Consistency-Aware Search for Word Alignment

Shiqi Shen[†], **Yang Liu**[†‡*], **Huanbo Luan**[†] and **Maosong Sun**[†‡]
[†]State Key Laboratory of Intelligent Technology and Systems
Tsinghua National Laboratory for Information Science and Technology
Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
[‡]Jiangsu Collaborative Innovation Center for Language Competence, Jiangsu, China
{vicapple22,liuyang.china,luanhuanbo}@gmail.com, sms@tsinghua.edu.cn

Abstract

As conventional word alignment search algorithms usually ignore the consistency constraint in translation rule extraction, improving alignment accuracy does not necessarily increase translation quality. We propose to use *coverage*, which reflects how well extracted phrases can recover the training data, to enable word alignment to model consistency and correlate better with machine translation. This can be done by introducing an objective that maximizes both alignment model score and coverage. We introduce an efficient algorithm to calculate coverage on the fly during search. Experiments show that our consistency-aware search algorithm significantly outperforms both generative and discriminative alignment approaches across various languages and translation models.

1 Introduction

Word alignment, which aims to identify the correspondence between words in two languages, plays an important role in statistical machine translation (Brown et al., 1993). Word alignment and translation rule extraction often constitute two consecutive steps in the training pipeline. Word-aligned bilingual corpora serve as a fundamental resource for translation rule extraction, not only for phrase-based models (Koehn et al., 2003; Och and Ney, 2004), but also for syntax-based models (Chiang, 2005; Galley et al., 2006). Dividing alignment and extraction into two separate steps significantly improves the efficiency and scalability of parameter estimation as compared with directly learning translation models from bilingual corpora (Marcu and Wong, 2002; DeNero and Klein, 2008; Cohn and Blunsom, 2009).

However, separating word alignment from translation rule extraction suffers from a major problem: maximizing the accuracy of word alignment does not necessarily lead to the improvement of translation quality. A number of studies show that alignment error rate (AER) only has a loose correlation with BLEU (Callison-Burch et al., 2004; Goutte et al., 2004; Ittycheriah and Roukos, 2005). Ayan and Dorr (2006) find that precision-oriented alignments result in better translation performance than recall-oriented alignments. Fraser and Marcu (2007) show that using AER and balanced F-measure can only partially explain the effect of alignment quality on BLEU for several language pairs.

We believe that the correlation problem arises from the discrepancy between word alignment and translation rule extraction. On one hand, aligners seek to find the alignment with the highest alignment model score, without regard to structural constraints. Consequently, sensible translation rules may not be extracted because they violate consistency constraints required by translation rule extraction (Och and Ney, 2004). Wang et al. (2010) find that the standard alignment tools are not optimal for training syntax-based models. As a result, they have to resort to re-aligning. On the other hand, the consistency constraint used in most translation rule extraction algorithms tolerate wrong links within consistent phrase pairs. Chiang (2007) uses the union of two unidirectional alignments, which usually has a low precision, for extracting hierarchical phrases. Therefore, it is important to include both alignment model score and the consistency constraint in the optimization objective of word alignment.

In this work, we propose to use *coverage*, which measures how well extracted phrases can

*Corresponding author: Yang Liu.

Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 1240–1249,
Lisbon, Portugal, 17-21 September 2015.

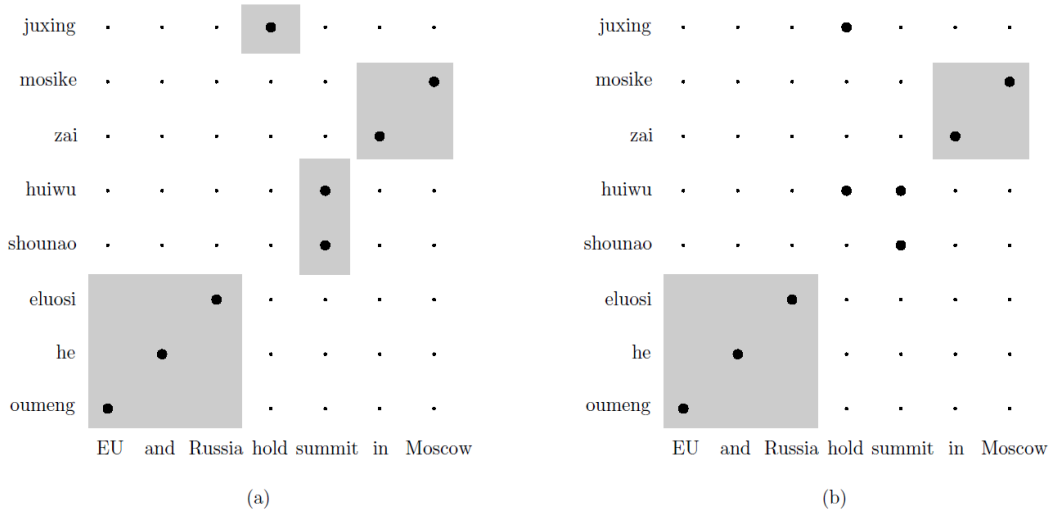

Figure 1: (a) An alignment resulting in a set of bilingual phrases (highlighted by shading) that can recover the training example, and (b) an alignment resulting in a set of bilingual phrases that fails to fully recover the training example. We assume the maximum phrase length $w = 3$. Our approach aims to avoid adding links that both have low posterior probabilities and hurt the recovery (e.g., the link between "huiwu" and "hold").

recover the training data, to bridge word alignment and (hierarchical) phrase-based translation. We introduce a new alignment search algorithm with an objective that maximizes both alignment model score and coverage while keeping the training algorithm unchanged. The coverage of an alignment is calculated on the fly during search using a local phrase extraction algorithm. Experiments show that our approach achieves significant improvements over state-of-the-art baselines across various languages and translation models.

2 Background

We begin by introducing the preliminaries of word alignment and phrase-based translation.

Definition 1 Given a source-language sentence $\mathbf{f} = f_1^J = f_1 \dots f_J$ and a target-language sentence $\mathbf{e} = e_1^I = e_1 \dots e_I$, an **alignment** $\mathbf{a}$ is a subset of the Cartesian product of the word positions of two sentences: $\mathbf{a} \subseteq \{(j, i) : j = 1, \dots, J; i = 1, \dots, I\}$.

Figure 1(a) shows an alignment for a Chinese sentence "oumeng he eluosi shounao huiwu zai mosike juxing" and an English sentence "EU and Russia hold summit in Moscow". We use black circles to denote links. The link $(1, 1)$ indicates that the first Chinese word "oumeng" and the first English word "EU" are translations of each other.

Definition 2 Given a training example $\langle \mathbf{f}, \mathbf{e}, \mathbf{a} \rangle$, a **bilingual phrase** B is a pair of source and target phrases: $B = (f_{j_1}^{j_2}, e_{i_1}^{i_2})$ such that $1 \leq j_1 \leq j_2 \leq J \wedge 1 \leq i_1 \leq i_2 \leq I$.

For example, ("zai mosike", "in Moscow") in Figure 1 can be denoted as a bilingual phrase $B = (f_6^7, e_6^7)$. For convenience, We use $B.j_1$ and $B.j_2$ to denote the beginning and ending positions of the source phrase in B, respectively. $B.i_1$ and $B.i_2$ are defined likewise for the target side.

Definition 3 A bilingual phrase $B = (f_{j_1}^{j_2}, e_{i_1}^{i_2})$ is said to be **tight** if and only if all boundary words (i.e., f_{j_1}, f_{j_2}, e_{i_1}, and e_{i_2}) are aligned. Otherwise, it is a **loose** bilingual phrase.

For example, in Figure 1, while (f_1^3, e_1^3) is a tight bilingual phrase, (f_1^4, e_1^4) is a loose bilingual phrase.

Definition 4 (Och and Ney, 2004) Given a training example $\langle \mathbf{f}, \mathbf{e}, \mathbf{a} \rangle$, a bilingual phrase $B = (f_{j_1}^{j_2}, e_{i_1}^{i_2})$ is said to be **consistent** with the word alignment $\mathbf{a}$ if and only if:

1. No words in the source phrase are aligned with words outside the target phrase and vice versa: $\forall (j, i) \in \mathbf{a} : j_1 \leq j \leq j_2 \leftrightarrow i_1 \leq i \leq i_2$,

2. At least one word in the source phrase is aligned with at least one word in the target

phrase: $\exists(j,i) \in \mathbf{a} : j_1 \leq j \leq j_2 \wedge i_1 \leq i \leq i_2$.

Alignment consistency forms the basis of translation rule extraction in modern SMT systems (Koehn and Hoang, 2007; Chiang, 2007; Galley et al., 2006; Liu et al., 2006). In Figure 1, (f_1^3, e_1^3) is consistent with the alignment because all words in "oumeng he eluosi" are aligned with all words in "EU and Russia". In contrast, in Figure 1(b), "huiwu shounao" and "hold summit" are not consistent with the alignment because "hold" is also aligned to a word "juxing" outside.

However, alignment consistency only defines a loose relationship between alignment and translation. A phrase pair consistent with alignment tolerates wrong inside links. For example, even if "oumeng" is aligned with "Russia", (f_1^3, e_1^3) is still consistent. This is one possible reason that maximizing alignment accuracy does not necessarily lead to improved translation performance.

3 Modeling Consistency in Word Alignment

Our intuition is that including the consistency constraint in word alignment can hopefully reduce the discrepancy between alignment and translation. While this idea has been suggested by a number of authors (e.g., (Deng and Zhou, 2009; DeNero and Klein, 2010)), our goal is to optimize arbitrary alignment models with respect to end-to-end translation in the search phase without labeled data (see Related Work for detailed comparison).

A natural way is to include consistency in the optimization objective as a regularization term. However, as consistency is only defined at the phrase level (see **Definition 4**), we need a sentence-level measure to reflect how well an alignment conforms to the consistency constraint. A straightforward measure is the number of bilingual phrases consistent with the alignment (*phrase count* for short), which is easy and efficient to calculate during search (Deng and Zhou, 2009). Unfortunately, optimizing with respect to *phrase count* is prone to yield alignments with very few links in a biased way, which result in a large number of bilingual phrases extracted from a small fraction of the training data. Another alternative is *reachability* (Liang et al., 2006a; Yu et al., 2013) that indicates whether there exists a full derivation to recover the training data. However, calculating reachability faces a major problem: a large portion of training data cannot be fully recovered due to noisy alignments and the distortion limit (Yu et al., 2013).

In this work, we propose *coverage*, which reflects how well extracted phrases can recover the training data, to measure the sentence-level consistency. In the following, we will introduce a number of definitions to facilitate the exposition.

Definition 5 A source word f_j is said to be **covered** by a bilingual phrase $B = (f_{j_1}^{j_2}, e_{i_1}^{i_2})$ if and only if $j_1 \leq j \leq j_2 : cov(f_j, B) = [\![j_1 \leq j \leq j_2]\!]$. Similarly, a target word e_i is covered by B if and only if $i_1 \leq i \leq i_2$.

The indicator function $[\![expr]\!]$ returns 1 if the boolean expression $expr$ is true and returns 0 otherwise. For example, in Figure 1(a), "oumeng" and "EU" are covered by the bilingual phrase $B = (f_1^3, e_1^3)$.

Definition 6 Given a set of bilingual phrases $\mathbf{B} = \{B^{(k)}\}_{k=1}^{K}$, a source word f_j is said to be **covered** by the bilingual phrase set $\mathbf{B}$ if and only if it is covered by at least one phrase in $B : cov(f_j, \mathbf{B}) = [\![\sum_{k=1}^{K} cov(f_j, B^{(k)}) > 0]\!]$. The definition for a target word is similar.

For example, in Figure 1(a), all source and target words are covered by the bilingual phrase set. In Figure 1(b), the source words "shounao", "huiwu", "juxing" and the target words "hold" and "summit" are not covered.

Definition 7 Given a sentence pair $\langle \mathbf{f}, \mathbf{e} \rangle$ and a phrase length limit w [1], the **hard coverage** of an alignment $\mathbf{a}$ is defined as a boolean value:

$$C_h(\mathbf{f}, \mathbf{e}, \mathbf{a}, w) = \Big[\!\Big[\delta\Big(\sum_{j=1}^{J} cov(f_j, \mathbf{B}), J \Big) \wedge \delta\Big(\sum_{i=1}^{I} cov(e_i, \mathbf{B}), I \Big) \Big]\!\Big] \quad (1)$$

where $\mathbf{B} = \textsc{Extract}(\mathbf{f}, \mathbf{e}, \mathbf{a}, 1, J, 1, I, w)$ is the set of consistent bilingual phrases extracted from the sentence pair using a standard phrase extraction algorithm (Och and Ney, 2004). The function δ returns true if the two parameters are same and returns false otherwise.

[1] The phrase length limit w is essential in defining coverage, restricting that the sentence pair must be covered by bilingual phrases no longer than w words. Otherwise, a very long bilingual phrase (e.g., the entire sentence pair) can achieve full coverage in a biased way.

Algorithm 1 A consistency-aware search algorithm for word alignment.

1: **procedure** ALIGN($\mathbf{f}, \mathbf{e}, \boldsymbol{\theta}, w, \beta, b, n$)
2: $open \leftarrow \emptyset$
3: $\mathcal{N} \leftarrow \emptyset$
4: $\langle \mathbf{a}, \mathbf{B} \rangle \leftarrow \langle \emptyset, \emptyset \rangle$
5: ADD($open, \langle \mathbf{a}, \mathbf{B} \rangle, \beta, b$)
6: **while** $open \neq \emptyset$ **do**
7: $closed \leftarrow \emptyset$
8: **for all** $\langle \mathbf{a}, \mathbf{B} \rangle \in open$ **do**
9: **for all** $l \in J \times I - \mathbf{a}$ **do**
10: $\mathbf{a}' \leftarrow \mathbf{a} \cup \{l\}$
11: $\mathbf{B}' \leftarrow$ UPDATE($\mathbf{f}, \mathbf{e}, \mathbf{a}, l, \mathbf{B}, w$)
12: **if** GAIN($\mathbf{f}, \mathbf{e}, \mathbf{a}, \mathbf{a}', w, \boldsymbol{\theta}$) > 0 **then**
13: ADD($closed, \langle \mathbf{a}', \mathbf{B}' \rangle, \beta, b$)
14: **end if**
15: ADD($\mathcal{N}, \langle \mathbf{a}', \mathbf{B}' \rangle, \beta, n$)
16: **end for**
17: **end for**
18: $open \leftarrow closed$
19: **end while**
20: **return** $\mathcal{N}$
21: **end procedure**

Depending on the tightness of extracted phrases (see **Definition 3**), we further distinguish between $C_{h+t}(\mathbf{f}, \mathbf{e}, \mathbf{a}, w)$ and $C_{h+l}(\mathbf{f}, \mathbf{e}, \mathbf{a}, w)$, which denote hard coverage calculated with tight and loose phrases, respectively.

Hard coverage denotes whether extracted phrases can fully recover the training data. For example, the values of hard coverage for Figures 1(a) and 1(b) are 1 and 0, respectively. As most training examples can hardly be fully recovered, we introduce *soft coverage* to better account for partially recoverable training data.

Definition 8 Given a sentence pair $\langle \mathbf{f}, \mathbf{e} \rangle$ and a phrase length limit w, the **soft coverage** of an alignment $\mathbf{a}$ is defined as

$$C_s(\mathbf{f}, \mathbf{e}, \mathbf{a}, w) = \frac{\sum_{j=1}^{J} cov(f_j, \mathbf{B}) + \sum_{i=1}^{I} cov(e_i, \mathbf{B})}{J + I} \quad (2)$$

Similarly, we also distinguish between C_{s+t} and C_{s+l} depending on the tightness of extracted phrases.

Definition 9 Given a word-aligned bilingual corpus $D = \{\langle \mathbf{f}^{(s)}, \mathbf{e}^{(s)}, \mathbf{a}^{(s)} \rangle\}_{s=1}^{S}$ and a phrase length limit w, the **corpus-level soft coverage** is defined as

$$C_s(D, w) = \frac{\sum_{j=1}^{|\mathbf{f}^{(s)}|} cov(f_j^{(s)}, \mathbf{B}^{(s)})}{\sum_{s=1}^{S} |\mathbf{f}^{(s)}| + |\mathbf{e}^{(s)}|} + \frac{\sum_{i=1}^{|\mathbf{e}^{(s)}|} cov(e_i^{(s)}, \mathbf{B}^{(s)})}{\sum_{s=1}^{S} |\mathbf{f}^{(s)}| + |\mathbf{e}^{(s)}|} \quad (3)$$

Algorithm 2 Updating the set of extracted bilingual phrases after adding a link.

1: **procedure** UPDATE($\mathbf{f}, \mathbf{e}, \mathbf{a}, l, \mathbf{B}, w$)
2: $\mathbf{B}' \leftarrow \mathbf{B}$
3: **for all** $B \in \mathbf{B}'$ **do**
4: **if** $B.j_1 \leq l.j \leq B.j_2 \vee B.i_1 \leq l.i \leq B.i_2$ **then**
5: $\mathbf{B}' \leftarrow \mathbf{B}' - \{B\}$
6: **end if**
7: **end for**
8: $j_1 \leftarrow l.j - w + 1$
9: $j_2 \leftarrow l.j + w - 1$
10: $i_1 \leftarrow l.i - w + 1$
11: $i_2 \leftarrow l.i + w - 1$
12: $\mathbf{a}' \leftarrow \mathbf{a} \cup \{l\}$
13: $\mathbf{B}'' \leftarrow$ EXTRACT($\mathbf{f}, \mathbf{e}, \mathbf{a}', j_1, j_2, i_1, i_2, w$)
14: $\mathbf{B}' \leftarrow \mathbf{B}' \cup \mathbf{B}''$
15: **return** $\mathbf{B}'$
16: **end procedure**

The corpus-level hard coverage is defined likewise.

4 Consistency-Aware Search

While Deng and Zhou (2009) focus on introducing an effectiveness function such as phrase count into alignment symmetrization, we are interested in guiding the search algorithms of arbitrary alignment models using coverage. Therefore, the objective of our search algorithm is defined as

$$score(\mathbf{f}, \mathbf{e}, \mathbf{a}, w, \boldsymbol{\theta}) = M(\mathbf{f}, \mathbf{e}, \mathbf{a}, \boldsymbol{\theta}) + \lambda C(\mathbf{f}, \mathbf{e}, \mathbf{a}, w) \quad (4)$$

where $M(\mathbf{f}, \mathbf{e}, \mathbf{a}, \boldsymbol{\theta})$ is alignment model score, $\boldsymbol{\theta}$ is a set of model parameters, $C(\mathbf{f}, \mathbf{e}, \mathbf{a}, w)$ is coverage (either hard or soft), and λ is a hyper-parameter that controls the preference between alignment model score and coverage. [2]

Therefore, the decision rule is given by

$$\hat{\mathbf{a}} = \underset{\mathbf{a} \in \mathcal{A}(\mathbf{f}, \mathbf{e})}{\operatorname{argmax}} \left\{ score(\mathbf{f}, \mathbf{e}, \mathbf{a}, w, \boldsymbol{\theta}) \right\} \quad (5)$$

where $\mathcal{A}(\mathbf{f}, \mathbf{e})$ is a set of all possible alignments for the sentence pair.

Algorithm 1 shows the consistency-aware search algorithm for word alignment. The input of the algorithm includes a source sentence $\mathbf{f}$, a target sentence $\mathbf{e}$, a set of model parameters $\boldsymbol{\theta}$, phrase length limit w, pruning parameters β and b, and the number of most likely alignments to be retaind n (line 1). Inspired by Liu et al. (2010),

[2] Note that training algorithms are unchanged. We only introduce a new search algorithm that takes coverage into consideration. We leave consistency-aware training algorithms for arbitrary alignment models for future work.

the algorithm starts with an empty alignment $\mathbf{a}$ together with an empty phrase set $\mathbf{B}$. We use $open$ to store active alignments during search and $\mathcal{N}$ to store top-n alignments after search (lines 2-4). The procedure ADD($open, \langle \mathbf{a}, \mathbf{B} \rangle, \beta, b$) adds $\langle \mathbf{a}, \mathbf{B} \rangle$ to $open$ and discards any alignment that has a score worse than β multiplied by the best score in the list or the score of the b-th best alignment (line 5). For each iteration (line 6), we use a list $closed$ to store promising alignments that have higher scores than the current alignment (line 8). For every possible link l (line 9), the algorithm produces a new alignment $\mathbf{a}'$ and updates the phrase set by calling a procedure UPDATE($\mathbf{f}, \mathbf{e}, \mathbf{a}, l, \mathbf{B}, w$) (lines 10-11). Then, the algorithm calls a procedure GAIN($\mathbf{f}, \mathbf{e}, \mathbf{a}, \mathbf{a}', w, \boldsymbol{\theta}$) to calculate the difference of model score after adding the link l:

$$score(\mathbf{f}, \mathbf{e}, \mathbf{a}', w, \boldsymbol{\theta}) - score(\mathbf{f}, \mathbf{e}, \mathbf{a}, w, \boldsymbol{\theta})$$

If $\mathbf{a}'$ has a higher score, it is added to $closed$ (line 13). We also update $\mathcal{N}$ to retain the top n alignment explored during the search (line 15). This process iterates until the model score does not increase.

Algorithm 2 describes how to update the set of extracted bilingual phrases after adding a link. Our idea is to only update the phrases near the added link l and keep other phrases unchanged. This strategy improves the efficiency by avoiding extracting phrases from the entire sentence pair. The algorithm first removes bilingual phrases that are either in the same row or in the same column with l (lines 2-7). For example, in Figure 1, the following bilingual phrases are removed after adding the link between "huiwu" and "hold" because the link breaks the consistency:

("shounao huiwu", "summit")
("juxing", "hold")

Other phrases out of the reach of the added link remain unchanged.

Then, the algorithm extracts bilingual phrases near l by calling the procedure EXTRACT. Note that the phrase extraction is restricted to a local region (j_1, j_2, i_1, i_2) by the phrase length limit w. We use $l.i$ and $l.j$ to denote the source and target positions of the link, respectively.

coverage	BLEU
C_{h+l}	24.89
C_{h+t}	23.16
C_{s+l}	24.69
C_{s+t}	25.41

Table 1: Comparison of different settings of coverage on the Chinese-English dataset using Moses. "h" denotes "hard", "s" denotes "soft", "l" denotes "loose", and "t" denotes "tight". The BLEU scores were calculated on the development set. For quick validation, we used a small fraction of the training data to train the phrase-based model.

5 Experiments

5.1 Setup

5.1.1 Languages and Datasets

We evaluated our approach in terms of alignment and translation quality on five language pairs: Chinese-English (ZH-EN), Czech-English (CS-EN), German-English (DE-EN), Spanish-English (ES-EN), and French-English (FR-EN). The evaluation metrics for alignment and translation are alignment error rate (AER) (Och and Ney, 2003) and case-insensitive BLEU (Papineni et al., 2002), respectively.

For Chinese-English, the training data consists of 1.2M pairs of sentences with 30.9M Chinese words and 35.5M English words. We used the SRILM toolkit (Stolcke, 2002) to train a 4-gram language model on the Xinhua portion of the English GIGAWORD corpus, which contains 398.6M words. For alignment evaluation, we used the Tsinghua Chinese-English word alignment evaluation data set (Liu and Sun, 2015). [3] For translation evaluation, we used the NIST 2006 dataset as the development set and the NIST 2002, 2003, 2004, 2005 and 2008 datasets as the test sets.

For other languages, the training data is Europarl v7. The English language model trained on the Xinhua portion of the English GIGAWORD corpus was also used for translation from European languages to English. For translation evaluation, we used the "news-test2012" dataset that contains 3,003 sentences as the development set and the "news-test2013" dataset that contains 3,000 sentences as the test set.

[3] http://nlp.csai.tsinghua.edu.cn/~ly/systems/TsinghuaAligner/TsinghuaAligner.html

method	# bp	# sp	# tp	# sw	# tw	C_{s+t}	C_{s+l}	AER	BLEU
C → E	49.8M	33.0M	14.1M	80.4K	89.5K	73.76	82.82	29.21	30.49
E → C	66.0M	14.9M	43.1M	80.0K	82.2K	74.58	86.49	33.04	29.76
Intersection	465.6M	64.3M	72.5M	133.1K	165.0K	72.46	98.52	28.01	29.90
Union	11.8M	7.4M	7.8M	51.3K	50.5K	53.17	54.34	32.80	30.24
GDF	15.9M	9.6M	10.0M	64.7K	62.2K	63.12	64.45	30.56	30.40
phrase count	388.7M	58.5M	63.7M	133.4K	164.5K	78.42	99.52	25.70	30.16
this work	46.0M	20.6M	21.7M	130.8K	159.6K	91.25	98.34	25.77	31.33

Table 2: Comparison of different alignment methods on the Chinese-English dataset. "GDF" denotes the grow-diag-final heuristic. "phrase count" denotes optimizing with respect to maximizing the number of extracted tight phrases. We used Moses to extract loose phrases from word-aligned training data for all methods. "# bp" denotes the number of extracted bilingual phrases, "# sp" denotes the number of source phrases, "# tp" denotes the number of target phrases, "# sw" denotes the source vocabulary size, "# tw" denotes the target vocabulary size. We report BLEU scores on the NIST 2005 test set.

alignment	NIST06	NIST02	NIST03	NIST04	NIST05	NIST08
generative	29.60	31.84	31.68	31.80	30.40	24.53
+*coverage*	30.63**	32.89**	32.77**	32.96**	31.33**	25.25**
discriminative	28.98	32.31	31.69	31.89	30.65	23.20
+*coverage*	29.98**	32.93**	32.45**	32.45**	31.10**	24.67**

Table 3: Translation evaluation on different alignment models. We apply our approach to both generative and discriminative alignment models. "generative" denotes applying the grow-diag-final heuristic to the alignments produced by IBM Model 4 in two directions. "discriminative" denotes the log-linear alignment model (Liu et al., 2010). Adding coverage leads to significant improvements. We use "**" to denote that the difference is statistically significant at $p < 0.01$ level.

5.1.2 Alignment Models

We apply our approach to both generative and discriminative alignment models. For generative models, we used GIZA++ (Och and Ney, 2003) to train IBM Model 4 in two directions. To calculate a model score for symmetrized alignments, we follow Liang et al. (2006b) to leverage link posterior marginal probabilities. For discriminative models, we used the open-source toolkit TsinghuaAligner (Liu and Sun, 2015) that implements the log-linear alignment model as described in (Liu et al., 2010). The model score for the log-linear model is also defined using link posteriors.

5.1.3 Translation Models

Two kinds of translation models, phrase-based (Koehn et al., 2003) and hierarchical phrase-based (Chiang, 2007), are used to evaluate whether our approach improves the correlation between alignment and translation. For the phrase-based model, we used the open-source toolkit Moses (Koehn and Hoang, 2007). For the hierarchical phrase-based model, we used an in-house re-implementation on par with state-of-the-art open-source decoders.

5.2 Comparison of Different Settings

We first investigate the optimal setting for coverage (hard vs. soft, tight vs. loose) on the Chinese-English dataset. For quick validation, we used a subset of the training data to train the phrase-based model using Moses. We used the development set to optimize the scaling factor λ (see Eq. (4)) and set it to 0.3 in our experiments.

Table 1 compares C_{h+l}, C_{h+t}, C_{s+l}, and C_{s+t}. We find that the "*soft + tight*" combination (i.e., C_{s+t}) yields the highest BLEU score on the development set. One possible reason is that tight phrases are usually of high quality and soft coverage allows for taking full advantage of the training data. On the contrary, C_{h+t} yields the lowest BLEU score because hard coverage fails to distinguish between partially recoverable training examples as it assigns zero to all partially recoverable data.

Then, we investigate the effect of the phrase length limit w in Algorithm 1 on translation quality. We find $w = 7$ achieves the best result,

which is consistent with the default setting in Moses. As a result, we used C_{s+t} and set $w = 7$ in the following experiments.

5.3 Comparison of Different Alignment Methods

We compare our approach with a number of alignment methods in terms of AER and BLEU, including IBM Model 4 in two directions (C $\rightarrow$ E and E $\rightarrow$ C), symmetrization heuristics (Intersection, Union, grow-diag-final), and consistency-aware models (tight phrase count and coverage). We used Moses to extract loose bilingual phrases from word-aligned bilingual corpora from all methods. Note that our approach uses C_{s+t} for finding alignments, from which Moses extracts loose phrases.

Table 2 lists the numbers of extracted bilingual phrases ("# bp"), source phrases ("# sp"), target phrases ("# tp"), source vocabulary size ("# sw"), and target vocabulary size ("# tw"). We find that a very large number of loose phrases can be extracted from the Intersection alignments, which also have the highest vocabulary sizes. However, a large portion of words in these phrases are actually unaligned, resulting in low translation quality.

We observe that adding consistency, either in terms of phrase count or coverage, significantly improves alignment accuracy by a large margin, suggesting that imposing structural constraint helps to reduce alignment errors. Our approach outperforms all methods in terms of BLEU significantly. Note that the coverage itself does not correlate well with BLEU. It is important to achieve a balance between model score and coverage. As mentioned in Section 5.2, we set $\lambda = 0.3$ in our experiments.

5.4 Translation Evaluation on Different Alignment Models

We apply our approach to both generative (Brown et al., 1993) and discriminative (Liu et al., 2010) alignment models. As shown in Table 3, we find that adding coverage to the optimization objective significantly improves the BLEU scores. All differences are statistically significant at $p < 0.01$ level. This finding suggests that our approach generalizes well to various alignment models.

5.5 Translation Evaluation on Different Translation Models

We also evaluated our approach on both phrase-based and hierarchical phrase-based models. As shown in Table 4, adding coverage to generative models leads to significant improvements for both models. All the differences are statistically significant at $p < 0.01$ level.

Although coverage is designed for extracting phrases, using coverage is still beneficial to hierarchical phrase-based models because hierarchical phrases are derived from phrases consistent with word alignment. [4]

5.6 Translation Evaluation on Different Language Pairs

Finally, we report BLEU scores across five language pairs in Table 5: Chinese-English (ZH-EN), Czech-English (CS-EN), German-English (DE-EN), Spanish-English (ES-EN), and French-English (FR-EN). ZH-EN uses four references and other language pairs only use single references.

We find that our approach outperforms the baseline statistically significantly at $p < 0.01$ for four language pairs and $p < 0.05$ for one language pair. Therefore, using coverage to bridge word alignment and machine translation can hopefully benefit more languages.

6 Related Work

Our work is inspired by three lines of research: (1) reachability in discriminative training of translation models, (2) structural constraints for alignment, and (3) learning with constraints.

6.1 Reachability in Discriminative Training of Translation Models

Discriminative training algorithms for statistical machine translation often need *reachable* training examples to find full derivations for updating model parameters (Liang et al., 2006a; Yu et al., 2013). Yu et al. (2013) report that only 32.1% sentences in the Chinese-English training data that contain 12.7% words are fully reachable

[4] We also tested our approach on syntax-based models (Galley et al., 2006; Liu et al., 2006) but failed to achieve significant improvements. The reason is that extracting syntactic translation rules often imposes an additional constraint: a phrase must be a constituent that is subsumed by a subtree. We believe that appending such constraint to the optimization objective will hopefully benefit syntax-based translation models. We leave this for future work.

translation	alignment	NIST06	NIST02	NIST03	NIST04	NIST05	NIST08
phrase	generative	29.60	31.84	31.68	31.80	30.40	24.53
	+*coverage*	30.63**	32.89**	32.77**	32.96**	31.33**	25.25**
hierarchical	generative	30.43	33.36	32.58	32.72	31.57	24.21
	+*coverage*	31.60**	34.67**	34.14**	34.24**	32.73**	24.89**

Table 4: Translation evaluation on different translation models. For translation, We used both phrase-based and hierarchical phrase-based models. For alignment, we used the generative model. "generative" denotes applying the grow-diag-final heuristic to the alignments produced by IBM Model 4 in two directions. Adding coverage leads to significant improvements. We use "**" to denote that the difference is statistically significant at $p < 0.01$ level.

alignment	ZH-EN	CS-EN	DE-EN	ES-EN	FR-EN
generative	30.40	19.89	21.13	26.39	26.22
+*coverage*	31.33**	20.04*	21.63**	26.79**	26.76**

Table 5: Translation evaluation on five language pairs. "generative" denotes applying the grow-diag-final heuristic to the alignments produced by IBM Model 4 in two directions. We use "*" and"**" to denote that the difference is statistically significant at $p < 0.05$ and $p < 0.01$, respectively. Note that ZH-EN uses four references and other language pairs only use single references.

due to noisy alignments and distortion limit. They find that most reachable sentences are short and generally literal.

We borrow the idea of measuring the degree of recovering training data from reachability but ignore the dependency between bilingual phrases for efficiency. To calculate reachability, one needs to figure out a full derivation, in which the bilingual phrases cover the training data and do not intersect with each other. Yu et al. (2013) indicate that using forced decoding to select reachable sentences with an unlimited distortion limit runs in $O(2^n n^3)$ time. In contrast, calculating coverage is much easier and more efficient by ignoring the dependency between phrases but still retains the spirit of measuring recovery.

6.2 Structural Constraints for Alignment

Modeling structural constraints in alignment has received intensive attention in the community, either directly modeling phrase-to-phrase alignment (Marcu and Wong, 2002; DeNero and Klein, 2008; Cohn and Blunsom, 2009) or intersecting synchronous grammars with alignment (Wu, 1997; Zhang and Gildea, 2005; Haghighi et al., 2009).

Our work is in spirit most close to (Deng and Zhou, 2009) and (DeNero and Klein, 2010). Deng and Bowen (2009) cast combining IBM Model 4 alignments in two directions as an optimization problem driven by an effectiveness function. They evaluate the impact of adding or removing a link with respect to phrase extraction using the effectiveness function of phrase count. The major difference is that we generalize their idea to arbitrary alignment models in the search phase rather than bidirectional alignment combination in the post-processing phase. In addition, we find that using coverage instead of phrase count results in better translation performance (see Table 2).

DeNero and Klein (2010) develop a discriminative model of extraction sets and optimize an extraction-based loss function with respect to translation. Their model is capable of predicting the extracted phrase set. While their approach relies on annotated data for training the discriminative model, our method only needs to tune the scaling factor λ on the development set. In addition, our approach is very general and can easily apply to arbitrary alignment models by appending a term to the optimization objective.

6.3 Learning with Constraints

Our work is also related to learning with constraints such as constraint-driven learning (Chang et al., 2007) and posterior regularization (Ganchev et al., 2010). The basic idea is to inject prior knowledge to the model as a regularization term. The major difference is that our coverage regularizer is independent of model parameters. As a result, alignment models can still be trained independently.

7 Conclusion

In this work, we have presented a general framework for optimizing word alignment with respect to machine translation. We introduce coverage to measure how well extracted bilingual phrases can recover the training data. We develop a consistency-aware search algorithm that calculates coverage on the fly during search efficiently. Experiments show the our approach is effective in both alignment and translation tasks across various alignment models, translation models, and language pairs.

In the future, we plan to apply our approach to syntax-based models (Galley et al., 2006; Liu et al., 2006; Shen et al., 2008) and include the constituency constraint in the optimization objective. It is also interesting to develop consistency-aware training algorithms for word alignment.

Acknowledgements

Yang Liu and Maosong Sun are supported by the 863 Program (2015AA011808) and the National Natural Science Foundation of China (No. 61331013 and No. 61432013). Huanbo Luan is supported by the National Natural Science Foundation of China (No. 61303075). This research is also supported by the Singapore National Research Foundation under its International Research Centre@Singapore Funding Initiative and administered by the IDM Programme. Many thanks go to Chunyang Liu and Meng Zhang for their discussions. We also thank the anonymous reviewers for their valuable feedback.

References

Necip Fazil Ayan and Bonnie J. Dorr. 2006. Going beyond aer: An extensive analysis of word alignments and their impact on mt. In *Proceedings of COLING-ACL 2006*, pages 9–16, Sydney, Australia, July.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Chris Callison-Burch, David Talbot, and Miles Osborne. 2004. Statistical machine translation with word- and sentence aligned parallel corpora. In *Proceedings of ACL 2004*.

Ming-Wei Chang, Lev Ratinov, and Dan Roth. 2007. Guiding semi-supervision with constraint-driven learning. In *Proceedings of ACL 2007*.

David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of ACL 2005*, pages 263–270, Ann Arbor, Michigan, June.

David Chiang. 2007. Hierarchical phrase-based translation. *Computational Linguistics*, 33(2):201–228.

Trevor Cohn and Phil Blunsom. 2009. A bayesian model of syntax-directed tree to string grammar induction. In *Proceedings of EMNLP 2009*.

John DeNero and Dan Klein. 2008. The complexity of phrase alignment problems. In *Proceedings of ACL 2008*.

John DeNero and Klein Klein. 2010. Discriminative modeling of extraction sets for machine translation. In *Proceedings of ACL 2010*.

Yonggang Deng and Bowen Zhou. 2009. Optimizing word alignment combination for phrase table training. In *Proceedings of ACL 2009*.

Alexander Fraser and Daniel Marcu. 2007. Measuring word alignment quality for statistical machine translation. *Computational Linguistics, Squibs and Discussions*, 33(3):293–303.

Michel Galley, Jonathan Graehl, Kevin Knight, Daniel Marcu, Steve DeNeefe, Wei Wang, and Ignacio Thayer. 2006. Scalable inference and training of context-rich syntactic translation models. In *Proceedings of COLING-ACL 2006*, pages 961–968, Sydney, Australia, July.

Kuzmann Ganchev, João Graça, Jennifer Gillenwater, and Ben Taskar. 2010. Posterior regularization for structured latent variable models. *Journal of Machine Learning Research*.

Cyril Goutte, Kenji Yamada, and Eric Gaussier. 2004. Aligning words using matrix factorisation. In *Proceedings of ACL 2004*.

Aria Haghighi, John Blitzer, John DeNero, and Dan Klein. 2009. Better word alignments with supervised itg models. In *Proceedings of ACL 2009*.

Abraham Ittycheriah and Salim Roukos. 2005. A maximum entropy word aligner for arabic-english machine translation. In *Proceedings of EMNLP 2005*.

Philipp Koehn and Hieu Hoang. 2007. Factored translation models. In *Proceedings of EMNLP-CoNLL 2007*, pages 868–876, Prague, Czech Republic, June.

Philipp Koehn, Franz J. Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of HLT-NAACL 2003*, pages 127–133, Edmonton, Canada, May.

Percy Liang, Alexandre Bouchard-Cote, Dan Klein, and Ben Taskar. 2006a. An end-to-end discriminative approach to machine translation. In *Proceedings of ACL 2006*.

Percy Liang, Ben Taskar, and Dan Klein. 2006b. Alignment by agreement. In *Proceedings of HLT-NAACL 2006*, pages 104–111, New York City, USA, June.

Yang Liu and Maosong Sun. 2015. Contrastive unsupervised word alignment with non-local features. In *Proceedings of AAAI 2015*.

Yang Liu, Qun Liu, and Shouxun Lin. 2006. Tree-to-string alignment template for statistical machine translation. In *Proceedings of COLING/ACL 2006*.

Yang Liu, Qun Liu, and Shouxun Lin. 2010. Discriminaitve word alignment by linear modeling. *Computational Linguistics*, 36(3):303–339.

Daniel Marcu and William Wong. 2002. A phrase-based, joint probability model for statistical machine translation. In *Proceedings of EMNLP 2002*.

Franz J. Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Franz J. Och and Hermann Ney. 2004. The alignment template approach to statistical machine translation. *Computational Linguistics*, 30(4):417–449.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a methof for automatic evaluation of machine translation. In *Proceedings of ACL 2002*.

Libin Shen, Jinxi Xu, and Ralph Weischedel. 2008. A new string-to-dependency machine translation algorithm with a target dependency language model. In *Proceedings of ACL 2008*.

Andreas Stolcke. 2002. Srilm - an extensible language modeling toolkit. In *Proceedings of ICSLP 2002*.

Wei Wang, Jonathan May, Kevin Knight, and Daniel Marcu. 2010. Re-structuring, re-labeling, and re-aligning for syntax-based machine translation. *Computational Linguistics*.

Dekai Wu. 1997. Stochastic inversion transaction grammars and bilingual parsing of parallel corpora. *Computational Linguistics*, 23:377–404.

Heng Yu, Liang Huang, Haitao Mi, and Kai Zhao. 2013. Max-violation perceptron and forced decoding for scalable mt training. In *Proceedings of EMNLP 2013*.

Hao Zhang and Danieal Gildea. 2005. Stochastic lexicalized inversion transduction grammars for alignment. In *Proceedings of ACL 2005*.