LAMP-TR-092
CS-TR-4399
UMIACS-TR-2002-81

August 2002

# Domain-Specific Term-List Expansion Using Existing Linguistic Resources

Bonnie Dorr, Tiejun Zhao

Language and Media Processing Labratory
Instititue for Advanced Computer Studies
College Park, MD 20742

## Abstract

This report describes a series of experiments involving expansion of a domain-specific human-generated "seed list" using available linguistic resources. The resources used for the expansion are intended to be general purpose: two large-scale Chinese-English dictionaries and a Chinese lexical knowledge base (HowNet). The methodology involves three steps: (1) hand extraction of head words from each entry in the human-generated seed list; (2) automatic comparison of these head words against entries in the linguistic resources-where an entry matches if the head word matches the entry exactly or is included in its the semantic definition; and (3) collection of any resulting matching entries into a larger term list. The terms extracted by this process were verified manually to confirm whether they were relevant to the topic of a specific domain. An important contribution of this work is the finding that the use of a bilingual term list for the expansion process does not provide a significant improvement over the use of a simpler, more easily produced, monolingual term list.

***The support of the LAMP Technical Report Series and the partial support of this research by the National Science Foundation under grant EIA0130422 and the Department of Defense under contract MDA9049-C6-1250 is gratefully acknowledged.

# Report Documentation Page

*Form Approved*
*OMB No. 0704-0188*

| | | |
|---|---|---|
| 1. REPORT DATE<br>**AUG 2002** | 2. REPORT TYPE | 3. DATES COVERED<br>**00-00-2002 to 00-00-2002** |

4. TITLE AND SUBTITLE
**Domain-Specific Term-List Expansion Using Existing Linguistic Resources**

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
**Institute for Advanced Computer Studies,Language and Media Processing Laboratory,College Park,MD,20742**

8. PERFORMING ORGANIZATION REPORT NUMBER

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

14. ABSTRACT
**see report**

15. SUBJECT TERMS

| 16. SECURITY CLASSIFICATION OF: | | | 17. LIMITATION OF ABSTRACT | 18. NUMBER OF PAGES | 19a. NAME OF RESPONSIBLE PERSON |
|---|---|---|---|---|---|
| a. REPORT<br>**unclassified** | b. ABSTRACT<br>**unclassified** | c. THIS PAGE<br>**unclassified** | **Same as Report (SAR)** | **12** | |

# Domain-Specific Term-List Expansion Using Existing Linguistic Resources

Bonnie Dorr and Tiejun Zhao
CLIP, UMIACS
September, 2002

## ABSTRACT

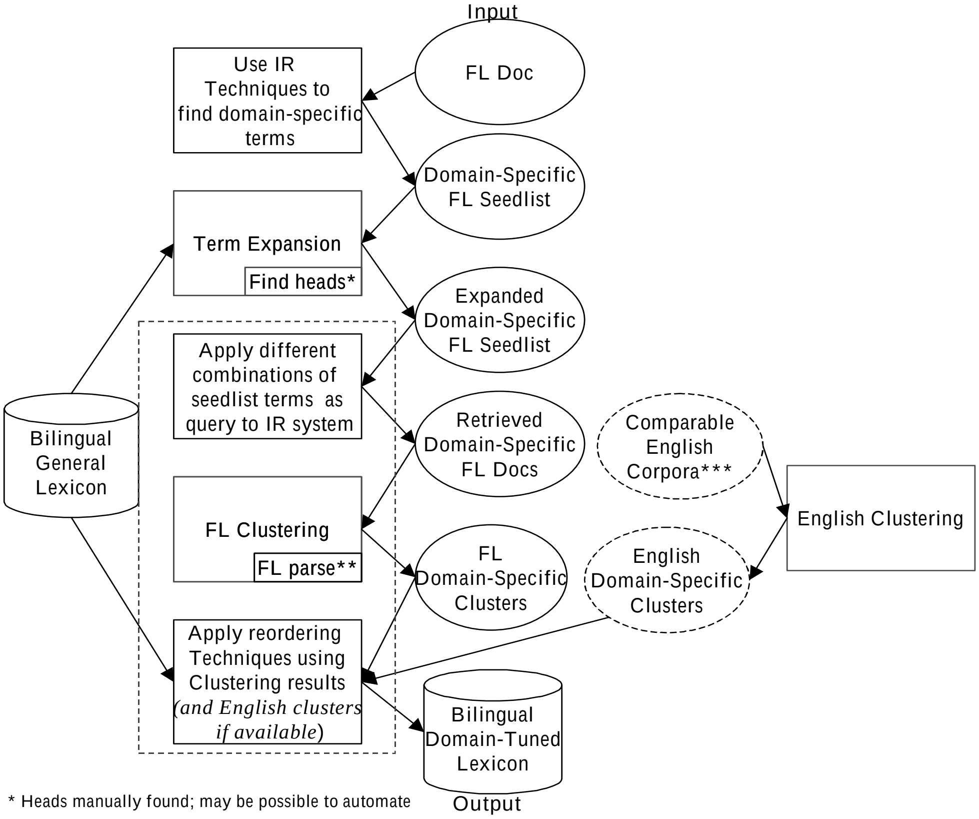
This report describes a series of experiments involving expansion of a domain-specific human-generated “seed list” using available linguistic resources. The resources used for the expansion are intended to be general purpose: two large-scale Chinese-English dictionaries and a Chinese lexical knowledge base (HowNet). The methodology involves three steps: (1) hand extraction of head words from each entry in the human-generated seed list; (2) automatic comparison of these head words against entries in the linguistic resources—where an entry matches if the head word matches the entry exactly or is included in its the semantic definition; and (3) collection of any resulting matching entries into a larger term list. The terms extracted by this process were verified manually to confirm whether they were relevant to the topic of a specific domain. An important contribution of this work is the finding that the use of a bilingual term list for the expansion process does not provide a significant improvement over the use of a simpler, more easily produced, monolingual term list.

## 1 Introduction

This report describes a series of experiments involving expansion of a domain-specific human-generated term list using available linguistic resources. The resources used for the expansion are intended to be general purpose: two large-scale Chinese-English dictionaries and a Chinese lexical knowledge base (HowNet). The methodology involves three steps: (1) hand extraction of head words from each entry in the human-generated seed set; (2) automatic comparison of these head words against entries in the linguistic resources—where an entry matches if the head word matches the entry exactly or is included in its the semantic definition; and (3) collection of any resulting matching entries into a larger term list. The terms extracted by this process were verified manually to confirm whether they were relevant to the topic of a specific domain.

An important contribution of this work is the finding that the use of a bilingual term list for the expansion process does not provide a significant improvement over the use of a simpler, more easily produced, monolingual term list. This finding is critical, given that our ultimate goal is to produce the “seed list” automatically from a monolingual input document—using automatic IR techniques rather than human labor—as part of a larger translation process. Our approach is to enhance existing general-purpose lexicons using domain-specific knowledge that is automatically detected from the words of the input document.

Figure 1 shows our overall plan for domain-tuning a general bilingual lexicon. Implemented boxes are outlined in red. In the current phase of our project (outside of the blue dashed lines), we assume the existence of a very small domain-specific foreign-language (FL) seed list, or a single large document from which such a seed list may be automatically extracted (possibly using monolingual or comparable corpora). We also assume the existence of a bilingual general lexicon to which we will ultimately apply the domain-tuning technique.


Input
FL Doc
Use IR Techniques to find domain-specific terms
Domain-Specific FL Seedlist
Term Expansion
Find heads*
Expanded Domain-Specific FL Seedlist
Apply different combinations of seedlist terms as query to IR system
Retrieved Domain-Specific FL Docs
Comparable English Corpora***
English Clustering
Bilingual General Lexicon
FL Clustering
FL parse**
FL Domain-Specific Clusters
English Domain-Specific Clusters
Apply reordering Techniques using Clustering results (and English clusters if available)
Bilingual Domain-Tuned Lexicon
Output

* Heads manually found; may be possible to automate

** Clustering can be done with or without parse

*** Possibly produced using STRAND-like s/w (?)

Figure 1: Proposal for Domain-Tuning of Bilingual Lexicon

In the next phase of this project (inside the blue dashed lines), we will build the domain-tuning module consisting of three components: (1) retrieval of a large set of monolingual (Chinese) documents using different combinations of the expanded domain-specific terms as a query; (2) application of a clustering algorithm to the retrieved monolingual document set; (3) reordering of English translations in each Chinese-English entry of our bilingual lexicon based on the clustering results (and possibly clustering from comparable English corpora, if available).

The first paper (to our knowledge) in which automatic domain-specific tuning of lexicons was implemented was that of Resnik and Melamed (1997). However, this work presupposes the existence of a very large parallel text (bitext) in the source and target language. More recent work (Chang et al., 2002) makes use of an existing (large) resource of pre-established domain tags—the Far East Dictionary—and is strictly monolingual in its application (enhancing WordNet entries with domain-specific tags).

In our approach, we produce a domain-tuned lexicon based on a bilingual lexicon and other monolingual or comparable corpora—but not necessarily parallel corpora. Although we are currently investigating the Chinese-English language pair only, we expect the techniques described herein to be applicable to other language pairs, provided there exists a general bilingual dictionary for those pairs.

In the next section, we describe the linguistic resources used in this phase of the project. Following this, we outline the techniques used in the term expansion process. Finally, we provide experimental results, an analysis, and a discussion of the results.

## 2 Linguistic Resources

It is frequently the case that general-purpose linguistic resources are more accessible than domain-specific lexicons. It is natural for us to make use of these available linguistic resources while gathering domain-specific data that can be used for lexicon tuning—e.g., prioritizing the translations of foreign-language terms according to their relevance to a particular domain. Our current goal is to expand a human-generated "seed list" of domain-specific terms through a comparison of head words against general-purpose Chinese-English dictionaries; ultimately, the expanded list will be used for document retrieval, clustering, and prioritization of English translations in each Chinese-English entry.

In our experiments, the general-purpose linguistic resources include two large-scale Chinese-English dictionaries and one Chinese lexical knowledge base with English translations. Figure 2 shows the characteristics of each resource. The first dictionary, CETA1, is the original Optilex dictionary obtained from MRM corporation. The second dictionary, CETA2, is the UMD parsed version of CETA1, but with additional usages.[1]

[1] CETA2 includes Chinese-English entries from both Optilex (using 20 sources extracted by John Kovarik, DoD) and the LDC Chinese-English bilingual term list v. 1.3 (IIRC). UMD performed subsequent cleanup of the file to remove punctuation and excessively verbose translations. In each Chinese-English entry, the English translations are ordered according to unigram frequency (without POS distinctions) in the Brown Corpus: First, single-word translations are organized in decreasing order of frequency; next, multiword translations are listed; and, finally, single-word translations with zero frequency in the Brown Corpus

An example of the distinction between CETA1 and CETA2 is that CETA2 includes more lexical variants, such as “syndicalism,” and “syndicalist”, where CETA1 contains “syndicalism”. In addition, CETA2 omits the grammatical categories and Chinese codes from the original dictionary; since these were not relevant to our study, we take CETA2 to be the more comprehensive dictionary. The third resource is HowNet, a Chinese-English semantic resource constructed by Dr. Zhendong Dong and colleagues (Dong and Dong, 2000).

| No. | File Name | Size | Contents |
|---|---|---|---|
| CETA1 | ceta.3col | 233,367 | Chinese words/English translations; some grammatical categories Chinese character codes |
| CETA2 | newest.lex.ordered | 341,366 | Chinese words/English translations |
| HowNet | hownet.txt | 116,533 | Chinese words/English translations; Chinese/English grammar categories; Semantic definitions |

Figure 2: Characteristics of Linguistic Resources

The human-generated domain-specific “seed list” contains 126 Chinese-English word pairs (file name: ChemWeaponsTermList.txt). Some terms in this list are inordinately long and essentially phrases unto themselves. Consider the following examples:

(2.1)
关于禁止在战争中使用窒息性、毒性或其他气体和细菌作战方法的议定书
Protocol for the Prohibition of the Use in War of Asphyxiating, Poisonous or Other Gases, and of Bacteriological Methods of Warfare

(2.2)
烷基（甲基、乙基、正丙基或异丙基）硫代膦酸烷基（氢或少于或等于 10 个碳原子的碳链，包括环烷基）-S-2-二烷（甲、乙、正丙或异丙）氨基乙酯及相应烷基化盐或质子化盐
O-Alkyl (H or <C10, incl. cycloalkyl) S-2-dialkyl(Me, Et, n-Pr or i-Pr)-aminoethyl alkyl (Me, Et, n-Pr or i-Pr) phosphonothiolates and corresponding alkylated or protonated salts

There are 17 terms (13.5% of the human-generated seed list) that are longer than 10 Chinese characters (phrases in English translation are slightly shorter). Among the other terms, 46 (36.5%) are names of special chemical products that have 4 or more Chinese

are listed. This resource was used for Chinese-English MT in a previous project (Dorr et al., 2002), but without the domain-specific lexicon tuning that we are investigating for the current project.

characters. If we were to apply an exact match algorithm, half of the terms in this “seed list” would not be found in our linguistic resources and our resulting list would not be complete enough to serve as a query set for our domain—many relevant texts would not be found in later document-retrieval experiments. Thus, our goal is to perform term expansion such that the overall recall is increased, but without significantly reducing precision.

## 3 Expansion Process

We now describe how we expanded the 126-entry seed list. Our approach is to add similar terms to the list using the following two techniques: (1) Match head words extracted from the original list; (2) Match semantic definitions of these head words. We refer to the former as head-word matching and the latter as semantic matching. If a term occurring in our general-purpose dictionaries matches a head word or a semantic definition associated with a head word, it will be added to the expanded set. Figure 3 illustrates the entire process of term expansion using the resources listed in Section 2.

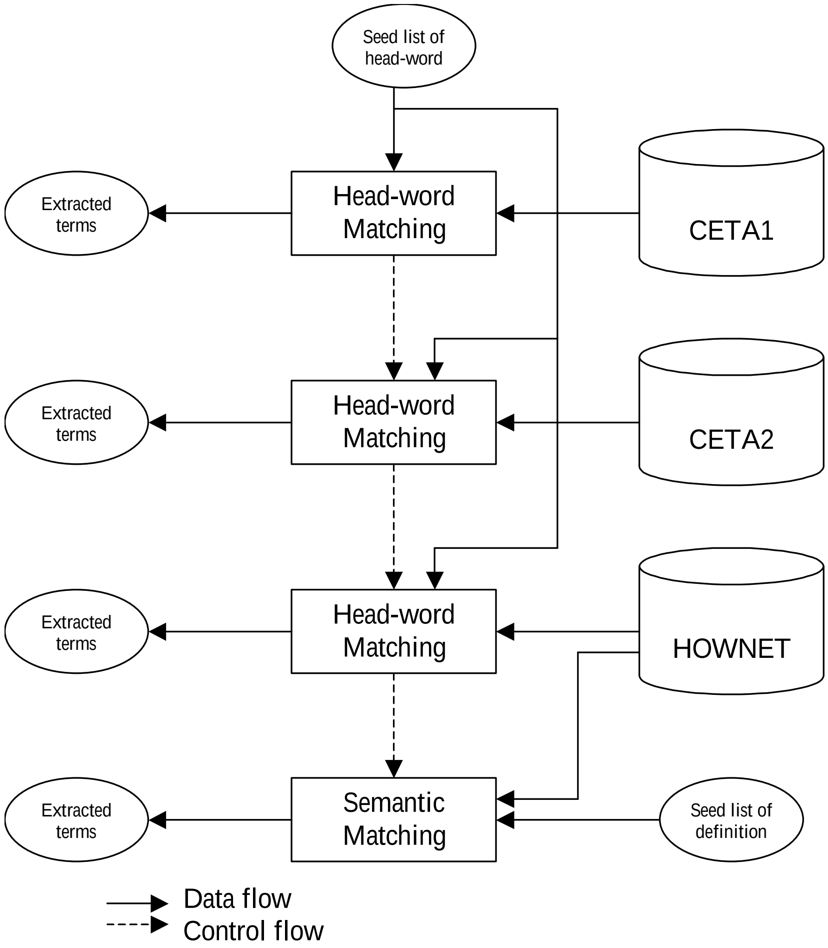


Figure 3: Term expansion process

Extraction of head words was done by hand, taking 35 minutes by a native Chinese speaker. The head word is the central noun (or verb) in a term or phrase. A total of 81 head-word Chinese/English pairs were extracted from the original seed list. The lengths of head-word pairs are no greater than 4 characters in Chinese; all English counterparts contain only one word.

Semantic definitions are taken from the Chinese-English HowNet database. However, not every seed list head is found in HowNet; only 49 head words are associated with semantic definitions. Moreover, some head words have more than one definition. The total number of definitions is 69, a few of which are not useful for term expansion because of their irrelevance to the specific domain. After hand-deleting irrelevant definitions, the final list of heads and associated definitions consists of 40 head words and 52 definitions.

Once the expanded term list is generated, a manual check is applied. A "duplicate check" is applied during each matching process, i.e., extracted terms are compared to the terms in the existing term set (and its current expansion) so that duplicates are not added.

## 4 Results of Experiments and Analysis

As mentioned above, the three linguistic resources used for our experiments are CETA1, CETA2, and HowNet. We apply a word-matching process to the two CETA dictionaries and we use two types of matching on the HowNet database. Four expanded lists are produced by this process. Following the expansion, we conducted a manual check, thus producing a set of "purified results," i.e., those terms that are judged to be in the domain of interest. Two additional sets of terms are described below.

Figure 4 displays the results of the experiments. CETA2* refers to the results of matching both Chinese head words and English head words. This can be compared with the two lines above it (CETA1 and CETA2), where only Chinese head words are matched.[2] HowNet+ refers to the result of matching semantic definitions before non-relevant definitions are deleted. Note that the precision in HowNet+ was the lowest of all the experiments (.092). Human inspection by a native speaker established the irrelevance of almost all 1,445 terms—except those corresponding to the 134 terms already found in the second HowNet experiment. Section 5 analyzes the CETA* and HowNet+ expansions further. It is clear from the table that the best results were obtained from CETA2, with a precision of .811 and a recall of .967.

[2] All HowNet expansions also used only the Chinese head words. The two types of matching used for HowNet are word matching, where the seed list heads are compared to all Chinese lexemes in HowNet; and semantic matching, where the semantic definitions whose lexemes are in the seed list are compared to semantic definitions in other HowNet entries.

| Lexicon | CETA1 | CETA2 | CETA2* | HowNet | HowNet | HowNet+ |
|---|---|---|---|---|---|---|
| Method | Match Chinese word only | Match Chinese word only | Match Chinese word & English translation | Match Chinese word only | Match sem definition | Match sem definition (incl irrelevant seedlist terms) |
| Runtime | 4m 52s | 6m 59s | 17m 51s | 2m 28s | 17s | 19s |
| New extractions | 841 | 5,381 | 6,583 | 36 | 701 | 1,445 |
| Cum. Extractions | 841 | 6,222<br>(841+5,381) | 7,424<br>(841+6,583) | 6,258<br>(36+6,222) | 6,959<br>(701+6,258) | 7,703<br>(1,445+6,258) |
| New purified results | 614 | 4,364 | 4,514 | 16 | 134 | 134 |
| Cum. purified results | 614 | 4,978<br>(614+4,364) | 5,128<br>(614+4,514) | 4,994<br>(16+4,978) | 5,128<br>(134+4,994) | 5128<br>(134+4,994) |
| Purify time (1 native Chinese speaker) | 25m | 30m | 1h 45m | 2m | 17m | 20m |
| Precision | .730<br>(614/841) | .811<br>(4364/5381) | .686<br>(4514/6583) | .444<br>(16/36) | .191<br>(134/701) | .092<br>(134/1445) |
| Recall | 1.0<br>(614/614) | .967<br>(4364/4514) | 1.0<br>(4514/4514) | 1.0<br>(16/16) | 1.0<br>(134/134) | 1.0<br>(134/134) |

Figure 4: Results of Experiments

Note that when terms are extracted, further “translation merging” is required, since it is often the case that a Chinese term occurs with more than one English translation in our resources. In such cases, the lines are automatically merged prior to matching.

## 5 Discussion

During expansion, we observed two interesting phenomena. First, using the English head-word for matching in the CETA* experiment introduces a significant quantity of noise in the extracted result. Second, manual deletion of irrelevant definitions—9 “seed list” heads—results in a 50% noise reduction between the two HowNet experiments involving matching of semantic definitions.

The first phenomenon is illustrated in Figure 5. Observe that there is an increase of more than 1,000 terms—22.3%—in the Chinese-English case (CETA2*) over the Chinese-only case (CETA2). However, most of these terms are relevant to the domain. Only 150 terms in the additional 1000 are suitable (3.4% more than CETA2). If we use only the Chinese head word, the precision increases from .6 (CETA2*) to .8 (CETA2) with only a slight drop in recall (.03). Given that the drop is so small, only the Chinese head word is used in the other experiments (CETA1 and HowNet).

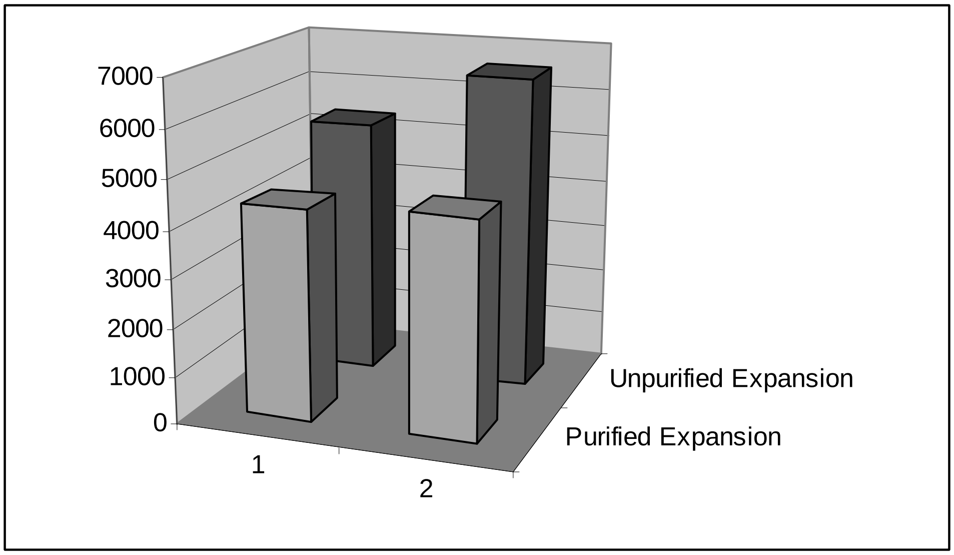


Figure 5: Comparison of expansion for CETA2 and CETA2*: Column 1 refers to the number of entries found in CETA2 (Chinese only) and Column 2 refers to the number of entries found in CETA2* (Chinese and English). Although the unpurified CETA2* expansion is 22.3% higher than the unpurified CETA2 expansion, the purified CETA2* expansion is only 3.4% higher than the purified CETA2 expansion. This result indicates that the use of English in the matching has a very low return for a significant increase in noise.

As for the second phenomenon, we found that semantic matching in our HowNet experiments gave rise to a significant amount of noise due to Chinese-English "translation fanout."[2] Such cases arise when a English|Chinese semantic definition in HowNet is associated with a Chinese seed-list head word and also a different entry that contains: (1) a domain-irrelevant Chinese word; or (2) an ancient Chinese word.

[2] The semantic matching process is based on a comparison between semantic definitions in HowNet, where each semantic definition is indicated by a "DEF=" symbol followed by a English|Chinese pair, such as sorrowful|悲哀. Because of the pairing of English and Chinese as a part of the semantic definition, the lexical entries in HowNet exhibit what has become known as "translation fanout" in standard bilingual lexicons. The ambiguity introduced by this fanout gives rise to the extraction of many irrelevant entries in the expansion process.

An example of the first case is shown in (5.1), where the seedlist head word 酸 is associated with the semantic definition “sorrowful|悲哀” in HowNet, which also appears in irrelevant entries such as 哀 (grieve) and 哀怨 (sad).

(5.1) (irrelevant)
Head Word: 酸……, DEF=sorrowful|悲哀
Extracted: 哀 grieve DEF=sorrowful|悲哀
……
哀怨 sad DEF=sorrowful|悲哀

This second case occurs when the semantic definitions of head words are too general, i.e., they express an “IS-A” relationship between the words and their definitions. Noise is introduced when these definitions are so general that they encompass too many terms. In addition, multiple semantic senses give rise to spurious definitions.

An example of this case is shown in (5.2), where the seedlist head word 化学 is associated with the semantic definition “knowledge|知识” in HowNet, which is too general a definition, showing up in entries that contain ancient Chinese words such as 八卦 (Eight Diagrams) and 中庸 (golden mean).

(5.2) (too general)
Head Word: 化学 DEF= knowledge|知识
Extracted: 八卦 Eight Diagrams DEF=knowledge|知识
……
中庸 golden mean DEF=knowledge|知识

Our HowNet+ experiment extracted a total of 135 words and phrases that were too general for the particular domain. To address this, we deleted 9 incorrect head definitions from the seed list, inducing a 50% reduction in extracted results, where all rejected terms were found to be irrelevant to the domain (as determined through human inspection).

## 6 Analysis of Deleted Words

This section presents a simple analysis of the human “purification” process, allowing us to characterize a standard for selecting or rejecting terms such that future automation may be possible. For this portion of our work, we present only the data rejected from the CETA1 experiment, which we take to be a representative sample.

Figure 6 shows a summary of this analysis on deleted words. The category “Proper noun” includes those words or phrases that are proper names of entities or events and “Special phrase in China” indicates special phrases only used in China (remember our dictionaries are Chinese-English lexicons). “Irrelevant Adj”, “Irrelevant N”, and “Irrelevant V” refer to common adjectives, nouns and verbs, respectively, that are not relevant to the domain. Finally, words with inappropriate format contain formatting errors. The reason these terms occurred in our original result set is that we employed a "relaxed matching" procedure that allowed for phrasal matching across certain (potentially erroneous) terms.

| Category | Amount (total 227) | Examples |
|---|---|---|
| Proper noun | 20 | 百年战争 hundred years war<br>中法战争 <pnt> the franco-chinese war |
| Special phrase in China | 29 | 长期共存互相监督 <ncu> long-term co-existence and mutual surveillance<br>中国人民抗日军事政治大学 <pn> the chinese people's resist-japan military and political university |
| Irrelevant Adj. | 32 | 不凉不酸儿 unconcernedly<br>穷酸臭美 <id> affected |
| Irrelevant N. | 93 | 生物模型 living mold<br>酸白菜 pickled celery cabbage<br>战争火坑 <ag> fiery furnace of war |
| Irrelevant V. | 15 | 拈酸吃醋 <id> to be jealous<br>腿酸脚麻 to become numb in the legs ; to be exhausted |
| Inappropriate format | 38 | 禁止记者◆访 <ag> barred to the press<br>作战技研 术 fighting technique, operational technique |

Figure 6: Analysis of deleted words (CETA1)

## 7 Conclusions and Future Research

This report describes the process and results of experiments involving domain-specific term expansion. Given a set of human-generated head words as a "seed list", we applied two expansion methods, head-word matching and semantic matching, to extract those relevant terms from available Chinese-English dictionaries. We then checked the expanded list by hand. An important contribution of this work is the finding that the use of a bilingual term list for the expansion process does not provide a significant improvement over the use of a simpler, more easily extracted, monolingual term list.

Our next step is to use the expanded list for further research in determining the optimal prioritization of English translations in each Chinese-English entry. We will also use the list to retrieve documents relevant to the domain for clustering and additional domain-tuning of our lexicons. Finally, our ultimate goal is to produce a "seed list" automatically using IR techniques based on terms in a document that is to be translated from Chinese to

English. We believe the experiments reported herein are the first step toward processing the resulting list, once it is extracted; that is, we intend to apply iterative bootstrapping, generating the "seed list" automatically—and then using the techniques described in this document for expansion and further information retrieval of documents that can assist us in domain-tuning our lexicons.

## Acknowledgements

This work has been supported by Mitre Contract 010418-771.

## References

[1] Bonnie Dorr, Gina-Anne Levow, Dekang Lin, "Construction of a Chinese-English Verb Lexicon for Embedded Machine Translation in Cross-Language Information Retrieval," to appear in Machine Translation, Special Issue on Embedded MT, 2002. ftp://ftp.umiacs.umd.edu/pub/bonnie/mtj-2002a.pdf

[2] Echa Chang, Chu-Ren Huang, Sue-Jin Ker, Chang-Hua Yang. 2002. "Induction of Classification from Lexicon Expansion: Assigning Domain Tags to WordNet Entries" in Proceedings of the First International WordNet Conference, Karnataka, India; Also: Poster at "SemaNet'02: Building and Using Semantic Networks," Workshop at COLING-2002, Taipei, Taiwan.

[3] Zhendong Dong and Qiang Dong. 2000. "HowNet Chinese-English Conceptual Database," http://www.keenage.com/zhiwang/e_zhiwang.html/, Online Software Database, released at the 38th Annual Meeting of the ACL (ACL-2000), Hong Kong.

[4] Philip Resnik and I. Dan Melamed. 1997. "Semi-Automatic Acquisition of Domain-Specific Translation Lexicons" in Proceedings of the 5th ANLP Conference.