Evaluation Techniques Applied to Domain Tuning of MT Lexicons

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

Abstract

We describe a set of evaluation techniques applied to domain tuning of bilingual lexicons for machine translation. Our overall objective is to translate a domain-specific document in a foreign language (in this case, Chinese) to English. First, we perform an *intrinsic* evaluation of the effectiveness of our domain-tuning techniques by comparing our domain-tuned lexicon to a manually constructed domain-specific bilingual termlist. Our results indicate that we achieve 66% recall and 95% precision with respect to a human-derived gold standard. Next, an *extrinsic* evaluation demonstrates that our domain-tuned lexicon improves the Bleu scores 50% over a statistical system—with a smaller improvement when the system is trained on a uniformly-weighted dictionary.

1 Introduction

This paper describes a set of evaluation techniques applied to domain tuning of bilingual lexicons for machine translation. Our overall objective is to translate a domain-specific document in a foreign language (FL)—in this case, Chinese—to English. Using automatically selected 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.

First, we demonstrate the effectiveness of our domain-tuning techniques in an *intrinsic* evaluation by comparing our domain-tuned lexicon to a manually constructed domain-specific bilingual termlist. This evaluation assesses the degree of *coverage* and *accuracy* of our domain-tuned lexicons. Our results indicate that we achieve 66% recall and 95% precision with respect to a human-derived gold standard.

Next, we discuss the results of an *extrinsic* evaluation that uses a well-known automated MT evaluation technique (Bleu). Our evaluation approach involves the addition of domain-tuned lexical entries to the training set of an IBM-style (statistical) MT system. The resulting MT system is then applied to a held-out set of test sentences. We demonstrate that our domain-tuned lexicon improves the Bleu scores 50% over the statistical system—with a smaller improvement when the same system is trained on a uniformly-weighted dictionary.

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 evaluation of the domain-tuning component. The next two sections provide the background and algorithm behind our approach to automatic domain-tuning of lexicons. Following this, we turn to the intrinsic and extrinsic evaluations of our approach. Finally, we compare our domain-tuned version to its un-tuned counterpart in a Bleu-style MT evaluation.

2 Background

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 meanings (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 pre-existence of large domain-specific resources such as sentence-aligned parallel corpora (Resnik

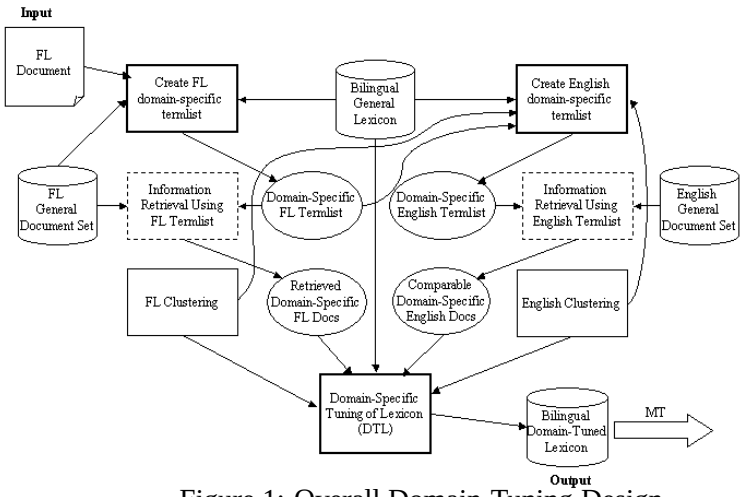

Figure 1: Overall Domain-Tuning Design

and Melamed, 1997; Melamed, 1997), hierarchically organized thesauri (Hulth et al., 2001), and pre-established domain tags (Chang et al., 2002). These resources are generally difficult to construct for a given language pair in a particular domain.

Our domain-tuning approach does not presuppose the existence of large domain-specific resources, but instead requires only: (1) the FL input document; (2) a general bilingual lexicon; and (3) a general-purpose clustering algorithm.[1] 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 exists a general bilingual dictionary for those pairs.

Figure 1 illustrates our overall approach. We implemented the components indicated with heavy borders. We borrow language-independent clustering software (LaTaT) to produce word clusters for the two languages (Pantel and Lin, 2002).[2] We also assume the existence of an IR system to produce 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 FL using standard tf.idf techniques.[3] 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.

The remainder of this paper focuses on the domain tuning algorithm and evaluation techniques applied to the second phase of the process, i.e., evaluation of the domain-tuning algorithm (described next).

3 Domain-Tuning Algorithm

Our domain-tuning algorithm relies on the pre-existence of the following resources:[4]

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 language.
3. A set of comparable, domain-specific documents in both languages.

The comparable, domain-specific documents are expected to be automatically selected 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.[5]

[1] By "general purpose," we mean that the similarity function applied by the algorithm should be derivable from any (large) training corpus of a given language.

[2] 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.

[3] For a general description of the well-known tf.idf technique, see (Manning and Schütze, 1999).

[4] The bilingual lexicon used for this effort is a large (600k entry) Chinese-English dictionary called Optilex, a machine-readable version of the CETA dictionary licensed from the MRM Corporation, Kensington, MD.

[5] We used 528 English documents and 352 Chinese documents from the domain of interest. Unfortunately, there were no links between comparable documents; thus we treated the document set in each language as one large document and assumed each one was comparable to the other.

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

Figure 2 shows the Domain-Tuning Algorithm. We use the following predicates and functions:

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, confidence values are set to 0.

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.

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.

The algorithm is further enhanced with *sub-phrase matching* mechanism for the handling of multiple word (or phrasal) translations; this mechanism 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.

Another domain-tuning enhancement involves the handling of translation pairs that do not occur in our comparable-document set. We apply *translational expansion*—analogous to the *query expansion* used in cross-language information retrieval (Ballesteros and Croft, 1997)—to assign ranks to such pairs. The highest ranked translation of each Chinese word is used to rank occurrences of the translation in other translation pairs that do not appear 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.[6]

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.

4 Experimental Set-Up for Evaluation

We generated 12 different DTLs using the above algorithm with different combinations of the three ex-

[6] If 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 $[e_1 e_2 \ldots e_n]$, the highest first-pass score associated with each e_i is used to compute the average of all $conf(c, e_i)$'s.

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

tensions: 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 – 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 between 10-20% for all lexicons, among 208K Chinese words or phrases. Figure 3 shows a sample entry for the first 6 DTLs to illustrate the format of the lexicons.

5 Evaluation of Domain-Tuned Lexicons

To measure the effectiveness of domain tuning, we conducted two different evaluations, one intrinsic and one extrinsic: (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).

5.1 Intrinsic Evaluation: Lexicon Coverage and Accuracy

In the first experiment, our purpose was to determine 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.

5.1.1 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 of a domain-specific term among 222 English translations. These 222 English translations were extracted from Chinese-English entries containing at least one English translation known to be relevant to the domain.[7] We generated the union of these two ground-truth sets as follows:[8]

1. If either annotator assigns *positive* to an English translation, the resulting annotation is *positive*.
2. Otherwise, the resulting annotation is *negative*.

	Ground Truth-1	Ground Truth-2	Union
Positive	186	179	196
Negative	36	143	26
Total	222	222	222

Table 2: Number of Instances in Ground-Truth Sets

The number of *positive* and *negative* instances and their union is given in Table 2. The agreement ratio between the two annotators using pairwise comparison (using an exact match of the labels) is 88%.

5.1.2 Evaluation Metrics

We evaluated accuracy and coverage using precision, recall, the averaged precision and recall (f-

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

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

measure)[9], 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.

5.1.3 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. 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 a fixed threshold, set at 0.1.

Table 4 presents the results for a baseline experiment[10] and all 12 DTLs. The **boldfaced** results are the best for the given settings. All domain-tuned lexicons outperform the baseline of 88.35% (precision), 49.82% (recall), 63.64% (f-measure), and 49.84% (correctness). The improvement of the DTLs over the baseline was as much as 21% in f-measure and 30% in correctness.

Threshold	Precision	Recall	F-Measure	Correct
Variable	93.64	52.02	66.88	54.46
0.1	92.79	52.02	66.67	54.02
0.2	96.05	36.87	53.28	42.86
0.3	96.36	26.77	41.90	34.38
0.4	97.50	19.70	32.77	28.57
0.5	97.22	17.68	29.91	26.79

Table 3: Evaluation Results for Different Thresholds for DTL 1

For the variable threshold, the precision is between 91.53% and 95.16%. DTL 11—which incorporates sub-phrase matching, translational expansion, and stopwords—scored highest for recall, f-measure and correctness under variable threshold. On the other hand, DTL 6 performed best in terms of recall, f-measure and correctness under the fixed threshold of 0.1 and the results for DTL 5 are very close to them.

The inclusion of stopwords in the lexicon generation leads to slight decreases in some cases and slight increases in others. For instance, for fixed threshold of 0.1, DTLs without using stopwords turned out to yield the best results. Sub-phrase matching mechanism increases the results slightly, in the range of 1-2%. The results indicate that using sub-phrase matching and translational expansion increases the performance in all measures. Overall, the results indicate that DTLs provide the information necessary to distinguish domain-specific vocabulary from other words.

5.2 Extrinsic Evaluation: MT Domain Coverage

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

5.2.1 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[11] using the CMU-Cambridge Toolkit (Clarkson and Rosenfeld, 1997). Since GIZA++ cannot accommodate a DTL directly, we designed a mechanism to incorporate each DTL into the translation

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

[10] For the baseline, we performed a random assignment of translations as 'positive' or 'negative'. Baseline figures represent an average over 1000 runs.

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

Lexicon	Precision		Recall		F-Measure		Correctness	
	Var.	T=0.1	Var.	T=0.1	Var.	T=0.1	Var.	T=0.1
Random	88.35		49.82		6 3.64		49.84	
DTL 1	93.64	92.79	52.02	52.02	66.88	66.67	54.46	54.02
DTL 2	94.07	**93.28**	56.06	56.06	70.25	70.03	58.04	57.59
DTL 3	94.07	**93.28**	56.06	56.06	70.25	70.03	58.04	57.59
DTL 4	94.64	90.70	53.54	59.09	68.39	71.56	56.25	58.48
DTL 5	**95.16**	91.55	59.60	**65.66**	73.29	76.47	61.61	64.29
DTL 6	92.86	92.20	59.09	**65.66**	72.22	**76.70**	59.82	**64.73**
DTL 7	92.66	91.96	51.01	52.02	65.80	66.45	53.12	53.57
DTL 8	93.16	92.50	55.05	56.06	69.21	69.81	56.70	57.14
DTL 9	93.16	92.50	55.05	56.06	69.21	69.81	56.70	57.14
DTL 10	91.94	90.40	57.58	57.07	70.81	69.97	58.04	56.70
DTL 11	92.65	91.30	**63.64**	63.64	**75.45**	75.00	**63.39**	62.50
DTL 12	91.53	92.42	54.55	61.62	68.35	73.94	55.36	61.61

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

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 on each DTL in independent experiments. We performed a Bleu evaluation (Papineni et al., 2002) on unigrams.[12]

5.2.2 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, first non-zero weighted entries are added proportionally to their confidence values and then each zero-weighted entry is added once to the training data (number of entries added to the data may be higher than N in this case). 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.

5.2.3 Results of MT Evaluation

Table 5 presents the unigram 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 between 33-51% when stopwords are ignored; the difference is much smaller when stopwords are used but it still makes a difference in the range of 10-15%. We also see that either kind of expansion improves the scores by 5-17% when stopwords are not used (DTL's 2,3 wrt 1, DTLs 5,6 wrt 4, and so on). Finally, the inclusion of stopwords (the last 6 DTL's) leads to an improvement of up to 32% (with a bigger impact when 0-score translations are excluded from training data) .

For comparison, we trained the un-tuned IBM-

[12]Because 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 indicates relative effectiveness of different systems; thus, we are interested not in the magnitude of the scores, but in their relative values. We did use the unigram Bleu scores because our major goal is to observe the effects of translational selection per word as opposed to fluency of the sentence.

	Bleu	
Lexicon	**Excl 0's**	**Incl 0's**
DTL 1	0.2158	0.3255
DTL 2	0.2264	0.3291
DTL 3	0.2251	0.3293
DTL 4	0.2244	0.3235
DTL 5	0.2425	0.3225
DTL 6	0.2441	0.3248
DTL 7	0.2849	0.3290
DTL 8	0.2914	**0.3301**
DTL 9	0.2923	**0.3301**
DTL 10	0.2952	0.3295
DTL 11	**0.3139**	0.3261
DTL 12	0.3107	0.3233

Table 5: MT Evaluation Results Using DTLs

Lexicon	**Training Data**	**Bleu**
No Dict	HKN	0.2193
No Dict	HKN & Chem Treaty	0.4625
Uniform Weight	HKN	0.3257
Uniform Weight	HKN & Chem Treaty	0.4794

Table 6: Evaluation Results Without Using DTLs

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 50% (with no dictionary) or 2% (with uniform-weighted dictionary). On the other hand, the un-tuned MT model trained on "Chem Treaty" outperforms our model by 40%. When we train on "Chem Treaty" using DTLs in our own model, our best DTL score is 0.4841 (not shown in the tables above)—slightly higher than that of the un-tuned variants.

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 0.2441 to 0.2489; but the 'Incl 0' counterpart decreased from 0.3248 to 0.3110. 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 $N = 10$.[13]

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.[14] Thus, we believe the DTL approach has the potential for assisting the process of building domain-specific MT systems in the face of limited resources, although further study is needed.

6 Conclusions and Future Work

We have presented intrinsic and extrinsic frameworks for evaluating the impact of domain-tuning on lexicon coverage and translation correctness. In our intrinsic evaluation, we measured coverage/accuracy based on recall and precision with respect to a human-produced gold standard. In our extrinsic evaluation, we measured the Bleu scores of a domain-tuned statistical MT system against systems using a uniformly-weighted dictionary or no dictionary at all. We take the view that these evaluation approaches are useful in cases where adequate training data does not exist (e.g., input-document translations)—which is the 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 leads to a high degree of noise in our 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).

Three other areas worthy of investigation are: (1) evaluation of the domain-tuning algorithm with respect to its impact on a purely symbolic MT system (e.g., Systran[15]); (2) application of the Bleu technique to a text larger than 155 sentences, either by providing multiple references or by using a larger in-domain text; (3) 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.

[13] 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.

[14] 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.)

[15] http://www.systransoft.com/

Acknowledgments

This work has been supported, in part, by Mitre/DoD Contract 010418-7712, Office of Naval Research MURI Contract FCPO.810548265, 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.

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.