Building Compact Lexicons for Cross-Domain SMT by Mining Near-Optimal Pattern Sets

Pankaj Singh, Ashish Kulkarni, Himanshu Ojha, Vishwajeet Kumar, and Ganesh Ramakrishnan

Computer Science and Engineering,
IIT Bombay, Mumbai, India
{pr.pankajsingh,kulashish}@gmail.com,
{himanshuojha.lko,vishwajeetkumar86}@gmail.com,
ganesh@cse.iitb.ac.in

Abstract. Statistical machine translation models are known to benefit from the availability of a domain bilingual lexicon. Bilingual lexicons are traditionally comprised of multiword expressions, either extracted from parallel corpora or manually curated. We claim that "patterns", comprised of words and higher order categories, generalize better in capturing the syntax and semantics of the domain. In this work, we present an approach to extract such patterns from a domain corpus and curate a high quality bilingual lexicon. We discuss several features of these patterns, that, define the "consensus" between their underlying multiwords. We incorporate the bilingual lexicon in a baseline SMT model and detailed experiments show that the resulting translation model performs much better than the baseline and other similar systems.

Keywords: submodular, pattern extraction, cross-domain SMT

1 Introduction

A statistical machine translation (SMT) model typically relies on the availability of a large parallel corpus, often collected from multiple sources and spanning different domains. While a domain-specific corpus might share some of its lexical characteristics with the cross-domain corpus, it often differs in its language usage and vocabulary. A cross-domain SMT model might, therefore, fail to reliably translate an in-domain text. While it is possible to train an in-domain translation model, domain-specific parallel corpus is either non-existent or scarce and expensive to generate. The problem of domain adaptation deals with augmenting a cross-domain translation model to reliably translate an in-domain text and poses an interesting research challenge [8].

Although in-domain parallel text might be difficult to obtain, in-domain bilingual lexicons are often readily available or could be manually curated. Typically, these are restricted to words or short phrases specific to the domain of interest. A medical domain bilingual lexicon, for instance, consists of technical and popular medical terminology covering the anatomy of body, certain diseases, medicines *etc.* In addition to these however, a domain corpus, due to its specific language structure, is often replete with redundant phrases. Consider for instance, the phrase "*...be given marketing authorisation*", appearing 218 times in the EMEA medical corpus [23]. These, if extracted and translated in a bilingual lexicon, might aid in-domain translation [21, 26]. In fact, repetition in a

Table 1: Examples of recurring patterns, sample snippets covered by them and the number of such covered snippets (in brackets) for the EMEA corpus

PATTERN: in patients with ⟨CAT1⟩ (568)	**contains ⟨CAT2⟩ mg of ⟨CAT3⟩ (91)**
in patients with HIT type II	capsule contains 25 mg of lenalidomide
in patients with CNS metastases	tablet contains 300 mg of maraviroc
in patients with ESRD	syringe contains 100 mg of anakinra
in patients with normal and impaired renal function	tablet contains 2.3 mg of sucrose
in patients with previous history of pancreatitis	capsule contains 200 mg of pregabalin
in patients with cirrhosis of the liver	vial contains 10 mg of the active substance
	tablet contains 30 mg of aripiprazole

domain corpus could be further exploited by observing that certain phrases, which might themselves be infrequent, tend to have "consensus" when generalized to higher-level patterns. Table 1 illustrates two patterns and corresponding sample phrases extracted from the EMEA medical domain corpus. These patterns are typically n-grams of tokens, domain-specific categories or higher-level phrase classes (noun phrase, verb phrase *etc.*).

Given a domain corpus, it is not obvious how to extract a set of such patterns to be manually translated. Moreover, in the absence of a parallel in-domain corpus, translation of these patterns requires manual effort, which poses other challenges. Specifically, syntactically well-formed patterns like *"the CAT5 of treatment"* might be easier for humans to translate than others like *"CAT4 condition has"*. Chen *et. al.* [3] present this and other quality criteria that every pattern must satisfy to be worth being translated in order to aid cross-domain SMT applications. We will refer to such patterns as *quality patterns*. In this work, we generalize the search space of patterns as well as the quality criteria that a pattern must meet.

More importantly, two or more quality patterns could have instances that significantly overlap in their spans in the corpus. Is translating each such *quality* pattern really necessary? We expect the human effort for translating patterns to have a budget constraint and therefore, a compact *set of patterns* is desirable. For example, it is desirable to extract a set of patterns (for bilingual lexicon), such that, the set maximally covers the corpus. We argue that some formulations of this problem are natural instances of submodular maximization. A set function f(.) is said to be submodular if for any element v and sets $A \subseteq B \subseteq V \setminus \{v\}$, where V represents the ground set of elements, $\mathrm{f}(A \cup \{v\}) - \mathrm{f}(A) \geq \mathrm{f}(B \cup \{v\}) - \mathrm{f}(B)$. This is called the diminishing returns property and states, informally, that adding an element to a smaller set increases the function value more than adding that element to a larger set. Submodular functions naturally model notions of coverage and diversity, and therefore, a number of subset selection problems can be modeled as forms of submodular optimization [7, 11].

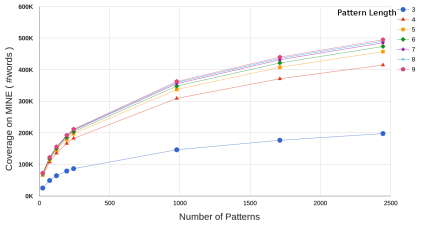
We illustrate the relevance of the submodular coverage function to pattern-subset selection in Figure 1. We plot the corpus coverage (in terms of number of words) with increasing number of patterns in the set, for pattern lengths varying from 3 to 9. In each case, while the coverage improves with increasing number of patterns, the gain in coverage progressively diminishes with growth in the size of the subset.

Our contribution is a framework to curate a high quality bilingual lexicon based on three key ideas. Our first two ideas generalize the approach of Chen *et al.* [3].

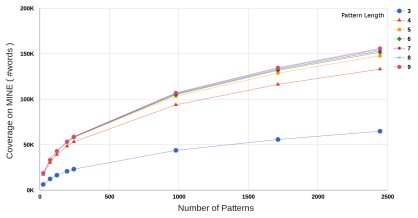
1. **Language of patterns:** A pattern could either be lexical, comprised of words alone, or it could be a combination of words and higher-level categories.

2. **Quality criteria for a pattern:** The quality (or cost) of every instance of a pattern is a function of several features including its frequency in the corpus and whether or not it is syntactically well-formed. The quality (or cost) of a pattern is then a simple (modular) aggregation of the instance costs.
3. **Quality criteria for a set of patterns:** We define the "goodness" of a set of patterns based on element-wise non-decomposable submodular costs.

We incorporate these patterns along with their translations, as entries in a bilingual lexicon and study [1] its effect on the translation accuracy for the domain adaptation of a baseline SMT model. While significantly improving over the baseline, we also show significant improvement over the modular setting of Chen *et al.*



(a) Corpus EMEA: corpus coverage vs. #patterns



(b) Corpus KDE4: corpus coverage vs. #patterns

Fig. 1: Gain in coverage shows diminishing returns with increasing number of patterns in the set

2 Related Work

Extraction of bilingual multi-word expressions (BMWE): SMT systems often use word-to-word alignment approaches for inferring translation probabilities from bilingual data [25, 17]. However, in some cases it might not be possible to perform word-to-word alignment between two phrases that are translations of each other [10]. This has motivated a body of work [10, 21, 18] on automatic extraction of multi-word expressions from bilingual corpora. Ren *et al.* [21] propose multiple techniques to integrate BMWE's into a phrase-based SMT system and show improvement over the baseline translation system. Recently, Liu *et al.* [12] proposed an approach to mine quality phrases from large text corpora. They use a phrasal segmentation-based approach for phrase mining and combine that with several phrase quality assessment metric in a scalable framework. While our approach is inspired by these works, we differ from them in that we aim to extract generalized patterns comprising words and categories. Also, we do not assume availability of a parallel corpus in the target domain.
Domain Adaptation: Typically, the application domain of a translation system might be different from the domain of the system's training data. In-domain parallel

[1] We release our code for optimal pattern-set identification, as well as the lexicons. `https://www.cse.iitb.ac.in/~ganesh/Publications.html`

corpus might either be non-existent or scarce, but, in-domain monolingual corpus is usually available. The problem of domain adaptation[2] has therefore been in focus and there has been work [26, 8, 16] to build in-domain translation lexicons and combine them with out-of-domain parallel corpus to achieve in-domain translation. Koehn and Schroeder [8] use limited in-domain parallel corpus to train a language model and a translation model and present techniques to integrate them with corresponding models trained on an out-of-domain corpus. Wu *et al.* [26] manually create an in-domain lexicon where the lexicon entries are restricted to words. They propose an algorithm to combine an out-of-domain bilingual corpus, an in-domain bilingual lexicon, and monolingual in-domain corpora in a unified framework for in-domain translation.

Pattern Mining: The other body of work most related to our approach comes from the area of pattern mining. While most earlier work [22] dealt with identifying consecutive word sequences, Joshi *et al.* [6] present an efficient approach to mine significant non-consecutive word sequences, where, *significance* is captured by the support measure. Contrary to mining patterns that satisfy pre-specified criterion, there has also been work on interactive pattern mining [27, 2, 1] that uses human feedback to identify a set of *interesting* patterns. Chen *et. al.* [3] proposed an English-Chinese medical summary translation system that adapts a baseline SMT model with significant patterns (of lexical as well as medical type tokens) learned from an English medical summary corpus. The quality of a pattern is assessed based on its frequency in the corpus and its linguistic completeness. While being closest to our work, we differ from them and the other aforementioned works in two ways. Firstly, we realize that the quality criterion for a set of patterns is not always a modular function of quality of the constituent patterns in the set. We define several quality criteria based on both element-wise decomposable (modular) costs and element-wise non-decomposable (non-modular) costs and combine them in a mathematical formalism for the task of significant pattern mining. Secondly, domain-specific classes often rely on the availability of corresponding term lexicons. Our framework also makes use of general phrase classes such as noun phrases (NP), verb phrases (VP), thereby extracting generic patterns whose instances themselves might not be frequent in a corpus (Refer to Figure 1). Moreover, the use of phrase classes allows for the induction of new instances in a class (type) lexicon.

3 Framework

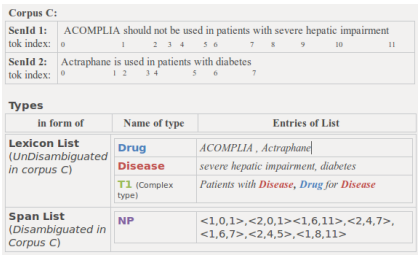
The task of lexicon curation finds applications in several NLP tasks including machine translation. We present a formulation of the problem and a solution framework that one could invoke based on underlying application requirements. The lexicon is composed of quality patterns extracted from a domain corpus and for the specific task of machine translation with low resource constraint, we then acquire translations of these patterns to create an in-domain *bilingual lexicon*.

3.1 Formal Problem Definition

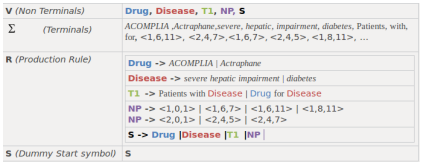
We are given a domain corpus $\mathcal{C}$ and optionally a set of "types" $\mathcal{T}$. A type might represent a *domain type*, like *disease* in medical domain, a *lexical type*, like *noun phrases*

[2] http://www.statmt.org/wmt07/shared-task.html

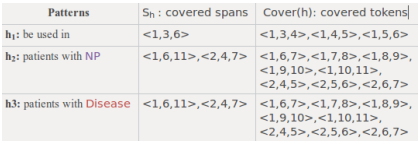
or a complex type involving a combination of these. The problem of lexicon curation is to extract from $\mathcal{C}$, a set H of quality patterns, as per a quality function $Q_{\mathcal{C}}(h)$ for the quality of a pattern $h \in H$ in the corpus and a quality function $Q_{\mathcal{C}}(H)$ for the quality of the set H.



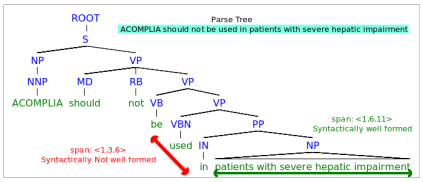
(a) Corpus and Types



(b) CFG Grammar G: (V, Σ, R, S)



(c) Patterns



(d) Syntactically well-formed span

Fig. 2: Examples of components of our solution framework

3.2 Solution Framework

We define and describe below the components of our solution framework.

1. **Context Free Grammar** G**:** A context free grammar (CFG) allows us to encode our types and is comprised of a set V of non-terminals, a set Σ of terminals, a start symbol $S \in V$ and a set P of productions $\alpha \rightarrow \beta$, where $\alpha \in V$ and $\beta \in (V \cup \Sigma)^*$. Our choice of CFG as a formalism to represent the types is motivated from the fact that the grammar can be directly consumed by a high-level grammar formalism like Grammatical Framework (GF) [20], which is type theoretic, multilingual, and modular and suits our downstream translation usecase. We define a grammar G, where, the set V of non-terminals corresponds to the set of types $\mathcal{T}$. Each type $T_i \in \mathcal{T}$ could be available as a lexicon list, comprising entries (undisambiguated), $T_i = \{t_{i1}, t_{i2}, ..., t_{ik}\}$, where, each entry $t_i \in T_i$ is a sequence of lexical tokens alone (in case of simple types) or a combination of lexical and type tokens (in case of complex types). Alternatively, a type could also be available as a set of spans in the corpus (disambiguated entries), $T_i = \{< s_{i1}, u_{i1}, v_{i1} >, < s_{i2}, u_{i2}, v_{i2} >, ..., < s_{ik}, u_{ik}, v_{ik} >\}$, obtained as an output from an annotator (for instance, Stanford NER *etc.*) (Refer to Figure 2). Here, a span is a 3-tuple of sentence id, start and end token index within the sentence. The set Σ of terminals then comprises:
   - in the presence of type lexicons, the set of lexical tokens in the entries of each type $T_i \in \mathcal{T}$;

- in the presence of an annotator, the set of spans $<s,u,v>$, encoded as productions of the form $T_i \rightarrow <s_{i1},u_{i1},v_{i1}> | <s_{i2},u_{i2},v_{i2}> | ... | <s_{ik},u_{ik},v_{ik}>$.

2. **Pattern Extractor:** A pattern extractor is a program that uses the context free grammar G, to extract from $\mathcal{C}$, a set $\mathcal{H}$ of patterns, where, each pattern $h \in \mathcal{H}$ is a sequence of tokens of words or types. A pattern could be thought of as a potential higher level domain type along with a set of spans in the corpus from which it is extracted. For a pattern h, let S_h be this set of spans. We say that the spans in S_h are covered by the pattern h. Consider a span $\mu_i = <s_i,u_i,v_i> \in S_h$. We define $tokens(\mu_i) = \{<s_i,u_i,u_{i+1}>,...,<s_i,v_{i-1},v_i>\}$ as the set of all tokens covered by the span μ_i. We then say that the *token coverage* of the pattern h is the set $cover(h) = \cup_{\mu \in S_h} tokens(\mu)$ (Refer to Figure 2c).
3. **Quality $Q_{\mathcal{C}}(h)$ of a pattern:** Quality of a pattern h is defined as a function $Q_{\mathcal{C}}(h): \mathcal{H} \rightarrow [0,1]$. Then the set $\mathcal{H}_Q = \{h \in \mathcal{H} | Q_{\mathcal{C}}(h) > r\}$, where $0 < r < 1$ is a quality threshold, is the set of all patterns in $\mathcal{H}$ that meet the quality criteria. Such a quality criterion of a pattern is typically a simple (modular) aggregation of the quality of its instances. Some examples of quality criteria are:
   (a) Pattern consensus: $|S_h|$, the number of spans covered by the pattern h;
   (b) Informativeness: For a set $\mathscr{C}$ of corpora, $\frac{|\mathscr{C}|}{|\{\mathcal{C} \in \mathscr{C} : |S_h^{\mathcal{C}}| > 0\}|}$, where, $S_h^{\mathcal{C}}$ is the set of spans covered by h in corpus $\mathcal{C}$;
   (c) Syntactic well-formedness: A span covered by a pattern is said to be syntactically well-formed if it forms a sub-tree in the parse tree of its corresponding sentence. A pattern is then syntactically well-formed if at least k of the spans covered by it are syntactically well-formed (Refer to Figure 2d).
   (d) Lexical rule-based consensus: Spans covered by a pattern should conform to a set of lexical rules. For instance, a pattern should not start or end with prepositions;
   (e) Semantic rule-based consensus: Enforces a semantic constraint among the tokens in $tokens(\mu)$, where, μ is a span covered by the pattern. For instance, while mining patterns specific to "mergers and acquisition" from a financial services transactions corpus, we might enforce a constraint on the semantic role of agents in the patterns to be either a *buyer* or a *seller*.
   (f) Model-based quality criteria: A trained classifier could be used to classify a pattern as interesting or not based on other criteria as features.
4. **Quality $Q_{\mathcal{C}}(H)$ of a patterns set H:** Quality of a set H of patterns, given a corpus $\mathcal{C}$, is defined as a function $Q_{\mathcal{C}}(H)$ from $2^{\mathcal{H}_Q} \rightarrow \mathbb{Z}$. Typically, $Q_{\mathcal{C}}(H)$ is either modular (*e.g.* $|H|$, $\sum_{h \in H} |cover(h)|$) or submodular (*e.g.* $|\cup_{h \in H} cover(h)|$).
5. **Pattern selection:** The problem of lexicon curation can now be posed as the problem of selecting an optimal subset H of $\mathcal{H}_Q$. Clearly, H is optimal quality set when $H = \mathcal{H}_Q$, however, in practice, selection of an optimal H often involves an optimization of conflicting requirements on the quality and the cost of the subset. Formally,

$$H^* = \arg\max_{H \subseteq \mathcal{H}_Q} Q^2_{\mathcal{C}}(H) \text{ s.t. } Q^1_{\mathcal{C}}(H) < c \tag{1}$$

Or

$$H^* = \arg\min_{H \subseteq \mathcal{H}_Q} Q^1_{\mathcal{C}}(H) \text{ s.t. } Q^2_{\mathcal{C}}(H) > d \tag{2}$$

where, c and d are thresholds on the cost and the quality of H respectively. It is known that this optimization has an efficient solution under the assumption that the cost function $Q^1_{\mathcal{C}}(H)$ be modular and the quality function $Q^2_{\mathcal{C}}(H)$ be submodular [5].

3.3 Our Approach

We implemented our framework to curate high quality compact lexicons for the cross-domain SMT task. Given a source language corpus, we pose the problem of curating an optimal set of high quality patterns, solve it using a greedy algorithm and use a human-in-the-loop approach to get their translation. More precisely, we follow the steps described below:

Context Free Grammar: We use Stanford parser to create a lexicon list of type Noun phrases (NP) present in the corpus. Refer to section 3.2 for details.

Pattern Extraction and Filtering: We use our grammar to index corpus and mine patterns. Our mining approach is inspired from Joshi *et al.* [6]. We first mine patterns for each sentence using their dynamic programming-based approach and then aggregate patterns across all sentences. Subsequently, we filter out bad patterns (we refer to this as pattern filtering), where, the quality of a pattern is judged based on two modular quality criteria—aggregated frequency of its instances and their syntactic well-formedness.

Pattern Selection: We formulate this as a subset selection problem (Refer to the formulations (1) and (2)). Although formulation (1) has a better approximation guarantee, both formulations performed equally well in our evaluation. Both formulations can be efficiently solved if the cost function $Q^1_{\mathcal{C}}(H)$ is modular and the quality function $Q^2_{\mathcal{C}}(H)$ is submodular [9]. In our experiments, we use as $Q^1_{\mathcal{C}}(H)$ the modular cardinality constraint $|H| < c$ and as $Q^2_{\mathcal{C}}(H)$ the submodular token coverage of corpus: $|\cup_{h \in H} cover(h)|$. Further, if $Q^2_{\mathcal{C}}(H)$ is a submodular and monotone function, that is, $A \sqsubseteq B$ then $Q^2_{\mathcal{C}}(A) \leq Q^2_{\mathcal{C}}(B)$, then this problem can be solved greedily with theoretical guarantee of $1 - \frac{1}{\epsilon}$ [14]. This is the best approximation result we can achieve efficiently [15]. Further, we use an accelerated version of this algorithm [13] which at every iteration lazily evaluates the function value to get the best item to add in the output set.

Pattern Translation: After mining a high quality set of patterns, we ask humans to provide translations of these patterns and thus create a *bilingual lexicon*. We leveraged Matecat [4] and MyMemory[3] to help human translators while using our interactive system for gathering translations.

4 Evaluation

4.1 Experimental Setup

We study the effect of curating a domain-specific bilingual lexicon using our approach on domain adaptation of pre-built cross-domain statistical machine translation (SMT) models[4] trained for three language pairs on the Europarl corpus [26]. We experimented with adapting the pre-built SMT model on three domain specific datasets [24, 23], each pertaining to a different domain: (i) JRC (legal), (ii) EMEA (medical) and (iii) KDE (technical). Each domain has specific language usage that differs from that of a large cross-domain corpus (*i.e.*, Europarl) typically used to train an SMT model. Moreover, the availability of parallel corpora for training SMT models is very limited in these domains. While each of these is a parallel corpus, we made use of the aligned target

[3] `https://mymemory.translated.net/`

[4] `http://www.statmt.org/moses/RELEASE-3.0/models/`

language corpus only for evaluation. The remaining source language data was used for mining patterns. We experimented with these datasets for three language pairs: English-French (en-fr), English-Spanish (en-es), and English-German (en-de). In Table 2, we present the number of sentences in each of these parallel corpora.

Table 2: Corpus statistics showing number of sentences in the domain-specific parallel corpora

Corpus	en-fr	en-es	en-de
JRC (legal)	814,167	805,756	537,850
EMEA (medical)	1,092,568	1,098,333	1,108,752
KDE4 (technical)	210,173	218,655	224,035

Table 3: Effect of filtering on the number of patterns extracted from the JRC corpus. *Filtered (F), Unfiltered(U)

Pattern length	3	4	5	6	7	8	9
F/U* %	15.9	11.1	9.2	8.3	7.4	6.5	6.1

For each dataset, we create a test set (TEST) for evaluation, by randomly sampling 3000 unique sentence pairs from the corpus. A different random sample of up to 100,000 source language sentences is used as the set for mining the patterns (MINE). Pattern extraction is performed using the sentences in the source language from the MINE set and for the set of quality patterns mined, we manually obtain their corresponding translations, while being guided by the aligned target language sentences from the MINE set. We evaluate the translation quality on the TEST set using the standard BLEU metric [19]. Our baseline corresponds to the pre-built SMT model. For domain adaptation, we incorporate the bilingual lexicon (curated using the MINE set) into the baseline model using the XML markup feature[5] available in the Moses tool.

The entire process of sampling TEST and MINE sets for each corpus and language pair is repeated thrice and the baseline and domain-adapted numbers are reported after averaging across the three runs. In the following sections, we present several intermediate results and ablation tests before presenting the final BLUE score comparisons. Owing to space issues we present select plots for select datasets and language pairs here.

4.2 Effect of Syntactic Completeness-based Consensus on Pattern Extraction

The pattern extraction step extracts all frequent patterns (frequency threshold = 2) of up to a certain length. This results in a large number of patterns, not all of which are syntactically well-formed. We filter out patterns whose instances do not conform to a phrasal structure as per the Stanford parser (inferred via the Grammar discussed earlier), thus leaving behind between 6% to 9% patterns for further processing for lexicon curation (*c.f.* Table 3).

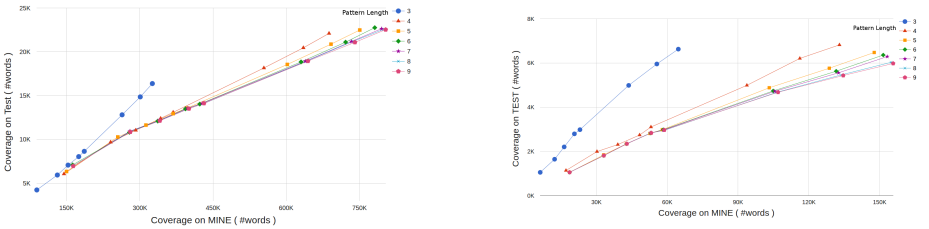
4.3 Effect of Varying the Lexicon Size

Pattern selection (Equation 1) allows to constrain the cardinality of the final set of quality patterns. The manual translation of these patterns requires human effort that

[5] inclusive and exclusive mode `http://www.statmt.org/moses/?n=Advanced.Hybrid`

is proportional to the number of patterns in this set. On the other hand, cardinality of this set might also affect the corpus coverage and thereby the translation accuracy. We study this effect by setting the cardinality of this set to various values: 25, 75, 125, 200, 250, 1000, 1750 and 2500.

Coverage on MINE versus TEST: In Figure 3, we present the corpus coverage (in terms of number of words) on the MINE and TEST data sets with varying number of patterns and for different pattern lengths. Patterns mined using the MINE split seem to generalize well and the coverage on both MINE and TEST increases as we increase the number of patterns. This observation holds true for other datasets and language pairs as well and the coefficient of correlation between MINE and TEST coverage is consistently above 0.99.

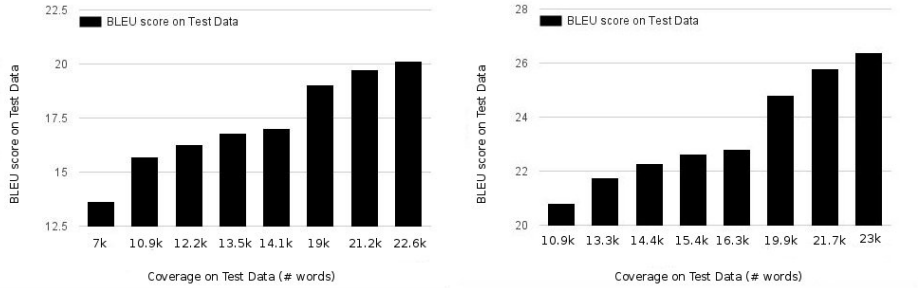


(a) JRC (en-de): Coverage on MINE vs. TEST

(b) KDE4 (en-es): Coverage on MINE vs. TEST

Fig. 3: Effect of varying the size of the set of quality patterns on corpus coverage

Coverage and translation accuracy on TEST: The patterns in the lexicon are translated and added to an in-domain bilingual lexicon. Figure 4 shows that as we increase the size of the lexicon, the TEST coverage improves and we see corresponding improvement in the translation accuracy.



(a) JRC (en-de): Coverage vs. BLEU on TEST (b) KDE4 (en-es): Coverage vs. BLEU on TEST

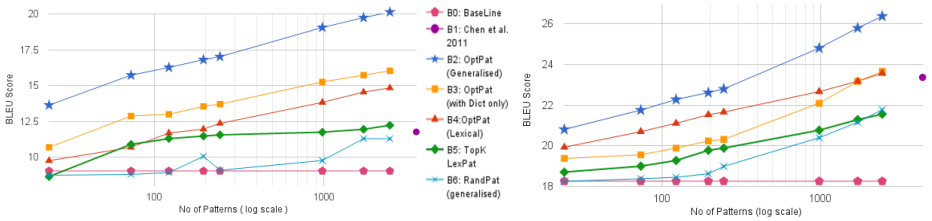
Fig. 4: Effect of varying the size of the set of quality patterns on translation accuracy.

4.4 Comparison of Different Approaches to Pattern-set Extraction for Cross-domain SMT

We compare different approaches to extracting a good set of patterns from a source language corpus and translating them for cross-domain SMT application. Figure 5 shows accuracy of these models for varying number of patterns in the lexicon.

Effect of bilingual lexicon in domain adaptation: The task of domain adaptation involves using a translation model trained on a large out-of-domain parallel corpus and adapting it to reliably translate an in-domain corpus. The pre-built SMT models trained on the out-of-domain Europarl corpus serve as baselines and are used to evaluate translations on the in-domain TEST splits (Refer to B0 in the figure). Next, we adapt the model for in-domain translation by incorporating our curated in-domain bilingual lexicons into the baseline model. The improvement in translation accuracy (Refer to B2) is quite evident. The lexicons capture significant in-domain patterns and provide their reliable translation, thereby, further aiding the baseline model already trained to translate common cross-domain phrases.

Effect of submodular optimization: Would we have got the same improvement in translation accuracy had we curated a bilingual lexicon from a random subset of patterns? We curated a bilingual lexicon using our set of quality patterns (B2) and another using a random subset of frequent patterns (B6) and compared their impact on translation accuracy. The translation model incorporating bilingual lexicon curated from random subset of frequent patterns does improve upon the baseline. However, the one with our high quality bilingual lexicon, obtained after submodular optimization, does much better in generating a high quality translation.



(a) JRC-corpus (en-de) (b) KDE Corpus (en-es)

Fig. 5: Comparison of different approaches for creating a bilingual lexicon for cross-domain SMT

Effect of pattern generalization: Since our patterns comprise words and phrase classes, phrases in a corpus that might otherwise be infrequent, turn out to be frequent when folded into patterns. In order to ascertain that this indeed positively affects our curated lexicon and the final translation model, we compared this with a lexicon curated from frequent lexical-only phrases (B4 and B5) in the corpus. The final set of patterns was obtained, in one case, by extracting the top-k frequent phrases (B5: modular criterion) and in the other case, by using the submodular quality criterion (Refer to B4). As can be seen in Figure 5, the modular frequency-based criterion does

much better on generalized patterns than on phrasal (lexical only) patterns and together with submodular optimization results in a much better bilingual lexicon.

Comparison with other work The system proposed by Chen *et al.* [3] comes closest to our work. We used publicly available domain lexicons to annotate our corpora with domain types and used their clustering-based approach to extract a set of significant patterns. The bilingual lexicon was created by sampling and manually translating one representative pattern from each cluster (Refer to B1). Next, we applied our submodular optimization-based approach on the same annotated corpora (Refer to B3). We observe that the pattern-set obtained using our quality criteria does better, even with patterns composed of domain types instead of the more general phrase classes.

5 Conclusion

We presented a novel framework for extraction of a high quality bilingual lexicon for domain specific translation. We defined several quality criteria that could be modeled as modular or submodular functions over the set of patterns mined from a domain specific corpus. The problem of pattern selection is then formulated as an optimization of these criteria and solved to produce a good set of representative in-domain patterns. Experimental results justify that a cross-domain SMT model indeed benefits from the availability of this high quality in-domain bilingual lexicon and does better in translating domain specific text.

Acknowledgments This research was supported by the Intranet Search project from IRCC at IIT Bombay.

References

1. Bhuiyan, M., Mukhopadhyay, S., Hasan, M.A.: Interactive pattern mining on hidden data: A sampling-based solution. In: Proceedings of the 21st ACM International Conference on Information and Knowledge Management. pp. 95–104. CIKM '12, ACM, New York, NY, USA (2012)
2. Bonchi, F., Giannotti, F., Mazzanti, A., Pedreschi, D.: Exante: Anticipated data reduction in constrained pattern mining. In: Knowledge Discovery in Databases: PKDD 2003, pp. 59–70. Springer (2003)
3. Chen, H., Huang, H., Tjiu, J., Tan, C., Chen, H.: Identification and translation of significant patterns for cross-domain smt applications. Proceedings of Machine Translation Summit XIII (2011)
4. Federico, M., Bertoldi, N., Cettolo, M., Negri, M., Turchi, M., Trombetti, M., Cattelan, A., Farina, A., Lupinetti, D., Martines, A., et al.: The matecat tool. In: Proceedings of COLING. pp. 129–132 (2014)
5. Iyer, R.K., Bilmes, J.A.: Submodular optimization with submodular cover and submodular knapsack constraints. In: Advances in Neural Information Processing Systems. pp. 2436–2444 (2013)
6. Joshi, S., Ramakrishnan, G., Balakrishnan, S., Srinivasan, A.: Information extraction using non-consecutive word sequences. In: Proceedings of TextLink 2007, The Twentieth International Joint Conference on Artificial Intelligence (2007)
7. Kempe, D., Kleinberg, J., Tardos, E.: Maximizing the spread of influence through a social network. In: SIGKDD (2003)

8. Koehn, P., Schroeder, J.: Experiments in domain adaptation for statistical machine translation. In: Proceedings of the Second Workshop on Statistical Machine Translation. pp. 224–227. StatMT '07, Association for Computational Linguistics, Stroudsburg, PA, USA (2007)
9. Krause, A., Golovin, D.: Submodular function maximization. Tractability: Practical Approaches to Hard Problems 3, 19 (2012)
10. Lambert, P.: Data inferred multi-word expressions for statistical machine translation. In: In MT Summit X (2005)
11. Lin, H., Bilmes, J.: Multi-document summarization via budgeted maximization of submodular functions. In: NAACL (2010)
12. Liu, J., Shang, J., Wang, C., Ren, X., Han, J.: Mining quality phrases from massive text corpora. In: Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data. pp. 1729–1744. SIGMOD '15, ACM, New York, NY, USA (2015)
13. Minoux, M.: Accelerated greedy algorithms for maximizing submodular set functions. In: Optimization Techniques, pp. 234–243. Springer (1978)
14. Nemhauser, G.L., Wolsey, L.A., Fisher, M.L.: An analysis of approximations for maximizing submodular set functions. Mathematical Programming 14(1), 265–294 (1978)
15. Nemhauser, G.L., Wolsey, L.A.: Best algorithms for approximating the maximum of a submodular set function. Mathematics of operations research 3(3), 177–188 (1978)
16. Nepveu, L., Lapalme, G., Qubec, M., Foster, G.: Adaptive language and translation models for interactive machine translation. In: In Proceedings of the Conference on Empirical Methods in Natural Language Processing (2004)
17. Och, F.J., Ney, H.: A systematic comparison of various statistical alignment models. Comput. Linguist. 29(1), 19–51 (Mar 2003)
18. Pal, S., Bandyopadhyay, S.: Handling multiword expressions in phrase-based statistical machine translation. Machine Translation Summit XIII pp. 215–224 (2011)
19. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting on association for computational linguistics. pp. 311–318. Association for Computational Linguistics (2002)
20. Ranta, A.: Grammatical framework. Journal of Functional Programming 14(02), 145–189 (2004)
21. Ren, Z., Lü, Y., Cao, J., Liu, Q., Huang, Y.: Improving statistical machine translation using domain bilingual multiword expressions. In: Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications. pp. 47–54. MWE '09, Association for Computational Linguistics, Stroudsburg, PA, USA (2009)
22. Tan, C.M., Wang, Y.F., Lee, C.D.: The use of bigrams to enhance text categorization. Inf. Process. Manage. 38(4), 529–546 (Jul 2002)
23. Tiedemann, J.: News from OPUS - A collection of multilingual parallel corpora with tools and interfaces. In: Recent Advances in Natural Language Processing, vol. V, pp. 237–248. John Benjamins, Amsterdam/Philadelphia, Borovets, Bulgaria (2009)
24. Tiedemann, J.: Parallel data, tools and interfaces in opus. In: Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12). European Language Resources Association (ELRA), Istanbul, Turkey (may 2012)
25. Vogel, S., Ney, H., Tillmann, C.: Hmm-based word alignment in statistical translation. In: Proceedings of the 16th Conference on Computational Linguistics - Volume 2. pp. 836–841. COLING '96, Association for Computational Linguistics, Stroudsburg, PA, USA (1996)
26. Wu, H., Wang, H., Zong, C.: Domain adaptation for statistical machine translation with domain dictionary and monolingual corpora. In: IEEE Signal Processing Magazine (2008)
27. Xin, D., Shen, X., Mei, Q., Han, J.: Discovering interesting patterns through user's interactive feedback. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 773–778. KDD '06, ACM, New York, NY, USA (2006)