A Data-Driven Approach for Extracting “the Most Specific Term” for Ontology Development

Guergana K. Savova[1], PhD, Marcelline Harris[1,2], R.N., PhD, Thomas Johnson[1], Serguei V. Pakhomov[1], PhD, Christopher G. Chute[1], MD, DrPH
[1]Division of Medical Informatics Research, Mayo Clinic, Rochester, MN
[2]Division of Nursing Research, Department of Nursing, Mayo Clinic, Rochester, MN

Abstract

We present a data-driven approach to extract the “most specific” terms relevant to an ontology of functioning, disability and health. The algorithm is a combination of statistical and linguistic approaches. The statistical filter is based on the frequency of the content words in a given text string; the linguistic heuristic is an extension of existing algorithms but goes beyond noun phrases and is formulated as a “complete syntactic node”. Thus, it can be applied to any syntactic node of interest in the particular domain. Two test sets were marked by three experts. Test set 1 is a well-constructed text from pain abstracts; test set 2 is actual medical reports. Results are reported as recall, precision, F-score and rate of valid terms in false positives. A limitation of the current research is the relatively small test set.

INTRODUCTION

Automated term extraction tools have been developed to achieve multiple purposes including information retrieval[1], machine translation[2] and for terminology and ontology building[3, 4]. In computational linguistics, the development of term extraction tools has been approached from several angles – using purely linguistic approaches, purely statistical techniques and a combination of both[5, 6, 7]. These research efforts have focused on nominal term extractions; i.e., nouns and noun phrases [8, 9, 6](NP’s).

Previous work in health terminology and ontology development has demonstrated the contributions when verb phrases (VP’s) as well as NP’s are extracted. For example, Grobe[10] used NLP approaches to extract terms from free text of nursing notes and, analyzing both VP’s and NP’s was able to construct an automated categorization algorithm with an overall agreement between the algorithm and human classifiers above 90%, a significant improvement over other approaches to the classification of nursing domain content. LeMoigno et al.[3] reported on building a surgical ontology, and similarly described a need to consider VP’s in the analysis of their textual corpora. Work at Mayo Clinic (Ruggieri and Pakhomov, personal communication) shows that for domains such as functional status, VP’s are critical to harness a full list of valid and necessary terms.

Furthermore, the domain of functioning, disability and health has long been plagued by ambiguous terminology that limits knowledge development[11, 12, 13]. The ICF coding system[11] provides needed direction in establishing the boundaries and classification of concepts within the domain. However, the list of terms within the ICF coding system is rather limited in scope.

Preliminary studies by our group (currently unpublished) indicate variability in term length when human experts are asked to identify the important terms related to concepts within the domain of functioning, disability and health. In a study of a small corpus (app. 500 words) marked for terms by 10 experts, we found that the average length of terms is 4.25 words (s.d. 1.72) and that all but 7 terms were complete (well-formed) syntactic phrases (syntactic nodes). The 7 terms that did not follow the syntactic node criterion are “distributed concepts” spanning over several non-adjacent nodes, e.g. in “The patient walks now with a cane”, a distributed concept is “walks with a cane” where “now” is omitted from the original source. Therefore, we were motivated to look at sequences of up to five content words forming unigrams, bigrams, trigrams, tetragrams and pentagrams. The results from these preliminary studies combined with the recognized importance of VP’s as well as NP’s motivate our work to go beyond shallow parsing of NP’s in term extraction tools.

Term extraction is the first step in ontology building. It identifies the linguistic representation of concepts. Our ultimate goal is to build an ontology for functioning, disability and health from large text corpora via feasible automated/semi-automated techniques. The proposed term extraction method is to be followed by establishing the relations among terms and clustering them into concept groups to fit into the ontology model.

RESEARCH AIMS

The goal of the project we are reporting on here was to devise an algorithm to extract the “most specific term candidate” from corpora relevant to the broad domain of functioning, disability and health. “The most specific term" is defined as the longest string of lexical items deemed important by experts for ontology inclusion. Thus, "the most specific term" can consist of one or more primitive terms and one or more words. "Most specific terms" and their dissection into primitive terms provide the basis for populating ontologies. The research aims for the current project were to:

- Develop a method/algorithm for extracting relevant syntactic nodes such as VPs and short sentences in addition to NP's (well-formed syntactic constituents).
- Limit the number of invalid terms from such an extraction by using term length statistic from our preliminary studies.
- Test the performance of such an algorithm on syntactically well-formed text (research abstracts) and syntactically fragmented text (medical reports)

ALGORITHM AND TOOL

Our algorithm relies on a fully parsed text, for which we use Charniak's parser[14] (processing speed of 2.5 sec/sent). Fully parsed text provides disambiguation. Currently, we are not using any semantic information and the context surrounding the term candidates. The error rate of Charniak's syntactic parser was estimated as 9% on the test corpora by expert inspection. Currently, we exclude terms from the test sets that come from sentences that were either skipped or not parsed by the syntactic parser.

Our term extraction algorithm relies on two filters. Filter 1 is the frequency of a content word and filter 2 is "syntactic node completeness". A "complete syntactic node" is defined as a NP or a VP at the uni-, bi- and tri-gram levels and as a NP, VP, and sentence (S) at the tetra- and penta-gram level. If a content word passes the threshold frequency at the unigram level only, then a check against an ngram database is done for all ngrams containing that high frequency word. The database contains unigram frequencies, possible ngrams and the parsed corpus. Then, filter 2 is applied. If the ngram is a complete syntactic node, then it is extracted as a term candidate. Going through the ngrams, the algorithm extracts not only phrases consisting of one node (e.g. NP consisting of an adjective and a noun like in "a basilar aneurysm") but also embedded phrases (e.g. NP consisting of a modified head noun followed by a PP like in "a coordination and dexterity program for her right arm").

Another motivator of the algorithm was to neutralize parsing errors; therefore the algorithm does not rely on a database of syntactic patterns for the candidate terms[5] as the linguistic filter, rather it checks for syntactic phrase/node completeness. An example of parse error tolerance is the term "require max assist" (the intended meaning is "[patients] require maximum assistance"):

```
<VP>
      <VB> require </VB>
      <S>
              <NP>
        <NNP> max </NNP>
              </NP>
              <VP>
        <VB> assist </VB>
               </VP>
       </S>
</VP>
```

Although the parse within the external VP is incorrect (the incorrect parse implies the meaning of "require [that] Max assist"), the string is extracted as it forms a complete syntactic node (VP). The database approach[5] relies on finding the syntactic pattern "VP->V NP" and because of the parse error, the term will not be extracted.

METHODS

Procedures

Our approach is data-driven and is based on NLP and statistical techniques. Three experts marked two test sets for the "most specific" term relevant to an ontology for functioning, disability, and health. Both test sets cover content identified within the ICF scope of the domain.

Table 1: Test sets by expert and length of the "most specific term"

Term length (in raw number of content words)	*Expert 1 (pain test set)*	*Expert 2 (pain test set)*	*Expert 3 (dismissal notes test set)*
1	33	73	85
2	134	230	266
3	150	175	269
4	91	78	190
5	60	36	103
>5	41	17	117
total	509	609	1030

Test set 1 consists of abstracts from the literature on pain (8,119 words). To limit potential "noise" that might be associated with irrelevant literature, we used a set of abstracts previously obtained from a national expert on pain. Two clinical experts then were independently asked to mark words within the abstracts that indicated the "most specific term" associated with relevant concepts (inter-rater agreement=0.81). Test set 2 consists of actual dismissal summaries from Mayo clinic records of patients discharged from a physical medicine and rehabilitation service who had given permission for the use of their records for research (30,607 words). One clinical expert marked terms from that test set. Table 1 shows the counts of terms by test set, expert and term length.

Analysis

The terms extracted by our algorithm were then compared against the test sets tagged by our clinical experts at every term length; recall, precision and F-measures were computed using the following statistical formulas[15]:

$$recall = \frac{valid\_computer\_derived\_terms}{expert\_derived\_terms}$$

$$precision = \frac{valid\_computer\_derived\_terms}{all\_computer\_derived\_terms}$$

$$F-measure = \frac{(b^2+1)PR}{b^2 P + R}$$

where P is precision and R is recall. Recall is reported for every term length and as an overall metric. Precision and F-measure are reported as overall scores. Additionally, in the F-measure statistic, when ß is one, precision and recall are given equal weights and we report the results by equal weight.

False positives are evaluated for valid terms by expert 1 and expert 3 to determine whether they are true errors. Expert 2 was not available for the additional validation. The formula we used to compute the rate of valid terms in false positives is:

$$rate\_valid\_terms\_in\_FP = \frac{valid\_terms\_in\_FP}{all\_FP}$$

where FP is false positives (non-matches, or terms not in the expert test set but in the computer derived set).

We evaluated the source sentence of the "most specific" term the expert and the algorithm derived in order to assure that our comparisons were from same sentences.

We extracted all NP's and VP's to check the syntactic node filter and the degree to which NP's and VP's provided coverage of the test sets, and report the percent of matches as "all NPs, VPs". The current algorithm was tested by applying several unigram frequency thresholds: 0, 2, 3, 5. A threshold of 7 was tested only on the dismissal notes corpus as it was the larger of the two test sets.

RESULTS

Detailed results at every term length are reported in Table 2. The summary results are in Table 3.

In all test sets, unigrams were the most influenced by the frequency threshold. This is expected as the frequency of all ngrams except unigrams relies on the frequency of their many constituents.

The error analysis shows that there are three groups of terms not entirely fitting the "complete syntactic node" filter (column 2 of the results tables). One group consists of NP's which have additional modifiers, e.g. in the sentence "pain was evaluated in FRAIL SENIORS", the expert marked "seniors" as the term, while the complete NP is "frail seniors". The second group is VP's that have several complimentizers, e.g. in the sentence "Food is cleared spontaneously with a spontaneous second swallow", the expert marked "cleared spontaneously", while the entire extracted VP is "is cleared spontaneously with a spontaneous second swallow". The third group not covered by extracting all NP's and VP's is short sentences (up to 7 words) and fragmented sentences (e.g. "Moderate assist for bathing.") and a small number of prepositional and adjectival phrases (about 1% of all terms).

Constraining the length to 5 content words lowers recall, mainly due to the exclusion of terms longer than 5 content words. Precision increases though. The combination of the length constraint and increased frequency cutoff threshold leads to a drop in recall but improved precision and F-measure.

Our algorithm extracts the primitive terms from the most specific terms as long as they pass the two filters. For example, the most specific term (in this example is a compound term) "fine motor coordination for upper extremities" was extracted along with its primitive components "fine motor coordination" and "upper extremities". The expert marked the most specific term thus not including its primitive constituents. That prompted the additional expert validation of the computer-derived term candidates originally reported as false positives (reported as rate_valid_terms_in_FP). That rate increases as the cutoff increases.

The best F-measure for expert 1 and expert 2 is with a cutoff frequency of 5 (cutoff = 7 was not done for them). The best F-measure for expert 3 is a cutoff frequency of 7. Additional analyses show that those cutoffs correspond roughly to the mean unigram frequencies in the corpora.

The algorithm was evaluated for source correctness to check whether the computer-derived terms are extracted from the same sentence as the expert indicated. 25 random sentences were checked against expert 1 database. In all cases, the source of the computer-derived term coincides with the source sentence of the expert derived term.

DISCUSSION

The current algorithm, (a combination of frequency and linguistic filters) has several advantages over purely linguistic approaches (deriving all NP's and VP's). First, it decreases the noise in the candidate terms and increases the F-measure. Its advantage over rule-based approaches (e.g. databases with syntactic patterns) is that it is less prone to parse errors within the node thus making the use of available syntactic parsers viable, and does not require database maintenance of syntactic rules.

Table 2: Detailed results by test set, "most specific" term length and expert.

	Linguistic Approach	*Current Approach (linguistic and statistical filters applied)*				
Term length (number of content words)	*RECALL method: all NPs, VPs*	*RECALL method: frequency = 0*	*RECALL method: frequency = 2*	*RECALL method: frequency = 3*	*RECALL method: frequency = 5*	*RECALL method: frequency = 7*
RESULTS: EXPERT 1						
1	100%	100%	67%	52%	52%	not extracted
2	99%	99%	95%	92%	81%	not extracted
3	97%	97%	96%	94%	89%	not extracted
4	91%	90%	88%	87%	87%	not extracted
5	85%	85%	82%	82%	60%	not extracted
> 5	63%	not extracted	not extracted	not extracted	not extracted	not extracted
RESULTS: EXPERT 2						
1	86%	86%	78%	70%	62%	not extracted
2	92%	91%	90%	85%	71%	not extracted
3	91%	91%	87%	85%	78%	not extracted
4	86%	82%	82%	81%	81%	not extracted
5	78%	72%	72%	72%	72%	not extracted
> 5	71%	not extracted	not extracted	not extracted	not extracted	not extracted
RESULTS: EXPERT 3						
1	94%	94%	85%	80%	71%	69%
2	87%	86%	84%	82%	76%	72%
3	80%	85%	83%	83%	82%	81%
4	72%	71%	68%	68%	68%	68%
5	58%	59%	59%	58%	58%	58%
> 5	63%	not extracted	not extracted	not extracted	not extracted	not extracted

Table 3: Summary results

	Linguistic Approach	*Current Approach (linguistic and statistical filters applied)*				
Term length (number of content words)	*MATCHES method: all NPs, VPs*	*MATCHES method: frequency = 0*	*MATCHES method: frequency = 2*	*MATCHES method: frequency = 3*	*MATCHES method: frequency = 5*	*MATCHES method: frequency = 7*
RESULTS: EXPERT 1						
recall (F-score)	93% (0.2)	87% (0.33)	83% (0.34)	80% (0.36)	73% (0.37)	not extracted
precision	11%	20%	22%	23%	25%	not extracted
rate_valid_terms_in_FP	44%	61%	67%	71%	80%	not extracted
RESULTS: EXPERT 2						
recall (F-score)	89% (0.22)	86% (0.38)	83% (0.40)	79% (0.40)	71% (0.41)	not extracted
precision	13%	24%	26%	27%	29%	not extracted
RESULTS: EXPERT 3						
recall (F-score)	77% (0.14)	71% (0.18)	69% (0.18)	68% (0.19)	65% (0.19)	64% (0.20)
precision	8%	10%	11%	11%	11%	12%
rate_valid_terms_in_FP	37%	30%	33%	33%	36%	39%

rate_valid_terms_in_FP = ratio of expert-determined valid terms from false positives (non-matches) over all false positives

Another advantage of the algorithm is that, to an extent, it takes care of low frequency multi-word combinations as long as they have at least one content word that passes the threshold. For example, if the candidate term "max assist" occurs only once in the corpus, but its component "assist" passes the frequency threshold, then "max assist" would be extracted regardless of its own singleton occurrence. However, the algorithm does not extract low frequency ngrams in which there is no single constituent that passes the frequency threshold.

In the current research, we asked the experts to mark the "most specific" term in the corpus. It was obvious that retaining the primitive terms of a compound term is very important. Primitive terms allow us to navigate up an ontology (e.g. "ability"). Compound terms, on the other hand, provide true combinations (e.g. "ability to walk", "ability to bathe independently", "ability to bathe max assist"). They provide the basis for going down and deep into an ontology.

A corollary is that creating test sets for term extraction for ontology development is a complex and use-case specific task. Compound and primitive terms need to be marked in those sets.

The current frequency thresholds are determined empirically. A more formal and more generic way for that will make the algorithm robust to processing bigger corpora. Other metrics besides frequency should be investigated, e.g. log likelihood, mutual information.

Test set 2 had many fragmented sentences that the expert marked as terms, e.g. "Bathes max assist." The proposed algorithm does extract sentences of length 4 and 5. If shorter sentences are also extracted, then a lot more noise will be introduced. Venues to be explored are dynamic linking between such sentences to already extracted terms thus separating them from sentences that are not term candidates. For example, if there is a VP term candidate "bathes max assist" already extracted from a well-formed sentence, then the sentence "Bathes max assist" can be linked to that term and extracted by association, not syntactic node completeness.

An outstanding issue is false positive (noise) reduction. For that, we are planning to experiment with several methods for term ranking once the candidate terms are extracted, e.g. C/NC value[7] which is a derivative of raw frequency. Other approaches are the exploration of the strength of the links within the ngrams in terms of log probabilities.

A limitation of the current study is the relatively small size of the test corpora. The algorithm has yet to be evaluated against larger corpora. However, the manual marking by experts is a time-consuming and expensive task. Another challenge for term retrieval for medical ontology building is how to extract term candidates from forms. Forms are widely used in the medical field and are rich in terms and concepts and could be considered mini-ontologies. Yet another issue is how to deal with distributed terms. However, it must be noted that their occurrence is extremely low.

CONCLUSIONS

Term extraction is the first step in building any ontology. It can be approached in two ways – asking domain experts about what needs to be included, or applying data-driven NLP techniques. The two are not exclusive; rather they are complementary. The paper presented an algorithm for and tests of extracting "the most specific" term candidate from medical text. The precision and recall scores are promising.

ACKNOWLEDGEMENTS

The research was partially supported by grant NLM 07453-01. We would like to thank the experts, Hyeoneui Kim, Linda Herrick and Lori Rhudy, for their patience and expertise in marking the test sets and the four anonymous reviewers for their very helpful comments.

REFERENCES

[1] Evans, D. A. and Zhai, C. 1996. "Noun-phrase analysis in unrestricted text for information retrieval". Proc. ACL, Santa Cruz, U Cal.,17-24

[2] Dagan, I. and Church, K. 1994. "Termight: identifying and translating technical terminology". Proc. 4th Conf. Appl. NLP, 34-40.

[3] Le Moigno, S.; Charlet, J.; Bourigault, D.; Degoulet, P.; Jaulent, M. 2002. Terminology extraction from text to build an ontology in surgical intensive care. Proc. AMIA. pp.430-434.

[4] Bourigault and Jacquemin. 1999. Term extraction and term clustering: an integrated platform for computer-aided terminology. Proc. EACL.

[5] Daille, B. 2001. Qualitative terminology extraction. In "Recent Advances in Computational Terminology". Eds. Bourigault, Jacquemin and L'Homme. John Benjamins Publishing Company. Amsterdam/Philadelphia.

[6] Justeson, J. S. and Katz, S. M. 1995. Technical terminology: some linguistic properties and an algorithm for identification in text. Natural Language Engineering, 1(1) pp. 9-27.

[7] Maynard, D.; Ananiadou, S. 2001. Term extraction using a similarity-based approach. In "Recent Advances in Computational Terminology". Eds. Bourigault, Jacquemin and L'Homme. John Benjamins Publishing Company. Amsterdam /Philadelphia.

[8] Nakagawa, H.; Mori, T. 2000. Automatic term recognition based on statistics of compound noun and its components. Terminology. http://www .r.dl.itc.utokyo.ac.jp/~nakagawa/

[9] Moldovan, D.; Girju, R.; Rus, V. 2000. Domain-specific knowledge acquisition from text. In 6th Applied NLP Conference.

[10] Grobe S.J. 1990. Nursing intervention lexicon and taxonomy study: language and classification methods. Adv Nurs Sci. 1990 Dec.13(2):22-33.

[11] International classification of functioning, disability and health: ICF. WHO. 2001.

[12] Verbruge, L.M. et al. 1994. The disablement process. Social sciences and Medicine. 1-14.

[13] Leidy, N.K. 1994. Functional status and the forward progress of merry-go-rounds: toward a coherent analytical framework. Nursing Research. 43: 196-202.

[14] Charniak, E. 1999. A maximum-entropy inspired parser. Technical Report CS-99-12, Brown University, August.

[15] Jurafsky, D.; Martin, J. 2000. Speech and Language Processing. Prentice Hall, Upper Saddle River, New Jersey. ISBN 0-13-105069-6