HMM Word-to-Phrase Alignment with Dependency Constraints

Yanjun Ma **Andy Way**
Centre for Next Generation Localisation
School of Computing
Dublin City University
{yma,away}@computing.dcu.ie

Abstract

In this paper, we extend the HMM word-to-phrase alignment model with syntactic dependency constraints. The syntactic dependencies between multiple words in one language are introduced into the model in a bid to produce coherent alignments. Our experimental results on a variety of Chinese–English data show that our syntactically constrained model can lead to as much as a 3.24% relative improvement in BLEU score over current HMM word-to-phrase alignment models on a Phrase-Based Statistical Machine Translation system when the training data is small, and a comparable performance compared to IBM model 4 on a Hiero-style system with larger training data. An intrinsic alignment quality evaluation shows that our alignment model with dependency constraints leads to improvements in both precision (by 1.74% relative) and recall (by 1.75% relative) over the model without dependency information.

1 Introduction

Generative word alignment models including IBM models (Brown et al., 1993) and HMM word alignment models (Vogel et al., 1996) have been widely used in various types of Statistical Machine Translation (SMT) systems. This widespread use can be attributed to their robustness and high performance particularly on large-scale translation tasks. However, the quality of the alignment yielded from these models is still far from satisfactory even with significant amounts of training data; this is particularly true for radically different languages such as Chinese and English.

The weakness of most generative models often lies in the incapability of addressing one to many (1-to-n), many to one (n-to-1) and many to many (m-to-n) alignments. Some research directly addresses m-to-n alignment with phrase alignment models (Marcu and Wong, 2002). However, these models are unsuccessful largely due to intractable estimation (DeNero and Klein, 2008). Recent progress in better parameterisation and approximate inference (Blunsom et al., 2009) can only augment the performance of these models to a similar level as the baseline where bidirectional word alignments are combined with heuristics and subsequently used to induce translation equivalence (e.g. (Koehn et al., 2003)). The most widely used word alignment models, such as IBM models 3 and 4, can only model 1-to-n alignment; these models are often called "asymmetric" models. IBM models 3 and 4 model 1-to-n alignments using the notion of "fertility", which is associated with a "deficiency" problem despite its high performance in practice.

On the other hand, the HMM word-to-phrase alignment model tackles 1-to-n alignment problems with simultaneous segmentation and alignment while maintaining the efficiency of the models. Therefore, this model sets a good example of addressing the tradeoffs between modelling power and modelling complexity. This model can also be seen as a more generalised

Proceedings of SSST-4, Fourth Workshop on Syntax and Structure in Statistical Translation, pages 101–109,
COLING 2010, Beijing, August 2010.

case of the HMM word-to-word model (Vogel et al., 1996; Och and Ney, 2003), since this model can be reduced to an HMM word-to-word model by restricting the generated target phrase length to one. One can further refine existing word alignment models with syntactic constraints (e.g. (Cherry and Lin, 2006)). However, most research focuses on the incorporation of syntactic constraints into discriminative alignment models. Introducing syntactic information into generative alignment models is shown to be more challenging mainly due to the absence of appropriate modelling of syntactic constraints and the "inflexibility" of these generative models.

In this paper, we extend the HMM word-to-phrase alignment model with syntactic dependencies by presenting a model that can incorporate syntactic information while maintaining the efficiency of the model. This model is based on the observation that in 1-to-n alignments, the n words bear some syntactic dependencies. Leveraging such information in the model can potentially further aid the model in producing more fine-grained word alignments. The syntactic constraints are specifically imposed on the n words involved in 1-to-n alignments, which is different from the cohesion constraints (Fox, 2002) as explored by Cherry and Lin (2006), where knowledge of cross-lingual syntactic projection is used. As a syntactic extension of the open-source MTTK implementation (Deng and Byrne, 2006) of the HMM word-to-phrase alignment model, its source code will also be released as open source in the near future.

The remainder of the paper is organised as follows. Section 2 describes the HMM word-to-phrase alignment model. In section 3, we present the details of the incorporation of syntactic dependencies. Section 4 presents the experimental setup, and section 5 reports the experimental results. In section 6, we draw our conclusions and point out some avenues for future work.

2 HMM Word-to-Phrase Alignment Model

In HMM word-to-phrase alignment, a sentence **e** is segmented into a sequence of consecutive phrases: $\mathbf{e} = v_1^K$, where v_k represents the k^{th} phrase in the target sentence. The assumption that each phrase v_k generated as a translation of one single source word is consecutive is made to allow efficient parameter estimation. Similarly to word-to-word alignment models, a variable a_1^K is introduced to indicate the correspondence between the target phrase index and a source word index: $k \rightarrow i = a_k$ indicating a mapping from a target phrase v_k to a source word f_{a_k}. A random process ϕ_k is used to specify the number of words in each target phrase, subject to the constraints $J = \sum_{k=1}^{K} \phi_k$, implying that the total number of words in the phrases agrees with the target sentence length J.

The insertion of target phrases that do not correspond to any source words is also modelled by allowing a target phrase to be aligned to a non-existent source word f_0 (NULL). Formally, to indicate whether each target phrase is aligned to NULL or not, a set of indicator functions $\varepsilon_1^K = \{\varepsilon_1, \cdots, \varepsilon_K\}$ is introduced (Deng and Byrne, 2008): if $\varepsilon_k = 0$, then NULL $\rightarrow v_k$; if $\varepsilon_k = 1$, then $f_{a_k} \rightarrow v_k$.

To summarise, an alignment **a** in an HMM word-to-phrase alignment model consists of the following elements:

$$\mathbf{a} = (K, \phi_1^K, a_1^K, \varepsilon_1^K)$$

The modelling objective is to define a conditional distribution $P(\mathbf{e}, \mathbf{a}|\mathbf{f})$ over these alignments. Following (Deng and Byrne, 2008), $P(\mathbf{e}, \mathbf{a}|\mathbf{f})$ can be decomposed into a phrase count distribution (1) modelling the segmentation of a target sentence into phrases ($P(K|J, \mathbf{f}) \propto \eta^K$ with scalar η to control the length of the hypothesised phrases), a transition distribution (2) modelling the dependencies between the current link and the previous links, and a word-to-phrase translation distribution (3) to model the degree to which a word and a phrase are translational to each other.

$$\begin{aligned} P(\mathbf{e}, \mathbf{a}|\mathbf{f}) &= P(v_1^K, K, a_1^K, \varepsilon_1^K, \phi_1^K|\mathbf{f}) \\ &= P(K|J, \mathbf{f}) \quad (1) \\ & \quad P(a_1^K, \varepsilon_1^K, \phi_1^K|K, J, \mathbf{f}) \quad (2) \\ & \quad P(v_1^K|a_1^K, \varepsilon_1^K, \phi_1^K, K, J, \mathbf{f}) \quad (3) \end{aligned}$$

The **word-to-phrase translation distribution** (3) is formalised as in (4):

$$P(v_1^K|a_1^K, \varepsilon_1^K, \phi_1^K, K, J, \mathbf{f}) = \prod_{k=1}^{K} p_v(v_k|\varepsilon_k \cdot f_{a_k}, \phi_k) \quad (4)$$

Note here that we assume that the translation of each target phrase is conditionally independent of other target phrases given the individual source words.

If we assume that each word in a target phrase is translated with a dependence on the previously translated word in the same phrase given the source word, we derive the bigram translation model as follows:

$$p_v(v_k|f_{a_k}, \varepsilon_k, \phi_k) = p_{t_1}(v_k[1]|\varepsilon_k, f_{a_k}) \prod_{j=2}^{\phi_k} p_{t_2}(v_k[j]|v_k[j-1], \varepsilon_k, f_{a_k})$$

where $v_k[1]$ is the first word in phrase v_k, $v_k[j]$ is the j^{th} word in v_k, p_{t_1} is an unigram translation probability and p_{t_2} is a bigram translation probability. The intuition is that the first word in v_k is firstly translated by f_{a_k} and the translation of the remaining words $v_k[j]$ in v_k from f_{a_k} is dependent on the translation of the previous word $v_k[j-1]$ from f_{a_k}. The use of a bigram translation model can address the coherence of the words within the phrase v_k so that the quality of phrase segmentation can be improved.

3 Syntactically Constrained HMM Word-to-Phrase Alignment Models

3.1 Syntactic Dependencies for Word-to-Phrase Alignment

As a proof-of-concept, we performed dependency parsing on the GALE gold-standard word alignment corpus using Maltparser (Nivre et al., 2007).[1] We find that 82.54% of the consecutive English words have syntactic dependencies and 77.46% non-consecutive English words have syntactic dependencies in 1-to-2 Chinese–English (ZH–EN) word alignment (one Chinese word aligned to two English words). For English–Chinese (EN–ZH) word alignment, we observe that 75.62% of the consecutive Chinese words and 71.15% of the non-consecutive Chinese words have syntactic dependencies. Our model represents an attempt to encode these linguistic intuitions.

[1] http://maltparser.org/

3.2 Component Variables and Distributions

We constrain the word-to-phrase alignment model with a syntactic coherence model. Given a target phrase v_k consisting of ϕ_k words, we use the dependency label r_k between words $v_k[1]$ and $v_k[\phi_k]$ to indicate the level of coherence. The dependency labels are a closed set obtained from dependency parsers, e.g. using Maltparser, we have 20 dependency labels for English and 12 for Chinese in our data. Therefore, we have an additional variable r_1^K associated with the sequence of phrases v_1^K to indicate the syntactic coherence of each phrase, defining $P(\mathbf{e}, \mathbf{a}|\mathbf{f})$ as below:

$$P(r_1^K, v_1^K, K, a_1^K, \varepsilon_1^K, \phi_1^K|\mathbf{f}) = P(K|J, \mathbf{f}) P(a_1^K, \phi_1^K, \varepsilon_1^K|K, J, \mathbf{f}) P(v_1^K|a_1^K, \varepsilon_1^K, \phi_1^K, K, J, \mathbf{f}) P(r_1^K|a_1^K, \varepsilon_1^K, \phi_1^K, v_1^K, K, J, \mathbf{f}) \quad (5)$$

The **syntactic coherence distribution** (5) is simplified as in (6):

$$P(r_1^K|a_1^K, \varepsilon_1^K, \phi_1^K, v_1^K, K, J, \mathbf{f}) = \prod_{k=1}^{K} p_r(r_k; \varepsilon, f_{a_k}, \phi_k) \quad (6)$$

Note that the coherence of each target phrase is conditionally independent of the coherence of other target phrases given the source words f_{a_k} and the number of words in the current phrase ϕ_k. We name the model in (5) the SSH model. SSH is an abbreviation of Syntactically constrained Segmental HMM, given the fact that the HMM word-to-phrase alignment model is a Segmental HMM model (SH) (Ostendorf et al., 1996; Murphy, 2002).

As our syntactic coherence model utilises syntactic dependencies which require the presence of at least two words in target phrase v_k, we therefore model the cases of $\phi_k = 1$ and $\phi_k \geq 2$

separately. We rewrite (6) as follows:

$$p_r(r_k;\varepsilon,f_{a_k},\phi_k) = \begin{cases} p_{\phi_k=1}(r_k;\varepsilon,f_{a_k}) & \text{if } \phi_k = 1 \\ p_{\phi_k\geq 2}(r_k;\varepsilon,f_{a_k}) & \text{if } \phi_k \geq 2 \end{cases}$$

where $p_{\phi_k=1}$ defines the syntactic coherence when the target phrase only contains one word ($\phi_k = 1$) and $p_{\phi_k\geq 2}$ defines the syntactic coherence of a target phrase composed of multiple words ($\phi_k \geq 2$). We define $p_{\phi_k=1}$ as follows:

$$p_{\phi_k=1}(r_k;\varepsilon,f_{a_k}) \propto p_n(\phi_k = 1;\varepsilon,f_{a_k})$$

where the coherence of the target phrase (word) v_k is defined to be proportional to the probability of target phrase length $\phi_k = 1$ given the source word f_{a_k}. The intuition behind this model is that the syntactic coherence is strong iff the probability of the source f_{a_k} fertility $\phi_k = 1$ is high.

For $p_{\phi_k\geq 2}$, which measures the syntactic coherence of a target phrase consisting of more than two words, we use the dependency label r_k between words $v_k[1]$ and $v_k[\phi_k]$ to indicate the level of coherence. A distribution over the values $r_k \in \mathcal{R} = \{\text{SBJ}, \text{ADJ}, \cdots\}$ ($\mathcal{R}$ is the set of dependency types for a specific language) is maintained as a table for each source word associated with all the possible lengths $\phi \in \{2, \cdots, N\}$) of the target phrase it can generate, e.g. we set $N = 4$ for ZH–EN alignment and $N = 2$ for EN–ZH alignment in our experiments.

Given a target phrase v_k containing $\phi_k(\phi_k \geq 2)$ words, it is possible that there are no dependencies between the first word $v_k[1]$ and the last word $v_k[\phi_k]$. To account for this fact, we introduce a indicator function φ as in below:

$$\varphi(v_k[1],\phi_k) = \begin{cases} 1 & \text{if } v_k[1] \text{ and } v_k[\phi_k] \text{have} \\ & \text{syntactic dependencies} \\ 0 & \text{otherwise} \end{cases}$$

We can thereafter introduce a distribution $p_\varphi(\varphi)$, where $p_\varphi(\varphi = 0) = \zeta$ $(0 \leq \zeta \leq 1)$ and $p_\varphi(\varphi = 0) = 1-\zeta$, with ζ indicating how likely it is that the first and final words in a target phrase do not have any syntactic dependencies. We can set ζ to a small number to favour target phrases satisfying the syntactic constraints and to a larger number otherwise. The introduction of this variable enables us to tune the model towards our different end goals. We can now define $p_{\phi_k\geq 2}$ as:

$$p_{\phi_k\geq 2}(r_k;\varepsilon,f_{a_k}) = p(r_k|\varphi;\varepsilon,f_{a_k})p_\varphi(\varphi)$$

where we insist that $p(r_k|\varphi;\varepsilon,f_{a_k}) = 1$ if $\varphi = 0$ (the first and last words in the target phrase do not have syntactic dependencies) to reflect the fact that in most arbitrary consecutive word sequences the first and last words do not have syntactic dependencies, and otherwise $p(r_k|\varphi;\varepsilon,f_{a_k})$ if $\varphi = 1$ (the first and last words in the target phrase have syntactic dependencies).

3.3 Parameter Estimation

The Forward-Backward Algorithm (Baum, 1972), a version of the EM algorithm (Dempster et al., 1977), is specifically designed for unsupervised parameter estimation of HMM models. The Forward statistic $\alpha_j(i,\phi,\varepsilon)$ in our model can be calculated recursively over the trellis as follows:

$$\alpha_j(i,\phi,\varepsilon) = \{\sum_{i',\phi',\varepsilon'} \alpha_{j-\phi}(i',\phi',\varepsilon')p_a(i|i',\varepsilon;I)\}$$
$$p_n(\phi;\varepsilon,f_i)\eta p_{t_1}(e_{j-\phi+1}|\varepsilon,f_i)$$
$$\prod_{j'=j-\phi+2}^{j} p_{t_2}(e_{j'}|e_{j'-1},\varepsilon,f_i)p_r(r_k;\varepsilon,f_i,\phi)$$

which sums up the probabilities of every path that could lead to the cell $\langle j,i,\phi\rangle$. Note that the syntactic coherence term $p_r(r_k;\varepsilon,f_i,\phi)$ can efficiently be added into the Forward procedure. Similarly, the Backward statistic $\beta_j(i,\phi,\varepsilon)$ is calculated over the trellis as below:

$$\beta_j(i,\phi,\varepsilon) = \sum_{i',\phi',\varepsilon'} \beta_{j+\phi'}(i',\phi',\varepsilon')p_a(i'|i,h';I)$$
$$p_n(\phi';\varepsilon',f_{i'})\eta p_{t_1}(e_{j+1}|\varepsilon',f_{i'})$$
$$\prod_{j'=j+2}^{j+\phi'} p_{t_2}(e_{j'}|e_{j'-1},\varepsilon',f_{i'})p_r(r_k;\varepsilon',f_{i'},\phi')$$

Note also that the syntactic coherence term $p_r(r_k;\varepsilon',f_{i'},\phi')$ can be integrated into the Backward procedure efficiently.

Posterior probability can be calculated based on the Forward and Backward probabilities.

3.4 EM Parameter Updates

The Expectation step accumulates fractional counts using the posterior probabilities for each parameter during the Forward-Backward passes, and the Maximisation step normalises the counts in order to generate updated parameters.

The E-step for the syntactic coherence model proceeds as follows:

$$c(r';f,\phi') = \sum_{(\mathbf{f},\mathbf{e})\in\mathbf{T}} \sum_{i,j,\phi,f_i=f} \gamma_j(i,\phi,\varepsilon=1) \\ \delta(\phi,\phi')\delta(\varphi_j(e,\phi),r')$$

where $\gamma_j(i,\phi,\varepsilon)$ is the posterior probability that a target phrase $t_{j-\phi+1}^{j}$ is aligned to source word f_i, and $\varphi_j(e,\phi)$ is the syntactic dependency label between $e_{j-\phi+1}$ and e_j. The M-step performs normalisation, as below:

$$p_r(r';f,\phi') = \frac{c(r';f,\phi')}{\sum_r c(r;f,\phi')}$$

Other component parameters can be estimated in a similar manner.

4 Experimental Setup

4.1 Data

We built the baseline word alignment and Phrase-Based SMT (PB-SMT) systems using existing open-source toolkits for the purposes of fair comparison. A collection of GALE data (LDC2006E26) consisting of 103K (2.9 million English running words) sentence pairs was firstly used as a proof of concept ("small"), and FBIS data containing 238K sentence pairs (8 million English running words) was added to construct a "medium" scale experiment. To investigate the intrinsic quality of the alignment, a collection of parallel sentences (12K sentence pairs) for which we have manually annotated word alignment was added to both "small" and "medium" scale experiments. Multiple-Translation Chinese Part 1 (MTC1) from LDC was used for Minimum Error-Rate Training (MERT) (Och, 2003), and MTC2, 3 and 4 were used as development test sets. Finally the test set from NIST 2006 evaluation campaign was used as the final test set.

The Chinese data was segmented using the LDC word segmenter. The maximum-entropy-based POS tagger MXPOST (Ratnaparkhi, 1996) was used to tag both English and Chinese texts. The syntactic dependencies for both English and Chinese were obtained using the state-of-the-art Maltparser dependency parser, which achieved 84% and 88% labelled attachment scores for Chinese and English respectively.

4.2 Word Alignment

The GIZA++ (Och and Ney, 2003) implementation of IBM Model 4 (Brown et al., 1993) is used as the baseline for word alignment. Model 4 is incrementally trained by performing 5 iterations of Model 1, 5 iterations of HMM, 3 iterations of Model 3, and 3 iterations of Model 4. We compared our model against the MTTK (Deng and Byrne, 2006) implementation of the HMM word-to-phrase alignment model. The model training includes 10 iterations of Model 1, 5 iterations of Model 2, 5 iterations of HMM word-to-word alignment, 20 iterations (5 iterations respectively for phrase lengths 2, 3 and 4 with unigram translation probability, and phrase length 4 with bigram translation probability) of HMM word-to-phrase alignment for ZH–EN alignment and 5 iterations (5 iterations for phrase length 2 with uniform translation probability) of HMM word-to-phrase alignment for EN–ZH. This configuration is empirically established as the best for Chinese–English word alignment. To allow for a fair comparison between IBM Model 4 and HMM word-to-phrase alignment models, we also restrict the maximum fertility in IBM model 4 to 4 for ZH–EN and 2 for EN–ZH (the default is 9 in GIZA++ for both ZH–EN and EN–ZH). "grow-diag-final" heuristic described in (Koehn et al., 2003) is used to derive the refined alignment from bidirectional alignments.

4.3 MT system

The baseline in our experiments is a standard log-linear PB-SMT system. With the word alignment obtained using the method described in

section 4.2, we perform phrase-extraction using heuristics described in (Koehn et al., 2003), Minimum Error-Rate Training (MERT) (Och, 2003) optimising the BLEU metric, a 5-gram language model with Kneser-Ney smoothing (Kneser and Ney, 1995) trained with SRILM (Stolcke, 2002) on the English side of the training data, and MOSES (Koehn et al., 2007) for decoding. A Hiero-style decoder Joshua (Li et al., 2009) is also used in our experiments. All significance tests are performed using approximate randomisation (Noreen, 1989) at $p = 0.05$.

5 Experimental Results

5.1 Alignment Model Tuning

In order to find the value of ζ in the SSH model that yields the best MT performance, we used three development test sets using a PB-SMT system trained on the small data condition. Figure 1 shows the results on each development test set using different configurations of the alignment models. For each system, we obtain the mean of the BLEU scores (Papineni et al., 2002) on the three development test sets, and derive the optimal value for ζ of 0.4, which we use hereafter for final testing. It is worth mentioning that while IBM model 4 (M4) outperforms other models including the HMM word-to-word (H) and word-to-phrase (SH) alignment model in our current setup, using the default IBM model 4 setting (maximum fertility 9) yields an inferior performance (as much as 8.5% relative) compared to other models.

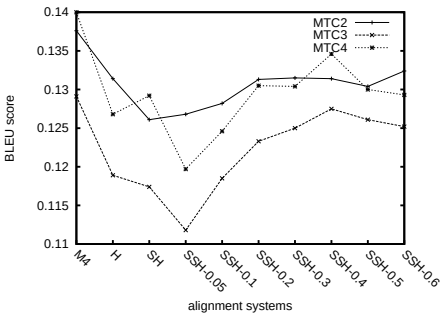

Figure 1: BLEU score on development test set using PB-SMT system

	PB-SMT		Hiero	
	small	medium	small	medium
H	0.1440	0.2591	0.1373	0.2595
SH	0.1418	0.2517	0.1372	0.2609
SSH	0.1464	0.2518	0.1356	0.2624
M4	0.1566	0.2627	0.1486	0.2660

Table 1: Performance of PB-SMT using different alignment models on NIST06 test set

5.2 Translation Results

Table 1 shows the performance of PB-SMT and Hiero systems using a small amount of data for alignment model training on the NIST06 test set. For the PB-SMT system trained on the small data set, using SSH word alignment leads to a 3.24% relative improvement over SH, which is statistically significant. SSH also leads to a slight gain over the HMM word-to-word alignment model (H). However, when the PB-SMT system is trained on larger data sets, there are no significant differences between SH and SSH. Additionally, both SH and SSH models underperform H on the medium data condition, indicating that the performance of the alignment model tuned on the PB-SMT system with small training data does not carry over to PB-SMT systems with larger training data (cf. Figure 1). IBM model 4 demonstrates stronger performance over other models for both small and medium data conditions.

For the Hiero system trained on a small data set, no significant differences are observed between SSH, SH and H. On a larger training set, we observe that SSH alignment leads to better performance compared to SH. Both SH and SSH alignments achieved higher translation quality than H. Note that while IBM model 4 outperforms other models on a small data condition, the difference between IBM model 4 and SSH is not statistically significant on a medium data condition. It is also worth pointing out that the SSH model yields significant improvement over IBM model 4 with the default fertility setting, indicating that varying the fertility limit in IBM model 4 has a significant impact on translation quality.

In summary, the SSH model which incorporates syntactic dependencies into the SH model achieves consistently better performance than

	ZH–EN		EN–ZH	
	P	R	P	R
H	0.5306	0.3752	0.5282	0.3014
SH	0.5378	0.3802	0.5523	0.3151
SSH	0.5384	0.3807	0.5619	0.3206
M4	0.5638	0.3986	0.5988	0.3416

Table 2: Intrinsic evaluation of the alignment using different alignment models

SH in both PB-SMT and Hiero systems under both small and large data conditions. For a PB-SMT system trained on the small data set, the SSH model leads to significant gains over the baseline SH model. The results also entail an observation concerning the suitability of different alignment models for different types of SMT systems; trained on a large data set, our SSH alignment model is more suitable to a Hiero-style system than a PB-SMT system, as evidenced by a lower performance compared to IBM model 4 using a PB-SMT system, and a comparable performance compared to IBM model 4 using a Hiero system.

5.3 Intrinsic Evaluation

In order to further investigate the intrinsic quality of the word alignment, we compute the Precision (P), Recall (R) and F-score (F) of the alignments obtained using different alignment models. As the models investigated here are asymmetric models, we conducted intrinsic evaluation for both alignment directions, i.e. ZH–EN word alignment where one Chinese word can be aligned to multiple English words, and EN–ZH word alignment where one English word can be aligned to multiple Chinese words.

Table 2 shows the results of the intrinsic evaluation of ZH–EN and EN–ZH word alignment on a small data set (results on the medium data set follow the same trend but are left out due to space limitations). Note that the P and R are all quite low, demonstrating the difficulty of Chinese–English word alignment in the news domain. For the ZH–EN direction, using the SSH model does not lead to significant gains over SH in P or R. For the EN–ZH direction, the SSH model leads to a 1.74% relative improvement in P, and a 1.75% relative improvement in R over the SH model. Both SH and SSH lead to gains over H for both ZH–EN and EN–ZH directions, while gains in the EN–ZH direction appear to be more pronounced. IBM model 4 achieves significantly higher P over other models while the gap in R is narrow.

Relating Table 2 to Table 1, we observe that the HMM word-to-word alignment model (H) can still achieve good MT performance despite the lower P and R compared to other models. This provides additional support to previous findings (Fraser and Marcu, 2007b) that the intrinsic quality of word alignment does not necessarily correlate with the performance of the resulted MT system.

5.4 Alignment Characteristics

In order to further understand the characteristics of the alignment that each model produces, we investigated several statistics of the alignment results which can hopefully reveal the capabilities and limitations of each model.

5.4.1 Pairwise Comparison

Given the asymmetric property of these alignment models, we can evaluate the quality of the links for each word and compare the alignment links across different models. For example, in ZH–EN word alignment, we can compute the links for each Chinese word and compare those links across different models. Additionally, we can compute the pairwise agreement in aligning each Chinese word for any two alignment models. Similarly, we can compute the pairwise agreement in aligning each English word in the EN–ZH alignment direction.

For ZH–EN word alignment, we observe that the SH and SSH models reach a 85.94% agreement, which is not surprising given the fact that SSH is a syntactic extension over SH, while IBM model 4 and SSH reach the smallest agreement (only 65.09%). We also observe that there is a higher agreement between SSH and H (76.64%) than IBM model 4 and H (69.58%). This can be attributed to the fact that SSH is still a form of HMM model while IBM model 4 is not. A similar trend is observed for EN–ZH word alignment.

	ZH–EN				EN–ZH			
	1-to-0	1-to-1	1-to-n		1-to-0	1-to-1	1-to-n	
			con.	non-con.			con.	non-con.
HMM	0.3774	0.4693	0.0709	0.0824	0.4438	0.4243	0.0648	0.0671
SH	0.3533	0.4898	0.0843	0.0726	0.4095	0.4597	0.0491	0.0817
SSH	0.3613	0.5092	0.0624	0.0671	0.3990	0.4835	0.0302	0.0872
M4	0.2666	0.5561	0.0985	0.0788	0.3967	0.4850	0.0592	0.0591

Table 3: Alignment types using different alignment models

5.4.2 Alignment Types

Again, by taking advantage of the asymmetric property of these alignment models, we can compute different types of alignment. For both ZH–EN (EN–ZH) alignment, we divide the links for each Chinese (English) word into 1-to-0 where each Chinese (English) word is aligned to the empty word "NULL" in English (Chinese), 1-to-1 where each Chinese (English) word is aligned to only one word in English (Chinese), and 1-to-n where each Chinese (English) word is aligned to n ($n \geq 2$) words in English (Chinese). For 1-to-n links, depending on whether the n words are consecutive, we have consecutive (con.) and non-consecutive (non-con.) 1-to-n links.

Table 3 shows the alignment types in the medium data track. We can observe that for ZH–EN word alignment, both SH and SSH produce far more 1-to-0 links than Model 4. It can also be seen that Model 4 tends to produce more consecutive 1-to-n links than non-consecutive 1-to-n links. On the other hand, the SSH model tends to produce more non-consecutive 1-to-n links than consecutive ones. Compared to SH, SSH tends to produce more 1-to-1 links than 1-to-n links, indicating that adding syntactic dependency constraints biases the model towards only producing 1-to-n links when the n words follow coherence constraint, i.e. the first and last word in the chunk have syntactic dependencies. For example, among the 6.24% consecutive ZH–EN 1-to-n links produced by SSH, 43.22% of them follow the coherence constraint compared to just 39.89% in SH. These properties can have significant implications for the performance of our MT systems given that we use the grow-diag-final heuristics to derive the symmetrised word alignment based on bidirectional asymmetric word alignments.

6 Conclusions and Future Work

In this paper, we extended the HMM word-to-phrase word alignment model to handle syntactic dependencies. We found that our model was consistently better than that without syntactic dependencies according to both intrinsic and extrinsic evaluation. Our model is shown to be beneficial to PB-SMT under a small data condition and to a Hiero-style system under a larger data condition.

As to future work, we firstly plan to investigate the impact of parsing quality on our model, and the use of different heuristics to combine word alignments. Secondly, the syntactic coherence model itself is very simple, in that it only covers the syntactic dependency between the first and last word in a phrase. Accordingly, we intend to extend this model to cover more sophisticated syntactic relations within the phrase. Furthermore, given that we can construct different MT systems using different word alignments, multiple system combination can be conducted to avail of the advantages of different systems. We also plan to compare our model with other alignment models, e.g. (Fraser and Marcu, 2007a), and test this approach on more data and on different language pairs and translation directions.

Acknowledgements

This research is supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (www.cngl.ie) at Dublin City University. Part of the work was carried out at Cambridge University Engineering Department with Dr. William Byrne. The authors would also like to thank the anonymous reviewers for their insightful comments.

References

Baum, Leonard E. 1972. An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. *Inequalities*, 3:1–8.

Blunsom, Phil, Trevor Cohn, Chris Dyer, and Miles Osborne. 2009. A gibbs sampler for phrasal synchronous grammar induction. In *Proceedings of ACL-IJCNLP 2009*, pages 782–790, Singapore.

Brown, Peter F., 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.

Cherry, Colin and Dekang Lin. 2006. Soft syntactic constraints for word alignment through discriminative training. In *Proceedings of the COLING-ACL 2006*, pages 105–112, Sydney, Australia.

Dempster, Arthur, Nan Laird, and Donald Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38.

DeNero, John and Dan Klein. 2008. The complexity of phrase alignment problems. In *Proceedings of ACL-08: HLT, Short Papers*, pages 25–28, Columbus, OH.

Deng, Yonggang and William Byrne. 2006. MTTK: An alignment toolkit for Statistical Machine Translation. In *Proceedings of HLT-NAACL 2006*, pages 265–268, New York City, NY.

Deng, Yonggang and William Byrne. 2008. HMM word and phrase alignment for Statistical Machine Translation. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(3):494–507.

Fox, Heidi. 2002. Phrasal cohesion and statistical machine translation. In *Proceedings of the EMNLP 2002*, pages 304–3111, Philadelphia, PA, July.

Fraser, Alexander and Daniel Marcu. 2007a. Getting the structure right for word alignment: LEAF. In *Proceedings of EMNLP-CoNLL 2007*, pages 51–60, Prague, Czech Republic.

Fraser, Alexander and Daniel Marcu. 2007b. Measuring word alignment quality for Statistical Machine Translation. *Computational Linguistics*, 33(3):293–303.

Kneser, Reinhard and Hermann Ney. 1995. Improved backing-off for m-gram language modeling. In *Proceedings of the IEEE ICASSP*, volume 1, pages 181–184, Detroit, MI.

Koehn, Philipp, Franz Och, and Daniel Marcu. 2003. Statistical Phrase-Based Translation. In *Proceedings of HLT-NAACL 2003*, pages 48–54, Edmonton, AB, Canada.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for Statistical Machine Translation. In *Proceedings of ACL 2007*, pages 177–180, Prague, Czech Republic.

Li, Zhifei, Chris Callison-Burch, Chris Dyer, Sanjeev Khudanpur, Lane Schwartz, Wren Thornton, Jonathan Weese, and Omar Zaidan. 2009. Joshua: An open source toolkit for parsing-based machine translation. In *Proceedings of the WMT 2009*, pages 135–139, Athens, Greece.

Marcu, Daniel and William Wong. 2002. A Phrase-Based, joint probability model for Statistical Machine Translation. In *Proceedings of EMNLP 2002*, pages 133–139, Philadelphia, PA.

Murphy, Kevin. 2002. Hidden semi-markov models (segment models). Technical report, UC Berkeley.

Nivre, Joakim, Johan Hall, Jens Nilsson, Atanas Chanev, Gülsen Eryigit, Sandra Kübler, Svetoslav Marinov, and Ervin Marsi. 2007. MaltParser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13(2):95–135.

Noreen, Eric W. 1989. *Computer-Intensive Methods for Testing Hypotheses: An Introduction*. Wiley-Interscience, New York, NY.

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

Och, Franz. 2003. Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of ACL 2003*, pages 160–167, Sapporo, Japan.

Ostendorf, Mari, Vassilios V. Digalakis, and Owen A. Kimball. 1996. From HMMs to segment models: A unified view of stochastic modeling for speech recognition. *IEEE Transactions on Speech and Audio Processing*, 4(5):360–378.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of Machine Translation. In *Proceedings of ACL 2002*, pages 311–318, Philadelphia, PA.

Ratnaparkhi, Adwait. 1996. A maximum entropy model for part-of-speech tagging. In *Proceedings of EMNLP 1996*, pages 133–142, Somerset, NJ.

Stolcke, Andreas. 2002. SRILM – An extensible language modeling toolkit. In *Proceedings of the International Conference on Spoken Language Processing*, pages 901–904, Denver, CO.

Vogel, Stefan, Hermann Ney, and Christoph Tillmann. 1996. HMM-based word alignment in statistical translation. In *Proceedings of ACL 1996*, pages 836–841, Copenhagen, Denmark.