ARMY RESEARCH LABORATORY

Semi-Automated Methods for Refining a Domain-Specific Terminology Base

by Gabriella Rose, Melissa Holland, Steve Larocca, and Robert Winkler

ARL-RP-0311 **February 2011**

A reprint from the *Volume I: Select Papers*, ARL-TM-2010, pp. 41–56, August 2010.

NOTICES

Disclaimers

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

Citation of manufacturer's or trade names does not constitute an official endorsement or approval of the use thereof.

Destroy this report when it is no longer needed. Do not return it to the originator.

Army Research Laboratory
Adelphi, MD 20783-1197

ARL-RP-0311 **February 2011**

Semi-Automated Methods for Refining a Domain-Specific Terminology Base

Gabriella Rose, Melissa Holland, Steve Larocca, and Robert Winkler
Computational and Information Sciences Directorate, ARL

A reprint from the *Volume I: Select Papers,* ARL-TM-2010, pp. 41–56, August 2010.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED (From - To)
February 2011	Reprint	

4. TITLE AND SUBTITLE
Semi-Automated Methods for Refining a Domain-Specific Terminology Base

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Gabriella Rose, Melissa Holland, Steve Larocca, and Robert Winkler

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
U.S. Army Research Laboratory
ATTN: RDRL-CII-T
2800 Powder Mill Road
Adelphi, MD 20783-1197

8. PERFORMING ORGANIZATION REPORT NUMBER
ARL-RP-0311

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited.

13. SUPPLEMENTARY NOTES
A reprint from the *Volume I: Select Papers*, ARL-TM-2010, pp. 41–56, August 2010.

14. ABSTRACT
A domain-specific term base may be useful not only as a resource for written and oral translation, but also for Natural Language Processing (NLP) applications, text retrieval, document indexing, and other knowledge management tasks. The objective of this investigation was to explore the use of alternative terminology extraction methods to refine and validate an existing military-specific bilingual dictionary. A series of semi-automatic methods was implemented to distill the existing term list by removing redundancies, resolving spelling variations, and separating individual expressions. Once the internal clean-up was completed, we compared two methods drawn from the terminology extraction literature in order to validate terms as military-specific and to propose a candidate list of non-specific terms for exclusion—term frequency calculations and terminology extraction lists. In this investigation, we wanted to find the best procedure to extract domain-specific terms for a low-resource domain; to demonstrate that terminology extraction methods can be used to validate and refine a domain-specific dictionary; and to provide the final, refined dictionary as a term base to support customization of machine translation systems for the military domain.

15. SUBJECT TERMS
Military-Specific Terminology

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Gabriella M. Rose
Unclassified	Unclassified	Unclassified	UU	22	19b. TELEPHONE NUMBER (*Include area code*) (301) 394-5627

Standard Form 298 (Rev. 8/98)
Prescribed by ANSI Std. Z39.18

U.S. Army Research Laboratory

SUMMER RESEARCH TECHNICAL REPORT

Semi-Automated Methods for Refining a Domain-Specific Terminology Base

GABRIELLA ROSE
MELISSA HOLLAND
STEVE LAROCCA
ROBERT WINKLER
MULILINGUAL COMPUTING BRANCH, CISD, ADELPHI

Contents

List of Figures

List of Tables

Abstract

A domain-specific term base may be useful not only as a resource for written and oral translation, but also for Natural Language Processing (NLP) applications, text retrieval, document indexing, and other knowledge management tasks. The objective of this investigation was to explore the use of alternative terminology extraction methods to refine and validate an existing military-specific bilingual dictionary. A series of semi-automatic methods was implemented to distill the existing term list by removing redundancies, resolving spelling variations, and separating individual expressions. Once the internal clean-up was completed, we compared two methods drawn from the terminology extraction literature in order to validate terms as military-specific and to propose a candidate list of non-specific terms for exclusion—term frequency calculations and terminology extraction lists. In this investigation, we wanted to find the best procedure to extract domain-specific terms for a low-resource domain; to demonstrate that terminology extraction methods can be used to validate and refine a domain-specific dictionary; and to provide the final, refined dictionary as a term base to support customization of machine translation systems for the military domain.

1. Introduction

Especially since the 2001 entrance of the United States into the war in Afghanistan, foreign language translation has become increasingly necessary yet still is not sufficiently resourced. Although human translators often provide high-quality work, that work can be costly and time consuming given that it is difficult to find qualified bilingual language experts across all needed domains. This lack of quick translation along with advances in the information technology field has prompted research into and use of semi-automatic machine translation (MT) methods to support human translators. Whereas word-to-word translation in specialized domains may be straightforward (e.g., stethoscope-*estetocopio*) given a language expert or a bilingual dictionary, the difficulty lies with multi-word expressions—with recognizing phrases that are in fact technical terms ("field of fire") and need to be treated as entities, and with finding their counterparts in the other language, where the phrase may or may not have the equivalent number of words.

Over the last 10 years, tools to enable automatic extraction of term bases have been developed, which speed the process of deriving term bases from a collection of documents in a domain of interest. A domain-specific term base may be useful not only as a resource for written and oral translation, but also for Natural Language Processing (NLP) applications, text retrieval (*1*), document indexing, and other knowledge management tasks. The National Virtual Translation Center (NVTC), an organization under the Federal Bureau of Investigation, was established in February 2003 for the exact purpose of "providing timely and accurate translations of foreign intelligence for all elements of the intelligence community (*2*)." In September of that year, an electronic compilation of 8953 terms with their translations was published by M. Green for the NVTC, under the title *Iraqi Military English-Arabic Arabic-English Dictionary*. While the sources of these translated terms and the purpose of the dictionary are unclear, it has been used successfully to support improved MT.

2. Examining the NVTC Bilingual Military Dictionary

Searching through the original term list, we found many internal discrepancies and inconsistencies that suggested that the term base may have been developed by several authors and provided rapidly to the field for urgent needs without opportunity for quality assurance. These internal issues would pose problems with its use in computational linguistics. The problems include the following:

1. Alignment and spacing errors
    a. White space preceding the expression alters its place when ordered alphabetically.
    b. White space trailing the expression can introduce two entries from the same expression:
        i. Example: One entry would be given as “Flank” while the other would be provided as “Flank” and they would have the same Arabic translation.
2. Thirty-three duplicate entries
    a. These entries are exactly the same in both Arabic and English; therefore, the duplicates can be removed.
3. Three variations of the same word
    a. The dictionary would include two non-identical English entries with the identical Arabic translation:
        i. Example: “Light antiaircraft” and “Light anti-aircraft” had the same Arabic translation “مقاومة طائرات خفيفة” and “مقاومة طائرات خفيفة”.
    b. For the purposes of this project, both entries were used, but at the end of the investigation, only the most commonly used, grammatically correct entry was included in the dictionary.
4. Five misspellings
    a. Example: “Airconditioned shelter” should be “Air-conditioned shelter”.
    b. When air-conditioned is listed as its own entry, it has the appropriate spelling, but when combined with another word, it is spelled incorrectly.
5. An unnecessary symbol, ◌ِ, was included after three English entries.
6. For computational linguistic purposes, tokenizations would have to be performed on the following collections: parentheses (622), ampersands (15), and slashes (166). A blank space was inserted where the original character was located.

Arabic experts looked at a random sample of the existing terminology that I proposed as representative and noted that (1) the terminologies were of many cultural dialects, but mainly Standard Modern Arabic, and (2) the Arabic translation of general English words did not have a military-specific connotation, suggesting that the term does not belong in the dictionary. Since we are simply focusing on the English portion of the term base, its bilingual nature does not really enter into the processes used to refine the dictionary at this time. Further research is needed for the Iraqi-Arabic portion.

3. Internal Clean-Up

In order to make the existing term base ready for computer intervention, several changes had to be made (noted in figure 1). Using a Perl script, we found that the original NVTC term base had 8953 entries with the following breakdown:

WPL: 1	AOL: 1832	WPL: 9	AOL: 3
WPL: 2	AOL: 4795	WPL: 10	AOL: 3
WPL: 3	AOL: 1591	WPL: 11	AOL: 3
WPL: 4	AOL: 440	WPL: 13	AOL: 2
WPL: 5	AOL: 182	WPL: 14	AOL: 1
WPL: 6	AOL: 83	WPL: 16	AOL: 2
WPL: 7	AOL: 24	WPL: 18	AOL: 1
WPL: 8	AOL: 15	WPL: 19	AOL: 1

WPL: Words per Line

AOL: Amount of Lines

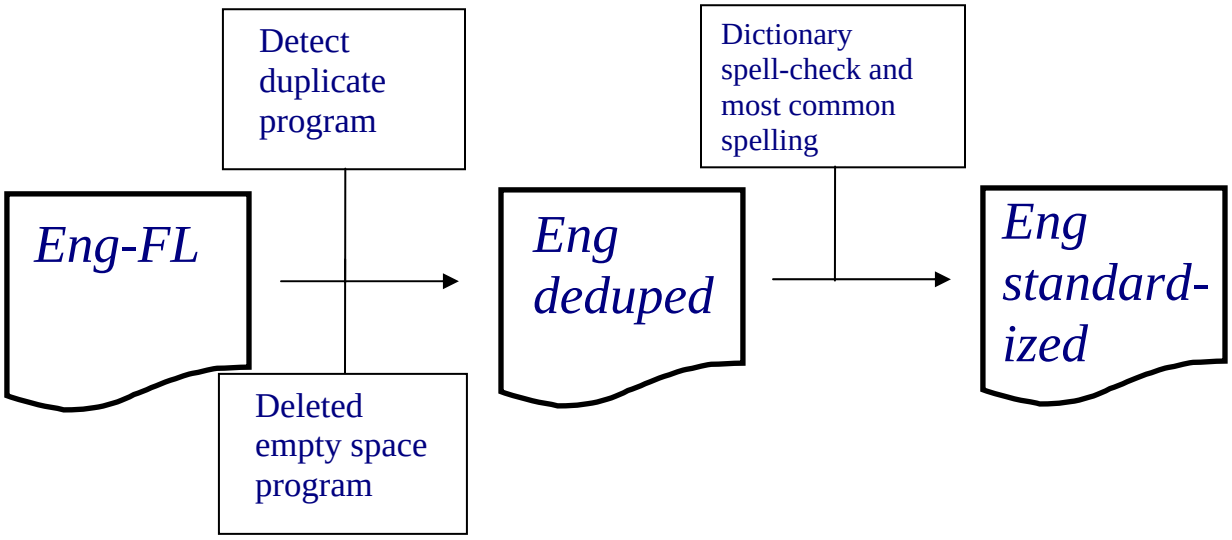

Figure 1. Internal correction process.

Once we became familiar with the term base, we determined that it had to be altered in order to accurately process the material. The list of problems identified in the introduction was used to refine existing text. First, the terms were alphabetized. Entries that had unnecessary preceding

white space were fixed. Microsoft Office was unable to remove trailing white spaces, so Perl was used for this purpose. The code removed all white space after each string in the text file and replaced the new entry in the dictionary.

Once the alignment and spacing errors were corrected, both a Perl script and Conditional Formatting within Microsoft Excel were used to identify all exact matches within the column of terms. Both methods identify a total of 331 duplicates in the English portion. Taking the entire dictionary into context, there were 33 duplicate entries (some entries were found three separate times); therefore, 37 entries were removed.

In response to the variations among words in the dictionary, we decided to include both entries to find the most common spelling in order to eliminate one of the entries later in the project. Misspellings were then corrected to help reinforce standardization of the term base. We also removed the unnecessary symbol following three of the entries.

Entries with two separate terms combined and submitted as one entry were noted (i.e., antiaircraft/artillery, director/directorate). These submissions should be separated into two entries for the purpose of accessibility in the field, and in our term frequency method, exact string matching is essential for accurate results. Therefore, all entries with gratuitous explanations and definitions following the term were removed. A Microsoft Excel macro was employed to eliminate all items within parentheses.

Once these alterations were completed, the new term base consisted of the following breakdown:

WPL: 1	AOL: 1832	WPL: 9	AOL: 3
WPL: 2	AOL: 4795	WPL: 10	AOL: 3
WPL: 3	AOL: 1591	WPL: 11	AOL: 3
WPL: 4	AOL: 440	WPL: 13	AOL: 2
WPL: 5	AOL: 182	WPL: 14	AOL: 1
WPL: 6	AOL: 83	WPL: 16	AOL: 2
WPL: 7	AOL: 24	WPL: 18	AOL: 1
WPL: 8	AOL: 15	WPL: 19	AOL: 1

WPL: Words per Line

AOL: Amount of Lines

4. Method One: Frequency Count

This proposed method to collection a set of domain-specific terminology is based on the principle of Term Frequency-Inverse Document Frequency (TF-IDF). As tested in *An Unsupervised Approach to Domain-Specific Term Extraction* (*3*), the principle behind frequency counting is the idea that certain terminology will generally occur with a higher frequency within domain-specific documents as opposed to in a general corpus. This theory, however, has its limitations. Single word terminology is much more difficult to access based on the occurrences of homographs. In the NVTC's dictionary for example, the entry "brief" could be found in several different contexts. In a military sense, the term can be used as a verb to summarize or give preparatory information to Soldiers, but in a general connotation, it could be used as an adjective or noun to describe duration and length.

4.1 Input

A domain-specific corpus of 2619 documents was then created by collecting various military documents from a variety of sources. The documents selected were chosen because of their translated nature; if a document was important enough to military use that it was translated into Arabic, then its extracted terminology is most likely vital to a bilingual dictionary. Thirteen items from the Ranger Handbook, one item from field manual 3-21.10, and five items from field manual 7-8 were selected, along with 93 documents from the Combating Terrorism Center's Harmony Database of Released Documents (CTC) and 2507 items from an Iraqi database from ARL's holdings. The CTC at West Point, dedicated to scholarly research and policy analysis to examine combat terrorism, published a series of letters, reports, and al-Qa'ida-related documents captured during the War on Terror for public access. This is important to our corpus as a first-hand account of events in Afghanistan, elucidating al-Qa'ida's actions and weaknesses. The Iraqi training material consists of PowerPoint training materials, scripts, and guides to a variety of field situations.

4.2 Output

The goal of this method was to take the internally cleaned dictionary and use exact string matching to search through the corpus for the number of occurrences of each term. Because of the extensive nature of the corpus, we used a Hadoop cluster, a programming framework designed for large-scale computational use, to expedite the process. Before processing the data, all the documents (Acrobat Reader, Microsoft Word, Microsoft Excel, and Microsoft PowerPoint) were converted into text files with the help of an online converter. The Iraqi training documents could not be easily converted, however, because of the high number of subfolders

within each main folder. Again using Perl, we renamed all documents, changing spaces to dashes and ampersands to underscores, and moved all documents to one large folder, which helped ease the conversion of the files.

Once all target files were converted, they were processed with the servers searching for exact string matches based on the dictionary's terms. The process resulted in two Excel files summarizing the findings. The first, "Word Count" (table 1), was a list of all keywords, the number of occurrences in the corpus, and on average how many times that keyword appeared per document. The second file, "Doc Count" (table 2), consisted of a list of each document, the number of key words in the document, and the average number of times a keyword appeared.

Table 1. Word Count chart excerpt.

Term	No. of Times Term Appears in Corpus
Map reconnaissance	16
Fallout	16
Psychological warfare	16
Stud	16
Barrel assembly	16
Medical unit	15

Table 2. Doc Count chart excerpt.

Document	No. of Terms in Dictionary that Appear in Corpus
Iraqi-Training-Disk_S3_MOUT_ Infantry-Rifleman-Course-Handout-Booklet-2003.txt	462
Iraqi-Training-Disk ca-documents_ instant-lessons-of-iraq-war.txt	458
AFGP-2002-600092-Trans-Meta.txt	448
Iraqi-Training-Disk_ca-documents_SASO-handbook.txt	434
AFGP-2002-600088-Trans-Meta.txt	371
AFGP-2002-600053-Trans-Meta.txt	361

The results from the TF method indicated that the most common terms were as follows:

One-word entries	Enemy	8622
	Support	5254
	Commander	4889
	Operations	4874
Two-word entries	First aid	448
	Armed forces	389
	Indirect fire	340
	Warning order	316
Three-word entries	Course of action	364
	Command and control	310
	Chain of command	306
	Concept of operations	216

The results support Zipf's Law (*4*) that term length is inversely proportional to its number of occurrences in a corpus. Zipf's Law will become an important factor in the term extraction process. We found 29.68% of all terms in the dictionary with a frequency of one or more in the corpus and 26.13% of those appeared more than once.

5. Method Two: Terminology Extraction

The goal of terminology mining or extraction is to collect a list of domain-pertinent terms from a given corpus. For the purposes of this investigation, the online extraction tool TermExtractor (*5*), developed by the Linguistic Computing Laboratory of the University of Roma, was used to determine what percentage of the extracted term list overlapped with the existing military bank.

The terms that appear in both corpora are then added to a proposed list of confirmed dictionary entries. Figure 2 shows the TermExtractor pipeline.

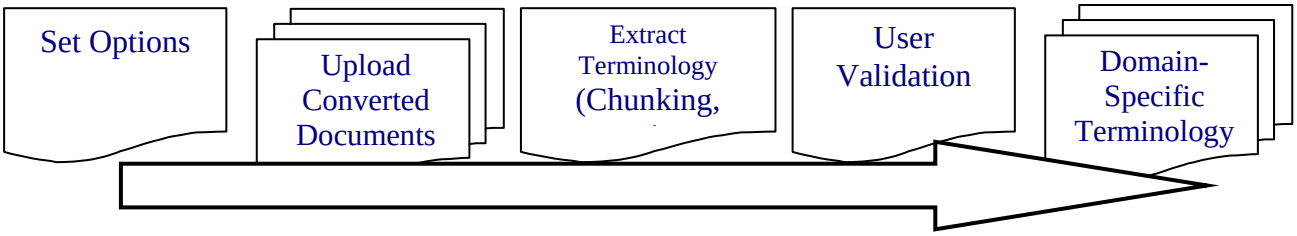

Figure 2. TermExtractor pipeline (*5*).

To ensure consistency in our results, we used the same corpus as a reference throughout the entire project. We submitted the same corpus of 2619 documents as in the TF method to be processed for specificity. TermExtractor uses input documentation to extract statistically relevant terminology through the use of chuncking and document parsing, as well as by filtering unecessary information. These filters eliminate stopwords such as "the, as, is, for" and general terminology that does not indicate domain-specificity. The extraction tool filters non-terminological strings through its evalution of the following:

- **Domain Pertinence**: High (numerical value) means a term is frequent in the domain of interest and is much less frequent in the other domains used for contrast (*6*):

  $DRDi\ (t) = -\sum P\hat{}(t/dk)\ \log(P\hat{}(t/dk)) = \sum norm\_freq(t,dk) \log(norm\_freq(t,dk))$

- **Lexical Cohesion:** The degree to which the terms adhere to one another within a string. This proved more effective than other measures of cohesion (*6*). The resulting numerical value is high if the words within a string occur more often with one another rather than alone in a corpus. The minimum was set to 0.05.

- **Structural Relevance:** When a title or subtitle is composed of domain-specific terms, then its importance is increased by some factor *x*. Highlighted, bolded, and italicized items are also included *(x*=5 for highlighted, capitalized, underlined, colored, smallcaps, italicized, and bolded terms, *and x*=10 for titles and abstract content*)*.

- **Miscellaneous:** A set of heuristics are applied to increase computational performance by removing generic articles and terminology, detecting misspellings, distinguishing part of speech, extracting unigram terminology, and detecting abbreviations.

The extraction tool also sets up contrastive corpora to eliminate common terminology that may be relevant to the specific domain but not entirely of that domain. These corpora include the following:

- Brown Corpus (3634 terms)
- Medicine (2281 terms)
- Computer Networks (16335 terms)
- Sports (1020 terms)

- Tourism (55590 terms)
- Wall Street Journal—Economy (3606 terms)

Although these terminology banks are not specifically indentified, it is important to set up some contrasting corpora to eliminate general terminology and possibly create a proposed list of terms for expulsion.

5.1 First Investigation

In the first investigation, the corpus was submitted without any restrictive measures to find the percentage of extracted terminology that would overlap with the existing term bank. Given Zipf's Law (*4*), the frequency distribution of word length is exponential; this means that, in accordance with a general corpus, a unigram (one word term) is far more likely to occur than a bigram and a trigram, and so forth. Due to time constraints, this law was employed, so any term that exceeded three words was considered domain-specific because of its exclusivity to a particular domain. For all one- to three-word terms, 3605 words occurred in both the term extraction list and the NVTC dictionary. This indicates that 40.27% of the dictionary is supported by this method; 43.87% of all unigrams, bigrams, and trigrams.

5.2 Second Investigation

For the second investigation, we entered the corpus and entered the existing term bank as a restrictive option. The extracted terminology from this trial excludes all terms in the dictionary in its proposed terminology list. At this point in the process, a human validator is required to identify the reliability of the extracted list. I randomly sampled 10% of the terms (648 items) and a subject matter expert evaluated this list, indicating whether the term was military-unique (18.06% of the sample) and highlighting the spelling errors (24.07%). Table 3 is an excerpt of the described process, with its proposed spelling corrections in column four.

Table 3. Methods comparison to dictionary.

Term	Military Specific	Spelling Error	Possible Correction
improvised sling	Yes		
include-ytank crewmembers	Yes	Yes	"including tank crewmembers"
includingthe regulationsandlaws		Yes	"including the regulations and laws"
indecision recklessness			
index contour line	Yes		

This list will be used later as a basis for what could be added to the dictionary. In order to refine the extracted list of terms, the same course of action can be taken as for the NVTC dictionary. The possible list of terms can be evaluated for its frequency in a new corpus and a new list of terms can be extracted and compared for its similarities.

6. Results

Although time constraints did not allow the full investigation to be executed, the original term base can be successfully modified and refined after comparing the dictionary with a general corpus and using IDF. The first portion of figure 3 indictates the overlap between the orignial NVTC dictionary and the results of the two methods. It appears that the TF method produces a better comparison to refining an existing military term base, but the term extraction method contributed as well. The second portion of figure 3 indicates the overlap between the TF method and the term extraction method.

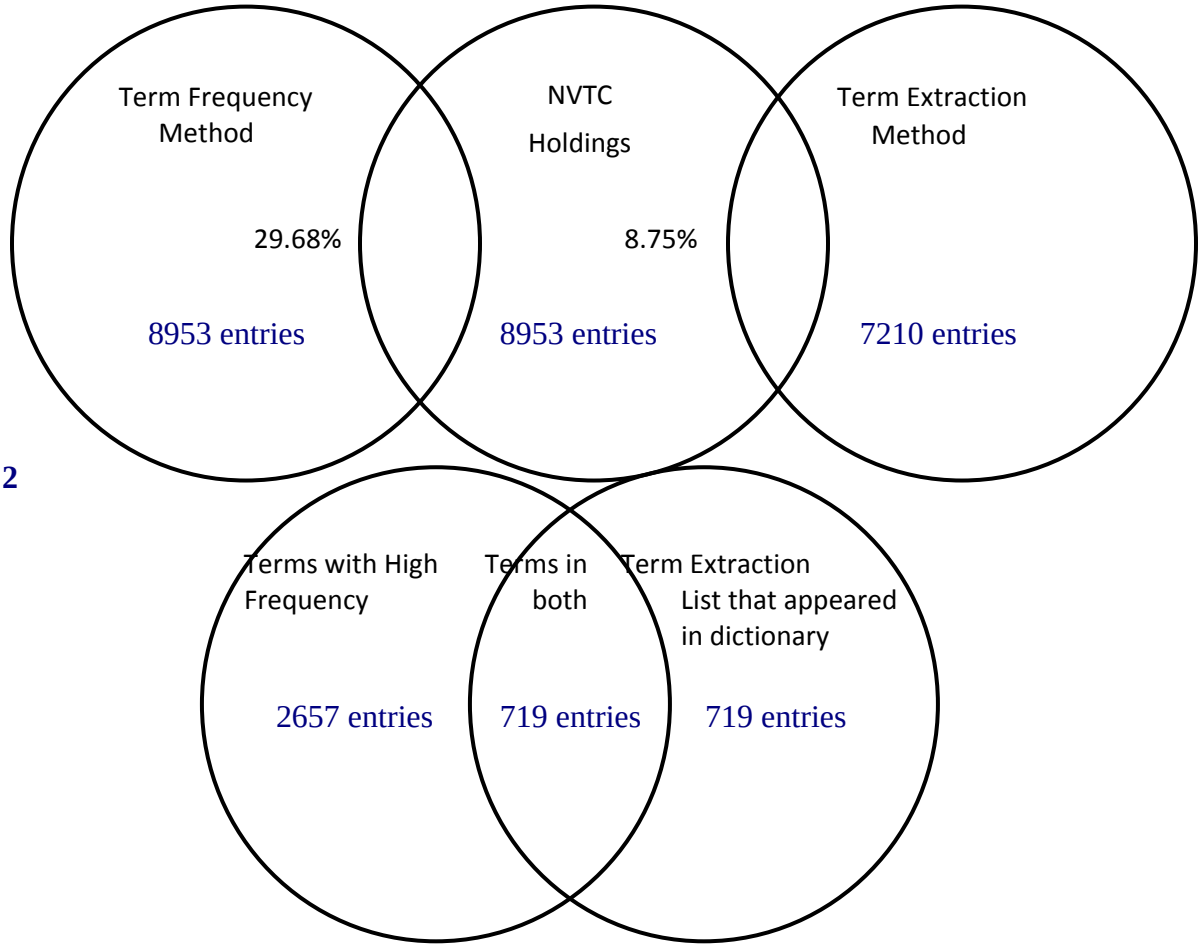

Figure 3. Comparison to dictionary.

In this study, 27.06% of terms that appeared with high frequency also appeared in the term extraction list.

In addition to assessing the term frequency of the dictionary when paired with a military-specific corpus, we also would like to compare the dictionary with a general corpus, such as English GigaWord. This process would not validate terms, but rather would propose a possible list for exclusion. By processing the dictionary with a general corpus, we would be able to eliminate general terms, but also single-word terms that occur frequently in both a general corpus and a military-corpus. These unigrams must be verfied with a human ground truth because of the appearance of homographs, as mentioned earlier.

The third proposed method that we plan to execute following this paper is IDF. The problem with TF measurements is that all documents and expressions are considered equally important in terms of assessing relevancy. IDF works to solve this problem along with TF by statistically identifying how important a word is to a corpus. If the TF-IDF is high, it indicates a rare term; it is considered low when terms occur frequently.

7. Conclusion

As of the moment, we have 46.70% of the dictionary accounted for as a result of the TF/term extraction methods, as well as a portion dedicated to Zipf's Law (8.27%). After all the previously mentioned methods have been executed, we hope to have a refined, efficient dictionary that will be useful in the field as well as for more computational research.

8. References

1. Avancini, H.; Lavelli, A.; Magnini, B.; Sebastiani, F.; Zanoli, R. Expanding Domain-Specific Lexicons by Term Categorization, *Proceedings of SAC*, Melbourne, FL, 2003.
2. Jordan, Everette E. Congressional Testimony. Federal Bureau of Investigation, 25 Jan 2007. [ONLINE]. http://www.fbi.gov/congress/congress07/jordan012507.htm (accessed 23 Jun 2010).
3. Kim, S. N.; Baldwin, T.; Kan, M.-Y. An Unsupervised Approach to Domain-Specific Term Extraction, ALTA Workshop, 2009.
4. Pierce, J. R. *Introduction to Information Theory: Symbols, Signals, and Noise*, 2nd rev. ed.; New York: Dover, 1980, pp 86–87, 238–239.
5. Sclano, F.; Velardi, P. TermExtractor: a Web Application to Learn the Common Terminology of Interest Groups and Research Communities. *9th Conf. on Terminology and Artificial Intelligence TIA 2007*, Sophia Antipolis, France, October 2007.
6. Park, Y.; R. J. Byrd, R. J.; Boguraev, B. K. Automatic glossaryextraction: Beyond terminology identification. *Proceedings of the 19th International Conference on Computational Linguistics*, Taipei, Taiwan, 26–30 August 2002, 772–778, Association for Computational Linguistics (ACL), http://www.aclweb.org/anthology/C/C02/C02-1142.pdf.

NO. OF COPIES	ORGANIZATION
1 ELEC	ADMNSTR DEFNS TECHL INFO CTR ATTN DTIC OCP 8725 JOHN J KINGMAN RD STE 0944 FT BELVOIR VA 22060-6218
1	US ARMY RSRCH LAB ATTN RDRL CIM G T LANDFRIED BLDG 4600 ABERDEEN PROVING GROUND MD 21005-5066
10	US ARMY RSRCH LAB ATTN IMNE ALC HRR MAIL & RECORDS MGMT ATTN RDRL CII B R WINKLER ATTN RDRL CII T S LAROCCA ATTN RDRL CII T V M HOLLAND (5 HCS) ATTN RDRL CIM L TECHL LIB ATTN RDRL CIM P TECHL PUB ADELPHI MD 20783-1197

TOTAL: 12 (1 ELEC, 11 HCS)

INTENTIONALLY LEFT BLANK.