Informatics Association*, 16(3), pp. 328–337.

A Task-Oriented Extension of the Chinese MeSH Concepts Hierarchy

Xinkai Wang and Sophia Ananiadou
National Centre for Text Mining
School of Computer Science
University of Manchester
United Kindom
E-mail: wangxa@cs.man.ac.uk, Sophia.Ananiadou@manchester.ac.uk

Abstract

The Medical Subject Headings (MeSH) thesaurus is used not only as a vocabulary for indexing and cataloguing biomedical articles, but also constitutes an important linguistic resource for text mining in the biomedical domain. As part of our research, the Chinese translation of English MeSH, Chinese Medical Subject Headings (CMeSH), will be used to provide essential terms and concepts to a Chinese-English cross-lingual information retrieval system. The original MeSH uses a small, accurate and concise set of terms designed for indexing and cataloguing. Such a set of terms does not, however, lend itself well to information retrieval applications, in which the ability to expand queries using term synonyms is important to maximise relevant search results. In this paper, we propose a new approach to extending the MeSH concept hierarchy (the MeSH Tree) based on an online version of the CMeSH term list. We use Google to collect synonyms for Chinese terms and calculate weights for each Chinese term and its synonyms. Our extension has been evaluated on a Chinese-English biomedical information retrieval application. The results indicate that the extension of CMeSH improves the performance of the information retrieval application. Furthermore, the extended resource should also be helpful in other related research.

1. Introduction

The Medical Subject Heading (MeSH) thesaurus, which was developed and released by the National Library of Medicine (http://www.nlm.nih.gov/mesh/), is widely accepted as the standard vocabulary used for indexing, cataloguing, and searching for biomedical and health-related information and documents. For example, Lowe and Banett (1994) report using MeSH to index medical literature. Cooper and Miller (1998) compare lexical and statistical methods used to extract a list of suggested MeSH terms from the narrative part of the electronic patient medical records. More recently, many researchers (Guo et al, 2004; Abdou and Savoy, 2007; Lu et al, 2008) have employed MeSH terms to evaluate or improve biomedical information retrieval applications. In addition, MeSH terms are treated as the standard vocabulary to which terms from other resources are mapped (Elkin et al, 1988; Shultz, 2006). MeSH vocabulary has also been employed in the construction of Chinese medical ontologies. Zhou et al. (2007) attempt to discover novel gene networks and functional knowledge of genes using a significant bibliographic literature database of traditional Chinese medicine. In their research, MeSH disease headings are applied to generate the index data for gene and disease MEDLINE literature.

As part of own research, we have made use of a MeSH-related resource, i.e., Chinese Medical Subject Headings (CMeSH) to evaluate and improve the performance and effectiveness of a Chinese-English biomedical cross-lingual information retrieval application.

CMeSH, which has been translated and is maintained by The Institute of Medical Information of the Chinese Academy of Medical Sciences, retains the terms and concepts of the English MeSH and their relations. Only a small number of studies have so far attempted to use CMeSH to improve the performance of natural language processing (NLP) applications, such as information retrieval or information extraction systems. Qin and Feng (1999) apply CMeSH terms to improve the indexing quality of Chinese abstracts from 1977 concerning family planning and gynecology, whilst Li et al. (2001) develop an information retrieval system with the help of CMeSH terms. The main reason why MeSH terms have not been more widely adopted in Chinese biomedical text processing lies in the philosophy of MeSH design. As MeSH terms are intended to index and catalogue the biomedical literature, they must be represented succinctly, concisely, and accurately. The Chinese translation of MeSH, CMeSH, inherits these features. This means that the original CMeSH terms have no synonyms - each English term has one and only one Chinese translation.

In order to expand the coverage of terms in CMeSH, and also to make it a more useful resource for our research, we have extended the original CMeSH with synonyms and term weights, and have integrated the extended set of terms into the MeSH concept hierarchy, i.e. the MeSH Tree. An online version of the CMeSH (http://www2.chkd.cnki.net/kns50/Dict/dict_list.aspx?firstLetter=A) is the starting point of our work. We have designed an algorithm which exploits the Google search engine to automatically collect synonyms of each original CMeSH term and calculate the frequency of each extracted term. Based on these frequencies, we have defined a formula to compute a weight for each Chinese term. The corresponding English term's weight is not computed, for the reasons discussed in Section 4. Finally, a Chinese-English cross-lingual information retrieval system (CLIR), based on the Lemur toolkit, has been employed to evaluate the extended CMeSH Tree.

2. Background

2.1 MeSH

MeSH consists of a controlled vocabulary, coupled with a hierarchical tree structure. The controlled vocabulary contains several types of concepts, namely Publication types, Geographics, Qualifiers, Descriptors, and Supplementary Concept Records.

- Publication types or Publication characteristics are used to indicate the genre of the indexed item, rather than its contents, e.g., ‘*Historical Article*’.
- Geographics include continents, regions, countries, and other geographic subdivisions; they are not used to characterize subject content but rather physical location.
- Descriptors are the main concepts or headings. They are used to index the catalogue, and to search biomedical documents. Examples include ‘*Dementia*’ and ‘*Carcinoma in Situ*’.
- Qualifiers, also known as Subheadings, are used for indexing and cataloging in conjunction with Descriptors. There are 95 Qualifiers (MeSH 2009), which provide a convenient means grouping together those documents which are concerned with a particular aspect of a subject. For example, ‘*Liver/drug effects*’ indicates that the article or book is not about the ‘*liver*’ in general, but about the effect of drugs on the liver.
- Supplementary Concept Records (SCRs) are used to index chemicals, drugs, and other concepts. Unlike Descriptors, SCRs have no tree numbers (see below).

The MeSH Tree is a hierarchy of MeSH descriptors, in which each descriptor is allocated a tree number, which represents the position of the node in the tree. The following example illustrates the structure and organization of the MeSH Tree.

```
……
Dementia;C10.228.140.380
AIDS Dementia Complex;C10.228.140.380.070
Alzheimer Disease;C10.228.140.380.100
Aphasia, Primary Progressive;C10.228.140.380.132
Creutzfeldt-Jakob Syndrome;C10.228.140.380.165
Dementia, Vascular;C10.228.140.380.230
CADASIL;C10.228.140.380.230.124
……
```

On each line, the text before the semi-colon constitutes a MeSH term. In the remainder of the paper, we refer to these as ‘English terms’. After the semi-colon, the string starting with Latin letter and followed by digits and dots represents a tree number, which encodes the term’s position within the tree. The version of the MeSH Tree used in this study is the MeSH Tree 2008, which has 24,763 unique terms and 48,442 tree nodes.

2.2 CMeSH

CMeSH is published by The Institute of Medical Information of the Chinese Academy of Medical Sciences, consisting of two different versions, i.e., a paper version and an electronic version. The official CMeSH contains three parts: a Chinese translation of MeSH, traditional Chinese medical subject headings and Special Classification for Medicine of China Library Classification. The usual usage of CMeSH is to index and catalogue biomedical literature in a library, or to provide standard keywords to describe journal articles and conference papers.

An online version of the CMeSH term list is available at: http://www2.chkd.cnki.net/kns50/Dict/dict_list.aspx?firstLetter=A. The example below illustrates the Chinese counterpart of the English MeSH example presented above.

Dementia	痴呆
AIDS Dementia Complex	艾滋病痴呆复合征
Alzheimer Disease	阿尔茨海默病
Aphasia, Primary Progressive	失语, 原发进行性
Creutzfeldt-Jakob Syndrome	克-亚综合征
Dementia, Vascular	痴呆, 血管性
CADASIL	大脑常染色体显性动脉病合并皮层下梗塞及脑白质病

In contrast to research achievements using the original MeSH, the usage of CMeSH is currently largely limited to acting as a gold standard for indexing and cataloging biomedical documents or for assigning indexing terms in IR systems. There is very little work that reports on evaluating cross-lingual information retrieval with CMeSH or on improving information extraction via CMeSH terms. Analyzing the social or economical factors which limit the usage of CMeSH is the duty of economists; our focus is on the deficiencies of the original CMeSH, based on our task-oriented requirements. From the above example, we can conclude that:

- There are no term weights for CMeSH terms.
- Each English term has one and only one Chinese translation.

Term weights are essential to text mining or NLP algorithms based on probabilistic and statistical models. Without term weights, CMeSH can thus function only as a traditional word list. In the cross-lingual information retrieval task, our experiments have shown the high degree to which term weights contribute towards the improvement of retrieval performance (see section 5.2). Another issue of the original CMeSH is that many Chinese translations are missing. Like other languages, the Chinese language can express a particular concept in multiple ways. For example, ‘Alzheimer Disease’ is translated as ‘阿尔茨海默病’ in the original CMeSH. However, it can also be written as ‘Alzheimer 病’, ‘阿滋海默症’, ‘老年性痴呆’, or ‘Alzheimer 氏病’. The original CMeSH thus lacks the ability to provide synonyms for a particular term. Our results have shown that the availability of such synonyms can also increase task performance.

The research described in this paper attempts to overcome the above-mentioned issues of the original CMeSH. Firstly, Google is used to collect web pages which may contain candidate translations. Following this, linguistic rules and a term extraction tool are applied to identify candidate terms from these web pages, and the frequency of each term is calculated. Finally, each term's weight is computed according to frequency of the term and that of its English equivalent.

3. Related Work on Ontology Evaluation

The extended CMeSH Tree constitutes an ontology. Evaluating the effectiveness of this ontology is a critical step of our research. In general, ontology evaluation cannot be compared to evaluation tasks in information retrieval or classic natural language processing tasks such as part-of-speech (POS) tagging, because the notion of precision and recall can not easily be defined. Methodologies used to evaluate ontologies generally fall under one of the following approaches:

- Testing the ontology in an application and evalutating the result (Porzel and Malaka, 2004); also called application-based evaluation;
- Comparing the ontology to a 'gold standard' (Maedche and Staab, 2002);
- Human evaluation of the ontology according to a set of predefined criteria, standards, requirements, etc. (Lozano-Tello and Gómez-Pérez, 2004);
- Comparing the ontology with a set of data (e.g., a collection of documents) from the domain to be covered by the ontology (Brewster et al., 2004); also called data-driven evaluation.

Evaluation of ontologies in general is carried out at three basic levels: vocabulary, taxonomy, and (non-taxonomic) semantic relations. We are not intending to evaluate the *isa* hierarchy (taxonomy) and the non-taxonomic relations (semantic relations) of the extended CMeSH Tree, because our work does not add new tree nodes to the MeSH concept hierarchy. Moreover, based on the fact that MeSH Tree, as a part of The Unified Medical Language System (UMLS), has been assessed by human experts against a set of criteria (Kumar and Smith, 2003; Smith, 2006), our evaluation of the extended CMeSH Tree will serve only to evaluate the enhanced ontology vocabulary. In order to do this, we have tested the extended CMeSH tree within a CLIR application.

4. Extension of CMeSH

The Figure 1 illustrates the workflow of the process of extending CMeSH. In this chart, the dashed lines represent the steps of obtaining the frequencies of English terms, while solid lines correspond to the steps of extending the original CMeSH, including computing the weights of Chinese translations. The details of the algorithm are explained in the subsections that follow.

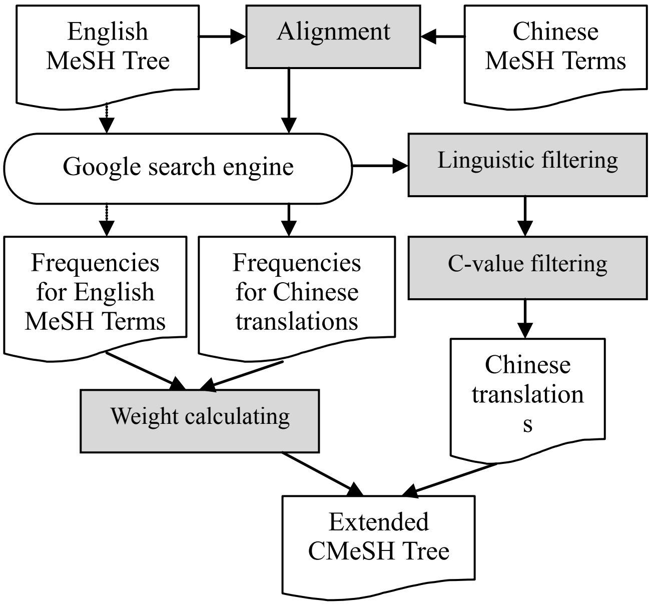


Figure 1: Extension workflow

4.1 Alignment

Alignment is the operation by which the English MeSH Tree terms are matched with the corresponding Chinese MeSH terms. In our experiments, we found that approximately 3% of English terms in the MeSH Tree had no translation in the online CMeSH term list, and that about 8.1% of Chinese terms had no matching MeSH tree terms. In order to resolve this issue, both English terms without Chinese translations, and Chinese CMeSH terms without English counterparts were ignored in the subsequent steps of processing. Following alignment, the MeSH Tree has the following appearance.

Dementia;C10.228.140.380 痴呆
AIDS Dementia Complex;C10.228.140.380.070 艾滋病痴呆复合征
Alzheimer Disease;C10.228.140.380.100 阿尔茨海默病
Aphasia, Primary Progressive;C10.228.140.380.132 失语, 原发进行性
Creutzfeldt-Jakob Syndrome;C10.228.140.380.16 克-亚综合征
Dementia, Vascular;C10.228.140.380.230 痴呆, 血管性
CADASIL;C10.228.140.380.230.124 大脑常染色体显性动脉病合并皮层下梗塞及脑白质病

4.2 Linguistic Filtering

Each Chinese term in the aligned MeSH tree was queried using Google. This caused a set of relevant documents to be returned, most of which were written in Chinese. In order to extract candidate terms from the returned documents, a set of two-level linguistic rules were applied to filter the document set and to discover candidate terms.

The rules take the form of regular expressions, which are based on the following linguistic characteristics in Chinese biomedical texts.

1). Most Chinese terms have suffixes. In the above example, ‘-复合征’ (meaning ‘complex’), ‘-综合征’ (meaning ‘syndrome’), and ‘-病’ (the general name for all kinds of disease) are all suffixes.

2). Some Chinese terms contains ‘inner’ keywords, which can help to identify terms. For example, in ‘失语, 原发进行性’ and ‘痴呆, 血管性’ (Their normal formats are ‘原发进行性失语’ and ‘血管性痴呆’ respectively.), ‘-性’ is an important character that can indicate, when used between two adjacent verbs and nouns, indicates that the first word describes the term after it, thus indicating a high probability of the presence of a term.

3). Compared to the clear suffixes of Chinese terms, it can be more difficult to recognize the start of a term from a stream of characters. Fortunately, terms are often followed by synonyms, which are often indicated using a particular set of phrases. For instance, in the sentence of ‘阿尔茨海默病(Alzheimer disease, AD), 又称早老性痴呆，……’, ‘又称’, which means ‘that is’ or ‘i.e.’, can be a good identifier to determine the beginning of the term. Other similar phrases are ‘简称’ (abbreviated as), ‘也叫’ (that is), ‘也称’ (that is), ‘还叫’ (that is), ‘叫做’ (named as or called), etc.

4). Some symbols (e.g. brackets and parentheses) can play the role of delimiters which define the boundaries of term. In the sentence of ‘…可以减少人们患早老性痴呆(阿尔茨海默氏症)的危险， … ’, the phrase between parentheses is a term, whose meaning is ‘Alzheimer Disease’. Such symbols, may, however, cause ambiguity. For example, the chemical term ‘1-(4-氟苯基)-1,3-二氢-5-异苯并呋喃腈’ (citalopram) contains brackets and comma. Without special rules, the extracted candidates should be ‘4-氟苯基’ and ‘1-(4-氟苯基)-1’, which are clearly incorrect and not terms. Thus, it is necessary to apply constraints to rules that exploit these symbols.

5). Many Chinese terms start with an English word. For example, ‘阿尔茨海默病’ can also be written as ‘Alzheimer 症’ or ‘AD 症’.

According to these linguistic features, we firstly define four word lists:

- **SUFF** This list contains 347 suffixes, such as ‘-复合征’, ‘-病’, ‘-腈’, ‘-烃’, and etc.
- **SYMB** This is a list of symbols which may function as delimiters, e.g. ‘(’, ‘)’, ‘[’, ‘]’, ‘,’, ‘，’, ‘。’, and etc.
- **PREF** This list defines the phrases which are considered as prefix indicators, like ‘又称’, ‘别名’, ‘还叫’, etc. Note that these phrases themselves cannot be one part of a term.
- **INPT** This is a list of the special Chinese characters or words whose appearance in a phrase indicates a high probability that the phrase is a term. For example, ‘-性-’, ‘-化-’, ‘-式-’, ‘-特发-’, and etc. are included in the list.

Twenty-three regular expression rules have been constructed based on these four sets of characters. These rules can be grouped into two levels: the first level rules, using maximum length matching strategy, are employed to extract the terms whose penultimate symbols are in SUFF or which are followed by the symbols in the SYMB list. The second level rules, based on the result of the first level rules, determine the start points of candidate terms.

4.3 C-value Filtering

C-value (Frantzi and Ananiadou, 2000) is a simple but effective tool to extract terms, especially cascaded terms, from free texts. We use C-value to discover such cascaded terms and also to filter high-scoring candidate terms. The C-value algorithm requires syntactic features. However, in this study, we do not apply any POS tagging to the results of linguistic filtering. The reasons are: 1) POS taggers trained on Chinese biomedical corpora are not currently available, and taggers trained on newswire are likely to introduce errors and thus affect the performance of the tool. 2) The output of linguistic filtering consists of short phrases, most of which have already been identified as terms or parts of terms. Therefore, in the current work, each candidate term resulting from the linguistic filtering step is assigned the noun phrase POS tag. The maximum number of terms selected from the list is 20. After implementing the C-value processing with the above parameters, the members of the resulting list are considered as the synonyms of the original CMeSH terms.

4.4 Term Weight Calculation

In this study, only Chinese term weights are calculated. This is because the purpose of the extended CMeSH tree is to provide an enhanced set of Chinese terms to improve a Chinese-English CLIR application. Queries are translated or/and expanded using CMeSH terms, and Chinese term weights are directly passed to the translated English query terms. The original weights of English terms, if assigned, were not used in this study.

The weight formula used is a variant of the one proposed by Lynam et al. (2001).

$$w_{ct} = \begin{cases} w' + 1.0, & \text{if } f_{ct} > f_{et} > 0 \\ w', & \text{otherwise} \end{cases}$$

$$w' = Exp\left(-Exp\left(-\frac{\log_{10}\left(\left(f_{ct}+0.5\right)/\left(f_{et}+0.5\right)\right)}{2}\right)\right)$$

where f_{ct} is the frequency of Chinese translation and f_{et} is the frequency of English term. Both frequencies correspond to the number of occurrences returned by Google. The weight of the Chinese translation can be computed by the sigmoid function *w'*. If the frequency of the Chinese translation is greater than that of English term (which means that Chinese translation is more popular than the English equivalents), then we increase its weight.

4.5 The Final CMeSH Resource

After merging the weight values with Chinese and English terms, the final CMeSH Tree has the follwing representation.

Dementia;C10.228.140.380 痴呆:0.343881162222 痴呆症:1.425371472485 失智:0.314097771253
AIDS Dementia Complex;C10.228.140.380.070 艾滋病痴呆综合征:0.099850335887 AIDS 痴呆综合征:0.050615806911 AIDS 痴呆症候群:0.018721486519 艾滋病痴呆复合征:0.080638707889 艾滋病痴呆复合症:0.00000853004 爱滋病痴呆复合症:0.00000461272
Alzheimer Disease;C10.228.140.380.100 阿尔茨海默病:0.097398853027 Alzheimer 病:0.04592354867 阿滋海默症:0.18626155397 老年性痴呆:0.244575613782 Alzheimer 氏病:0.073412383701 早老性痴呆:0.074511778 Alzheimer 氏症:0.041794385
Aphasia, Primary Progressive;C10.228.140.380.132 失语, 原发性进行性:0.000000211796 原发性进行性失语:0.317609856764 原发性进行性失语症:0.208916619538
Creutzfeldt-Jakob Syndrome;C10.228.140.380.165 克-亚综合征:0.014768214 Creutzfeldt-Jakob 病:0.324075159688 Creutzfeldt-Jakob 综合征:0.001337330099 早老痴呆症:0.264388092697 克-雅氏综合征:0.005840557652 克-雅氏病:0.119031301389 克雅氏病:0.119031301389 库贾氏病:0.203074988059 牛海绵状脑病:1.398253302721 疯牛病:1.431304403242 皮质-纹体-脊髓变性:0.006199180668 克鲁兹弗得-雅柯病:0.002491919204 库雅氏症:0.029140724567
Dementia, Vascular;C10.228.140.380.230 痴呆, 血管性:0.226335681686 血管性痴呆:0.36594277919 血管梗塞型痴呆症:0.000230068283 血管型失智症:0.081200049502
CADASIL;C10.228.140.380.230.124 大脑常染色体显性动脉病合并皮层下梗塞及脑白质病:0.00000043303 常染色体显性遗传病合并皮质下梗死和白质脑病:0.00000043303 CADASIL 病:0.02736615094 遗传性多发梗死痴呆病:0.003007920526 伴皮层下梗死和白质脑病的常染色体显性遗传性脑动脉病:0.014594966461 伴皮质下梗死和白质脑病的常染色体显性遗传性脑动脉病:0.01608595828 显性遗传脑动脉病伴皮层下梗死及脑白质病:0.00005469028

The extended CMeSH enriches the original resource with both synonyms and term weights. Moreover, Chinese terms and their synonyms are mapped to English MeSH terms with their tree number.

5. Evaluation

A CLIR system has been employed to evaluate the scope of vocabulary in the extended CMeSH Tree. The difference between cross-lingual information retrieval and mono-lingual information retrieval is that CLIR requires a stage in which either the queries are translated from the source language into the target language (in which the document set is written), or else the document set is translated into the language in which the queries are expressed. In this study, we translate Chinese queries into English and carry out information retrieval on English documents. The CMeSH Tree is used to translate or/and expand the Chinese query terms.

5.1 Experimental description

The toolkit used for constructing the CLIR system is Lemur (http://www.lemurproject.org/). The document collections are the TREC Genomics data from 2006 and 2007 (Hersh et al., 2006, 2007), which contain a total of 162,259 biomedical papers (11.9 GB). The indexing and retrieving algorithm is Okapi BM25; the parameters used for Okapi BM25 are the system default values. All documents are indexed for document level retrieval. Indexing of the TREC Genomics documents does not involve stemming. Stopwords are removed using the PubMed stop list (http://www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book=helppubmed&part=pubmedhelp&rendertype=table&id=pubmedhelp.T43).

TREC Genomics 2006 and 2007 tasks provide 64 sentences as queries and a gold standard relevance judgement, i.e., each query has been associated with a set of relevant documents in TREC Genomics data set by human experts, except Query 173 and Query 180. There are also scripts and utilities to calculate retrieval performance. In order to obtain the Chinese queries, we make use of the same strategy reported in Levow et al. (2004), in which the original queries are manually translated into other languages. Here, we manually translate TREC Genomics queries from English into Chinese. For instance, Query 161 in the 2006 task is:

‘**What is the role of IDE in Alzheimer’s disease**’.

The corresponding Chinese translation is as follows:

‘在阿尔茨海默病中 IDE 的作用是什么’.

The Chinese query sentences need to be segmented before translating into them into English queries. As our work is intended to evaluate the effectiveness of CMeSH Tree terms rather than the performance of the information retrieval, and also that there is no high-performance POS tagger for the Chinese biomedical domain, we manually segment the Chinese queries into word sequence. This process ensures that errors are not introduced by automatic word segmentation. In the above example, the segmentation result is as follows:

‘在 阿尔茨海默病 中 IDE 的 作用 是 什么’.

We then remove words from the query whose grammatical categories do not correspond to one of the following: nouns or noun phrases, verbs (except link verbs and auxiliary verbs), and adjectives. Words without Chinese characters, like ‘IDE’, are also retained in the query. By carrying out a preliminary experiment, (i.e. performing mono-lingual information retrieval on TREC genomics data to compare the performances of two word selection policies: all words and selected words), we found that the above categories of words play a more important role on retrieving relevant documents than prepositions, adverbs, postpositions, interrogatives, etc. Following this filtering strategy, the above-mentioned example now contains the following words:

‘阿尔茨海默病 IDE 作用’.

All the following experiments were carried out on queries which were segmented and filtered according to the above description.

The segmented and filtered Chinese queries are translated into English queries using either a domain dictionary or CMeSH Tree, depending on the experiment being undertaken (see next section). English queries are represented as *indri queries* using ‘Indri query language’ (Strohman, 2005), which is based on ‘Inquery query language’. Finally, Lemur’s Indri search engine retrieves relevant documents and computes the performance parameters, such as mean average precision (MAP) and average precision (AP).

5.2 Experiments

CMeSH Tree terms were applied to translate Chinese terms into English equivalents. The quality of CMeSH Tree is thus reflected in the performance of CLIR. To fully evaluate the extended CMeSH tree, we designed four experiments. The baseline experiment makes use of a domain dictionary to translate queries. The other three experiments are aimed to evaluating a) the CMeSH Tree terms themselves, b) term weights within the CMeSH Tree, and c) the CMeSH Tree hierarchy.

1. Baseline

For the baseline experiment, we employed a free domain dictionary, ‘谷歌金山词霸 2.0’ (Google and Kingsoft Dictionary 2.0) (http://g.iciba.com/), to translate Chinese query terms into English counterparts. The policy for out-of-vocabulary (OOV) is to ignore all unknown words.

2. CMeSH Tree term translation

For this experiment, the extended CMeSH Tree terms were used to translate Chinese queries. Terms or words which were not present in the CMeSH Tree were ignored during translation. In the following discussion, this experiment is referred to as ‘term_t’.

3. CMeSH Tree term translation with weights

This experiment, subsequently referred to as ‘term_w’, had the aim of evaluating the effectiveness of our term weighting algorithm. Whenever a Chinese term was found in the CMeSH Tree, the weight of that Chinese term was passed to the English translation. The Indri query consists of these translations and their weights. Here, we inherit the OOV processing policy used in the ‘term_t’ experiment.

4. CMeSH Tree terms translation with hierarchy expansion

In this experiment, we expanded Chinese queries according to the hierarchical structure of the CMeSH Tree. Spasić and Ananiadou (2005) refer to an algorithm to compute the tree similarity (TS).

$$ts(C_1, C_2) = \frac{2 \cdot common(C_1, C_2)}{depth(C_1) + depth(C_2)}$$

where C_1 and C_2 are the classes related to Term 1 and Term 2 respectively, $common(C_1, C_2)$ denotes the number of common classes in the paths leading from the root to the given classes, and $depth(C)$ is the number of classes in the path connecting the root and the given class. $common(C_1, C_2)$ is subject to the following conditions in this study: Given that C_1 and C_2 denote classes of Term 1 and Term 2 respectively, if the ‘common’ function value is the depth of C_2, then Term 2 is the parent node of Term 1; the second condition indicates that Term 2 is the sibling of Term 1.

$$common(C_1, C_2) = \begin{cases} depth(C_2) \\ depth(C_1) - 1, \ where\ depth(C_1) = depth(C_2) \end{cases}$$

Using this constraint, the TS algorithm expands a Chinese query term only with its siblings and parent in the CMeSH Tree.

After expanding the original Chinese query terms, the CMeSH term list is used to translate the expanded queries into English. For this experiment, term weights are ignored, because it is intended to evaluate the ontology hierarchy. We name this experiment ‘term_h’.

5.2 Results and Discussion

Table 1 illustrates the Mean Average Precision (MAP) of each experiment. This is the mean value of all queries’ average precision.

	baseline	term_t	term_w	term_h
2006	0.2622	0.2857	**0.3014**	0.2706
2007	0.1735	0.1813	**0.1899**	0.1712

Table 1: MAPs for the four experiments

According to the results of experiment term_t, we can conclude that the extended CMeSH has a positive effect on IR results. Compared with baseline, the MAPs have been increased by 2.53% (for the 2006 task) and 0.78% (for the 2007 task) with the help of extended CMeSH terms. The experiment term_w proves that the term weighting algorithm can improve the performance of Chinese-English CLIR greatly – from baseline’s results of 26.22% (2006 task) and 17.35% (2007 task) to 30.14% (2006 task) and 18.99% (2007 task) respectively. The reasons for these improvements are 1). Our CMeSH extension provides more Chinese terms than the dictionary; 2). Our term weighting algorithm succeeds in assigning terms with reasonable weight value.

The results of the term_h experiment are not as expected. For the 2006 task, the MAP is only slightly better than that of baseline experiment, whilst for the 2007 task, it is the worst result of all four experiments. The reason for this bad performance is that our simple query expansion technique introduces too many terms into queries, which reduces the precision of the search engine.

Figure 1 shows the Average Precision (AP) for each query in all four experiments.

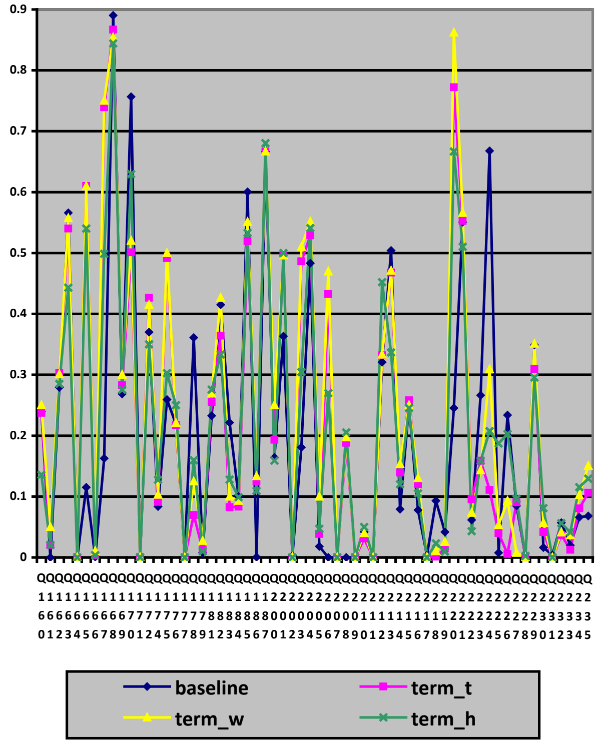


Figure 1: AP of each query in four experiments

Figure 1 provides another perspective on the CMeSH Tree extension. It illustrates how well the CMeSH Tree extension performs on each query. These results are consistent with those presented in Table 1. For example, in most cases, the curve of experiment term_w is higher than other curves, which indicates that CMeSH Tree term translation with term weight is the best result among experiments. This chart also gives the details of performances of each query. For instance, Query 224 obtains the best AP in baseline experiment; CMeSH terms highly decrease its AP by 55.67% (for term_t), 35.87% (for term_w), and 46.02% (for term_h), compared with the result of baseline. Meanwhile, Query 220 is greatly improved by CMeSH term translation with term weight, but the term_h strategy reduces its performance.

6. Conclusions

In this paper, we have proposed a task-oriented extension of the Chinese MeSH Tree. We have employed the Google search engine to collect Chinese synonyms and calculate weights for them. We have evaluated our extension to the tree using a Chinese-English cross-lingual information retrieval system in the biomedical domain. We have analysed the scope and effectiveness of the extended CMeSH Tree terms and their weights. The results of our experiments illustrate that the extended CMeSH Tree significantly improves the performance of the CLIR application. The enhanced performance of the CLIR serves to demonstrate the quality of the extended CMeSH Tree. It is intended that this new linguistic resource will also help others in future research.

Future work will include the following:

1. Ensuring that all terms in the original English MeSH tree have corresponding Chinese translations (a small number are still missing);
2. Computing the weights of English as well as Chinese terms;
3. Finding English synonyms of English MeSH terms;
4. Evaluating the performance of the extended CMeSH Tree in other NLP applications.

7. Acknowledgements

We sincerely thank Mr. Paul Thompson of the National Centre for Text Mining for his help with proof-reading and comments regarding this paper. We also thank the two anonymous reviewers for their helpful comments.

8. References

Abdou, Samir, and Savoy, Jacques, (2007). Searching in MEDLINE: Query expansion and manual indexing evaluation. *Information Processing and Management.* Volume 44, Issue 2, 2008, pp. 781--789.

Brewsster, C., Alani, H., Dasmahapatra, S., and Wilks, Y., (2004). Data driven ontology evaluation. In *Proceedings of International Conference on Langauge Resources and Evaluation, Lisbon*, pp. 24--30.

Cooper, Gregory F., and Miller, Randolph F., (1998). An Experiment comparing Lexical and Statistical Methods for Extracting MeSH Terms from Clinic Free Text. *Journal of the American Medical Informatics Association.* Volume 5, Issue 1, 1998, pp. 62--75.

Elkin, Peter L.; Cimino, James J.; Lowe, Henry J.; Aronow, David B.; Payne, Tom H.; Pincetl, Pierre S.; and Barnett, G. Octo, (1988). Mapping to MeSH: The Art of Trapping MeSH Equivalence from within Narrative Text. In *Proceedings of the Twelfth Annual Symposium on Computer Applications in Medical Care*, IEEE Comput Soc Press (1988), pp. 185--190.

Frantzi, Katerina, Ananiadou, Sophia, and Mima, Hideki, (2000). Automatic recognition of multi-word terms: the C-value/NC-value method. *International Journal of Digital Library*, 3(2), pp. 117--132.

Guo, Y., Harkema, H., and Gaizauskas, R., (2004). Sheffield University and the TREC 2004 genomics track: Query expansion using synonymous terms. In *Proceedings of the Thirteenth Text REtrieval*

Conference. Gaithersburg, MD: Department of Commerce, National Institute of Standards and Technology, pp. 16--19.

Hersh, W., Cohen, A. M., Roberts, P., and Rekapalli, H. K. (2006), TREC 2006 genomics track, In *Proceedings of the Fifteenth Text REtrieval Conference*, Gaithersburg, MD: Department of Commerce, National Institute of Standards and Technology.

Hersh, W., Cohen, A. M., Roberts, P., and Rekapalli, H. K. (2007), TREC 2007 genomics track overview, In *Proceedings of the Sixteenth Text REtrieval Conference*, Gaithersburg, MD: Department of Commerce, National Institute of Standards and Technology.

Kumar, A., and Smith, B., (2003). The Unified Medical Language System and the Gene Ontology: Some Critical Reflections. In *KI2003: Advances in AI (2003)*, pp. 135--148.

Levow, Gina-Anne, Oard, Douglas W., and Resnik, Philip, (2004). Dictionary-Base Techniques for Cross-Language Information Retrieval, *Information Processing & Management* 41(3), pp. 523--547.

Li, Danya, Hu, Tiejun, Zhu, Wenyan, Qian, Qing, Ren, Huiling, Li, Junlian, and Yang, Bin, (2001). Retrieval System for the Chinese Medical Subject Headings (in Chinese). *Chinese Journal of Medical Library*, Issue 4, 2001.

Lozano-Tello, A., and Gómez-Pérez, A., (2004). Ontometric: A method to choose the appropriate ontology. *Journal of Database Management*, 15(2), pp. 1--18.

Lowe, H. J., and Barnett, G., O., (1994). Understanding and using the medical subject headings (MeSH) vocabulary to perform literature searches. *Journal of the American Medical Association*, 271 (14): pp. 1103--1108.

Lu, Zhiyong, Kim, Won, and Wilbur, W. John, (2008). Evaluation of query expansion using MeSH in PubMed, *Information Retrieval*, Volume 12, Issue 1, 2009, pp. 69--80.

Lynam, T. R., Clarke, C. L. A., and Cormack, G. V., (2001). Information Extraction with Term Frequencies. In *Proceedings of the First International Conference on Human Language Technology Research*, pp. 1--4.

Maeche, A., and Staab, S., (2002). Measuring similarity between ontologies. In *Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management. Ontologies and the Semantinc Web*, pp. 251--263.

Porzel, R., and Malaka, R., (2004). A task-based approach for ontology evaluation. In *Proceedings of ECAI 2004 Workshop on Ontology Learning and Population*.

Qin, Yayun, and Feng, Qichang, (1999). 中文医学主题词表(机读版)在文献标引中的应用(in Chinese), *Journal of Medical Intelligence*, Issue 5, 1999.

Smith, B., (2006). From concepts to clinical reality: an essay on the benchmarking of biomedical terminologies, *Journal of Biomedical Informatics*, 39(3), pp. 288--298.

Shultz, M., (2006). Mapping of medical acronyms and initialisms to medical subject headings (mesh) across selected systems. *Journal of the Medical Library Association*, Volume 94, Issue 4, pp. 410--414.

Spasić, I. and Ananiadou, S., (2005). A Flexible measure of contextual similarity for biomedical terms. *Pacific Symposium on Biocomputing 10*, pp. 197--208.

Strohman, T., Metzler, D., Turtle, H., and Croft, W. B., (2005). Indri: A language-model based search engine for complex queries. In *proceedings of International Conference on Intelligence Analysis*, May 2-6, extended paper.

Zhou, Xuezhong, Liu, Baoyan, Wu, Zhaohui, and Feng, Yi, (2007). Integrative mining of traditional Chinese medicine literature and MEDLINE for functional gene networks. *Artificial Intelligence in Medicine*, 41(2), pp. 87--104.

Structuring of Status Descriptions in Hospital Patient Records

Svetla Boytcheva[1], Ivelina Nikolova[2], Elena Paskaleva[2], Galia Angelova[2],
Dimitar Tcharaktchiev[3] and Nadya Dimitrova[4]

[1]State University of Library Studies and Information Technologies, Sofia, Bulgaria
[2]Institute for Parallel Processing, Bulgarian Academy of Sciences, Sofia, Bulgaria
[3]University Specialized Hospital for Active Treatment of Endocrinology, Medical University, Sofia, Bulgaria
[4]National Oncological Hospital, Sofia, Bulgaria
E-mail: svetla.boytcheva@gmail.com, {iva, hellen, galia}@lml.bas.bg, dimitardt@gmail.com,
dimitrova.nadia@gmail.com

Abstract

In this article we present a text analysis system designed to extract key information from clinical text in Bulgarian language. The extracted and structured features will be further employed for classification of patient cases, effective information retrieval and further processing in different medical tasks. Using shallow analysis within an Information Extraction (IE) approach, the system builds structured descriptions of patient status and complications. We discuss some particularities of the medical language of Bulgarian patient records, functionality of our current prototype, and evaluation results regarding the IE tasks we tackle at present.

1. Introduction

Patient Records (PRs) contain much textual information. When studying hospital PRs in Bulgaria, which discuss a single hospital episode, one discovers that the most important findings, opinions, summaries and recommendations are stated as free text while clinical data usually supports the textual statements or provides clarification of particular facts. Thus the essence of patient-related information is communicated as unstructured text message of one medical expert with addressee another medical expert. In addition, the clinical documents present only partial information about the patients, so some kind of aggregation is needed to provide a complex view to the patient health status. Automatic analysis of biomedical text is a complex task which requires various linguistic and conceptual resources (Spasic et al., 2005). Despite the difficulties and challenges, however, there are industrial systems and research prototypes in many natural languages, which aim at the information extraction of features from patient-related texts. So the application of language technologies to medical PRs is viewed as an advanced but standard task which is a must in health informatics.

In this paper we present an IE prototype which extracts important facts from hospital PRs of patients diagnosed with different types of diabetes. The extracted and structured features will be employed for classification of patient cases, effective information retrieval and further processing in different medical tasks. The system, called EVTIMA, is under development in a running project for medical text processing in Bulgarian language. It should be classified as an ontology-driven IE system, following the classification in (Spasic at al., 2005). The article is structured as follows: section 2 briefly overviews related approaches; section 3 presents the specific settings of the IE and the medical texts characteristics, section 4 describes the prototype and its functionality; section 5 deals with the evaluation and section 6 contains the conclusion.

2. Related Work

When designing our system, its rules for shallow analysis and the training corpus, we have studied carefully the CLEF site (CLEF 2008). Other systems which process patient symptoms and diagnosis treatment data are: the Medical Language Extraction and Encoding System which was designed for radiology reports and later extended to other domains such as discharge summaries (Friedman, 1997); (caTIES, 2006) which processes surgical pathology reports; the open-source NLP system Health Information Text Extraction (HITEx, 2006), and the Clinical Text Analysis and Knowledge extraction system cTAKES (Savova et al., 2008). Interesting and useful ideas about processing of medical terminology and derivation of terminological variants are given in (Valderrábanos et al., 2002). Negative statements in Bulgarian patient-related texts are studied in (Boytcheva et al., 2005).

3. IE in the EVTIMA System

The system presented here deals with anonymised PRs supported by the Hospital Information System (HIS) of the University Specialised Hospital for Active Treatment of Endocrinology “Acad. I. Penchev” at Medical University, Sofia. The current HIS facilitates PR structuring since the diagnosis, encoded in ICD-10 (the International Classification of Diseases v. 10), is selected via menu. The drugs prescribed to the patient, which treat the illness causing the particular hospital stay, are also supported via the so-called Computerised Provider Order Entry. In this way some information is structured and easy to find.

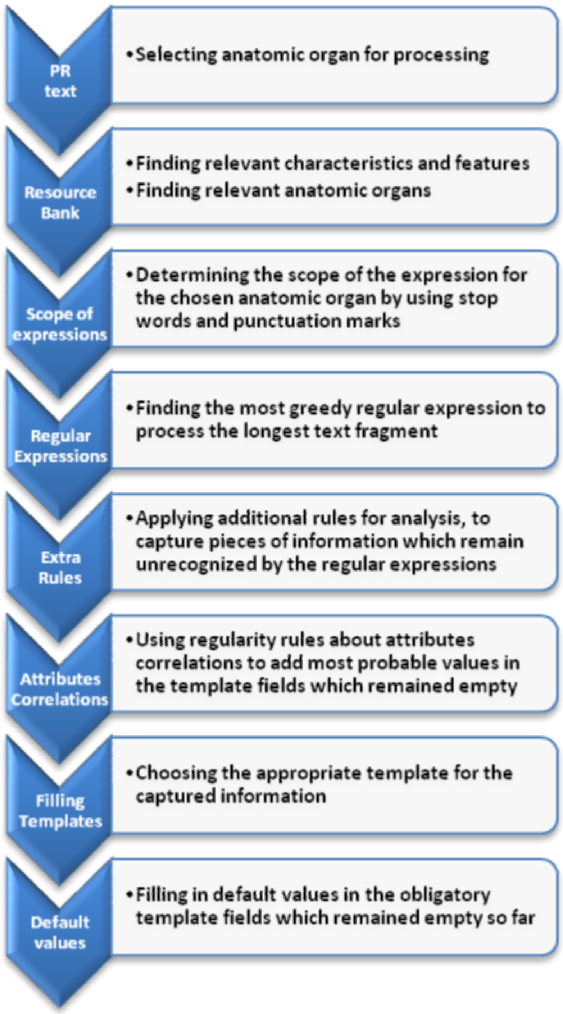

PR text
•Selecting anatomic organ for processing
Resource Bank
•Finding relevant characteristics and features
•Finding relevant anatomic organs
Scope of expressions
•Determining the scope of the expression for the chosen anatomic organ by using stop words and punctuation marks
Regular Expressions
•Finding the most greedy regular expression to process the longest text fragment
Extra Rules
•Applying additional rules for analysis, to capture pieces of information which remain unrecognized by the regular expressions
Attributes Correlations
•Using regularity rules about attributes correlations to add most probable values in the template fields which remained empty
Filling Templates
•Choosing the appropriate template for the captured information
Default values
•Filling in default values in the obligatory template fields which remained empty so far

Figure 1: Major IE steps for PR processing

However, in the PR discussion of case history, the previous diseases and their treatments are described as unstructured text only.

In addition, in the hospital archive we find PRs as separated text files, and these PRs consist of free text. Therefore our prototype needs to recognise the ICD terms, drug names, patient age and sex, family risk factors and so on. To do so we perform the steps shown on Figure 1.

3.1 Medical Language in Hospital PRs

The length of PR texts in Bulgarian hospitals is usually 2-3 pages. The document is organised in the following sections: (*i*) personal details; (*ii*) diagnoses of the leading and accompanying diseases; (*iii*) anamnesis (personal medical history), including current complains, past diseases, family medical history, allergies, risk factors; (*iv*) patient status, including results from physical examination; (*v*) laboratory and other tests findings; (*vi*) progress notes; (*vii*) discussion; (*viii*) treatment; (*ix*) recommendations. Despite the clear PR fragmentation, there are various problems for automatic text processing.

Bulgarian medical texts contain a specific mixture of terminology in Latin (about 1%) and Cyrillic letters, Latin terms transcribed with Cyrillic letters. The terms occur in the text with a variety of wordforms which is typical for the highly-inflectional Bulgarian language. The major part of the text consists of short declarative sentences and sentence phrases without agreement, often without proper punctuation marks. There are various kinds of typos in the original text which cannot be properly dealt with within our research project. In the evaluation we consider only normalised texts with standard abbreviations and without spelling errors, because we aim at a research study. Our present experimental corpus is formed of 1150 PRs with some 6400 words, about 2000 of them being medical terms.

Another PR text particularity is that descriptions are often missing. It is common for the Bulgarian clinicians to not describe in the PR the organs where no pathological deviations have been found. As reported in (Boytcheva et al., 2009), only 86% of the PRs in our corpus discuss explicitly the patient status regarding skin colour, 63% - fat tissue, about 42% - skin turgor and elasticity, and 13% - skin hydration. In order to fill in the maximum number of slots in our IE template, default values have to be assigned. The presence of these values will be important in a following case classification.

The PRs contain also a lot of numerical values of analyses and clinical test data (about 16%) which are subject of additional studies.

3.2 Available Resources

A lexicon of 30 000 Bulgarian lexemes, which is part of a large general-purpose lexical database with 70 000 lexemes, serves as basic dictionary for morphological analysis of Bulgarian medical text. The International Classification of Diseases (ICD-10) contains 10 970 terms. The Bulgarian version of ICD-10 has no clinical extension, i.e. some medical terms had to be extracted from the PR corpus. So far we have extended the basic lexicon by medical terminology containing 5 288 terms. Constructing the dictionaries is a continuous process; the Latin terms have to be taken into consideration as well.

Within our project, biomedical NLP for Bulgarian texts starts from scratch due to the lack of language resources as well as ontological resources and tools. For instance, no Named Entity Recognition module has been implemented for Bulgarian entities in the medical domain; the syntactic rules for partial analysis are constructed for the first time; there are no conceptual resources labelled by Bulgarian medical vocabulary – except ICD-9 and ICD-10. It is curious to note that the list of drugs and medications is supported with Latin names by the Bulgarian Drug Agency, even for drugs produced in Bulgaria (BDA 2010), but in the PR texts the medications are predominantly referred to in Bulgarian language, so the drug-related vocabulary is also compiled on the fly in two languages.

4. Present Functionality

At present our system performs the following tasks: text segmentation and normalisation; and text analysis. Here we discuss the latter one.

In order to extract the status of some patient organ the IE system has to find in the PR text a term referring to some anatomic organ, which is important for diabetes, e.g. *thyroid gland, skin, limbs/legs, eyes, neck* etc. For this purpose the system first automatically segments the text into several semantic zones as listed above (See section 3.3). The following text analysis is applied separately on the epicrisises' parts in order to extract structured knowledge from the free text. Up to now the anamnesis and the patient status areas of the epicrisis are being processed. From the anamnesis we extract features like age, sex, sickness name, type and duration and from the patient status - the organs' condition. Of major interest are the limbs, skin and thyroid gland, because many of the diabetes complications are especially related to aggravations in these organs.

The status section contains mostly short declarative sentences in present tense which describe several organs; in the current corpus we find most often 20-30 different anatomic organs and status conditions. So it is important to identify the description boundaries for each particular organ. The scope is determined by using domain knowledge concerning anatomic organs and their parts, which shows where an organ description ends and another one begins. Domain knowledge helps significantly to design text analysis rules.

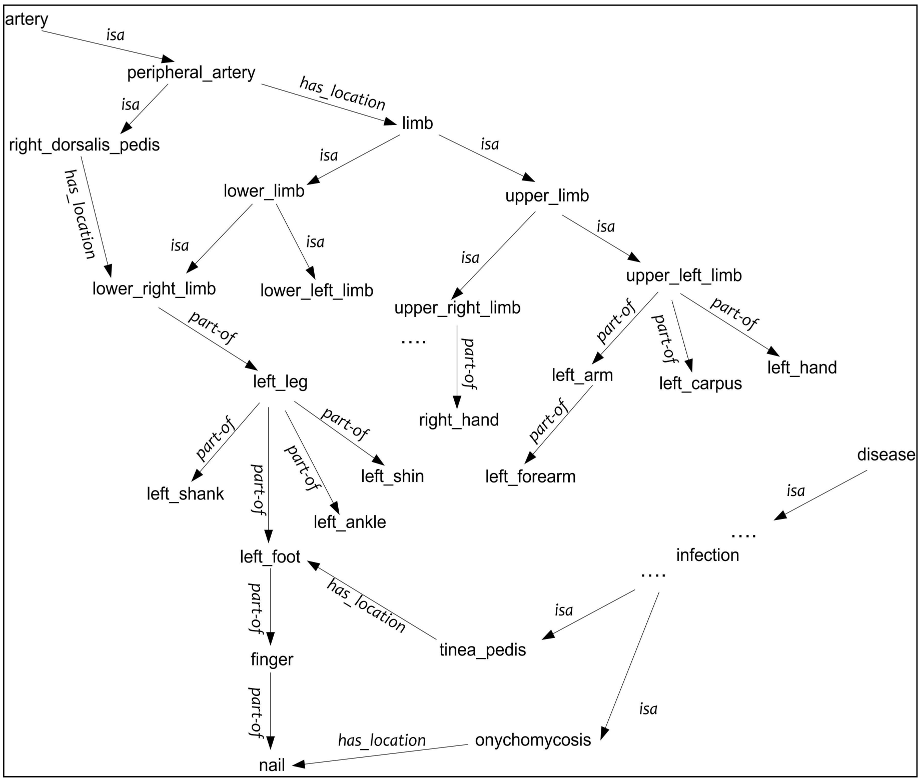


Figure 2. Fragment of conceptual model of anatomic organs, their parts, and some diseases.

Figure 2 shows an excerpt of semantic network which reflects important relations between organs: *isa, part-of* and *has-location*. Knowledge about limbs and their parts helps to scope the description of limb status. Let us consider the following example:

> *Крайници – отслабени пулсации на a. dorsalis pedis двустранно. Претибиални и перималеоларни отоци. Онихомикоза, tinea pedis. Сукусиореналис – (-) отр. двустранно.*
>
> *Limbs – reduced dorsal pedal pulse on both feet. Pretibial and perimaleolar edema. Onychomycosis, tinea pedis. Succusio renalis – bilateral negative (-).*

Here the IE system finds the term *limbs* in the first sentence and runs the IE process in order to extract *limbs'* status. The second sentence contains adjectives, which are attributes related to *limbs* - "*претибиални*" (*leg*) and "*перималеоларни*" (*ankle*). The third sentence contains the terms "*онихомикоза*" (*onychomycosis*) - fungal infection of the nails and "*tinea pedis*", denoting fungal foot infection. Both entities denote diseases of lower legs. Mapping these terms to the concepts and relations at Fig. 2, the IE system considers sentences 1-3 as a continuous status description of *limbs*. The fourth sentence contains "*succusio renalis*", which refers to kidney percussion, and this is a signal that the limbs description is completed. Moreover, "*succusio renalis*" occurs in the next sentence, as we know empirically that new organ descriptions start in another sentence despite the fact that all statements are mixed into one paragraph. Another example is:

> *Крайници - без отоци, варикозни промени. Запазени пулсации на периферните артерии, запазени повърхностна, термо и вибрационна чувствителност. Затруднена и болезнена походка, използва помощни средства."*
>
> *Limbs – without oedema, varicose changes. Palpable peripheral arteries pulse, preserved tactile, thermo and vibratory sensation. Walks with difficulty, algetic gait, uses assistive devices".*

The status of the peripheral arteries is important for patients with diabetes (which we know due to domain communication knowledge provided by the medical experts), so the second sentence is considered as continuation of limbs discussion. Here the occurrence of the word "*gait*" in the third sentence signals the completion of the limbs description, which in this case is split into the first and second sentence. Another aspect of domain communication knowledge is the default that under *limbs* the medical experts usually understand *lower limbs*; parts of upper limbs should be mentioned explicitly to denote arms, forearms etc.

About 96% of all PRs in our training corpus contain organ descriptions in this format (Boytcheva et. al., 2009). In this way the shallow analysis by cascades of regular expressions, which is proposed in (Boytcheva et. al., 2009), proves to be successful approach for structuring the statements concerning patient status. For each particular organ, there is a predefined template, where text units are stored as conceptual entities during domain interpretation phase.

In our corpus there are four manners to present the organs and their features:

1. General – by giving some default value, e.g. "*без патологични промени, без особености*" (*without pathological changes, without specifics*), "*със запазена/нормална характеристика*" (*with preserved/present/normal characteristics*) etc. Figure 3 presents the obligatory IE template fields for limbs. If the PR contains only general statements, the template can store the obligatory fields for the four limbs;

2.Explicit – the PR text contains particular specific values. The characteristic name might be missing since the attribute is sufficient to recognise the feature: e.g. "*preserved peripheral pulsations*" instead of "*preserved pulsations of the peripheral arteries*". The attributes are described by a variety of expressions, e.g. for the "*volume of the thyroid gland*" the value "*normal*" can be represented as "*not enlarged, not palpable enlarged, not palpable*". These explicit statements are processed at the fourth and fifth steps of the algorithm. Depending on the depth of the body parts ontology with related anatomic organs (AO) for the processed **AO**, and according to the details encountered in the PR text, the dynamically generated templates can consists of different levels of nested fields (Figure 4 and Figure 5);

3. Partial – The text contains descriptions about organ parts, not about the main **AO**. For instance, the limbs status can be expressed like e.g. "*atrophic changes of the legs skin with pretibial oedema*". Then the IE system adds additional characteristics to the template at steps four-six of the algorithm shown in Figure 1, with final refinement at step eight when the default values are filled in.

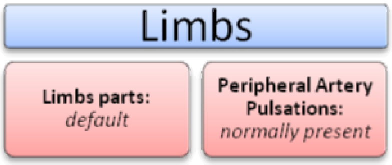


Figure 3: General Template for Limbs

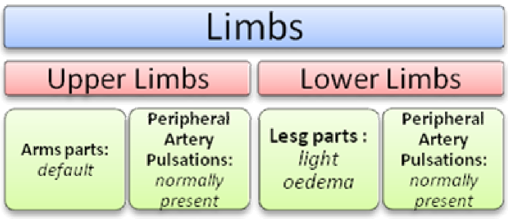

Limbs
Upper Limbs
Lower Limbs
Arms parts: *default*
Peripheral Artery Pulsations: *normally present*
Lesg parts : *light oedema*
Peripheral Artery Pulsations: *normally present*

Figure 4: Template with specific fields about lower limbs

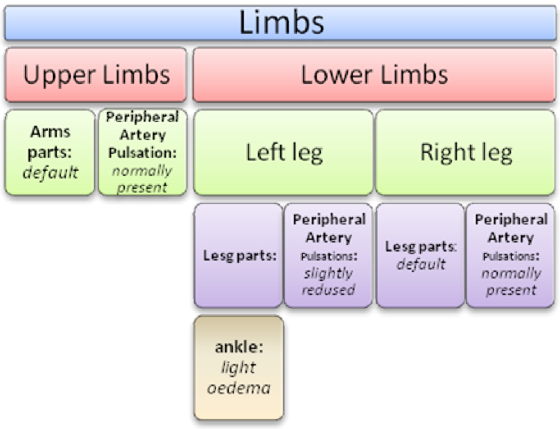

Limbs
Upper Limbs
Lower Limbs
Arms parts: *default*
Peripheral Artery Pulsation: *normally present*
Left leg
Right leg
Lesg parts:
Peripheral Artery Pulsations: *slightly redused*
Lesg parts: *default*
Peripheral Artery Pulsations: *normally present*
ankle: *light oedema*

Figure 5: Template with fields about the lower left limb

Figure 4 illustrates the case when the respective PR contains specific information concerning only the lower limbs - *light swelling*. Then for the upper limbs, which are not explicitly described, the IE system sets the default value. The same decision is made for the field "*Peripheral Artery Pulsations*" which is set to "*preserved*" both for the upper and lower limbs.

The PR behind the template at Figure 5 contains specific information about the "*left ankle*"; then the IE system infers that the missing values should be set to the default limb attribute values;

4. By diagnosis – sometimes a diagnosis is given instead of the organ description, e.g. "*onychomycosis, tinea pedis*". This information is captured at the fourth and fifth steps of the algorithm at Figure 1.

5. Evaluation

In a previous paper we have evaluated the recognition of skin characteristics (Boytcheva et al., 2009). These results are presented here to allow for a more complete assessment of our IE results. The extraction of attributes was evaluated using a corpus of 197 PRs as a training set (set1) and another set of 1500 PRs as a test set (set2). The evaluation is made organ by organ since each **AO** description is analysed independently. There are PRs which do not contain any descriptions relevant to the tested **AO**s but they are removed from the evaluation figures.

Usually the Information Extraction performance is assessed in terms of three measures. The *precision* is calculated as the number of correctly extracted **AO** descriptions, divided by the number of all recognised **AO** descriptions in the test set. The *recall* is calculated as the number of correctly extracted **AO** descriptions, divided by the number of all available **AO** descriptions in the test set (some of them may remain unrecognised by the particular IE module). Thus the precision measures the success and the recall – the recognition ability and "sensitivity" of the algorithms. The *F-measure* (harmonic mean of precision and recall) is defined as

F = 2 x *Precision* x *Recall* / (*Precision* + *Recall*).

Table 1 summarises the precision, recall and the F-measure of correctly extracted descriptions for *sex*, *age*, *disease*, *diabetes type* and the **AO** *skin*, *neck*, *thyroid gland* and *limbs*. For each test the extraction algorithm, shown at Figure 1, uses organ-specific regular expressions.

The cases of incorrect analysis are due to more complex syntactic structures in the PR text which need to be analysed by a syntactic analyser (parser) and deeper approach to sentence analysis.

Since we are using nested rules for capturing features, spelling errors in (or lack of) expressions which are to be matched on an upper level may prevent the further matching of the rules in the nested fields. We have noticed that this was one of the reasons for the comparatively low recall value of the diabetes duration (its recognition rule is nested in the diabetes acronym and diabetes type recognition rules). The recall for the feature *sex* is surprisingly low. This can be explained with the fact that there were too few samples in the anonymised training corpus and when testing on a new dataset the available rules could not capture the relevant features. The main reason is the new author styles encountered in the enlarged corpus. These inconsistencies are covered by continuous update and adjustment of the IE rules for shallow analysis.

Feature	Precision %	Recall %	F-measure %
#1	88.89	90.00	89.44
#2	80.00	50.00	61.54
#3	98.28	96.67	97.47
#4	96.00	83.33	89.22
#5	95.65	73.82	81.33
#6	95.65	88.00	91.67
#7	94.94	90.36	92.59
#8	93.41	85.00	89.01

Table 1: IE precision, recall and f-measure evaluation (#1 - age; #2 - sex; #3 - diagnose; #4 - diabetes duration; #5 - skin; #6 - neck; #7 - thyroid gland; #8 - limbs).

Test set3	Number of PRs with explicit text description	Percentage PRs with explicit characteristics
Ankle	201	99,5%
Leg	201	99,5%
Peripheral artery	201	99,5%
Feet	8	3,98%
Skin	11	5,47%
Nails	15	7,46%
Others	21	10,45%
Gait	3	1,49%

Table 2: Availability of status statements in 201 randomly-selected PRs

Template filled in at:	ankle	leg	peripheral Artery	Feet	skin	nails	Others
Step 4 by regular expressions	150	148	199	6	9	14	18
Step 5 by extra rules	40	40	1	0	0	0	0
Step 6 by correlations	0	0	0	1	1	0	2
Step 8 by defaults	9	11	0	0	0	0	0
Total:	**199**	**199**	**200**	**7**	**10**	**14**	**20**

Table 3: IE performance steps 4-8 for limbs descriptions: number of PRs from where status statements are extracted at each step

We see that the simple regular expressions work relatively well and produce enough input for statistical observations of patient status data. It is also clear that further efforts are required for proper development of the extraction algorithms in order to cover sophisticated language expressions. The automatic recognition of the scope of quantifiers, negation, and temporal qualifications is a challenge to be met in the future.

Regarding the limbs, an additional in-depth evaluation was done on a set of 201 PRs (set3) which were randomly selected from set2. Some PRs in set3 contain no explicit discussions about certain limb characteristics. Table 2 gives the numbers of available or missing descriptions for each attribute; it turns out that the status of ankles, legs and the peripheral arteries is explicitly mentioned in almost all PRs.

The IE algorithm shown at Figure 1 was run on set3 and we have counted manually how the processing at steps 4-8 contributes to the final success rates. The results are summarised at Table 3. It is clear that an essential share of the successful recognitions is achieved at steps four and five of the algorithm. The regular expressions, applied at step 4, recognise the majority of the descriptions. Extra rules are applied at step 5, to extract phrasal units including the negative descriptions.

Some text descriptions of limbs status contain complex statements including embedded sentences and clinical tests; due to this reason the IE recall is relatively low (85%, see Table 1 row 8). The skin feature extraction has also obtained a comparatively low recall, which could be explained by the fact that there is a large variety of lexical expressions describing the values of some skin characteristics, combined with a variety of the authors' expression manners.

In section 4 we have shown that the occurrence of the word "*gait*" signals the end of the limbs description. Please note that according to Table 2, this is helpful for 1,49% of all PRs to be analysed, so many other "hints" need to be acquired and declared in the domain ontology.

6. Conclusion

Our system is the first one which supports medical text mining for Bulgarian. So far we have evaluated its performance on several IE tasks dealing with the anamnesis and patient' status zones of the PRs. It shows the complexity of medical text processing which is due to the complexity of the medical domain and the particularities of the medical texts written in specific, well-established style. The role of explicitly-declared domain knowledge is shown; it supports the information extraction algorithms since by providing constraints and inference mechanisms. At the same time the article illustrates the obstacles to build semantic systems in the medical domain: this requires much effort for construction of the conceptual resources as well as the lexicons and grammatical knowledge in case of text

Despite the difficulties, the paper shows that certain facts can be extracted relatively easy. These promising results support the claim that the Information Extraction approach is helpful for the obtaining of specific medical statements which are described in the PR texts. We plan to develop algorithms for discovering more complex relations and other dependences, which is a target for our future work.

7. Acknowledgements

The research work presented in this paper is supported by grant DO 02-292/December 2008 "Effective search of conceptual information with applications in medical informatics", funded by the Bulgarian National Science Fund in 2009-2011.

8. References

Boytcheva, S., A. Strupchanska, E. Paskaleva, and D. Tcharaktchiev (2005). *Some Aspects of Negation Processing in Electronic Health Records*. In Proc. of the Int. Workshop *Language and Speech Infrastructure for Information Access in the Balkan Countries*, held in conjunction with RANLP-05, Borovets, Bulgaria, pp. 1-8.

Boytcheva, S., I. Nikolova, E. Paskaleva, G. Angelova, D. Tcharaktchiev and N. Dimitrova (2009). Extraction and Exploration of Correlations in Patient Status Data. In: Savova, G., V. Karkaletsis and G. Angelova (Eds). Biomedical Information Extraction, Proc. of the Int. Workshop held in conjunction with RANLP-09, Borovets, Bulgaria, pp. 1-7.

BDA Bulgarian Drug Agency (2010). See the site http://www.bda.bg/index.php?lang=en

caTIES Cancer Text Information Extraction System (2006). See https://cabig.nci.nih.gov/tools/caties.

CLEF Clinical E-Science Framework (2008), University of Sheffield. See http://nlp.shef.ac.uk/clef/.

Friedman C (1997). Towards a comprehensive medical language processing system: methods and issues. Proc. AMIA Annual Fall Symposium, pp. 595-599.

HITEx Health Information Text Extraction (2006). See https://www.i2b2.org/software/projects/hitex/hitex_manual.html.

Savova, G. K., K. Kipper-Schuler, J. D. Buntrock, and Ch. G. Chute (2008). UIMA-based Clinical Information Extraction System. LREC 2008 Workshop W16: Towards enhanced interoperability for large HLT systems: UIMA for NLP.

Spasic, I., S. Ananiadou, J. McNaught, and A. Kumar (2005). Text mining and ontologies in biomedicine: Making sense of raw text. Oxford University Press, Briefings in Bioinformatics 2005, Vol. 6(3), pp. 239-251.

Valderrábanos, A., A. Belskis, and L. I. Moreno (2002). Multilingual Terminology Extraction and Validation. In Proc. LREC 2002 (3rd Int. Conf. on Language Resources and Evaluation), Gran Canaria.

Annotation of all coreference in biomedical text: Guideline selection and adaptation

K. Bretonnel Cohen[1,2], Arrick Lanfranchi[2], William Corvey[2],
William A. Baumgartner Jr.[1], Christophe Roeder[1], Philip V. Ogren[1,3],
Martha Palmer[2], Lawrence E. Hunter[1]

1: Center for Computational Pharmacology
University of Colorado School of Medicine
Aurora, Colorado, USA
2: Department of Linguistics
University of Colorado at Boulder
Boulder, Colorado, USA
3: Department of Computer Science
University of Colorado at Boulder
Boulder, Colorado, USA

Abstract

This paper describes an effort to build a corpus of full-text journal articles in which every co-referring noun phrase is annotated. The identity and appositive relations were marked up. Several annotation schemas were evaluated and are described here; the OntoNotes guidelines were selected. Biomedical journal articles required a number of adaptations to the OntoNotes guidelines—mainly doing away with the notion of generics, which also had implications for the handling of nominal modifiers. Domain experts and linguists were evaluated with respect to their ability to function as annotators, and both were found to be effective. Progress is reported with about one third of the project done; inter-annotator agreement at this stage is 0.684 by the MUC metric.

1. Introduction

The Colorado Richly Annotated Full Text (CRAFT) corpus is a set of 97 full-text journal articles that is currently being annotated in a joint project between the University of Colorado School of Medicine and the Linguistics Department of the University of Colorado at Boulder. The corpus contains about 597,000 words of text and is being annotated for a number of types of linguistic information, including part of speech and treebanking, and semantic information, including a variety of types of named entities. Additionally, we have included some discourse structure annotation in the form of coreference.

This project is unlike any other that we are aware of in that it includes marking *all* coreferential relations of identity and apposition between *all* noun phrases in the full text of a large body of publicly available biomedical publications. Other work (Gasperin, 2006; Gasperin et al., 2007) has done annotation of full biomedical text, but only of noun phrases referring to biological entities. In contrast, we mark up coreference between any and all noun groups in the documents. IIR has done full coreference and appositive annotation of full-text journal articles, but on a smaller set, and has not made them publicly available. The CRAFT corpus is about twice the size of the IIR document set and will be made publicly available. This is the most ambitious project of its kind of which we are aware.

2. Methods and Results

2.1. Selection of annotation guidelines

One of the desiderata of the project was to contribute to the development of a standard for coreference annotation by adopting a pre-existing set of annotation guidelines, rather than developing our own de novo. To this end, a number of publicly available annotation guidelines were evaluated. (We did not consider projects that only tackled pronominal anaphora, to the exclusion of full noun phrase coreference, or projects that tackled clinical data.) We were initially relatively agnostic as to desiderata, other than that we wanted the guidelines to include a full range of coreferential and bridging anaphoric relations, as well as part/whole relationships.

2.1.1. OntoNotes

OntoNotes (Hovy et al., 2006) is a large, multi-center project to create a multi-lingual, multi-genre corpus annotated at a variety of linguistic levels, including coreference (Pradhan et al., 2007). As part of the OntoNotes project, the BBN Corporation prepared a set of annotation guidelines. They are not publicly available. (The version currently available online at the Linguistic Data Consortium is a full numbered version out of date, compared to the version that we used.) However, (Pradhan et al., 2007) gives the flavor of the approach.

The OntoNotes guidelines include events, pronominal and full anaphora and coreferents, and verbs as markables. Predicative nouns are not treated as coreferential. There is a separate relation for appositives. Nominal premodifiers are markables, with some restrictions that we discuss below in the section on domain-specific changes to the guidelines. The guidelines only include the identity and appositive relations, with set membership and part/whole not included, but this turned out not to be relevant to our project, since funding constraints precluded annotating these relations.

2.1.2. Gasperin

Gasperin (Gasperin, 2006; Gasperin et al., 2007) is the only previous attempt that we are aware of to annotate the full text of biomedical journal articles. Her project involved annotating five such articles. The annotation guidelines reflect a nuanced attempt to reflect the semantics of the biological domain. They do this in part by only selecting biomedical entities as markables, but more importantly by defining a domain-relevant set of relations. In addition to coreference, the relations are three types of associative relations: homology; related "biotype" (e.g. a gene and its protein or a gene and a subsequence of that gene); and the set/member relation.

Appositives and predicative nouns are both treated as coreferential. Premodifiers are not specifically addressed in the guidelines.

Predicative nouns are treated as coreferential. Pronouns are excluded completely. Probably the most striking aspect of Gasperin's guidelines is its definition of markables; they are limited to "bio-entities." This was a major mismatch with our goals, which included annotating all coreferential entities.

2.1.3. GENIA project at IIR

The Institute for Infocomm Research in Singapore and the Department of Computer Science spearheaded an annotation project involving the 2,000 abstracts in the GENIA corpus (Yang et al., 2004a; Yang et al., 2004b) as well as 43 full papers. They annotated four relations: identity, appositives, pronouns (considered separately from other identity relations), and relative pronouns. The annotation of markables included a minimal string that would suffice for identification. As in the case of the other projects besides OntoNotes, they annotated only nouns, not verbs; the guidelines allowed for these, but they did not find any while preparing the guidelines. Premodifiers were not considered markables. Predicative nouns were marked when they were definite.

2.1.4. MUC7

The groundbreaking MUC7 guidelines (Hirschman, 1997) have underlaid a number of subsequent coreference annotation projects. The scheme covered only a single relation, identity. The annotation of markables included a minimal string that would suffice for identification of the coreferring element. Noun phrases and pronouns were included as markables. Gerunds were excluded as markables. Appositives and predicate nominals were both marked, as coreferential. Prenominal modifiers were annotated only in the case where they could be linked to something other than another prenominal modifier.

A number of authors have critiqued various aspects of the MUC7 coreference annotation guidelines, e.g. (van Deemter and Kibble, 2001). A strong point of the MUC7 guidelines is that they make an attempt to deal with changing numerical values, as in *The results of this analysis showed that the statistical support for the linkage of the C57BL/6 locus on Chromosome 3 for ANA increased from logarithm of odds (LOD) 5.4 to LOD 6.4*, which other guidelines do not, although even the MUC7 organizers were not happy with their approach to this. We note that a number of other problems pointed out with the MUC7 guidelines, including situations where the referential status of the markable is unclear, are present in all of the other guidelines that we examined as well.

2.2. Domain-specific changes to the guidelines

After reviewing the pre-existing guidelines, senior annotators marked up a sample full-text article, following the OntoNotes guidelines. We found the OntoNotes guidelines to be a good match to our conception of how coreference should be annotated, and noted that they have responded to a number of critiques of earlier guidelines. For example, compared to the MUC-7 guidelines, the treatment of appositives in terms of heads and attributes rather than separate mentions is an improvement in terms of referential status, as is the handling of predicative nouns. The inclusion of verbs and events is a desirable increase in scope. The guidelines are more detailed, as well. They were also attractive from a political point of view—we wanted to adopt a widely accepted set of guidelines, and they seemed like a good candidate for this due to their association with a large project. We adopted them; however, the nature of the biomedical domain required a major adaptation of the guidelines.

2.2.1. Generics

The OntoNotes guidelines make crucial reference to a category of nominal that they refer to as a *generic*. Generics include:

- bare plurals
- indefinite noun phrases
- abstract and underspecified nouns

The status of generics in the annotation guidelines is that they cannot be linked to each other via the IDENTITY relation. They can be linked with subsequent non-generics, but never to each other, so every generic starts a new IDENTITY chain (assuming that it does corefer with subsequent markables).

The notion of a generic is problematic in the biomedical domain. The reason for this is that any referring expression in a biomedical text is or should be a member of some biomedical ontology, be it in the set of Open Biomedical Ontologies, the Unified Medical Language System, or a nascent ontology. As such, it has the status of a named entity. To take an example from BBN, consider the status of *cataract surgery* in the following:

> Allergan Inc. said it received approval to sell the PhacoFlex intraocular lens, the first foldable silicone lens available for **cataract surgery**. The lens' foldability enables it to be inserted in smaller incisions than are now possible for **cataract surgery**.

According to the OntoNotes guidelines, *cataract surgery* is a generic, by virtue of being abstract or underspecified, and therefore the two noun phrases are not linked to each other via the IDENTITY relation. However, *cataract surgery* is a concept within the Unified Medical Language System (Concept Unique Identifier C1705869), where it occurs as part of the SNOMED Clinical Terms. As such, it is a named entity like any other biomedical ontology concept, and should not be considered generic. Indeed, it is easy to find examples of sentences in the biomedical literature in which we would want to extract information about the term *cataract surgery* when it occurs in contexts in which the OntoNotes guidelines would consider it generic:

- *Intravitreal administration of 1.25 mg bevacizumab at the time of cataract surgery was safe and effective in preventing the progression of DR and diabetic maculopathy in patients with cataract and DR.* (PMID 19101420)
- *Acute Endophthalmitis After Cataract Surgery: 250 Consecutive Cases Treated at a Tertiary Referral Center in the Netherlands.* (PMID 20053391)
- *TRO can present shortly after cataract surgery and lead to serious vision threatening complications.* (TRO is thyroid-related orbitopathy; PMID 19929665).

In these examples, we might want to extract an IS_ASSOCIATED_WITH relation between <bevacizumab, cataract surgery>, <acute endophthalmitis, cataract surgery>, and <thyroid-related orbitopathy, cataract surgery>. This makes it important to be able to resolve coreference with them.

Thus, our project's guidelines do not consider there to be generics as such in the genre and domain that it is concerned with[1].

2.2.2. Prenominal modifiers

A related issue concerned the annotation of prenominal modifiers. The OntoNotes guidelines call for prenominal modifiers to be annotated only when they are proper nouns. However, since we considered all entities to be named entities, our guidelines called for annotation of prenominal modifiers regardless of whether or not they were proper nouns, as such.

2.3. The annotation schema

2.3.1. Noun groups

The basic unit of annotation in the project is the base noun phrase. We defined this as one or more nouns and any sequence of leftward determiners, adjectives, and conjunctions not separated by a preposition or other noun-phrase-delimiting part of speech; and rightward modifiers such as relative clauses and prepositional phrases.

Thus, all of the following would be considered base noun phrases:

- *striatal volume*
- *neural number*
- *striatal volume and neural number*
- *the structure of the basal ganglia*
- *It*

These were not pre-annotated—the annotators selected their spans themselves. This is one potential source of lack of interannotator agreement. Base noun phrases were annotated only when they participated in one of the two relationships that we targetted.

2.3.2. Definitions of the two relations

The two relations that are annotated in the corpus are the IDENTITY relation and the APPOSITIVE relation. The identity relation holds when two units of annotation refer to the same thing in the world. The appositive annotation holds when two noun phrases or a noun phrase and an appositive are adjacent and not linked by a copula or other linking word.

2.3.3. Details of the annotation schema

More specifically, the annotation schema is defined as:

IDENTITY chain An IDENTITY chain is a set of base noun phrases and/or appositives that refer to the same thing in the world. It can contain any number of elements.

Base noun phrase Discussed above.

APPOSITIVE relation An appositive instance has two elements, a head and a set of attributes. The set of attributes may contain just a single element (the prototypical case). Either the head or the attributes may themselves be appositives.

Nonreferential pronoun All nonreferential pronouns are included in this single class.

Thus, an example set of annotations would be:

All brains analyzed in this study are part of [the Mouse Brain Library]$_a$ ([MBL]$_b$). [The MBL]$_c$ is both a physical and Internet resource. (PMID 11319941)

- APPOSITIVE chain: *The Mouse Brain Library$_a$, MBL$_b$*
- IDENTITY chain: *Mouse Brain Library$_a$, The MBL$_c$*

2.4. Training of the annotators

We hired and trained biologists and linguists as a group. Annotators were given a lecture on the phenomenon of coreference and on how to recognize coreferential and appositive relations, as well as nonreferential pronouns. They were then given a non-domain-specific practice document. Following a separate session on the use of the annotation tool, they were given an actual document to annotate. This document is quite challenging, and exercised all of the necessary annotation skills. We began with paired annotation, then introduced a second document for each annotator to mark up individually. Once annotators moved on to individual training annotation, they met extensively with a senior annotator to discuss questions and review their final annotations.

During the initial training phase, we paired biologists with linguists and had them work on the same article independently, then compare results. This turned out to be an unnecessary step, and we soon switched to having annotators work independently from the beginning.

[1] Note that in our guidelines, as in the OntoNotes project, indefinite noun phrases are used to start new IDENTITY chains, and are not linked with previous markables, but this is because they are discourse-new, not because we consider them to be generics.

2.5. Two populations of annotators

We hired two very different types of annotators—linguistics graduate students, and biologists at varying levels of education and with varying specialties. Impressionistically, we did not notice any difference in their performance. The biologists were able to grasp the concept of coreference, and the linguists did not find their lack of domain knowledge to be an obstacle to annotation. Both the biologists and the linguists had opportunities to clarify issues via email and meetings with the senior annotators.

2.6. The annotation process

Most articles are single-annotated, but a subset of fifteen will be double-annotated by random pairs of annotators to calculate inter-annotator agreement.

The length of the articles means that a single IDENTITY chain can extend over an exceptionally long distance. To cope with this, annotators typically marked up single paragraphs as a whole, and then linked entities in that paragraph to earlier mentions in the document.

In the case of questions, annotators had access to senior annotators via email and meetings.

Annotation was done using Knowtator, a Protégé plug-in (Ogren, 2006a; Ogren, 2006b).

2.7. Progress to date

Table 1 shows the total numer of documents, IDENTITY chains, APPOSITIVE chains, nonreferential noun phrases, and base noun phrases with about one third of the project done. (In calculating these numbers, when a document had been double-annotated, we took the average of the two annotators for that document.) These numbers give some idea of the scale of the task of annotating all coreferential noun phrases in full-text scientific journal articles, along with the data in Table 2, where we see the averages. There we note especially that the average time to annotate a single article is twenty hours. This number should be useful for estimating the funding needed for future annotation projects of this sort.

The low number of nonreferential pronouns is striking, but accords with the work of Gasperin, who reported numbers of nonreferential pronouns in her full-text articles that were so low that she omitted them from the annotation schema.

In contrast, the high number of base noun phrases is quite striking. Recall that our process includes the annotators marking the boundaries of the base noun phrases themselves; this high number suggests that a significant time savings might be realized by pre-marking the syntactic constituents of the papers. We also note from early work by Hirschman et al. that this would likely increase our inter-annotator agreement.

Average inter-annotator agreement over a set of ten articles is .684 by the MUC metric. We give a number of other metrics in Table 3 (MUC, (Vilain et al., 1995), B3, (Bagga and Baldwin, 1998), CEAF, (Luo, 2005), and Krippendorff's alpha (Passonneau, 2004; Krippendorff, 1980)). We note that the value for Krippendorff's alpha is lower than the 0.67 that Krippendorff indicates must be obtained before values can be conclusive, but no other IAA values for projects using the OntoNotes guidelines have been published to compare these numbers to. Note again that these are preliminary numbers representing progress to date only and do not represent the inter-annotator agreement values for the completed project.

Documents annotated	31
Total IDENT chains	6,650
Total APPOS chains	1,230
Total nonreferential pronouns	296
Total base noun phrases	27,870

Table 1: **Total documents and markables to date.** Base noun phrase count includes only those base noun phrases included within an IDENT or APPOS chain.

Average time per document	20 hours
Average IDENT chains per document	215
Average APPOS chains per document	40
Average nonreferential pronouns per document	10
Average base noun phrases per document	899

Table 2: **Average markables to date.** Base noun phrase count includes only those base noun phrases included within an IDENT or APPOS chain.

3. Discussion

We found that published and unpublished annotation guidelines for coreference differ widely with respect to the definitions of markables, handling of appositives, handling of predicative noun phrases, handling of nominal premodifiers, and the inclusion of other relations in addition to coreference. A number of them are compared and contrasted here.

We also found that a candidate for a standard set of coreference guidelines, the OntoNotes guidelines, required considerable adaptation to fit the nature of the biomedical domain, particularly with respect to the notion of generics, which play a large role in the OntoNotes guidelines. The description of our annotation process describes how these guidelines can be incorporated into an active annotation project, and our report on progress to date gives an early indication of the scale of the data that will result from annotation to these guidelines.

Acknowledgements

We thank the BBN Corporation for sharing the OntoNotes coreference annotation guidelines with us, and the other

Metric	Average
MUC	.684
Class-B3	.858
Entity-B3	.750
Mention-based CEAF	.644
Entity-based CEAF	.480
Krippendorff's alpha	.619

Table 3: **Inter-annotator agreement values for a sample of ten documents.** A variety of metrics is reported.

writers of guidelines for making their guidelines publicly available. We also thank Guergana Savova and Jiaping Zheng of the Mayo Clinic for sharing their inter-annotator agreement code with us. Two anonymous reviewers improved the paper.

4. References

Amit Bagga and Breck Baldwin. 1998. Algorithms for scoring coreference chains. In *Proceedings of the Linguistic Coreference Workshop at The First International Conference on Language Resources and Evaluation (LREC '98)*, pages 563–566.

Caroline Gasperin, Nikiforos Karamanis, and Ruth Seal. 2007. Annotation of anaphoric relations in biomedical full-text articles using a domain-relevant scheme. In *Proceedings of DAARC 2007*.

Caroline Gasperin. 2006. Semi-supervised anaphora resolution in biomedical texts. In *Linking natural language processing and biology: towards deeper biological literature analysis*, pages 96–103. Association for Computational Linguistics.

L. Hirschman. 1997. MUC-7 coreference task definition.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. OntoNotes: The 90% solution. In *Human Language Technology Conference of the NAACL Companion Volume*, pages 57–60.

Klaus Krippendorff. 1980. *Content Analysis: An Introduction to Its Methodology (Commtext Series)*. SAGE Publications, September.

Xiaoqiang Luo. 2005. On coreference resolution performance metrics. In *Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP)*, pages 25–32.

Philip Ogren. 2006a. Knowtator: a Protege plugin for annotated corpus construction. In *HLT-NAACL 2006 Companion Volume*.

Philip Ogren. 2006b. Knowtator: a plug-in for creating training and evaluation data sets for biomedical natural language systems. In *The International Protege conference*, pages 73–76.

Rebecca J. Passonneau. 2004. Computing reliability for coreference annotation. In *Proceedings of the Language Resources and Evaluation Conference*.

Sameer S. Pradhan, Lance Ramshaw, Ralph Weischedel, Jessica MacBride, and Linnea Micciulla. 2007. Unrestricted coreference: identifying entities and events in OntoNotes. In *ICSC '07: Proceedings of the International Conference on Semantic Computing*, pages 446–453.

Kees van Deemter and Rodger Kibble. 2001. On coreferring: Coreference in MUC and related annotation schemes. *Computational Linguistics*, 26(4):629–637.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. In *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, pages 45–52.

Xiao Feng Yang, Jian Su, Guo Dong Zhou, and Chew Lim Tan. 2004a. A NP-cluster based approach to coreference resolution. In *Proceedings of 20th International Conference on Computational Linguistics (COLING 2004)*, pages 226–232.

Xiaofeng Yang, Guodong Zhou, Jian Su, and Chew Lim Tan. 2004b. Improving noun phrase coreference resolution by matching strings. In *IJCNLP04*, pages 326–333.

Contribution of Syntactic Functions to Single Word Term Extraction

Xing Zhang[1] and Alex Chengyu Fang[2]

Dialogue Systems Group
Department of Chinese, Translation and Linguistics, City University of Hong Kong
Tat Chee Avenue, Kowloon, Hong Kong SAR
[1]zxing2@student.cityu.edu.hk
[2]acfang@cityu.edu.hk

Abstract

This paper intends to investigate what contributions syntactic functions can make towards single word term extraction. It examines the probabilistic relations between medical terms and their syntactic functions. By probabilistic relations, it means the relations between term occurrence ratios and different paths of syntactic functions. An Automatic Term Extraction (ATE) system on the basis of such extended syntactic information is built up to find out which paths of syntactic functions are good indicators for terms after training on a large medical corpus drawn from MEDLINE. Accordingly, term candidates occurring in these syntactic paths will be assigned higher weights for better probabilistic estimates. As a result, the most helpful syntactic paths in identifying terms will be found out. One linguistic motivated method, SF-Value, is proposed to weight termhood of term candidates. Results of experiments show that single word terms are extracted dominantly at a fairly good recall besides multiword terms. In this way, syntactic behaviors of single word terms prove to be especially effective in selecting single word terms. All in all, this work studies the actual usage of terms in real texts rather than a static description of their internal structures. It dynamically characterizes patterns of term usage to a much deeper degree. And this information will in turn contribute to practical ATE system.

1. Introduction

Terms usually refer to the linguistic manifestation of concepts in a specific domain. More specifically, terms are the linguistic expression of the concepts of special communication and are organized into systems of terms which ideally reflect the associated conceptual system (Ananiadou, 1994). Generally, terms are divided into single word terms and multi word terms. Past works have different opinions on levels of challenges these two types of terms pose. Wermter and Hahn (2005) think the recognition of single-word terms usually does not pose any particular challenge; it's multiword terms that are much more difficult. While other researchers believe single term words are much more difficult to recognize because semantic information is needed to distinguish between the general usage of a word and its terminological usage (Eumeridou et al., 2004) and statistically is very difficult to capture domain-specific single-word terms (Sclano & Velardi, 2007).

With respect to application for ATE, lots of works have been devoted to multiword extraction. Methods used include morpho-syntactic properties (Daille, 1996), and classic statistical measures as TF·IDF (Salton, 1988), Mutual Information (Church, 1989), Log-Likelihood Ratio (Dunning, 1993), Dice Factor (Smadja et al., 1996), etc. C-Value (Frantzi & Ananiadou, 1998) measure is widely considered as the state-of-the-art model for ATE, which can also perform well on other languages such as Japanese (Mima &Ananiadou, 2001), Slovene (Vintar, 2004).

As for single word term extraction, limited works have been done. TerMight (Dagan & Church, 1994) just define Single-word candidates by taking the list of all words that occur in the document and do not appear in a standard stop-list of "noise" words. Xu et al. (2002) designed a TFIDF-based single word term classifier. Bernhard (2006) presents a pattern-based technique to extract single word term, which is based on some classical word-forming unit, e.g. prefixes (extra-, anti-), initial combining forms (hydro-, pharmaco-) and suffixes (-ism). Corpora comparison method was used in Rayson & Garside (2000), Baroni & Bernardini (2004), Kit & Liu (2008).

This paper aims to investigate what contributions syntactic functions can make towards single word term extraction. It presents a linguistic-grounded method to measure termhood of single word term. The intuition of this work is that terms tend to play certain kinds of syntactic functions more prominently. And this kind of syntactic behavior of terms can be captured as termhood by computation of term ratios in different syntactic paths. Syntactic path in this work refers to a path of syntactic functions of one NP. Specifically, it is defined as concatenation of elementary syntactic functions tagged by

Survey Parser. Term ratios are defined as the frequencies of term occurrences in each syntactic path over all term occurrence frequencies in all syntactic paths. This work studies the correlation between terms and syntactic functions through careful analysis of real experiments with theoretical insights. The corpus it uses is built from MEDLINE[1] and its performance is compared with two existing term extractor, TerMine and TermExtractor (Sclano & Velardi, 2007).

2. Methods and Experiments

This study proposes a method to measure the probabilistic relations between terms and their syntactic functions. This method is a weighting scheme based on term ratios in syntactic paths of term candidates in parsed texts. Overall architecture of this system includes three major modules. The first module is to get abstracts from MEDLINE. It will create experimental corpus from MEDLINE database. The second module creates a term list from MeSH, and annotates the corpus according to this term list. Another major function of the second module is to compute term ratios in different syntactic paths. The third module uses the knowledge of term ratios in different syntactic paths to assign different weights to different syntactic paths and then compute the Syntactic Function Value (SF-Value) of each term candidate using the following formula:

$$SFValue = \sum_{i=1}^{n} FSS_i \times WSS_i$$

This formula contains two parameters: FSS_i is the frequency of syntactic $path_i$, WSS_i is the weight of syntactic $path_i$, n is the count of how many syntactic paths this term candidate occurs in. WSS_i is computed previously from training on a corpus that are annotated with MeSH terms. It is computed on the basis of the proportion of term occurrence frequency in this syntactic path among the total term occurrence frequencies of all syntactic paths. Therefore, WSS_i is higher for syntactic paths that are more likely to be filled by terms than other syntactic paths. And SF-Value is higher for terms that are present in syntactic functions with a higher WSS_i. Moreover, SF-Value is higher for NPs that occur more often in syntactic paths that are themselves more often occupied by terms than others.

2.1 Resource Building and Processing

2.1.1 Corpora Building up

This study built a small subset of MEDLINE abstracts based on the controlled search of the database using the keyword *internal medicine*. This search produces 252,033 abstracts (until 17 July 2008). Each abstract consists of a single title and a number of sentences. One sub-corpus of 360 abstracts was manually checked by human professionals for possible tagging mistakes of syntactic functions after they were parsed by the Survey Parser (Fang, 1996). A list of medical terms was created from Medical Subject Headings (MeSH 2009) beforehand. This MeSH term list consists of 602,436 terms, 430,848 are multi word terms, and 171,588 are single word terms. These corpora then will be terminologically annotated: noun phrases that match the term list are tagged as terms.

This manually checked sub-corpus is of 82, 055 words and further divided into ten subsets randomly, 36 abstracts each. Each time, nine out of ten subsets is used as training, and the one left out is used as testing set. And the whole procedure is repeated 10 times. The advantages of such ten-fold cross validation enable the greatest possible amount of data used for training in each iteration. And we can also predict accuracy for unseen data sets.

Gold standard used in this work is the number of true MeSH terms in testing corpora. In order to get all true MeSH terms in testing corpora, N-grams (N is from 1 to 10) will be extracted from each corpus at first and matched against MeSH term list. N-gram matching method is employed in order to avoid effects from parsers because different parsers will output different NP lists, which will lead to different term candidates at the beginning. Therefore, the absolute number of MeSH terms had better to be parser independent. The following table (Table 1) presents basic statistics of testing subsets.

2.1.2 Survey Parser

Survey Parser was first designed to complete the syntactic annotation of the International Corpus of English. It effectively parses sentences in many layers with detailed syntactic functions. The unique feature of the parsing scheme is that it analyzes the syntactic functions of the constituent paths and represents them in the form of a parsing tree. Survey Parser also classifies syntactic functions into two major kinds: one is phrasal functions, and the other is clausal functions, which correspond to the basic elements of English sentences such as subject, verb, object, complement and adverbial. For example, if *cells* is tagged as a term, and the syntactic path for it is recorded as "NPHD-N%PC-NP%A-PP", which means that it is a noun of the function NP head, which is a part of larger NP of the function preposition complement, and which is part of a preposition phase of the function adverbial.

2.1.3 Stop List

In this experiment, a stop word list is created, which consist of a few frequent grammatical words, such as definite articles, demonstrative and possessive adjectives, and indefinite articles. Words in stop list are uninformative for terminology extraction. The aim of using a 'stop word' list is to remove very frequent words which are not considered to carry terminological meanings.

[1] MEDLINE is the National Library of Medicine's premier bibliographic database. MeSH is the U.S. National Library of Medicine's controlled vocabulary used for indexing articles for MEDLINE.

Testing corpora	# of words	# of parsed sentences	All MeSH terms	Multi MeSH terms	Single MeSH terms
subset 1	7,929	356	466	199	267
subset 2	7,576	313	429	120	309
subset 3	8,355	374	447	117	330
subset 4	7,693	325	479	152	327
subset 5	8,562	391	490	137	353
subset 6	8,562	357	494	138	356
subset 7	8,562	334	525	151	374
subset 8	8,353	381	515	140	375
subset 9	8,675	375	475	134	341
subset 10	7,788	343	524	157	367

Table 1: Basic Statistics of Testing Corpora

2.2 Experimental Setup

After parsing training and testing texts with Survey parser, the experiment is realized in the second and third module introduced earlier. The second module mainly annotates the training corpora with MeSH terms and computes term ratios of syntactic paths. The procedure includes following steps:

- Match NPs in training corpora with terms in MeSH term list.
- Record syntactic paths in which MeSH terms are identified.
- Calculate term occurrence frequencies of such syntactic paths and compute term ratios of them;
- Compute different weights (WSS) respectively for those syntactic paths.

The third module is to compute SF-Value for each term candidate:

- Input testing texts and extract all NPs with their frequencies from it.
- Use stop list to filter those NPs; and delete those with a length larger than 10 words.
- Produce a NP list.
- For each NP, compute frequencies of syntactic paths where this NP has occurred in.
- Compute SF-Value for each NP, and arrange them in descending order.
- Set a threshold value for SF-Value and NPs with SF-Value above this threshold are considered as terms.

This study adopts all the symbols used in the Survey Parser, for example, SU stands for subject, PC stands for prepositional complement and so on (see Appendix 1). And '%' is used to indicate a node of higher level. '+' is used to indicate two nodes of the same level.

2.2.1 Results of Term Ratios of Syntactic Paths

From the first module, around three hundred kinds of syntactic paths are recorded if taking clausal functions as ending nodes. However, there are only a few paths accounting most prominently, such as SU-NP, PC-NP%A-PP, while the rest has quite a low frequency each. Therefore, in order to deal with sparse data and meanwhile keep distinctive features of these syntactic functions to a great extent, this research conflates those syntactic paths with the same beginning nodes and ending nodes. This conflation promotes the ranking of some syntactic paths by means of grouping these syntactic paths with extremely low term occurrences.

2.2.2 Syntactic Paths Ending in Clausal Functions

The following table is a list of syntactic paths ranked higher in training corpora (see Table 2).

Syntactic Paths	Frequency	Ratio
SU-NP	2515	17.57%
OD-NP	1657	11.58%
DEFUNC-NP	491	3.43%
A-PP	315	2.20%
VB-VP	280	1.96%
PC-NP	241	1.68%
APPOS-NP	205	1.43%
NPPR-AJP+NPHD-N%PC-NP%A-PP	156	1.09%
CS-NP	145	1.01%
NPPR-AJP+NPHD-N%SU-NP	91	0.64%

Table 2: Top Ten Term Ratios in Syntactic Paths Ending in Clausal Functions

From above table, we can see that terms take the function SU (subject) most frequently, followed by these taking the function of OD (direct object). The third ranking is the function DEFUNC (detached function*)*, followed by function A (adverbial). And the accumulative ratios of these ten kinds of syntactic paths total around 45%, which indicates nearly half of the terms occurring in these ten paths.

2.2.3 Conflation of Syntactic Paths with the Same Beginning Node and Ending Node

Besides these paths discussed earlier, there are other 570 kinds of paths with a term ratio below 0.5%. Therefore, these paths are conflated before allocating weights to them. The principle is that syntactic paths with the same beginning syntactic functions and the same ending clausal functions are conflated into a single group. For example:

Syntactic Paths	Term Ratios
NPPR-AJP+NPHD-N%PC-NP%NPPO-PP%SU-NP	0.199%
NPPR-AJP+NPHD-N%PC-NP%NPPO-PP%PC-NP%NPPO-PP%SU-NP	0.003%
NPPR-AJP+NPHD-N%NPPO-PP%NPPO-PP%NPPO-PP%SU-NP	0.001%

Table 3: Conflation of Syntactic Paths

In this table, the starting syntactic functions of these three syntactic paths are all NPPR-AJP+NPHD-N, which means a node of NPPR-AJP together with a node of NPHD-N. And the ending syntactic functions are all SU-NP, therefore, these three paths are conflated as NPPR-AJP+NPHD-N%SU-NP, and the term ratios of them is added up as 0.203% after conflation.

Syntactic Paths	Frequency	Ratio
PC-NP%A-PP	2855	25.29%
SU-NP	1996	17.68%
OD-NP	1296	11.48%
PC-NP%SU-NP	711	6.30%
PC-NP%NPPO-PP	431	3.82%
PC-NP%OD-NP	426	3.77%
DEFUNC-NP	371	3.29%
VB-VP	300	2.66%
NPPR-AJP+NPHD-N%A-PP	233	2.06%
A-PP	230	2.04%

Table 4: Top 10 Syntactic Paths after Conflation

From above table, the syntactic paths PC-NP%A-PP is promoted to be the first, while SU-NP and OD-NP ranking next. And the number of syntactic paths is reduced to 128 all together. And meanwhile, term ratio of each syntactic path is increased, which means effectiveness of the weighting strength of each path is enhanced.

3. Results Analysis

3.1 Comparison with Existing Term Extraction Systems

TerMine is the online service provided by National Centre for Text Mining of University of Manchester. It mainly employs C-Value to extract terms. As C-Value is designed for multiword terms, TerMine extracts multi word only.

TermExtractor is the online service provided by the Linguistic Computing Laboratory of the University of Roma "La Sapienza". It uses domain relevance, domain consensus and lexical cohesion, to weight term candidates. It can let the users set word lengthen for terms to be extracted. Therefore, if the minimum number is set as 1, single word terms would be extracted.

For comparison, the ten testing corpora will be uploaded to TerMine and TermExtractor separately. And the results given back will be matched against the MeSH term list from 2009 MeSH files. As we can see from Table 5, ATE system using SF Value can extract much more single MeSH terms than either TerMine or TermExtractor. Take testing subset 1 as example, TerMine output 1239 terms, 110 are multi MeSH terms; TermExtractor output 266 terms, 51 of them are MeSH terms, and only 2 are single. Comparatively, all NPs extracted by SF Value are 2350, 1023 are single NPs. Among all these NPs, 419 of them are MeSH terms and single MeSH terms is 309, accounting for 73 percent.

3.2 Evaluation

For evaluation of SF Value, an automatic method is implemented in this ATE system. Within the interval of the minimum SF Value to maximum SF Value, a set of threshold values will be set automatically. Each time, the threshold value will be increased on the basis of a preset amount. Precision and recall will be computed with respect to each threshold. And F-score will be computed as:

$$F - score = 2\frac{(precision \times recall)}{precision + recall}$$

From Table 6, we can see the recall of single MeSH term is very high, with an average value of 0.86. And average F-score of all 10 testing corpora is 0.30 before conflation of syntactic paths. It is worthy of noticing that F-scores of these 10 testing corpora are significantly improved after conflating syntactic paths (p<0.05). And the average F-score reaches 0.36 after conflation of syntactic paths. Based on these results, we can find that SF-Value is especially effective in measuring termhood of single word terms which are an important part in term extraction.

Testing Corpora	Term Candidates by TerMine			Term Candidates by TermExtractor			Term Candidates by ATE System based on SF-Value					
	All Term Candidates	Multi MeSH Terms	Single MeSH Terms	All Term Candidates	All MeSH Terms	Single MeSH Terms	All NPs	Multi NPs	Single NPs	All MeSH Terms	Multi MeSH Terms	Single MeSH Terms
subset 1	1239	110	0	266	51	2	2350	1327	1023	419	110	309
subset 2	1179	114	0	261	46	1	2138	1196	942	374	112	262
subset 3	1296	119	0	277	65	4	2361	1379	982	391	122	269
subset 4	1384	132	0	277	68	4	2399	1386	1013	400	137	263
subset 5	1407	120	0	291	59	1	2577	1560	1017	423	134	289
subset 6	1265	127	0	305	60	3	2432	1459	979	438	130	308
subset 7	1388	132	0	273	62	1	2432	1428	1004	455	147	308
subset 8	1327	126	0	316	72	4	2568	1533	1035	451	132	319
subset 9	1387	123	0	218	49	1	2471	1450	1021	425	124	301
subset 10	1355	153	0	257	61	5	2417	1437	980	463	163	300

Table 5: Term Candidates Produced by TerMine, TermExtractor and ATE System based on SF-Value

Testing Corpora	Single MeSH Terms by N-gram	Single MeSH Terms by SF Value	Recall	F-score before Conflation	Threshold Value for F-score before Conflation	F-score after Conflation	Threshold Value for F-score after Conflation
subset 1	267	309	1	0.34	0.03	0.39	0.13
subset 2	309	262	0.85	0.31	0.01	0.35	0.10
subset 3	330	269	0.82	0.29	0.06	0.37	0.10
subset 4	327	263	0.80	0.30	0.06	0.35	0.13
subset 5	353	289	0.82	0.28	0	0.33	0.08
subset 6	356	308	0.87	0.30	0	0.34	0.08
subset 7	374	308	0.82	0.31	0	0.35	0.13
subset 8	375	319	0.85	0.29	0.01	0.38	0.13
subset 9	341	301	0.88	0.29	0.01	0.34	0.13
subset 10	367	300	0.82	0.33	0.06	0.39	0.13
Average			0.86	0.30		0.36	

Table 6: Performance of ATE System based on SF-Value

4. Conclusion and Future Work

This research shows there is a probabilistic correlation between syntactic functions of a NP in sentences and the termhood of this NP. One weighting measure, SF- Value, is implemented to compute termhood of terms. This measure incorporates linguistic knowledge about syntactic properties of terms and can assign effective values to term candidates. By setting threshold values, both multi word terms and single word terms can be selected effectively. In particular, single word terms can be selected at a dominant rate. The syntactic properties of single word terms can be measured by such a simple statistical value, which can be considered as a practical indicator of single word term.

The most innovative aspect of this research is the exploring the contribution of syntactic functions in recognizing and extracting single terms from texts. This approach represents a novel, linguistically motivated perspective in the area of terminology extraction. Most importantly, unlike other ATE systems that include various terminology extraction techniques, this system lies mainly on syntactic properties of terms with an aim to study the relations between terms and their syntactic functions. This feature promotes us to obtain reliable statistics on term occurrences. Therefore, this research can be fairly valuable in that it shows the direct correlation between term occurrence and syntactic functions of an NP. And it also proves the effectiveness of such a method in distinguishing single terms and non-terms.

In the future, this method should be validated for different domains. What's more, during linguistic processing of these corpora, it is found that different parsers present linguistic information of different granularities, which will subsequently affect the accuracy of term recognition. Therefore, different parsing results may be tested to find out how the performance of ATE can be influenced.

5. Acknowledgements

Research described in this article was supported in part by research grants from City University of Hong Kong (Project No. 7002387, 7008002, and 9610126).

6. References

Ananiadou, S. (1994). A methodology for automatic term recognition. In *Proceedings of COLING 94*, pp. 1034-1038.

Baroni, M. and Bernardini, S. (2004). Boot-CaT: Bootstrapping Corpora and Terms from the Web. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC)*, pp. 1313–1316.

Bernhard, D. (2006). Multilingual Term Extraction from Domain-specific Corpora Using Morphological Structure. In *Proceedings the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL-06): Posters and Demonstrations*, pp. 171–174.

Church, K. (1989). Word association norms, mutual information and lexicography. In *Proceedings of the 27th annual meeting of the ACL.* Vancouver, pp. 76-83.

Dagan, I. & Church, K. (1994). Termight: identifying and translating technical terminology. In *Proceedings of the fourth conference on Applied Natural Language Processing*, pp. 34-40.

Daille, B. (1996). Study and Implementation of Combined Techniques for Automatic Extraction of Terminology. In J. Klavans & P. Resnik (Eds.), *The Balancing Act: Combining Symbolic and Statistical Approaches to Language,* Cambridge, MA: The MIT Press, pp. 49–66.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics* 19(1), pp. 61-74.

Eumeridou, E., Nkwenti-Azeh, B. & McNaught, J. (2004). An Analysis of Verb Subcategorization Frames in Three Special Language Corpora with View towards AutomaticTerm Recognition. *Computers and the Humanities* 38, pp. 37-60.

Fang, A.C. (1996). The Survey Parser: Design and Development. In S. Greenbaum (Ed.), *Comparing English World Wide: The International Corpus of English*, Oxford: Oxford University Press, pp. 142-160.

Frantzi, K., Ananiadou, S. & Tsujii, J. (1998) The C-value/NC-value Method of Automatic Recognition of Multi-word Terms. In *Proceedings of the Second European Conference on Research and Advanced Technology for Digital Libraries*, pp. 585-604.

Sclano F. and Velardi P. (2007). TermExtractor: A Web application to learn the common terminology of interest groups and research communities. In *Proceedings of the 9th Conference on Terminology and Artificial Intelligence*, Sophia Antinopolis, France.

Mima, H. and Ananiadou, S. (2001). An Application and Evaluation of the C/NC-Value Approach for the Automatic Term Recognition of Multi-Word Units in Japanese. *International Journal on Terminology* 6(2), pp. 175-194

Kit, C. & Liu, X. (2008). Measuring Mono-word Termhood by Rank Difference via Corpus Comparison. *Terminology* 14(2), pp. 204-229.

Medical Subject Headings: http://www.nlm.nih.gov/pubs/factsheets/mesh.html.

Rayson P. and Garside R. (2000). Comparing Corpora using Frequency Profiling. In *Proceedings of the ACL Workshop on Comparing Corpora*, pp. 1–6.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing and Management: an International Journal*, 24(5), pp. 513-523.

Smadja, F., McKeown, K. R., & Hatzivassiloglou, V. (1996). Translating collocations for bilingual lexicons: a statistical approach. *Computational Linguistics*, 22(1), pp. 1-38.

Vintar S. (2004). Comparative Evaluation of C-value in the Treatment of Nested Terms. Memura 2004 – Methodologies and Evaluation of Multiword Units in Real-World Applications. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation*, pp. 54-57.

TerMine. (2009). NaCTeM website (http://nactem.ac.uk) .

Xu, F., Kurz D., Piskorski J., and Schmeier S. (2002). Term extraction and mining term relations from free-text documents in the financial domain. In *Proceedings of the 5th International Conference on Business Information Systems (BIS'02)*, Poznan, Poland.

Wermter J. and Hahn, U. (2005). Massive Biomedical Term Discovery. In *Proceedings of the 8th International Conference on Discovery Science*, pp. 281- 293.

Appendix

1. Survey Parser Symbols

The Phrasal Categories and Functions: Adverb (AVP), Premodifier (AVPR), Head (AVHD), Postmodifier (AVPO), Adjective (AJP) Premodifier (AJPR), Head (AJHD), Postmodifier (AJPO), Determiner (DTP), Premodifier (DTPR), Predeterminer (DTPE), Central determiner (DTCE), Postdeterminer (DTPS), Postmodifier (DTPO), Noun (NP) Determiner (DT), Premodifier (NPPR), Head (NPHD), Postmodifier (NPPO, Prepositional (PP) Modifier (PMOD), Prepositional (P), Complement (PC) , Subordinator (SUBP) Modifier (SUBMO), Head (SUBHD),Verb (VP), Operator (OP), Auxiliary verb (AVB), Main verb (MVB).

The Clausal Categories and Functions: Subject (SU), Provisional subject (PRSU), Notional subject (NOSU),Verb (VB), Predicate (PRED), Object Direct object (OD), Indirect object (OI), Provisional object (PROD), Notional object (NOOD), Complement Subject complement (CS), Object complement (CO),Transitive complement (CT), Focus complement (CF), Adverbial (A), Cleft operator (CLOP), Existential operator (EXOP), Imperative operator (IMPOP), Interrogative operator (INTOP), Inversion operator (INVOP), Coordinator (COOR), Detached function (DEFUNC0,Discourse marker (DISMK), Clause element (ELE), Focus (FOC), Linker (LK), Punctuation (PUNC),Subordinator (SUB).

2. Top Single Term candidates ranked by SF Value in descending order (from testing subset 10).

The leftmost value has two values, 1 or 0. 1 indicates this NP is a true MeSH term, 0 indicates it is not.

```
0 0.170745 background
0 0.132059 glanzmann1 0.132059 thrombasthenia
0 0.179134 agt
0 0.197231 disorder1 0.171540 alloantibodies
1 0.286158 autoantibodies
1 0.132059 paraproteins
1 0.887126 diagnosis
0 0.069618 assays
0 0.025571 consuming
0 0.823444 tests
1 0.457820 time
0 1.888486 we
0 0.632440 case
0 0.233235 detection
1 0.327183 methods
1 0.276609 male
0 0.000121 count
1 0.000121 lymphoma
0 1.743777 results0 0.078183 normal
1 0.294236 ristocetin
0 0.077491 response
```