

Contribution of Syntactic Functions to Single Word Term Extraction

Xing Zhang¹ and Alex Chengyu Fang²

Dialogue Systems Group
Department of Chinese, Translation and Linguistics, City University of Hong Kong
Tat Chee Avenue, Kowloon, Hong Kong SAR
¹zxing2@student.cityu.edu.hk
²acfang@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 MEDLINE¹ 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
<u>NPPR-AJP+NPHD-N</u> % <u>PC-NP</u> % <u>NPPO-PP</u> % <u>SU-NP</u>	0.199%
<u>NPPR-AJP+NPHD-N</u> % <u>PC-NP</u> % <u>NPPO-PP</u> % <u>PC-NP</u> % <u>NPPO-PP</u> % <u>SU-NP</u>	0.003%
<u>NPPR-AJP+NPHD-N</u> % <u>NPPO-PP</u> % <u>NPPO-PP</u> % <u>SU-NP</u>	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 Automatic Term 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.
- Sciano 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