Using Information about Multi-word Expressions for the Word-Alignment Task

Sriram Venkatapathy[1]
Language Technologies Research Center,
Indian Institute of
Information Technology,
Hyderabad, India.
sriramv@linc.cis.upenn.edu

Aravind K. Joshi
Department of Computer and
Information Science and Institute for
Research in Cognitive Science,
University of Pennsylvania, PA, USA.
joshi@linc.cis.upenn.edu

Abstract

It is well known that multi-word expressions are problematic in natural language processing. In previous literature, it has been suggested that information about their degree of compositionality can be helpful in various applications but it has not been proven empirically. In this paper, we propose a framework in which information about the multi-word expressions can be used in the word-alignment task. We have shown that even simple features like point-wise mutual information are useful for word-alignment task in English-Hindi parallel corpora. The alignment error rate which we achieve (AER = 0.5040) is significantly better (about 10% decrease in AER) than the alignment error rates of the state-of-art models (Och and Ney, 2003) (Best AER = 0.5518) on the English-Hindi dataset.

1 Introduction

In this paper, we show that measures representing compositionality of multi-word expressions can be useful for tasks such as Machine Translation, word-alignment to be specific here. We use an on-line learning framework called MIRA (McDonald et al., 2005; Crammer and Singer, 2003) for training a discriminative model for the word alignment task (Taskar et al., 2005; Moore, 2005). The discriminative model makes use of features which represent the compositionality of multi-word expressions.

Multi-word expressions (MWEs) are those whose structure and meaning cannot be derived from their component words, as they occur independently. Examples include conjunctions such as '*as well as*' (meaning 'including'), idioms like '*kick the bucket*' (meaning 'die') phrasal verbs such as '*find out*' (meaning 'search') and compounds like '*village community*'. They can be defined roughly as idiosyncratic interpretations that cross word boundaries (Sag et al., 2002).

A large number of MWEs have standard syntactic structure but are semantically non-compositional. Here, we consider the class of verb based expressions (verb is the head of the phrase), which occur very frequently. This class of verb based multi-word expressions include verbal idioms, support-verb constructions, among others. The example '*take place*' is a MWE but '*take a gift*' is not.

In the past, various measures have been suggested for measuring the compositionality of multi-word expressions. Some of these are mutual information (Church and Hanks, 1989), distributed frequency (Tapanainen et al., 1998) and Latent Semantic Analysis (LSA) model (Baldwin et al., 2003). Even though, these measures have been shown to represent compositionality quite well, compositionality itself has not been shown to be useful in any application yet. In this paper, we explore this possibility of using the information about compositionality of MWEs (verb based) for the word alignment task. In this preliminary work, we use simple measures (such as point-wise mutual information) to measure compositionality.

The paper is organized as follows. In section 2, we discuss the word-alignment task with respect to the class of multi-word expressions of interest in this paper. In section 3, we show empirically,

---

[1]At present visiting Institute for Research in Cognitive Science, University of Pennsylvania, PA, USA.

the behavior of verb based expressions in a parallel corpus (English-Hindi in our case). We then discuss our alignment algorithm in section 4. In section 5, we describe the features which we have used in our training model. Section 6 discusses the training algorithm and in section 7, the results of our discriminative model for the word alignment task. Related work and conclusion follow in section 8 and 9 respectively.

2 Task: Word alignment of verbs and their dependents

The task is to align the verbs and their dependents (arguments and adjuncts) in the source language sentence (English) with words in the target language sentence (Hindi). The dependents of the verbs in the source sentence are represented by their head words. Figure 1. shows an example of the type of multi-word expressions which we consider for alignment.
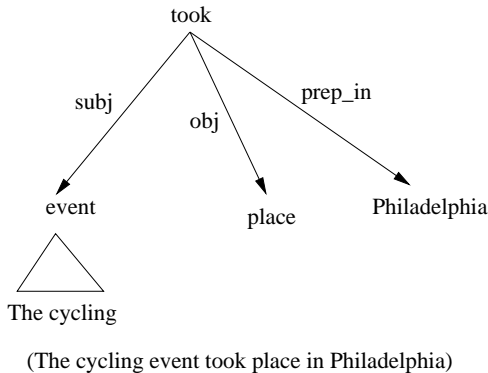

(The cycling event took place in Philadelphia)

Figure 1: Example of MWEs we consider

In the above example, the goal will the to align the words 'took', 'event', 'place' and 'Philadelphia' with corresponding word(s) in the target language sentence (which is not parsed) using a discriminative approach. The advantage in using the discriminative approach for alignment is that it lets you use various compositionality based features which are crucial towards aligning these expressions. Figure 2. shows the appropriate alignment of the expression in Figure 1. with the words in the target language. The pair (take place), in English, a verb and one of its dependents is aligned with a single verbal unit in Hindi.

It is essential to obtain the syntactic roles for dependents in the source language sentence as they are required for computing the compositionality value between the dependents and their verbs. The
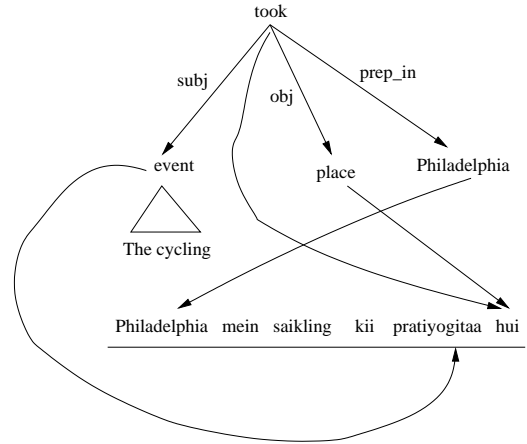

Figure 2: Alignment of Verb based expression

syntactic roles on the source side are obtained by applying simple rules to the output of a dependency parser. The dependency parser which we used in our experiments is a stochastic TAG based dependency parser (Shen, 2006). A sentence could have one or more verbs. We would like to align all the expressions represented by those verbs with words in the target language.

3 Behavior of MWEs in parallel corpora

In this section, we will briefly discuss the complexity of the alignment problem based on the verb based MWE's. From the word aligned sentence pairs, we compute the fraction of times a source sentence verb and its dependent are aligned together with the same word in the target language sentence. We count the number of times a source sentence verb and its dependent are aligned together with the same word in the target language sentence, and divide it by the total number of dependents. The total size of our word aligned corpus is 400 sentence pairs which includes both training and test sentences. The total number of dependents present in these sentences are 2209. Total number of verb dependent pairs which aligned with same word in target language are 193. Hence, the percentage of such occurrences is 9%, which is a significant number.

4 Alignment algorithm

In this section, we describe the algorithm for aligning verbs and their dependents in the source language sentence with the words in the target language. Let V be the number of verbs and A be the number of dependents. Let the number of words in

the target language be N. If we explore all the ways in which the $V + A$ words in the source sentence are aligned with words in the target language before choosing the best alignment, the total number of possibilites are N^{V+A}. This is computationally very expensive. Hence, we use a Beam-search algorithm to obtain the K-best alignments.

Our algorithm has three main steps.

1. Populate the Beam : Use the local features (which largely capture the co-occurence information between the source word and the target word) to determine the K-best alignments of verbs and their dependents with words in the target language.

2. Re-order the Beam: Re-order the above alignments using more complex features (which include the global features and the compositionality based feature(s)).

3. Post-processing : Extend the alignment(s) of the verb(s) (on the source side) to include words which can be part of the verbal unit on the target side.

For a source sentence, let the verbs and dependents be denoted by s_{ij}. Here i is the index of the verb ($1 <= i <= V$). The variable j is the index of the dependents ($0 <= j <= A$) except when $j = 0$ which is used to represent the verb itself. Let the source sentences be denoted as $S = \{s_{ij}\}$ and the target sentences by $T = \{t_n\}$. The alignment from a source sentence S to target sentence T is defined as the mapping $\bar{a} = \{a_{ijn} \mid a_{ijn} \equiv (s_{ij} \rightarrow t_n), \forall i, j\}$. A beam is used to store a set of K-best alignments between a source sentence and the target sentence. It is represented using the symbol B where B_k ($0 <= k <= K$) is used to refer to a particular alignment configuration.

4.1 Populate the Beam

The task in this step is to obtain the K-best candidate alignments using local features. The local features mainly contain the coccurence information between a source and a target word and are independent of other alignment links or words in the sentences. Let the local feature vector be denoted as $f_L(s_{ij}, t_k)$. The score of a particular alignment link is computed by taking the dot product of the weight vector W with the local feature vector (of

words connected by the alignment link). Hence, the local score will be

$$score_L(s_{ij}, t_k) = W.f_L(s_{ij}, t_k)$$

The total score of an alignment configuration is computed by adding the scores of individual links in the alignment configuration. Hence, the alignment score will be

$$score_{La}(\bar{a}, S, T) = \sum score_L(s_{ij}, t_k)$$

$$\forall s_{ij} \in S \ \& \ s_{ij} \rightarrow t_k \in \bar{a}$$

We propose an algorithm of order $O((V + A)N log(N) + K)$ to compute the K-best alignment configurations. First, the local scores of each verb and its dependents are computed for each word in the target sentence and stored in a local beam denoted by b_{ij}. The local beams corresponding to all the verbs and dependents are then sorted. This operation has the complexity $(V + A) N log(N)$.

The goal now is to pick the K-best configurations of alignment links. A single slot in the local beam corresponds to one alignment link. We define a boundary which partitions each local beam into two sets of slots. The slots above the boundary represent the slots which have been explored by the algorithm while slots below the boundary have still to be explored. The figure 3. shows the boundary which cuts across the local beams.
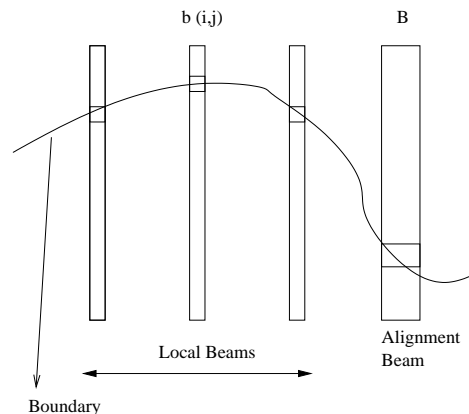

Figure 3: Boundary

We keep on modifying the boundary untill all the K slots in the Alignment Beam are filled with the K-best configurations. At the beginning of the algorithm, the boundary is a straight line passing through the top of all the local beams. The top slot of the alignment beam at the beginning represents

the combination of alignment links with the best local scores.

The next slot $b_{ij}[p]$ (from the set of unexplored slots) to be included in the boundary is the slot which has the least difference in score from the score of the slot at the top of its local beam. That is, we pick the slot $b_{ij}[p]$ such that $score(b_{ij}[p]) - score(b_{ij}[1])$ is the least among all the unexplored slots (or alignment links). Trivially, $b_{ij}[p-1]$ was already a part of the boundary.

When the slot $b_{ij}[p]$ is included in the boundary, various configurations, which now contain $b_{ij}[p]$, are added to the alignment beam. The new configurations are the same as the ones which previously contained $b_{ij}[p-1]$ but with the replacement of $b_{ij}[p-1]$ by $b_{ij}[p]$. The above procedure ensures that the the alignment configurations are K-best and are sorted according to the scores obtained using local features.

4.2 Re-order the beam

We now use global features to re-order the beam. The global features look at the properties of the entire alignment configuration instead of alignment links locally.

The global score is defined as the dot product of the weight vector and the global feature vector.

$$score_G(\bar{a}) = W.f_G(\bar{a})$$

The overall score is calculated by adding the local score and the global score.

$$score(\bar{a}) = score_{La}(\bar{a}) + score_G(\bar{a})$$

The beam is now sorted based on the overall scores of each alignment. The alignment configuration at the top of the beam is the best possible alignment between source sentence and the target sentence.

4.3 Post-processing

The first two steps in our alignment algorithm compute alignments such that one verb or dependent in the source language side is aligned with only one word in the target side. But, in the case of compound verbs in Hindi, the verb in English is aligned to all the words which represent the compound verb in Hindi. For example, in Figure 3, the verb "lost" is aligned to both 'khoo' and 'dii'.

Our alignment algorithm would have aligned "lost" only to 'khoo'. Hence, we look at the window of words after the word which is aligned to
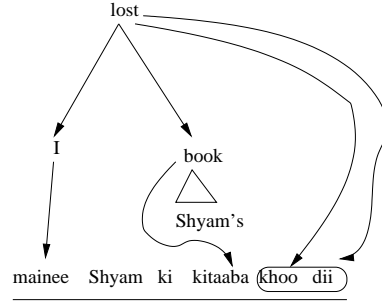

Figure 4: Case of compound verb in Hindi

the source verb and check if any of them is a verb which has not been aligned with any word in the source sentence. If this condition is satisfied, we align the source verb to these words too.

5 Parameters

As the number of training examples (294 sentences) is small, we choose to use very representative features. Some of the features which we used in this experiment are as follows,

5.1 Local features (F_L)

The local features which we consider are mainly co-occurence features. These features estimate the likelihood of a source word aligning to a target word based on the co-occurence information obtained from a large sentence aligned corpora[1].

1. **DiceWords**: Dice Coefficient of the source word and the target word

   $$DCoeff(s_{ij}, t_k) = \frac{2 * Count(s_{ij}, t_k)}{Count(s_{ij}) + Count(t_k)}$$

   where $Count(s_{ij}, t_k)$ is the number of times the word t_k was present in the translation of sentences containing the word s_{ij} in the parallel corpus.

2. **DiceRoots**: Dice Coefficient of the lemmatized forms of the source and target words. It is important to consider this feature because the English-Hindi parallel corpus is not large and co-occurence information can be learnt effectively only after we lemmatize the words.

3. **Dict**: Whether there exists a dictionary entry from the source word s_{ij} to the target word

t_k. For English-Hindi, we used a dictionary available at IIIT - Hyderabad, India.

4. **Null**: Whether the source word s_{ij} is aligned to nothing in the target language.

5.2 Global features

The following are the four global features which we have considered,

- **AvgDist**: The average distance between the words in the target language sentence which are aligned to the verbs in the source language sentence . AvgDist is then normalized by dividing itself by the number of words in the target language sentence. If the average distance is small, it means that the verbs in the source language sentence are aligned with words in the target language sentence which are located at relatively close distances, relative to the length of the target language sentence.

  This feature expresses the distribution of predicates in the target language.

- **Overlap**: This feature stores the count of pairs of verbs in the source language sentence which align with the same word in the target language sentence. Overlap is normalized by dividing itself by the total pairs of verbs.

  This feature is used to discourage overlaps among the words which are alignments of verbs in the source language sentence.

- **MergePos**: This feature can be considered as a compositionality based feature. The part of speech tag of a dependent is essential to determine the likelihood of the dependent to align with the same word in the target language sentence as the word to which its verb is aligned.

  This binary feature is active when the alignment links of a dependent and its verb merge. For example, in Figure 5., the feature 'merge_RP' will be active (that is, merge_RP = 1).

- **MergeMI**: This is a compositionality based feature which associates point-wise mutual information (apart from the POS information) with the cases where the dependents which have the same alignment in the target
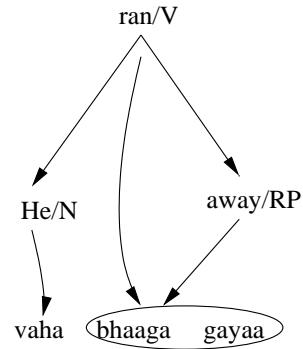

Figure 5: Example of MergePos feature

language as their verbs. This features which notes the the compositionality value (represented by point-wise mutual information in our experiments) is active if the alignment links of dependent and its verb merge.

The mutual information (MI) is classified into three groups depending on its absolute value. If the absolute value of mutual information rounded to nearest integer is in the range 0-2, it is considered LOW. If the value is in the range 3-5, it is considered MEDIUM and if it is above 5, it is considered HIGH.

The feature "merge_RP_HIGH" is active in the example shown in figure 6.
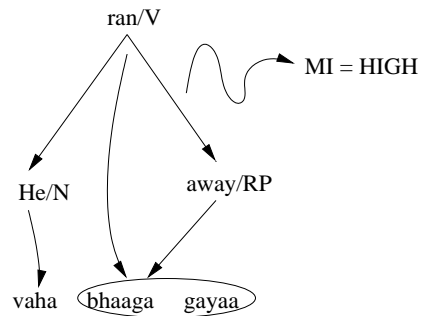
Figure 6: Example of MergeMI feature

6 Online large margin training

For parameter optimization, we have used an online large margin algorithm called MIRA (McDonald et al., 2005) (Crammer and Singer, 2003). We describe the training algorithm that we used very briefly. Our training set is a set of English-Hindi word aligned parallel corpus. We get the verb based expressions in English by running a dependency parser (Shen, 2006). Let the number of sentence pairs in the training data be m. We have

$\{S_q, T_q, \hat{a}_q\}$ for training where $q <= m$ is the index number of the sentence pair $\{S_q, T_q\}$ in the training set and $\hat{a}_q$ is the gold alignment for the pair $\{S_q, T_q\}$. Let W be the weight vector which has to be learnt, W_i be the weight vector after the end of ith update. To avoid over-fitting, W is obtained by averaging over all the weight vectors W_i.

A generic large margin algorithm is defined follows for the training instances $\{S_q, T_q, \hat{a}_q\}$,

1. Initialize W_0, W, i

2. for p:1 to NIterations

3.   for q:1 to m

4.   Get K-Best predictions $\alpha_q = \{a_1, a_2...a_k\}$ for the training example $(S_q, T_q, \hat{a}_q)$ using the current model W^i and applying step 1 and 2 of section 4. Compute W^{i+1} by updating W^i based on $(S_q, T_q, \hat{a}_q, \alpha_q)$.

5.   i = i + 1

6. $W = W + W^{i+1}$

7. $W = \frac{W}{NIterations*m}$

The goal of MIRA is to minimize the change in W^i such that the score of the gold alignment $\hat{a}$ exceeds the score of each of the predictions in α by a margin which is equal to the number of mistakes in the predictions when compared to gold alignment. While computing the number of mistakes, the mistakes due to the mis-alignment of head verb could be given greater weight, thus prompting the optimization algorithm to give greater importance to verb related mistakes and thereby improving overall performance.

Step 4 in the algorithm mentioned above can be substituted by the following optimization problem,

$$\text{minimize } \|(W^{i+1} - W^i)\|$$
$$\text{s.t. } \forall k, score(\hat{a}_q, S_q, T_q) - score(a_{q,k}, S_q, T_q)$$
$$>= Mistakes(a_k, \hat{a}_q, S_q, T_q)$$

The above optimization problem is converted to the Dual form using one Lagrangian multiplier for each constraint. In the Dual form, the Lagrangian multipliers are solved using Hildreth's algorithm. Here, prediction of α is similar to the prediction of $K - best$ classes in a multi-class classification problem. Ideally, we need to consider all the possible classes and assign margin constraints based on every class. But, here the number of such classes is exponential and thus we restrict ourselves to the $K - best$ classes.

7 Results on word-alignment task

7.1 Dataset

We have divided the 400 word aligned sentence pairs into a training set consisting of 294 sentence pairs and a test set consisting of 106 sentence pairs. The source sentences are all dependency parsed (Shen, 2006) and only the verb and its dependents are considered for both training and testing our algorithm. Our training algorithm requires that the each of the source words is aligned to only one or zero target words. For this, we use simple heuristics to convert the training data to the appropriate format. For the words aligned to a source verb, the first verb is chosen as the gold alignment. For the words aligned to any dependent which is not a verb, the last content word is chosen as the alignment link. For test data, we do not make any modifications and the final output from our alignment algorithm is compared with the original test data.

7.2 Experiments with Giza

We evaluated our discriminative approach by comparing it with the state-of-art Giza++ alignments (Och and Ney, 2003). The metric that we have used to do the comparison is the Alignment Error Rate (AER). The results shown below also contain Precision, Recall and F-measure.

Giza was trained using an English-Hindi aligned corpus of 50000 sentence pairs. In Table 1., we report the results of the GIZA++ alignments run from both the directions (English to Hindi and Hindi to English). We also show the results of the intersected model. See Table 1. for the results of the GIZA++ alignments.

	Prec.	Recall	F-meas.		AER
Eng $\rightarrow$ Hin	0.45	0.38	0.41		0.5874
Hin $\rightarrow$ Eng	0.46	0.27	0.34		0.6584
Intersected	0.82	0.19	0.31		0.6892

Table 1: Results of GIZA++ - Original dataset

We then lemmatize the words in both the source and target sides of the parallel corpora and then run Giza++ again. As the English-Hindi dataset

of 50000 sentence pairs is relatively small, we expect lemmatizing to improve the results. Table 2. shows the results. As we hoped, the results after lemmatizing the word forms are better than those without.

	Prec.	Recall	F-meas.		AER
Eng → Hin	0.52	0.40	0.45		0.5518
Hin → Eng	0.53	0.30	0.38		0.6185
Intersected	0.82	0.23	0.36		0.6446

Table 2: Results of GIZA++ - lemmatized set

7.3 Experiments with our model

We trained our model using the training set of 294 word aligned sentence pairs. For training the parameters, we used a beam size of 3 and number of iterations equal to 3. Table 3. shows the results when we used only the basic local features (Dice-Words, DiceRoots, Dict and Null) to train and test our model.

	Prec.	Recall	F-meas.		AER
Local Feats.	0.47	0.38	0.42		0.5798

Table 3: Results using the basic features

When we add the the global features (AvgDist, Overlap), we obtain the AER shown in Table 4.

	Prec.	Recall	F-meas.		AER
+ AvgD., Ove.	0.49	0.39	0.43		0.5689

Table 4: Results using the features - AvgDist, Overlap

Now, we add the transition probabilities obtained from the experiments with Giza++ as features in our model. Table 5. contains the results.

The compositionality related features are now added to our discriminative model to see if there is any improvement in performance. Table 6. shows the results by adding one feature at a time.

We observe that there is an improvement in the AER by using the compositionality based features, thus showing that compositionality based features aid in the word-alignment task in a significant way (AER = 0.5045).

8 Related work

Various measures have been proposed in the past to measure the compositionality of multi-word ex-

	Prec.	Recall	F-meas.		AER
+ Giza++ prob.	0.54	0.44	0.49		0.5155

Table 5: Results using the Giza++ probabilities

	Prec.	Recall	F-meas.		AER
+ MergePos	0.54	0.45	0.49		0.5101
+ MergeMI	0.55	0.45	0.50		0.5045

Table 6: Results using the compositionality based features

pressions of various types. Some of them are Frequency, Point-wise mutual information (Church and Hanks, 1989), Distributed frequency of object (Tapanainen et al., 1998), Distributed frequency of object using verb information (Venkatapathy and Joshi, 2005), Similarity of object in verb-object pair using the LSA model (Baldwin et al., 2003), (Venkatapathy and Joshi, 2005) and Lexical and Syntactic fixedness (Fazly and Stevenson, 2006). These features have largely been evaluated by the correlation of the compositionality value predicted by these measures with the gold standard value suggested by human judges. It has been shown that the correlation of these measures is higher than simple baseline measures suggesting that these measures represent compositionality quite well. But, the compositionality as such has not been used in any specific application yet.

In this paper, we have suggested a framework for using the compositionality of multi-word expressions for the word alignment task. State-of-art systems for doing word alignment use generative models like GIZA++ (Och and Ney, 2003; Brown et al., 1993). Discriminative models have been tried recently for word-alignment (Taskar et al., 2005; Moore, 2005) as these models give the ability to harness variety of complex features which cannot be provided in the generative models. In our work, we have used the compositionality of multi-word expressions to predict how they align with the words in the target language sentence.

For parameter optimization for the word-alignment task, Taskar, Simon and Klein (Taskar et al., 2005) used a large margin approach by factoring the structure level constraints to constraints at the level of an alignment link. We cannot do such a factorization because the scores of alignment links in our case are not computed in a completely isolated manner. We use an online large margin approach called MIRA (McDonald et al.,

2005; Crammer and Singer, 2003) which fits well with our framework. MIRA has previously been used by McDonald, Pereira, Ribarov and Hajic (McDonald et al., 2005) for learning the parameter values in the task of dependency parsing.

It should be noted that previous word-alignment experiments such as Taskar, Simon and Klein (Taskar et al., 2005) have been done with very large datasets and there is little word-order variation in the languages involved. Our dataset is small at present and there is substantial word order variation between the source and target languages.

9 Conclusion and future work

In this paper, we have proposed a discriminative approach for using the compositionality information about verb-based multi-word expressions for the word-alignment task. For training our model, use used an online large margin algorithm (McDonald et al., 2005). For predicting the alignment given a model, we proposed a K-Best beam search algorithm to make our prediction algorithm computationally feasible.

We have investigated the usefulness of simple features such as point-wise mutual information for the word-alignment task in English-Hindi bilingual corpus. We have show that by adding the compositionality based features to our model, we obtain an decrease in AER from 0.5155 to 0.5045. Our overall results are better than those obtained using the GIZA++ models (Och and Ney, 2003).

In future, we will experiment with more advanced compositionality based features. But, this would require a larger dataset for training and we are working towards buidling such a large dataset. Also, we would like to conduct similar experiments on other language pairs (e.g. English-French) and compare the results with the state-of-art results reported for those languages.

References

Timothy Baldwin, Colin Bannard, Takaaki Tanaka, and Dominic Widdows. 2003. An empirical model of multiword expression decomposability. In Diana McCarthy Francis Bond, Anna Korhonen and Aline Villavicencio, editors, *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 89–96.

P. Brown, S. A. Pietra, V. J. Della, Pietra, and R. L. Mercer. 1993. The mathmatics of stastistical machine translation. In *Computational Linguistics*.

Kenneth Church and Patrick Hanks. 1989. Word association norms, mutual information, and lexicography. In *Proceedings of the 27th. Annual Meeting of the Association for Computational Linguistics, 1990*.

Koby Crammer and Yoram Singer. 2003. Ultraconservative online algorithms for multiclass problems. In *Journal of Machine Learning Research*.

Afsaneh Fazly and Suzanne Stevenson. 2006. Automatically constructing a lexicon of verb phrase idiomatic combinations. In *Proceedings of European Chapter of Association of Computational Linguistics*. Trento, Italy, April.

Ryan McDonald, Fernando Pereira, Kiril Ribarov, and Jan Hajic. 2005. Non-projective dependency parsing using spanning tree algorithms. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 523–530, Vancouver, British Columbia, Canada, October. Association of Computational Linguistics.

Robert C. Moore. 2005. A discriminative framework for bilingual word alignment. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 81–88, Vancouver, British Columbia, Canada, October. Association of Computational Linguistics.

F. Och and H. Ney. 2003. A systematic comparisoin of various statistical alignment models. In *Computational Linguistics*.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multi-word expressions: a pain in the neck for nlp. In *Proceedings of CICLing , 2002*.

Libin Shen. 2006. *Statistical LTAG Parsing*. Ph.D. thesis.

Pasi Tapanainen, Jussi Piitulaine, and Timo Jarvinen. 1998. Idiomatic object usage and support verbs. In *36th Annual Meeting of the Association for Computational Linguistics*.

Ben Taskar, Locoste-Julien Simon, and Klein Dan. 2005. A discriminative machine learning approach to word alignment. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 73–80, Vancouver, British Columbia, Canada, October. Association of Computational Linguistics.

Sriram Venkatapathy and Aravind Joshi. 2005. Measuring the relative compositionality of verb-noun (v-n) collocations by integrating features. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 899–906. Association of Computational Linguistics, Vancouver, British Columbia, Canada, October.