English to Urdu
Hierarchical Phrase-based Statistical Machine Translation

Nadeem Khan [1] Waqas Anwar[1] Usama Ijaz Bajwa [1] Nadir Durrani [2]
[1]COMSATS Institute of Information Technology, Abbottabad Pakistan
[2]University of Stuttgart
[1] {nadeemk,waqas,usama}@ciit.net.pk, [2]durrani@ims.uni-stuttgart.de

Abstract

This paper addresses the Hierarchical Phrase-based (HPB) models which are used in development of different Statistical Machine Translation (SMT) Systems for many modern languages.
Any SMT System needs large parallel corpora for accurate performance. Therefore, availability of a large parallel corpus is a pre-requisite for designing a reliable, robust SMT system between any two languages. The HPB models have shown strong capability of generalization and reordering, which in turn gets improved results for the sparse resourced languages.
This paper considers English as Source and Urdu as target language for experiments. For this study, Hierarchical phrase-based Baseline SMT system is used for English to Urdu translation. At the end automatic evaluation of system is performed by using BLEU and NIST as evaluation metrics. Average BLEU evaluation score the developed system got is 13% which is a good competitive score for any sparse resourced language.

Keywords: HPB model, Statistical Machine Translation, Parallel corpus, Natural Language Processsing.

1 Introduction

Urdu is the national language of Pakistan, and also one of the spoken languages in India and is written in Perso-Arabic script. Urdu consists of many words which come from several languages, including Arabic, Persian and Turkish. English and Urdu, although both belong to the Indo-European language family, word order and morphology have very different characteristics for these two languages.

Both languages have different morphological and syntactic features. Urdu is highly inflectional language hence is rich in morphology. In Urdu the verbs are inflected according to gender, number, and person. English is a fixed word order language and follows the SVO (subject verb object) structure while Urdu is a free word order language and the most common sentence structure used by the native speakers is SOV (subject object verb). Also, instead of English prepositions, Urdu nouns and verbs are followed by postpositions. The above discussion emphasizes the differences between source language English and the target language Urdu.
Hierarchical phrase-based machine translation (Chiang, 2007; Watanabe et al., 2006) belongs to the current dominant and promising statistical machine translation approaches influenced by (Brown et al., 1993). Although the model captures, global reordering SCFG (Synchronous context-free grammar), the reordering does not explicitly introduces the model to restrict word order. On the contrary, the lexicalized reordering models (Tillman, 2004; Nagata et al., 2006) are often used for the translation on the Phrase-based models. These lexicalized reordering models cannot be applied directly to hierarchical phrase-based translation, because the representation of the hierarchical phrase translation uses the nonterminal symbols.
Hierarchical phrase-based statistical machine translation (Chiang, 2007) has shown the competition between phrase-based and syntax-based models for different language pairs. An important issue with hierarchical set-based translation, the size of the model on which training is carried out, which is usually several times larger than the trained phrase-based counterpart from the same dataset. This leads to over generation search errors and a slow decoder (de Gispert et al., 2010). In this work, the main focus is on hierarchical models

The 4th Workshop on South and Southeast Asian NLP (WSSANLP), International Joint Conference on Natural Language Processing, pages 72–76,
Nagoya, Japan, 14-18 October 2013.

usage in SMT, focused on the expression of (Chiang,.2007), which is officially based on the syntax, and always consult the term SCFG.
HPB SMT (Chiang, 2005) is a tree-based model that automatically extracts a synchronous CFG from the training corpus. HPB SMT extracted hierarchical rules, the basic units of translation in the HPB model phrases extracted according to the PB model (Koehn et al., 2003). So, hierarchical rules have the strengths of the statistically extracted continuous sets plus the ability to translate interrupted sentences too and learn sentence reordering without a separate reordering model. The HPB SMT model has two kinds of rules: hierarchical rules and glue grammar rules. Hierarchical set-based translation (Chiang, 2005; Chiang,2007) expand highly lexicalized-Based models of sentence translation systems lexicalized rearrangement model and disjoint sets. A major drawback with this approach compared to set-based systems is that the total number of rules that are learned are several orders of magnitude larger than standard tables, which leads to over-generation rate and help search error and too much longer decoding time. Chiangs hierarchical phrase-based (HPB) translation model uses SCFG for translation free derivation (Chiang, 2005; Chiang, 2007) and has been widely accepted in SMT.

2 Previous Work

Various work with different approaches have been proposed for many Subcontinent Languages in the field of machine translation when translating from English into anyone of these languages or when discussing the divergence in translation of anyone of these languages specially for Hindi.
(Chiang et.al.2005) proposed Hierarchical Phrase-Based Model for SMT for different European languages, that can be made applicable for sparse resourced languages by having some tuning on the model.
(Jawaid et.al,.2011) investigated phrase-based statistical machine translation between English and Urdu, two Indo-European languages that differ significantly in their word-order preferences. They showed reordering of words and phrases is a necessary part of the translation process. While local reordering is modeled nicely by phrase-based systems. Again long-distance reordering is known to be a hard problem. They performed experiments using the Moses SMT system. They also presented an Urdu aware approach based on reordering phrases in syntactic parse tree of the source English sentence.
(Singh et.al. 2004) proposed an approach for English to Bangla MT that uses syntactic transfer of English sentences to Bangla with optimal time complexity. In generation stage they used a dictionary to identify subject, object and also other information like person, number and generate target sentences.
(Singh,.2012) presented a Phrase based model approach to English-Hindi translation. In this work they discussed the simple implementation of default phrase-based model for SMT for English to Hindi and also give an overview of different Machine translation application that are in use nowadays
(Sharma et.al.2011) presented a baseline Phrase-based system for English to Hindi Translation with a pretty small amount of data. They used human evaluation metrics as their evaluation measures. These evaluations cost higher than the already available automatic evaluation metrics.
(Islam et al.2010) proposed a phrase-based Statistical Machine Translation (SMT) system that translates English sentences to Bengali. They added a transliteration module to handle OOV (Out of Vocabulary) words. A preposition handling module is also incorporated to deal with systematic grammatical differences between English and Bangla. To measure the performance of their system, they used BLEU, NIST and TER scores as their evaluation metrics. The dataset collected from KDE and EMILLE corpus.
By having a look at the work above, It is clear that there is not a single proposed work when talking about one of the important language of South Asia namely Urdu.

3 Evaluation

This section discusses the training, tuning and testing of different model components. The evaluation was carried on Ubuntu 11.10 running on Intel Core i3 machine with 4GB of RAM and 500GB of Hard disk space.

3.1 Dataset

We used the EMILLE (Enabling Minority Language Engineering) corpus. EMILLE was a 3 year EPSRC project at Lancaster University and Sheffield University. Its final output was an elec-

tronic corpus of 97 million words of South Asian languages and is becoming a standard data repository for the languages of this region.
The parallel corpus consists of 200,000 words of text in English and its accompanying translations in Hindi, Bengali, Punjabi, Gujarati and Urdu. Its bilingual resources consists of roughly above 12,500 sentences for all the available languages.
We were able to do sentence alignment and extract over 8,000 sentence for our required language pair i.e. English and Urdu using the sentence alignment algorithm given by (Moore.,2002).In any SMT development project making of parallel corpus is the most complicated task.
EMILLE corpus is the most rough available corpus ever seen for languages of this region. Cleaning of this corpus for making it completely compatible and sentence aligned, is the very first step and also the toughest one that is used in the development of any SMT system. Further details about parallel data are given in Table 1.

Total Sentences	Training	Tuning	Testing
8245	6596	825	824

Table 1: Complete Statistics of Parallel Corpus

The monolingual resources used in the development of language model for this study work consists of overall 40,000 above segments in which there are target parallel corpus of Urdu with corpus of Quran and Bible that is available for free on web.

3.2 Experimental Setup

The k-fold cross validation method was used for sampling of the corpus. Here k=5 was selected by taking 4/5 of the total corpus as training and 1/5 as tuning and test set for experiment on all folds. Each fold comprises roughly above 800 tuning and same number of sentences for testing along with above 6500 sentences for training. After sampling of data, tokenization of training set, tuning set and test set is done for all folds dataset followed by lowercasing of datasets. All this is done by the scripts being supplied with the Moses decoder (Koehn et al., 2007). This lowercased training data is used for word alignment. As the overall training data is very sparse, so there is no need to use clean-out script for cleaning long sentences from the training data before training the system.
We ran Moses (Koehn et al., 2007) using Koehn's training scripts. In our work additional switches like hierarchical and glue-grammar are also used in training command as the experiments are being done with the HPB model. For the other parameters, the default values were used i.e. 3-gram language model and maximum phrase-length= 6. The word alignments for the system are extracted by using GIZA++ (Och and Ney, 2003) that is linked with the training script of Moses.
Language Model is built on the available monolingual Urdu corpus. This language model is implemented as an n-gram model using the SRILM Toolkit. For all the experiments, the same language model is used for all folds as translation is being performed from English into Urdu. System tuning for the extraction of optimized feature weights that would be used in testing of the developed system is done by using MERT (Och, 2003). When it comes to testing, the testing phase was completed by using the Moses decoder. Note that the command used here is moses-chart instead of moses. This is because the work is being carried out with hierarchical-phrase based model. This testing is done in the same way for all fold test sets.

3.3 Results

All the evaluation scores and some sample translations from the developed SMT system are given in this section:
As working on a sparse resourced language, we have achieved much better BLEU scores with a mean of 0.132 and a Standard deviation of 0.09 for the entire given test sets. Overall results of BLEU and NIST evaluation score for all folds are given below in Table: 2 and 3 respectively.
A comparison of the developed Hierarchical Phrase-based translation system with the traditional Phrase-based system was also carried out. It can be noted from Table 2 and 3 the traditional Phrase-based model system got better BLEU and NIST scores as compared to the Hierarchical Phrase-based model approach for this language pair. This is because of the morphological richness and the divergence of Urdu with English. Another important factor that is causing this small difference between the results of two models is the sparseness of corpus for this language pair as we got such a small size of corpus in our experiments. From Table 2, It can be noted the sudden change occurred in BLEU score for fold-5 of both the tra-

ditional Phrase-based system as well as our Hierarchical Phrase-based system, this change is due to the relevance in domains of test and training dataset.

Folds	f1	f2	f3	f4	f5
Phrase-based	0.11	0.08	0.12	0.13	0.40
Hierarchical	0.09	0.06	0.10	0.10	0.29

Table 2: Comparison of BLEU evaluation score with traditional Phrase-based model

Folds	f1	f2	f3	f4	f5
Phrase-based	3.52	3.25	3.92	4.27	7.36
Hierarchical	3.42	3.26	3.80	4.16	6.36

Table 3: Comparison of NIST evaluation score with traditional Phrase-based model

From the BLEU and NIST evaluation score results above, It can be noted clearly how much sparseness and difference in domain of datasets within all 5 folds.In BLEU score, fold-5 got the highest percentage of result i.e. 29%, and in other fold there is not a single result greater than 10% for our Hierarchical Phrase-based system, this is because of the difference in domains of training and testing corpus. In fold-5 the test data become pretty much closer to the domains of dataset on which system is actually trained.
Figure 1 shows how the decoder perform translations of the test dataset using the chart decoder for Hierarchical Phrase-based model.

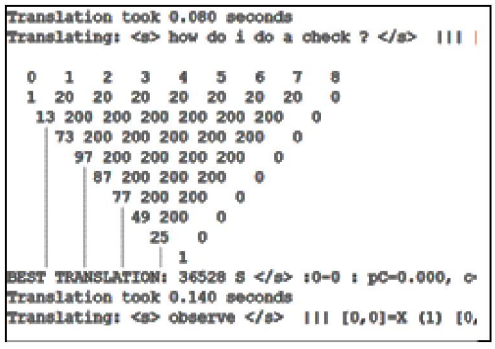

Figure 1: Working of Hierarchical Phrase-based decoder.

Some example of translation generated by our developed English-Urdu translation system are given in Table 4. There is much difference in ordering of words in some sentences like in third sentence from the Table 4, the decoder did not changed the order of words while translating in Urdu, it just translated the words one by one from English side and give us completely out of order output on Urdu side.

Input English Sentence	Output URdu Sentence
Parent Partnership Service	والدین شراکت سروس
you are aged 60 or over	آپ کی عمر 60 سال یا اس سے زیادہ
Not working,or working on average less than 16 hours a week?	کام نہیں، یا ہفتے میں اوسطاً 16 گھنٹے ؟
How to claim see page 30.	کلیم کیسے کیا جائے دیکھیے صفحہ30۔
Working with the court	عدالت کے ساتھ کام

Table 4: Translations generated by our system.

4 Conclusion

In this work we explored one of the important but relatively less addressed research problems. As Urdu has got rich morphology and its inflectional behavior is also very variable, therefore, an extensively sparse corpus for experiments was used. Also because of the sparseness of the corpus, the evaluation results are not that much impressive as compared to those for some of the European languages.
In this work Hierarchical Phrase-based model for training was employed. We carried out a set of experiments by choosing the training, tuning and test sets from parallel corpus using the k-fold cross validation method to cater for the fact that we had only a small amount of parallel data. We found that our targeted language Urdu has got pretty much divergence when translating from English and that is the reason for that much difference in obtained MT evaluation scores on our given testsets.

References

Andreas Stolcke 2002. *SRILM - an extensible language modeling toolkit.*. In International Conference Spoken Language Processing, Denver, Colorado.

Adri a de Gispert et.al 2010. *Hierarchical phrase-based translation with weighted finite-state transducers and shallow-n grammars.*. In Proceedings of Association of Computational Linguistics.

Brown. P et.al 1993. *The mathematics of statistical*

machine translation: Parameter estimation.. In Proceedings of Association of Computational Linguistics.

Bushra Jawaid et.al 2011. *Word-Order Issues in English-to-Urdu Statistical Machine Translation.*. The Prague Bulletin of Mathematical Linguistics.

Chiang, D 2005. *A hierarchical phrase-based model for statistical machine translation.*. In Proceedings of Association of Computational Linguistics.

Chiang, D 2007. *Hierarchical phrase-based model for statistical machine translation.*. In Proceedings of Association of Computational Linguistics.

Dharvendra Singh et al 2012. *Modelling Phrase-Based Statistical Machine Translation for English-Hindi Language.*. IJRREST: International Journal of Research Review in Engineering Science and Technology

Franz J. Och 2003. *Minimum Error Rate Training in Statistical Machine Translation*. In 41st Annual Meeting of the Association for Computational Linguistics (ACL 2003), Sapporo, Japan.

Franz J. Och and Hermann Ney 2003. *A systematic comparison of various statistical alignment models.*. In Proceedings of Association of Computational Linguistics.

Maxim Roy. 2009. *A Semi-supervised Approach to Bengali-English Phrase-Based Statistical Machine Translation,* . Proceedings of the 22nd Canadian Conference on Artificial Intelligence

Nagata et al 2006. *A clustered global phrase reordering model for statistical machine translation.*. In Proceedings of Association of Computational Linguistics.

Nakul Sharma, Parteek Bhatia and Varinderpal Singh 2011. *English to Hindi Statistical Machine Translation.*. International Journal in Computer Networks and Security (IJCNS).

Philip Koehn, Franz J. Och and D.Marcu . 2003. *Statistical Phrase-Based Translation* . In HLT-NAACL 2003: conference combining Human Language Technology conference series and the North American Chapter of the Association for Computational Linguistics conference series, Edmonton, AB, 2003.

Philip Koehn et.al 2005. *Edinburgh system description for 2005 IWSLT speech translation evaluation.*. In Proceedings. the 2nd IWSLT.

Philip Koehn et.al 2007. . *Moses: Open Source Toolkit for Statistical Machine Translation.*. Annual Meeting of the Association for Computational Linguistics (ACL).

Robert C. Moore 2002. *Fast and accurate sentence alignment of bilingual corpora,* . In Conference of the Association for Machine Translation in the Americas (AMTA).

Sajib Dasgupta, Abu Wasif and Sharmin Azam 2004. *An Optimal Way towards Machine Translation from English to Bengali*. Proceedings of the 7th International Conference on Computer and Information Technology (ICCIT).

Tillman. C 2004. *A unigram orientation model for statistical machine translation.*. In Proc. HLT-NAACL.

Watanabe, T. H. Tsukada, and H. Isozaki 2006. *Left-to-right target generation for hierarchical phrase-based translation.*. In Proceedings of Association of Computational Linguistics.

Zahurul Islam, Jrg Tiedemann and Andreas Eisele 2010. *English to Bangla Phrase Based Machine Translation.*. In the Proceedings of the 14th Annual Conference of The European Association for Machine Translation.