Dictionary Generalization Across Languages

Vishwajeet Kumar[1], Ashish Kulkarni[1], Alan Akbik[2], Dr. Ganesh Ramakrishnan[1]

[1]IIT Bombay, Mumbai, India [2]IBM Research, Almaden, CA

Motivation

- Statistical machine translation models trained on large amount of cross domain corpora fails to reliably translate in-domain text.
- Any in-domain sentence aligned parallel corpus is almost non-existent.
- 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.
- Domain corpus is highly redundant and phrases, which might themselves be infrequent, tend to have "consensus" when generalized to higher-level patterns.
- Annotation projection based on parallel corpus has shown great promise in creating proposition banks for languages for which high quality parallel corpora and syntactic parsers are available.

Contributions

An approach to extract such patterns from a domain corpus and curate a high quality bilingual dictionary and a technique to create proposition banks for low resource languages.

- An approach to extract high quality patterns that are: *frequent, syntactically well-formed, and provide maximum corpus coverage*
- An interactive system that gathers human feedback on the translation of these patterns;
- An approach to create proposition banks for low resource languages using bilingual dictionary.

Problem of Mining Quality Patterns

We are given a domain corpus **C** and optionally a set of "types" **T**. The problem of lexicon curation is to extract from **C**, a set **H** of quality patterns, as per a quality function $Q_C(h)$ for the quality of a pattern $h \in H$ in the corpus and a quality function $Q_C(H)$ for the quality of the set **H**.

Extraction of Quality Patterns

Pattern Extraction: Mines frequent patterns from an in-domain source language corpus. An algorithm that uses context free grammar $\mathcal{G}$ to extract from corpus $\mathcal{C}$, a set $\mathcal{H}$ of patterns.
Pattern Selection: Selects a minimal set of quality patterns that are syntactically well-formed and provide maximum corpus coverage.

$$H^* = \underset{H \subseteq \mathcal{H}_Q}{\arg\max} Q_C^2(H) s.t. Q_C^1(H) < c \qquad (1)$$

where c is threshold on modular cost function $Q_C^1(H)$ and $Q_C^2(H)$ is submodular quality function.

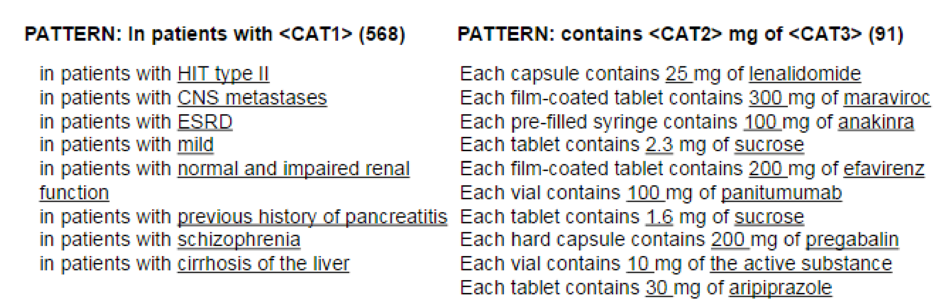


Figure 1: Examples of patterns

Adaptation of Annotation Projection

We adapted annotation projection using bilingual dictionaries to create proposition banks for low resource languages as follows:

- **Target Language Predicates:** only target language verbs that are aligned to literal source language translations are labeled as frames.
- **Target language arguments:** project not only the role label of source language arguments heads but entire argument dependency structure.

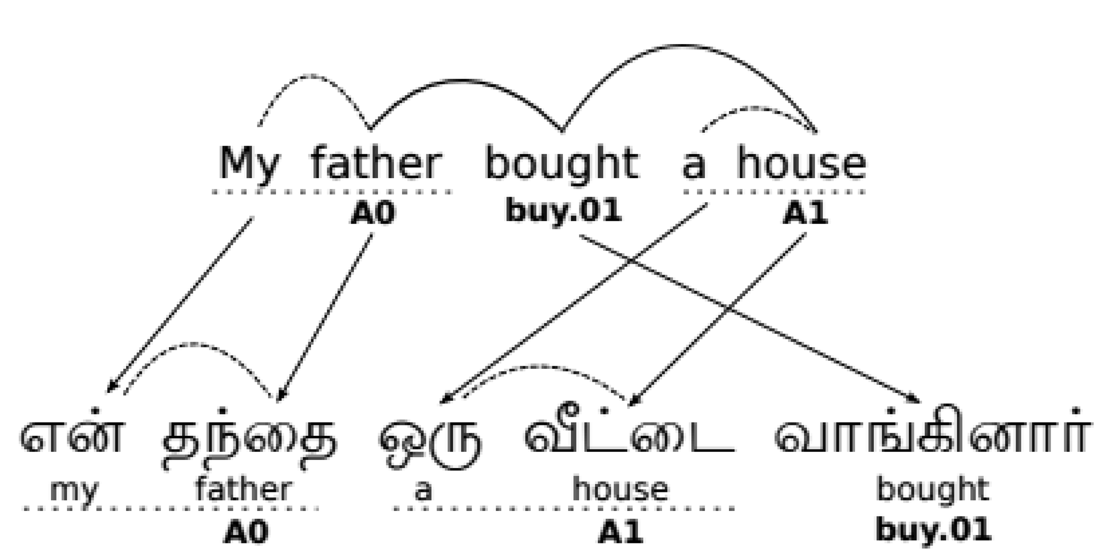


Figure 2: Annotation projection on a pair of simple sentences

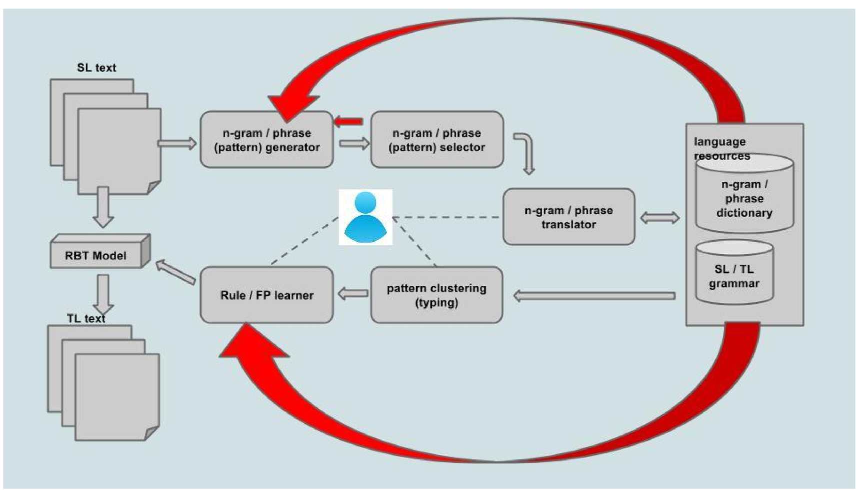


Figure 3: System Architecture

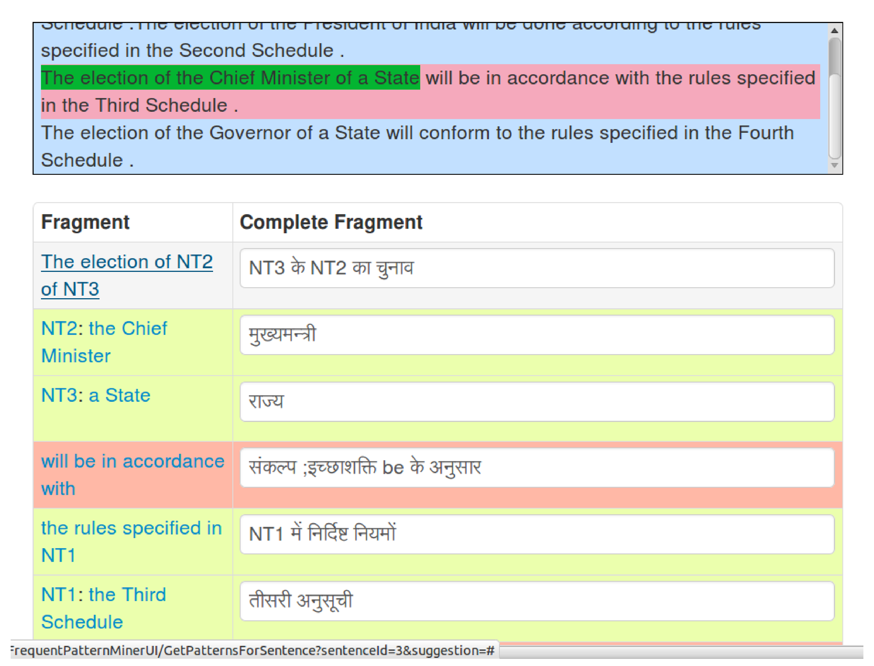


Figure 4: User Interface

Human Translation Framework

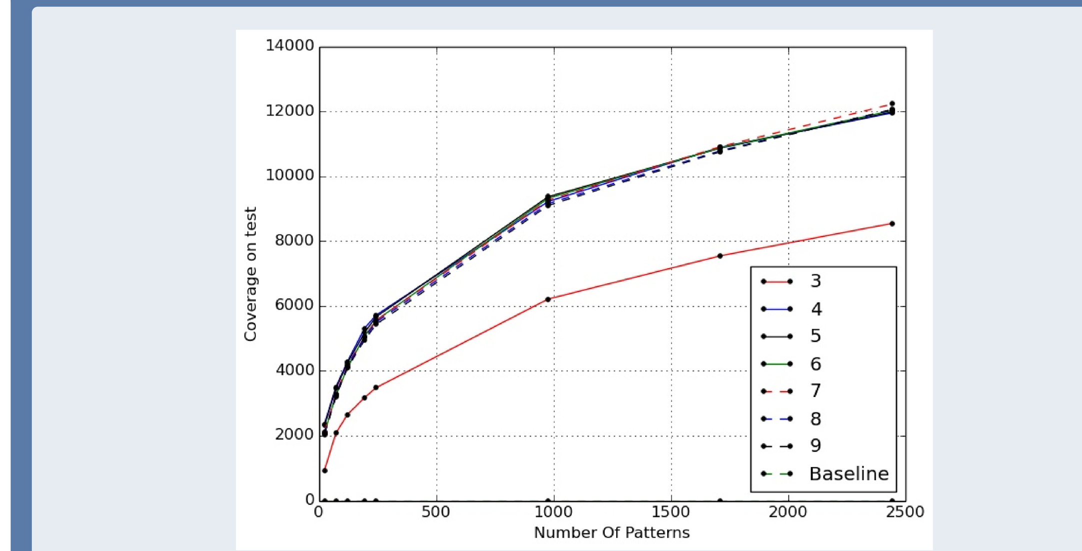
The system for gathering human feedback on pattern translation has these key features:

- A pattern, along with its context, is presented to users;
- Translators can view all instances of non-terminals constituting a pattern;
- Translation suggestions (which users can edit) are obtained from various sources including our rule-based translator, and external statistical and memory-based translators;
- Users can reorder translation of sentence fragments and post-edit the final sentence translation.

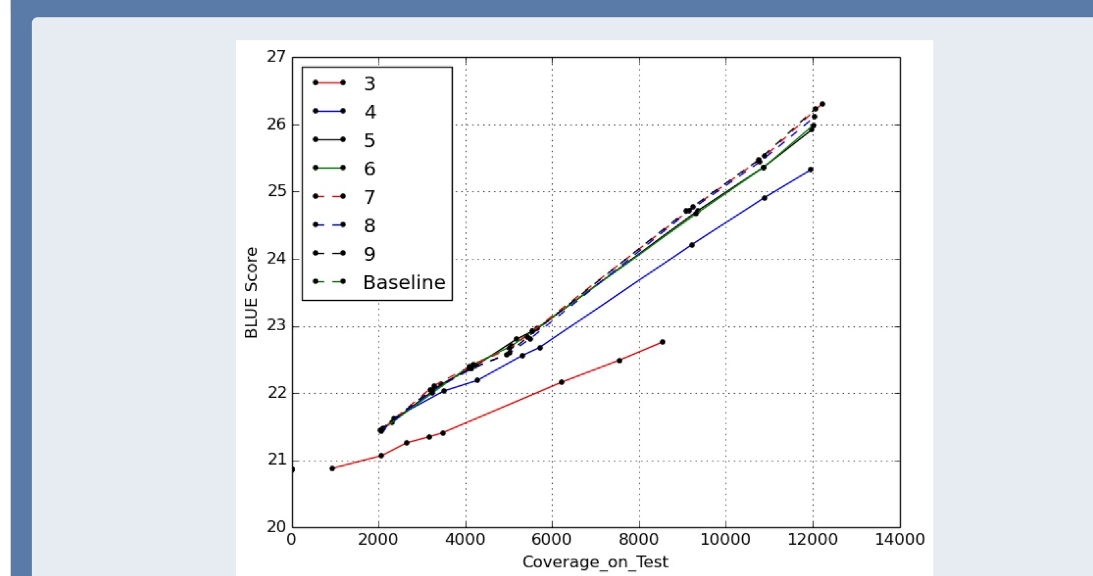
Evaluation

- Split the datasets into MINE and TEST.
- MINE for extracting patterns and TEST for evaluating their coverage.
- Three fold cross validation for pattern extraction.
- Pattern length and frequency threshold varied from 2 to 6.

Effect of varying dictionary size on EMEA(en-fr) corpus coverage



Effect of corpus coverage on translation accuracy of TEST



Conclusion

Given an in-domain corpus, we presented an approach to extract high quality pattern that maximally cover the corpus, a system to leverage humans for high quality translation of patterns. We also presented use of bilingual dictionary in adaptation of annotation projection for creating proposition banks for low resource languages.

References

[1] Pankaj Singh, Ashish Kulkarni, Himanshu Ojha, Vishwajeet Kumar, and Ganesh Ramakrishnan.
Building compact lexicons for cross-domain smt by mining near-optimal pattern sets.
In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 290–303. Springer, 2016.

[2] Vishwajeet Kumar, Ashish Kulkarni, Pankaj Singh, Ganesh Ramakrishnan, and Ganesh Arnaal.
A machine assisted human translation system for technical documents.
In *Proceedings of the 8th International Conference on Knowledge Capture*, page 33. ACM, 2015.

[3] Alan Akbik, Vishwajeet Kumar, and Yunyao Li.
Towards semi-automatic generation of proposition banks for low-resource languages.
In *EMNLP*, pages 993–998, 2016.