Three models for discriminative machine translation using Global Lexical Selection and Sentence Reconstruction

Sriram Venkatapathy
Language Technologies Research
Centre, IIIT-Hyderabad
Hyderabad - 500019, India.
sriram@research.iiit.ac.in

Srinivas Bangalore
AT&T Labs - Research
Florham Park, NJ 07932
USA
srini@research.att.com

Abstract

Machine translation of a source language sentence involves selecting appropriate target language words and ordering the selected words to form a well-formed target language sentence. Most of the previous work on statistical machine translation relies on (local) associations of target words/phrases with source words/phrases for lexical selection. In contrast, in this paper, we present a novel approach to lexical selection where the target words are associated with the entire source sentence (global) without the need for local associations. This technique is used by three models (Bag–of–words model, sequential model and hierarchical model) which predict the target language words given a source sentence and then order the words appropriately. We show that a hierarchical model performs best when compared to the other two models.

1 Introduction

The problem of machine translation can be viewed as consisting of two subproblems: (a) lexical selection, where appropriate target language lexical items are chosen for each source language lexical item and (b) lexical reordering, where the chosen target language lexical items are rearranged to produce a meaningful target language string. Most of the previous work on statistical machine translation, as exemplified in (Brown et al., 1993), employs word–alignment algorithm (such as GIZA++ (Och et al., 1999)) that provides local associations between source words and target words. The source–to–target word–alignments are

sometimes augmented with target–to–source word alignments in order to improve the precision of these local associations. Further, the word–level alignments are extended to phrase–level alignments in order to increase the extent of local associations. The phrasal associations compile some amount of (local) lexical reordering of the target words—those permitted by the size of the phrase. Most of the state–of–the–art machine translation systems use these phrase–level associations in conjunction with a target language model to produce the target sentence. There is relatively little emphasis on (global) lexical reordering other than the local re-orderings permitted within the phrasal alignments. A few exceptions are the hierarchical (possibly syntax–based) transduction models (Wu, 1997; Alshawi et al., 1998; Yamada and Knight, 2001; Chiang, 2005) and the string transduction models (Kanthak et al., 2005).

In this paper, we present three models for doing discriminative machine translation using *global lexical selection* and *lexical reordering*.

1. **Bag–of–Words model** : Given a source sentence, each of the target words are chosen by looking at the entire source sentence. The target language words are then permuted in various ways and then, the best permutation is chosen using the language model on the target side. The size of the search space of these permutations can be set by a parameter called the permutation window. This model does not allow long distance re-orderings of target words unless a very large permutation window chosen which is very expensive.

2. **Sequential Lexical Choice model** : Given a source sentence, the target words are predicted in an order which is faithful to the or-

der of words in the source sentence. Now, the number of permutations that need to be examined to obtain the best target language strings are much less when compared to the Bag–of–Words model. This model is expected to give good results for language pairs such as English–French for which only local word order variations exist between sentences.

3. **Hierarchical lexical association and re-ordering model** : For language pairs such as English–Hindi or English–Japanese where there is a high degree of global reordering (Figure 1), it is necessary to be able to handle long distance movement of words/phrases. In this approach, the target words predicted through global lexical selection are associated with various nodes of the source dependency tree and then, hierarchical reordering is done to obtain the order of words in the target sentence. Hierarchical reordering allows phrases to distort to longer distances than the previous two models.

Figure 1: Sample distortion between English–Hindi

The outline of the paper is as follows. In Section 2, we talk about the global lexical selection. Section 3 describes three models for global lexical selection and reordering. In Section 4, we report the results of the translation models on English–Hindi language pair and contrast the strengths and limitations of the models.

2 Global lexical selection

For global lexical selection, in contrast to the local approaches of associating target words to the source words, the target words are associated to the entire source sentence. The intuition is that there may be lexico–syntactic features of the source sentence (not necessarily a single source word) that might trigger the presence of a target word in the target sentence. Furthermore, it might be difficult to exactly associate a target word to a source sentence in many situations - (a) when

translations are not exact but paraphrases (b) the target language does not have one lexical item to express the same concept that is expressed in the source word. The extensions of word alignments to phrasal alignments attempt to address some of these situations in additional to alleviating the noise in word–level alignments.

As a consequence of the global lexical selection approach, we no longer have a tight association between source language words/phrases and target language words/phrases. The result of lexical selection is simply a bag of words(phrases) in the target language and the target sentence has to be reconstructed using this bag of words.

The target words in the bag, however, might be enhanced with rich syntactic information that could aid in the reconstruction of the target sentence. This approach to lexical selection and sentence reconstruction has the potential to circumvent the limitations of word–alignment based methods for translation between significantly different word order languages. However, in this paper, to handle large word order variations, we associate the target words with source language dependency structures to enable long distance reordering.

3 Training the discriminative models for lexical selection and reordering

In this section, we present our approach for a global lexical selection model which is based on discriminatively trained classification techniques. Discriminant modeling techniques have become the dominant method for resolving ambiguity in speech and natural language processing tasks, outperforming generative models for the same task. We expect the discriminatively trained global lexical selection models to outperform generatively trained local lexical selection models as well as provide a framework for incorporating rich morpho–syntactic information.

Statistical machine translation can be formulated as a search for the best target sequence that maximizes $P(T \mid S)$, where S is the source sentence and T is the target sentence. Ideally, $P(T \mid S)$ should be estimated directly to maximize the conditional likelihood on the training data (discriminant model). However, T corresponds to a sequence with a exponentially large combination of possible labels, and traditional classification approaches cannot be used directly. Although

97

Conditional Random Fields (CRF) (Lafferty et al., 2001) train an exponential model at the sequence level, in translation tasks such as ours the computational requirements of training such models are prohibitively expensive.

3.1 Bag-of-Words Lexical Choice Model

This model doesn't require the sentences to be word aligned in order to learn the local associations. Instead, we take the sentence aligned corpus as before but we treat the target sentence as a bag–of–words or BOW assigned to the source sentence. The goal is, given a source sentence S, to estimate the probability that we find a given word (t_j) in its translation ie.., we need to estimate the probabilities $P(true|t_j, S)$ and $P(false|t_j, S)$. To train such a model, we need to build binary classifiers for all the words in the target language vocabulary. The probability distributions of these binary classifiers are learnt using maximum entropy model (Berger et al., 1996; Haffner, 2006). For the word t_j, the training sentence pairs are considered as positive examples where the word appears in the target, and negative otherwise. Thus, the number of training examples for each binary classifier equals the number of training examples. In this model, classifiers are training using n–gram features (BOgrams(S)).

During decoding, instead of producing the target sentence directly, what we initially obtain is the target bag of words. Each word in the target vocabulary is detected independently, so we have here a very simple use of binary static classifiers. Given a sentence S, the bag of words ($BOW(T)$) contains those words whose distributions have the positive probability greater than a threshold (τ).

$$BOW(T) = \{t \mid P(true \mid t, BOgrams(S)) > \tau\}$$
(1)

In order to reconstruct the proper order of words in the target sentence, we consider various permutations of words in $BOW(T)$ and weight them by a target language model. Considering all possible permutations of the words in the target sentence is computationally not feasible. But, the number of permutations examined can be reduced by using heuristic forward pruning or by constraining the permutations to be within a local window of adjustable size (also see (Kanthak et al., 2005)). We have chosen to constrain permutations here. Constraining the permutation using a local window can provide us some very useful local re-

orderings.

The bag–of–words approach can also be modified to allow for length adjustments of target sentences, if we add optional deletions in the final step of permutation decoding. The parameter τ and an additional word deletion penalty δ can then be used to adjust the length of translated outputs.

3.2 Sequential Lexical Choice Model

The previous approach gives us a predetermined order of words initially which are then permuted to obtain the best target string. Given that we would not be able to search the entire space, it would be a helpful if we could start searching various permutations using a more definite string. One such definite order in which the target words can be placed is the order of source words itself. In this model, during the lexical selection, we try to place the target words in an order which is faithful to the source sentence.

This model associates sets of target words with every position in the source sentence and yet retains the power of global lexical selection. For every position (i) of the source sentence, a prefix string is formed which consists of the sequence of words from positions 1 to i. Each of these prefix strings are used to predict bags of target words using the global lexical selection. Now, these bags generated using the prefix strings are processed in the order of source positions. Let T_i be the bag of target words generated by prefix string i (Figure 2).
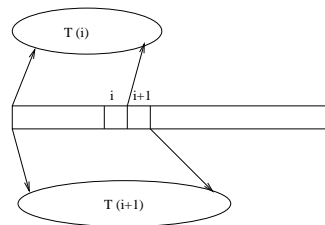

Figure 2: The generation of target bags associated with source sentence position

The goal is to associate a set of target words with every source position. A target word t is attached to the ith source position if it is present in T_i but not in T_{i-1} and the probability $P(true|t, T_i) > \tau$. The intuition behind this approach is that a word t is associated with a position i if there was some information present at the ith source position that triggered the probability of the t to exceed the threshold τ.

Hence, the initial target string is the sequence of target language words associated with the sequence of source language positions. This string is now permuted in all possible ways (section 3.1) and the best target string is chosen using the language model.

3.3 Hierarchical lexical association and reordering model

The *Sequential Lexical Choice Model* presented in the last section is expected to work best for language pairs for which there are mostly local word order variations. For language pairs with significant word order variation, the search for the target string may still fail examine the best target language string given the source sentence. The model proposed in this section should be able to handle such long distance movement of words/phrases.

In this model, the goal is to search for the best target string T which maximizes the probability $P(T|S, D(S))$, where S is the source sentence and $D(S)$ is the dependency structure associated with the source sentence S. The probabilities of the target words given the source sentence are estimated in the same way as the bag–of–words model. The only main difference during the estimation stage is that we consider the dependency tree based features apart from the n-gram features.

The decoding of the source sentence S takes place in three steps,

1. Predict the bag–of–words : Given a source sentence S, predict the bag of words BOW(T) whose distributions have a positive probabilities greater than a threshold (τ).

2. Attachment to Source nodes : These target words are now attached to the nodes of source dependency trees. For making the attachments, the probability distributions of target words conditioned on features local to the source nodes are used.

3. Ordering the target language words : Traverse the source dependency tree in a bottom-up fashion to obtain the best target string.

3.3.1 Predict the bag–of–words

Given a source sentence S, all the target words whose positive probability distributions are above τ are included in the bag.

$$BOW(T) = \{t \mid P(true|t, f(S))\} \quad (2)$$

In addition to the n–gram features, this model uses cues provided by the dependency structure to predict the target bag–of–words.
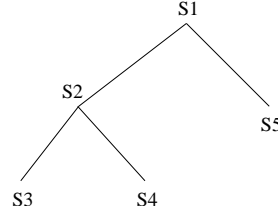

Figure 3: Dependency tree of a source sentence with words s1, s2, s3, s4 and s5

Hence, the features that we have considered in the model are (Figure 3),

1. N-grams. For example, in Figure 2, 's1', 's2 s3 s4', 's4 s5' etc.

2. Dependency pair (The pair of nodes and its parents). Example in Figure 2., 's2 s1', 's4 s2' etc.

3. Dependency treelet (The triplet of a node, it's parent and sibling). For example, 's3 s2 s4', 's2 s1 s5' etc.

3.3.2 Attachment to Source nodes

For every target word t_j in the bag, the most likely source nodes are determined by measuring the positive distribution of the word t_j given the features of the particular node (Figure 4). Let $S(t_j)$ denote the set of source nodes to which the word t_j can be attached to, then $S(t_j)$ is determined as,
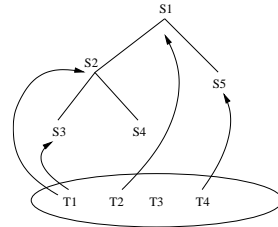

Figure 4: Dependency tree of a source sentence with words S1, S2, S3, S4 and S5

$$S(t_j) = argmax_s(P(true|t_j, f(s)) \quad (3)$$

where $f(s)$ denotes the features of S in which only those features are active which contain the

lexical item representing the node s. The target words are in the global bag are processed in the order of their global probabilities $p(t|S)$. While attaching the target words, it is ensured that no source node had more than ρ target words attached to it. Also, a target word should not be attached to more to more than σ number of times. There is another constraint that can be applied to ensure that the ratio of the total target words (which are attached to source nodes) to the total number of words in the source sentence does exceed a value (μ).

3.4 Ordering the target language words

In this step, the source sentence dependency tree is traversed in a bottom–up fashion. At every node, the best possible order of target words associated with the sub-tree rooted at the node is determined. This string is then used as a cohesive unit by the superior nodes.
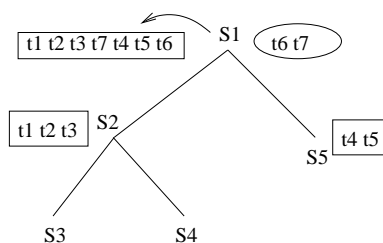
Figure 5: The target string associated with node S1 is determined by permuting strings attached to the children (in rectangular boxes, to signify that they are frozen) and the lexical items attached to S1

For example, in Figure 5, let 't1 t2 t3', 't4 t5' be the best strings associated with the children of nodes s2 and s3 respectively. Let t6 and t7 be the words that are attached to node s1. The best string for the node s1 is determined by permuting the strings 't1 t2 t3', 't4 t5', 't6' 't7' in all possible ways and then choosing the best string using the language model.

4 Dataset

The language pair that we considered for our experiments are English–Hindi. The training set consists of 37967 sentence pairs, the development set contains 819 sentence pairs and the test set has 699 sentence pairs. The dataset is from the newspaper domain with topics ranging from politics to tourism. The sentence pairs have a maxi-

mum source sentence length of 30 words. The average length of English sentences is 18 while that of Hindi sentences is 20.

The source language vocabulary is 41017 and target sentence vocabulary is 48576. The token/type ratio of English in the dataset is 16.70 and that of Hindi is 15.64. This dataset is relatively sparse. So, the translation accuracies on this dataset would be relatively less when compared to those on much larger datasets. In the target side of the development corpus, the percentage of unseen tokens is 13.48%(3.87% types) while in the source side, the percentage of unseen tokens is 10.77%(3.20% types). On furthur inspection of a small portion of the dataset, we found that the maximum percentage of the unseen words on the target side are the named entities.

5 Results

5.1 Bag-of-Words model

The quality of the bag–of–words obtained is governed by the parameter τ (probability threshold). To determine the best τ value, we experiment with various values of τ and measure the lexical accuracies (F-score) of the bags generated on the development set (See Figure 6). The total number of features used for training this model are 53166 (with count-cutoff of 2).
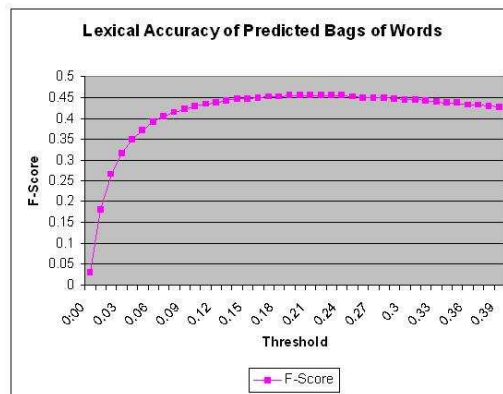
Figure 6: Lexical Accuracies of the Bags-of-words

Now, we order the bags of words obtained through global selection to get the target language strings. While reordering using the language model, some of the noisy words from the bag can be deleted by setting a deletion cost (δ). We experimented with various deletion costs, and tuned it according to the best BLEU score that we

obtained on the development set. Figure 7 shows the best BLEU scores obtained by reordering the bags associated with various threshold values.
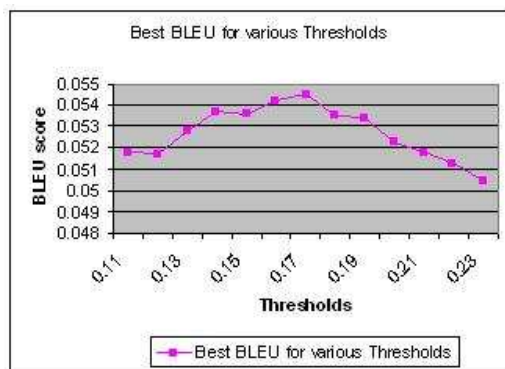

Figure 7: Lexical Accuracies of the Bags-of-words

We can see that we obtained the best BLEU when we choose a threshold of 0.17 to obtain the bag–of–words, when the deletion cost is set to 19.

The reference target strings of the development set has 15986 tokens. So, while tuning the parameters, we should ensure that the bags (obtained using the global lexical selection) that we consider have more tokens than 15986 to allow some deletions during reordering, and in effect obtain the target strings whose total token count is approximately equal to 15986. Figure 8 shows the variation in BLEU scores for various deletion costs by fixing the threshold at 0.17.
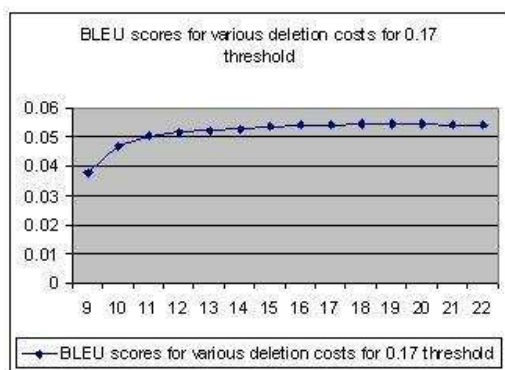

Figure 8: BLEU scores for various deletion costs when the threshold for global lexical selection is set to 0.17

On the test set, we now fix the threshold at 0.17 (τ) and the deletion cost (δ) at 19 to obtain the target language strings. The BLEU score that we obtained for this set is 0.0428.

5.2 Sequential Lexical Choice Model

The lexical accuracy values of the sequence of words obtained by the sequential lexical choice model are comparable to those obtained using the bag–of–words model. The real difference comes for the BLEU score. The best BLEU score obtained on the development set was **0.0586** when τ was set to 0.14 and deletion cost was 15. On the test set, the BLEU score obtained was 0.0473.

5.3 Tree based model

The lexical accuracy values of the words obtained in this model are comparable to the lexical accuracy values of the bag of words model. The total number of features used for training this model are 118839 (with count-cutoff of 2). On the development set, we obtained a BLEU score of **0.0650** for τ set at 0.17 and the deletion cost set at 20. On the test set, we obtained a BLEU score of 0.0498. We can see that the BLEU scores are now better than the ones obtained using any of the other models discussed before. This is because the Tree based model has both the strengths of the global lexical selection that ensures high quality lexical items in the target sentences and that of an efficient reconstruction model which takes care of long distance reordering. The table summarizes the BLEU scores obtained by the three models on the development and test sets.

	Devel. Set	Test. Set
Bag-of-Words	0.0545	0.0428
Sequential	0.0586	0.0473
Hierarchical	0.0650	0.0498

Table 1: Summary of the results

6 Conclusion

In this paper, we present a novel approach to lexical selection where the target words are associated with the entire source sentence (global) without the need for local associations. This technique is used by three models (Bag–of–words model, sequential model and hierarchical model) which predict the target language words given a source sentence and then order the words appropriately. We show that a hierarchical model performs best when compared to the other two models. The hierarchical model presented in this paper has both the strengths of the global lexical selection and efficient reconstruction model.

In the future, we are planning to improve the hierarchical model by making two primary additions

- Handling cases of structural non-isomorphism between source and target sentences.

- Obtaining K-best target string per node of the source dependency tree instead of just one per node. This would allow us to explore more possibilities without having to compromise much on computational complexity.

References

Hiyan Alshawi, Srinivas Bangalore, and Shona Douglas. 1998. Automatic acquisition of hierarchical transduction models for machine translation. In *Proceedings of the 36th Annual Meeting Association for Computational Linguistics*, Montreal, Canada.

A.L. Berger, Stephen A. D. Pietra, D. Pietra, and J. Vincent. 1996. A Maximum Entropy Approach to Natural Language Processing. *Computational Linguistics*, 22(1):39–71.

P. Brown, S.D. Pietra, V.D. Pietra, and R. Mercer. 1993. The Mathematics of Machine Translation: Parameter Estimation. *Computational Linguistics*, 16(2):263–312.

David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 263–270, Ann Arbor, Michigan, June. Association for Computational Linguistics.

P. Haffner. 2006. Scaling large margin classifiers for spoken language understanding. *Speech Communication*, 48(iv):239–261.

S. Kanthak, D. Vilar, E. Matusov, R. Zens, and H. Ney. 2005. Novel reordering approaches in phrase-based statistical machine translation. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, pages 167–174, Ann Arbor, Michigan.

J. Lafferty, A. McCallum, and F. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of ICML*, San Francisco, CA.

Franz Och, Christoph Tillmann, and Herman Ney. 1999. Improved alignment models for statistical machine translation. In *In Proc. of the Joint Conf. of Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 20–28.

Dekai Wu. 1997. Stochastic Inversion Transduction Grammars and Bilingual Parsing of Parallel Corpora. *Computational Linguistics*, 23(3):377–404.

K. Yamada and K. Knight. 2001. A syntax-based statistical translation model. In *Proceedings of 39th ACL*.