Domain Tuning of Bilingual Lexicons for MT

Necip Fazıl Ayan, Bonnie J. Dorr, Okan Kolak
Institute for Advanced Computer Studies &
Department of Computer Science
University of Maryland
College Park, 20742, USA
{nfa, bonnie, okan }@umiacs.umd.edu

Domain Tuning of Bilingual Lexicons for MT

Abstract

Our overall objective is to translate a domain-specific document in a foreign language (in this case, Chinese) to English. Using automatically induced domain-specific, comparable documents and language-independent clustering, we apply domain-tuning techniques to a bilingual lexicon for downstream translation of the input document to English. We will describe our domain-tuning technique and demonstrate its effectiveness by comparing our results to manually constructed domain-specific vocabulary. Our coverage/accuracy experiments indicate that domain-tuned lexicons achieve 88% precision and 66% recall. We also ran a Bleu experiment to compare our domain-tuned version to its un-tuned counterpart in an IBM-style MT system. Our domain-tuned lexicons brought about an improvement in the Bleu scores: 9.4% higher than a system trained on a uniformly-weighted dictionary and 275% higher than a system trained on no dictionary at all.

1 Introduction

Knowledge of *domain-specific vocabulary*—a set of words or terms from a document that indicate the topic or primary content of the text—is necessary for many NLP tasks. In monolingual processing, domain specificity is a key issue in the retrieval of relevant documents from large document collections: the degree of domain specificity impacts the accuracy of text classification (Sakurai, 1999). In multilingual processing, appropriate translation choices cannot be made without knowledge of domain-specific meaning distinctions (Ahmad, 1995).

To address this need, several researchers have applied domain-tuning procedures to bilingual lexicons. However, those who have investigated techniques for automatic acquisition of bilingual terms do not distinguish between domain-specific and general terms, thus reporting relatively low accuracy for extraction of domain-specific terminology: 40% in (Dagan and Church, 1994), 70% in (Daille, 1994), and 73% in (Smadja et al., 1996). More recently, researchers have developed approaches that achieve higher accuracy—but these rely heavily on the preexistence of large domain-specific resources such as sentence-aligned parallel corpora (Resnik and Melamed, 1997; Melamed, 1997), hierarchically organized thesauri (Hulth et al., 2001), and preestablished domain tags (Chang et al., 2002). These resources are generally difficult to construct for a given language pair in a particular domain.

In this paper, we investigate the effectiveness of a new approach to automatic domain-tuning of lexicons for translation a foreign-language (FL) document. We do not presuppose the existence of large domain-specific resources, but instead require only: (1) the FL input document; (2) a general bilingual lexicon; and (3) a general-purpose clustering algorithm. Although we are currently investigating the Chinese-English language pair, we expect the techniques described herein to be applicable to other language pairs (and other domains), provided there ex-

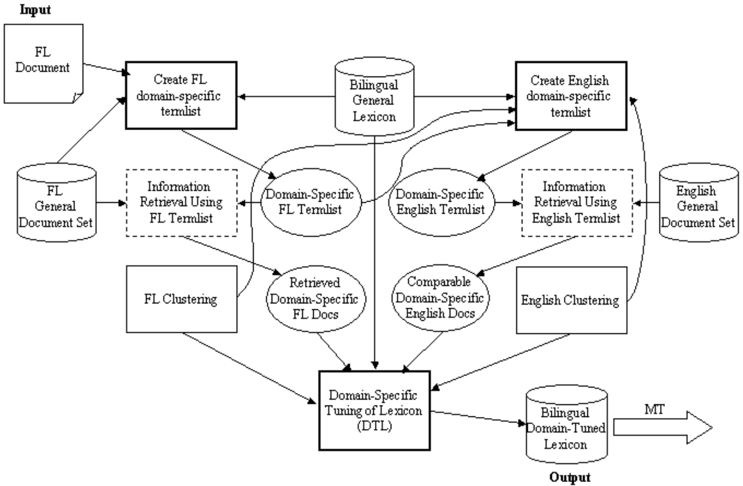
Input
FL Document
Create FL domain-specific termlist
Bilingual General Lexicon
Create English domain-specific termlist
FL General Document Set
Information Retrieval Using FL Termlist
Domain-Specific FL Termlist
Domain-Specific English Termlist
Information Retrieval Using English Termlist
English General Document Set
FL Clustering
Retrieved Domain-Specific FL Docs
Comparable Domain-Specific English Docs
English Clustering
Domain-Specific Tuning of Lexicon (DTL)
Bilingual Domain-Tuned Lexicon
MT
Output

Figure 1: Overall Domain-Tuning Design

ists a general bilingual dictionary for those pairs.

Figure 1 illustrates the overall design of our alternative approach. Our three implemented components are indicated as heavy borders. We borrow language-independent clustering software (LaTaT) to produce word clusters for the two languages (Pantel and Lin, 2002).[1] We also assume the existence of an IR system to produce the comparable, domain-specific documents from a set of automatically-extracted query terms.

The entire process consists of two phases. The first (roughly the top half of Figure 1) builds the resources necessary for the domain-tuning process. This phase includes work that is closely related to research in cross-language information retrieval (Davis and Dunning, 1995), (Oard, 1997). We start with a FL input document for which we desire a translation. From this, we produce a set of domain-specific query terms for the foreign language using standard tf.idf techniques.[2] These query terms (along with the bilingual lexicon and general clusters) are fed into the process that produces the English domain-specific query terms. The foreign-language and English terms serve as input to information retrieval, which must produce comparable, domain-specific documents in each language.

The second phase (roughly the bottom half of Figure 1) transforms the general bilingual lexicon into a domain-tuned lexicon (DTL) for translating the input document. This phase is also closely related to research in cross-language information retrieval, most notably, in its use of techniques that are analogous to *query expansion* (Ballesteros and Croft, 1997) for handling words that are not found in the comparable-document set.

While our ultimate goal is to translate a document from a foreign language (currently Chinese) into English, the emphasis of this paper is on the domain-tuning component—the second phase—of the overall process. We will describe and evaluate our approach to domain-tuning of a bilingual lexicon. The next section describes the algorithm (and its variants). After this, we demonstrate the effectiveness of our approach by comparing our results to manually constructed domain-specific vocabulary. Finally, we use Bleu to compare our domain-tuned version to its un-tuned counterpart in an IBM-style MT system.

2 Domain-Tuning Algorithm

Our domain-tuning algorithm (and its extensions) rely on the pre-existence of the following resources:

1. A bilingual lexicon for L_1 (the foreign language) and L_2 (English): Each word in L_1 is listed with one or more translations in L_2.
2. A set of word clusters in each of the two languages.
3. A set of comparable, domain-specific documents in both languages.

As mentioned above, the clusters are produced by a general, language-independent clustering algorithm. The comparable, domain-specific documents are expected to be automatically induced by applying information-retrieval techniques to the input document. As a stand-in for the unincorporated IR component, we use a human-verified set of comparable, domain-specific documents in the two languages.[3]

We now give a formal description of the Domain Tuning Algorithm (DTA).

[1] The corpus used in the Chinese clustering was 800MB of Chinese novels from a web site (www.mypcera.com). The English clusters were created using the AQUAINT corpus from the TREC QA track in 2002, which contains 3GB of newspaper text. The English clustering includes 2243 clusters. The Chinese version includes 1316 clusters. A specific word may be a member of more than one cluster. The Chinese clusters were created without parsing; the English ones used parsing. However, an experiment run by Dekang Lin (pc.) indicates that parsing has a very small, insignificant impact on the effectiveness of word-similarity determination for clustering.

[2] An evaluation of the domain-specific effectiveness of our automatically extracted query terms is reported elsewhere—we omit the citation here to preserve anonymity. For a general description of the well-known tf.idf technique, see (Manning and Schütze, 1999).

[3] We used 528 English documents and 352 Chinese documents from the domain of interest. Unfortunately, because there were no links between documents in this comparable corpus, we treated the document set in each language as one large document and assumed each one was comparable to the other. In this case, every word in one language is assumed to be in a comparable document with each and every word in the other language. We expect better performance with smaller, individually comparable documents, perhaps extracted using web-mining tools such as STRAND (Resnik, 1999).

2.1 Formal Description of DTA

The DTA takes $\{c_1, \ldots, c_n\}$ and $\{e_1, \ldots, e_m\}$ to be Chinese and English words, respectively. The following predicates/functions are used to describe the algorithm in Figure 2:

1. $comparable(c_i, e_j)$: TRUE if c_i and e_j are in comparable documents (i.e. if there is at least one Chinese document D_c and English document D_e such that D_c contains c_i, D_e contains e_j, and D_c and D_e are comparable to each other); FALSE otherwise.
2. $same\_c(e_i, e_j)$: TRUE if e_i and e_j are in the same cluster, FALSE otherwise.
3. $conf(c_i, e_j)$: Indicates the confidence of c_i and e_j as translational equivalents in a particular domain. Initially, all confidence values are set to 0.

```
For each Chinese word $c$ in the bilingual lexicon
  Let $T = \{e_1, e_2, \ldots, e_n\}$, i.e., the translations of $c$
  For each $e_i \in T$
    Case 1: $comparable(c, e_i)$.
      Set $conf(c, e_i) = 2$.
    Case 2a: $\exists e_x : comparable(c, e_x) \wedge same\_c(e_i, e_x)$.
      Set $conf(c, e_i) = 1$.
    Case 2b: $\exists c_x : comparable(c_x, e_i) \wedge same\_c(c, c_x)$.
          Set $conf(c, e_i) = 1$.
    Case 3: Neither case 1 nor case 2 applies.
      Set $conf(c, e_i) = 0$.
```

Figure 2: Domain Tuning Algorithm (DTA)

For each word in L_1 (henceforth, Chinese) the algorithm attempts to assign a confidence value to each translation in L_2 (henceforth, English) using the comparable-document set and word clusters. The confidence value assigned by the algorithm depends primarily on the occurrence of a word and its translation in the set of comparable documents. Thus, the algorithm relies most heavily on the comparability of a Chinese term and its English translation—but some weight is also given for comparability between terms that appear in the same cluster. While assigning the confidence values, each step is taken to be *applicable* or *not applicable*. If it is applicable, no account is taken of the number of documents to which it applies.

The final step is to normalize the confidence values assigned by the algorithm. For this purpose, the confidence values are mapped to a weight between 0 and 1 such that the sum of the weights for all English translations of a Chinese word is equal to 1.

2.2 Extensions

The domain-tuning algorithm outlined in Figure 2 has certain deficiencies. In this section, we examine issues concerning the handling of: (1) Multi-word (phrasal) translations; (2) Missing words; and (3) Stopwords.

2.2.1 Multi-Word Translations

The most critical issue regarding the algorithm is its inability to reliably assign confidence values to multiple word (or phrasal) translations. The comparable documents are indexed on single words; thus, if a Chinese word c has a phrasal translation $[e_1 e_2 \ldots e_n]$, that translation will not be found in the English document (which means Case 1 fails). Even when a multi-word translation occurs in the comparable documents, it is ranked lower than a single-word translation that occurs in the comparable documents. This means that the highest possible rank of a multi-word English translation is given only when the multi-word translation is in a cluster with another (non-phrasal) word that appears in the comparable documents—case 2 of the algorithm above. An informal inspection of the comparable documents reveals an abundance of multi-word translations relevant to the domain, e.g., *nuclear bomb* or *chemical treaty*.[4]

To overcome this, we implemented a sub-phrase matching mechanism which assigns a confidence value to a multi-word translation $[e_1 e_2 \ldots e_n]$ of a Chinese word c as follows:

1. For each English word e_i in the multi-word translation, assign a confidence value to (c, e_i) using the algorithm in Figure 2.
2. Take the average of all $conf(c, e_i)$'s to assign an overall confidence value to the translation $[e_1 e_2 \ldots e_n]$.

In our evaluation, we examined variants of our algorithm where sub-phrase matching is turned on and turned off. If the sub-phrase matching is turned off, all multi-word translations are treated as if they were single words.

2.2.2 Missing Words

Because our comparable-document set is not likely to include every relevant Chinese and English word in our bilingual lexicon, we are faced with the standard "word not found" problem raised frequently in the cross-language information retrieval community. Our solution is to use a technique we call *translational expansion*—analogous to the *query expansion* used in cross-language information retrieval (Ballesteros and Croft, 1997). We implement this as a second pass over the lexicon, whereby we find relevant entries that were missed during the

[4] Our test domain was nuclear-biological-chemical weapons.

"first-pass domain-tuning." The general idea is to find the highest rank assigned to each translation and use that rank for other occurrences of the translation in the lexicon—even if that translation is associated with a Chinese-English pair that does not occur in the comparable documents.

There are two different approaches to this translational expansion:

1. **Expand Zero Score Translations (ExpZero)**: Apply expansion only to translations that were assigned a zero score in the first pass.
2. **Expand All Translations (ExpAll)**: Apply expansion to all translations processed in the first pass.

Expansion is designed to assign the highest possible rank associated with a translation to every occurrence of that translation. We apply expansion prior to normalization of the confidence scores to avoid spurious effects of other ranked translations on an individual score. If the sub-phrase matching is turned on, sub-phrases are treated accordingly: rather than computing $conf(c, e_i)$ for each individual word in a particular multi-word translation $[e_1e_2 \ldots e_n]$, the highest first-pass score associated with each e_i is used to compute the average of all $conf(c, e_i)$'s.

2.2.3 Stopwords

Since the objective of the domain-tuning algorithm is to identify the words that are specific to the given domain, it is worthwhile to test out a variant of the algorithm where stopwords are ignored in the dictionary for the purpose of ranking. In our evaluations, we examine the impact of inclusion or exclusion of the stopwords during the lexicon generation.

3 Experimental Set-Up

We generated 12 different DTLs. For each one, we changed three parameters of the algorithm: sub-phrase matching or not, inclusion of stopwords or not, and translational expansion (one of two different variants) or not. Table 1 lists the settings for all 12 lexicons, DTL 1 through DTL 12.

Each entry in the lexicon consists of a Chinese word and its translations, where each translation is accompanied by a confidence value. The percentage of the Chinese words with at least one non-zero score translation is nearly 10% for all lexicons, among 200K Chinese words or phrases. Figure 3 shows a sample entry for the first 6 DTLs to illustrate the format of the lexicons.

Lexicon	Include Stopwords	Sub-phrase Matching	Translational Expansion
DTL 1	No	No	None
DTL 2	No	No	ExpZero
DTL 3	No	No	ExpAll
DTL 4	No	Yes	None
DTL 5	No	Yes	ExpZero
DTL 6	No	Yes	ExpAll
DTL 7	Yes	No	None
DTL 8	Yes	No	ExpZero
DTL 9	Yes	No	ExpAll
DTL 10	Yes	Yes	None
DTL 11	Yes	Yes	ExpZero
DTL 12	Yes	Yes	ExpAll

Table 1: Settings for 12 DTLs

DTL 1: 乙醇 [ethanol:0.00] [ethyl alcohol:0.00]
DTL 2: 乙醇 [ethanol:1.00] [ethyl alcohol:0.00]
DTL 3: 乙醇 [ethanol:1.00] [ethyl alcohol:0.00]
DTL 4: 乙醇 [ethanol:0.00] [ethyl alcohol:0.00]
DTL 5: 乙醇 [ethanol:0.50] [ethyl alcohol:0.50]
DTL 6: 乙醇 [ethanol:0.50] [ethyl alcohol:0.50]

Figure 3: A Sample Entry from 6 DTLs

4 Experiments and Results

For measuring the effectiveness of our domain-tuning algorithm, we conducted two different experiments: (1) We compared the coverage and accuracy of our DTLs against a gold-standard—using standard information-retrieval metrics (e.g., *recall* and *precision*); (2) We compared the result of our lexicon-enhanced MT model against un-tuned versions in an IBM-style MT system—using Bleu (Papineni et al., 2002).

4.1 Evaluation of Lexicon Coverage and Accuracy

In the first experiment, our purpose was to observe the quality of the generated lexicons by comparing some subset of them against a human-produced ground truth. All experiments were done using our domain-tuned Chinese-English lexicons. The same comparison may be applied to any FL-English pair, without having any knowledge of the foreign language.

4.1.1 The Gold Standard

The gold standard is a subset of the lexicon where each entry was human-judged for relevance to the domain. An English translation of a Chinese word is annotated *positive* (+) if it is one of the most possible translations of that word in the given domain.

Otherwise, it is a *negative* (-) instance. For the experiments, we take the corresponding set of words from the DTL and compare them, pairwise, against the gold standard.

We generated two different ground-truth sets by two human subjects. The subjects were native English speakers and the task was to identify whether a translation was a *positive* or *negative* instance among 2244 English translations.[5] The 2244 English translations were extracted from Chinese-English entries containing at least one English translation known to be relevant to the domain.[6] We generated the union of these two ground-truth sets as follows:[7]

1. If both annotators assign *positive* to an English translation, the resulting annotation is *positive*.
2. Otherwise, if either annotator assigns *maybe* to an English translation, the resulting annotation is *maybe*.
3. Otherwise, the resulting annotation is *negative*.

	Ground Truth-1	**Ground Truth-2**	**Union**
Positive	313	273	402
Negative	1690	1853	1578
Maybe	241	118	264
Total	2244	2244	2244

Table 2: Number of Instances in Ground-Truth Sets

The number of *positive*, *negative*, and *maybe* instances their union is given in Table 2. The agreement ratio between the two annotators using pairwise comparison is:

1. 79.77%, if the agreement is on an exact match of labels (Positive-positive, negative-negative, and maybe-maybe).
2. 93.27%, if maybe is a dummy label (which matches *positive* or *negative*).

4.1.2 Evaluation Metrics

We evaluated accuracy and coverage using precision, recall, the averaged precision and recall (f-measure)[8], and "correctness." Precision is the ratio of the number of correctly identified *positive* instances to the number of all instances identified. Recall is the ratio of the number of *positive* instances identified correctly to the number of *positive* instances in the ground truth. Correctness takes into account *negative* instances, i.e., it is the ratio of the number of correctly identified *positive* and *negative* instances to the total number of instances identified.

Threshold	**Precision**	**Recall**	**F-Measure**	**Correct**
Variable	80.00	48.35	60.27	48.21
0.1	82.88	50.55	62.80	51.34
0.2	86.84	36.26	51.16	43.75
0.3	87.27	26.37	40.51	37.05
0.4	85.00	18.68	30.63	31.25
0.5	88.89	17.58	29.36	31.25

Table 3: Evaluation Results for Different Thresholds for DTL 1

4.1.3 Results of Coverage/Accuracy Evaluation

To compare the DTLs to the ground-truth set, we need to transform confidence values into a measure that reflects the notion of positivity/negativity. The simplest way to do this is to use a threshold for confidence values, whereby all translations with a confidence value higher than the threshold are taken as *positive* instances. In our experiments, we demonstrate the impact of different algorithmic variants by presenting different threshold values and measuring the quality of the lexicons using the metrics. In addition to fixed threshold values (0.1, 0.5, etc.), we also apply a variable threshold value for each word depending on the number of translations associated with the word. In this case, the threshold is set to $1/n$ where n is the number of translations of the word evaluated. This will be shown as *Variable* in our result tables.

We compared all the entries in the termlist constructed by the domain expert, using the corresponding part of the lexicon. The translations that were marked as *maybe* in the ground-truth set were assumed to be *positive* for the evaluation results presented below. For all the results, we include multi-word translations in the calculation of precision, recall, f-measure and the correctness.

To illustrate the effect of different thresholds, we present the precision, recall, f-measure and correctness values using different thresholds for only DTL 1 in Table 3. All other DTLs exhibit similar behavior: f-measure and correctness results begin to drop drastically for thresholds greater than 0.1. Thus, in the remainder of this paper, we will use only the variable and fixed (0.1) thresholds.

Table 4 presents the results for all 12 DTLs. The ones in **boldface** are the best for the given settings. For the variable threshold, the precision is between 80% and 85.48% and DTL 5 (with subphrase matching and translational expansion) gives

[5] Subjects were allowed to mark the translations as *maybe* when they are not sure about the label.

[6] The relevant English translations were manually generated independently by a different native English domain expert.

[7] This version of ground truth is intended to be an approximation to post-annotation inter-annotator discussion, which traditionally results in agreement.

[8] The f-measure = $\frac{2 \times Precision \times Recall}{Precision + Recall}$.

the best results; recall is in the range of 48.35%–63.19%. DTL 11—which incorporates sub-phrase matching, translational expansion, and stopwords—scored highest for f-measure and correctness.

Surprisingly, the inclusion of stopwords in the lexicon generation does not degrade the results: in some cases, there is a slight increase; in others, a slight decrease. With a fixed threshold of 0.1, the results for precision, recall and f-measure increase by 2–3%. The best precision (87.88% for DTL 12), recall (66.48% for DTL 6), and f-measure (75% for DTL 11) are achieved when sub-phrase matching and translational expansion are used. The correctness increases up to 8% (DTL 12). Overall, the results indicate that DTLs provide the information necessary to distinguish domain-specific vocabulary from other words.

4.2 Machine Translation Evaluation

We incorporated the DTLs into an IBM-style statistical machine translation framework (Brown et al., 1990); we then evaluated the results using Bleu.

4.2.1 MT System

A statistical MT system has 3 basic components, a language model, a translation model, and a decoder. The language model is a monolingual component that characterizes only the target language. Our language model is trained on the (parallel) Hong Kong News[9] using the CMU-Cambridge Toolkit (Clarkson and Rosenfeld, 1997). The translation model, which bridges the source and target languages, is trained by GIZA++ (Och and Ney, 2000) on different DTLs. Since GIZA++ cannot accommodate a DTL directly, we designed a mechanism to incorporate each DTL into the translation model. The decoder generates and ranks translation candidates using the language and translation models; we used the ReWrite decoder by ISI (Marcu and Germann, 2001).

We translated 155 lines of a domain-specific input document which we refer to as the "Chem Treaty." All the modules were identical across all experiments, with the exception of the translation model, which was trained by incorporating information from each of the DTLs in independent experiments, as explained in the following section. We performed a Bleu evaluation (Papineni et al., 2002) using the NIST MT Evaluation Toolkit (Doddington, 2002).[10]

4.2.2 Incorporation of DTLs into the Translation Model

Our approach to incorporating DTLs into the translation model is to append 0 or more copies of each lexicon pair to the training data. The number of copies inserted for each pair is an indication of the importance of that translation pair to the domain, i.e., a high confidence value for a pair dictates a high number of appended copies of the pair. We picked a fixed number of entries, N, to be appended to the training data for each Chinese word in the DTL. Consider this example:

c_1 (e_1:0.60) (e_2:0.40) (e_3:0.0)
c_2 (e_4:0.0) (e_5:0.0)

If we take $N = 10$, then we add (c_1,e_1) 6 times and (c_1,e_2) 4 times to the training data. We performed another set of experiments where we accommodated translations with zero weight: (1) If all translations of a Chinese word are zero-weighted, each one is added N/X times, where X is the number of translations for that word; (2) If only some of the entries are zero-weighted, each zero-weighted entry is added once to the training data and the remaining translations are distributed proportionally to their confidence values. In the example above, this scheme would add (c_1,e_1) 6 times, (c_1,e_2) 4 times, (c_1,e_3) 1 time, (c_2,e_4) 5 times, and (c_2,e_5) 5 times to the training data. In the experiments reported below, we used $N = 10$. Once the initial set of experiments were completed, we experimented with different N values to investigate its impact.

4.2.3 Results of MT Evaluation

Table 5 presents the Bleu scores for our 12 DTLs, using training data both with and without zero-weighted entries. From these results, we see that including zero-weighted entries improves the scores nearly 100% when stopwords are ignored; the difference is smaller when stopwords are used. We also see that either kind of expansion improves the scores by 5–17% when stopwords are not used (DTL's 2,3,5,6). Finally, the inclusion of stopwords (the last 6 DTL's) leads to an improvement of up to 100%

[9] Available from LDC at http://www.ldc.upenn.edu/.

[10] Because of there is only 1 reference translation per sentence (for a total of 155), the scores are lower than would be the case if we had multiple translations of each sentence, as has been acknowledged previously (Doddington, 2002). However, the Bleu score indication of relative effectiveness of different systems; thus, we are interested not in the magnitude of the scores, but in their relative values.

Lexicon	Precision		Recall		F-Measure		Correctness	
	Var.	**T=0.1**	**Var.**	**T=0.1**	**Var.**	**T=0.1**	**Var.**	**T=0.1**
DTL 1	80.00	82.88	48.35	50.55	60.27	62.80	48.21	51.34
DTL 2	81.36	84.03	52.75	54.95	64.00	66.45	51.79	54.91
DTL 3	81.36	84.03	52.75	54.95	64.00	66.45	51.79	54.91
DTL 4	83.93	82.95	51.65	58.79	63.95	68.81	52.68	56.70
DTL 5	**85.48**	84.51	58.24	65.93	69.28	74.07	58.04	62.50
DTL 6	84.92	85.82	58.79	**66.48**	69.48	74.92	58.04	63.84
DTL 7	80.73	82.14	48.35	50.55	60.48	62.59	48.66	50.89
DTL 8	82.05	83.33	52.75	54.95	64.21	66.23	52.23	54.46
DTL 9	82.05	83.33	52.75	54.95	64.21	66.23	52.23	54.46
DTL 10	83.06	85.60	56.59	58.79	67.32	69.71	55.36	58.48
DTL 11	84.56	86.96	**63.19**	65.93	**72.33**	**75.00**	**60.71**	**64.29**
DTL 12	84.75	**87.88**	54.95	63.74	66.67	73.89	55.36	63.39

Table 4: Coverage/Accuracy Evaluation with Variable Threshold (Var.) and Fixed Threshold (T=0.1)

and sub-phrase matching with stopwords (DTL's 10,11,12) seems to improve performance 7–24%.

	Bleu	
Lexicon	**Excl 0's**	**Incl 0's**
DTL 1	0.0266	0.0576
DTL 2	0.0280	0.0594
DTL 3	0.0279	0.0598
DTL 4	0.0254	0.0570
DTL 5	0.0296	0.0556
DTL 6	0.0298	0.0575
DTL 7	0.0476	0.0596
DTL 8	0.0490	**0.0615**
DTL 9	0.0491	0.0602
DTL 10	**0.0592**	0.0589
DTL 11	0.0525	0.0580
DTL 12	0.0550	0.0581

Table 5: MT Evaluation Results Using DTLs

Lexicon	Training Data	Bleu
No Dict	HKN	0.0223
No Dict	HKN & Chem Treaty	0.1609
Uniform Weight	HKN	0.0562
Uniform Weight	HKN & Chem Treaty	0.1508

Table 6: Evaluation Results Without Using DTLs

For comparison, we trained the un-tuned IBM-style system using different dictionary inputs (no dictionary vs. uniformly weighted dictionary) and training data (Hong Kong News (HKN) vs. HKN supplemented with a non-test portion of "Chem Treaty"). The results are shown in Table 6. Without training on "Chem Treaty", our best system (DTL 8) outperforms the un-tuned version by 275% (with no dictionary) or 9.4% (with uniform-weighted dictionary). On the other hand, the un-tuned MT model trained on "Chem Treaty" outperforms our model by 261%. Even if we train on "Chem Treaty" in our own model, our best DTL score 0.1581 (not shown in the tables above)—not significantly different from that of the un-tuned variants.

We conclude that—given a foreign-language document to translate—if the translations already exist for a portion of that document, these should be used for training rather than expending resources on domain-tuning. However, it is unrealistic to expect that a portion of an input document will already be translated.[11] Thus, the DTL approach is an important step toward the successful translation of domain-specific documents in the face of limited resources.

We also examined the impact of choosing different values of N, the number of copies of each domain-tuned entry appended to the training data. With $N = 100$ the 'Excl 0' version of DTL 6 increased from 0.0298 to 0.0329—a significant improvement; but the 'Incl 0' counterpart decreased from 0.0575 to 0.0525. In general, when we increased the value of N to 100 for all of our DTLs, the top-performing ones were still lower than those with $N = 10$.[12]

5 Conclusions and Future Work

The aim of this project is produce automatic domain-tuning techniques for translating a domain-specific document. We have demonstrated the effectiveness of our DTLs by showing a high degree of recall and precision with respect to a human-produced gold standard. Our Bleu experiments indicate a significant improvement when measured against systems using a uniformly-weighted dictionary or no dictionary at all. In summary, our approach has proven superior when adequate training data does not exist (e.g., input-document translations)—which is the

[11] In fact, it would have to be a very significant portion of the input document in order to be useful. (The test/training split is generally 1 to 3.)

[12] It is possible that there is more noise than signal when we combine the addition of 100 entries with the inclusion of 0-weighted entries.

most likely scenario for any given domain and language pair.

In the experiments, we viewed our comparable corpora as one large document for each of the two languages. The implication is that each FL word has as many translations as the number of unique English words in the comparable document—an overgeneralization that may lead to a high degree of noise in our final results. If we were to use multiple (smaller) comparable documents, the number of translation pairs would be significantly reduced, potentially improving the performance of our algorithm. A future area of research is the incorporation of alternative tools for building domain-specific comparable corpora using tools, e.g., STRAND (Resnik, 1999).[13]

Two other areas worthy of investigation are: (1) examining the impact of clustering on domain-tuning, e.g., whether there is a difference between general-purpose clustering and domain-specific clustering; (2) experimenting with new methods for assigning confidence values to our lexical entries, e.g., using the tf.idf technique once we add multi-document comparable corpora to our system—or at least testing out confidence values other than 1 or 2 for the single-document case.

Acknowledgments

This work has been supported, in part, by ONR MURI Contract FCPO.810548265, Mitre Contract 010418-7712, and NSF CISE Research Infrastructure Award EIA0130422. We would like to thank all the annotators who participated in the evaluation of our term lists.

References

Khurshid Ahmad. 1995. Language Engineering and the Processing of Specialist Terminology. Technical Report http://www.computing.surrey.ac.uk/ai/pointer/paris, University of Surrey, Guildford, Surrey, UK.

Lisa Ballesteros and W. Bruce Croft. 1997. Phrasal Translation and Query Expansion Techniques for Cross-Language Information Retrieval. In *Proceedings of the 20th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 84–91, July.

Peter F. Brown, John Cocke, Stephen A. DellaPietra, Vincent J. DellaPietra, Frederick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. 1990. A Statistical Approach to Machine Translation. *Computational Linguistics*, 16(2):79–85, June.

Echa Chang, Chu-Ren Huang, Sue-Jin Ker, and Chang-Hua Yang. 2002. Induction of Classification from Lexicon Expansion: Assigning Domain Tags to WordNet Entries. In *Proceedings of the First International WordNet Conference (also: Poster at SemaNet'02: Building and Using Semantic Networks, Workshop at COLING-2002)*, Karnataka, India.

Philip Clarkson and Ronald Rosenfeld. 1997. Statistical Language Modeling Using the CMU-Cambridge Toolkit. In *Proceedings of the ESCA Eurospeech Conference*, Rhodes, Greece.

Ido Dagan and Ken W. Church. 1994. TERMRIGHT: Identifying and Translating Technical Terminology. In *Proceedings of the Fourth ACL Conference on Applied NLP*, Stuttgart, Germany.

Beatrice Daille. 1994. Study and Implementation of Combined Techniques for Automatic Extraction of Terminology. In *Proceedings of the 32th Annual Meeting of ACL, Workshop on The Balancing Act: Combining Symbolic and Statistical Approaches to Languages*, Las Cruces, Nouveau Mexique.

Mark Davis and Ted Dunning. 1995. A TREC Evaluation of Query Translation Methods for Multilingual Text Retrieval. In *Proceedings of the Fourth Text Retrieval Conference (TREC-4)*, NIST, Gaithersburg, MD.

George Doddington. 2002. The NIST Automated Measure and Its Relation to IBM's BLEU. In *Proceedings of LREC-2002 Workshop on Machine Translation Evaluation: Human Evaluators Meet Automated Metrics*, Gran Canaria, Spain, June.

Anette Hulth, J. Karlgren, A. Jonsson, H. Bostrom, and L. Asker. 2001. Automatic Keyword Extraction Using Domain Knowledge. In *Proceedings of Second International Conference on Computational Linguistics and Intelligent Text Processing*, Mexico City, February.

Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, Massachusetts.

Daniel Marcu and Ulrich Germann. 2001. The ISI ReWrite Decoder Release 0.7.0b. Technical Report http://www.isi.edu/˜germann/software/ReWrite-Decoder/, Information Sciences Institute, University of Southern California.

I. Dan Melamed. 1997. A Scalable Architecture for Bilingual Lexicography. Technical Report MS-CIS-9701, Dept. of Computer and Information Science, University of Pennsylvania.

Douglas W. Oard. 1997. Cross-Language Text Retrieval Research in the USA. In *Proceedings of the Third DELOS Workshop; Cross-Language Information Retrieval, number 97-W003in Ercim Workshop Proceedings*, European Research Consortium for Informatics and Mathematics.

[13] In an initial attempt to produce a multiple-document set, we converted a set of comparable documents into a sentence-aligned resource using Translation Equivalence Detection (Smith, 2002); however, the resulting domain-tuned lexicon was inferior to ones induced directly from raw comparable resources (a 3–5% lower f-measure). On the other hand, this may have happened because the original comparable corpus was not conducive to the creation of a parallel resource—or because the equivalence detection tool is not suited to producing comparable corpora.

Franz J. Och and Hermann Ney. 2000. Improved Statistical Alignment Models. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL'00)*, pages 440–447, Hongkong, China, October.

Patrick Pantel and Dekang Lin. 2002. Discovering Word Senses from Text. In *Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 613–619, Edmonton, Canada.

K. Papineni, S. Roukos, T. Ward, and W. Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of Association of Computational Linguistics*, Philadelphia, PA.

Philip Resnik and I. Dan Melamed. 1997. Semi-Automatic Acquisition of Domain-Specific Translation Lexicons. In *Proceedings of the 5th ANLP Conference*, Washington, DC.

Philip Resnik. 1999. Mining the Web for Bilingual Text. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*, University of Maryland, College Park, Maryland, June.

Yuu Sakurai. 1999. Automatic Generation of the Domain-specific Dictionary for Text Classification. Master's thesis, School of Information Science, Japan Advanced Institute of Science and Technology. http://www.jaist.ac.jp/library/thesis/is-master-1999/paper/yskr/abstract.ps.

F. Smadja, K. R. McKeown, and V. Hatzivassiloglu. 1996. Translating Collocations for Bilingual Lexicons: A Statistical Approach. *Computational Linguistics*, 22(1):1–38.

Noah A. Smith. 2002. From Words to Corpora: Recognizing Translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, Philadelphia, Pennsylvania.