Mach Translat (2013) 27:139–166
DOI 10.1007/s10590-013-9136-6

Substring-based machine translation

Graham Neubig · Taro Watanabe · Shinsuke Mori · Tatsuya Kawahara

Received: 7 December 2012 / Accepted: 16 February 2013 / Published online: 14 March 2013

Abstract Machine translation is traditionally formulated as the transduction of strings of words from the source to the target language. As a result, additional lexical processing steps such as morphological analysis, transliteration, and tokenization are required to process the internal structure of words to help cope with data-sparsity issues that occur when simply dividing words according to white spaces. In this paper, we take a different approach: not dividing lexical processing and translation into two steps, but simply viewing translation as a single transduction between character strings in the source and target languages. In particular, we demonstrate that the key to achieving accuracies on a par with word-based translation in the character-based framework is the use of a many-to-many alignment strategy that can accurately capture correspondences between arbitrary substrings. We build on the alignment method proposed in Neubig et al. (Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics. Portland, Oregon, pp. 632–641, 2011), improving its efficiency and accuracy with a focus on character-based translation. Using a many-to-many aligner imbued with these improvements, we demonstrate that the traditional framework of phrase-based machine translation sees large gains in accuracy over character-based translation with more naive alignment methods, and achieves comparable results to word-based translation for two distant language pairs.

G. Neubig (✉)
Nara Institute of Science and Technology, 8916-5 Takayama-cho, Ikoma, Nara, Japan
e-mail: neubig@is.naist.jp

T. Watanabe
National Institute of Information and Communications Technology, 3-5 Hikari-dai, Seika-cho,
Soraku-gun, Kyoto, Japan

S. Mori · T. Kawahara
Kyoto University, Yoshida Honmachi, Sakyo-ku, Kyoto, Japan

Keywords Character-based translation · Alignment · Inversion transduction grammar

1 Introduction

Statistical machine translation (SMT) is generally treated as the task of translating a source-language sentence $\boldsymbol{f}_1^J$ to a target-language sentence $\boldsymbol{e}_1^I$, where each element of f_j and e_i is assumed to be a word in the source and target languages. However, the definition of "word" is often problematic. The most obvious example of this is in unsegmented languages such as Chinese, Japanese, or Thai, where word segmentation is a necessary step prior to translation, and it has been noted that the segmentation standard has a large effect on translation accuracy (Chang et al. 2008). Even for languages with explicit word boundaries, all MT systems perform at least some cursory form of tokenization, splitting punctuation and words to prevent the sparsity that would occur if punctuated and non-punctuated words were treated as different entities. Sparsity also manifests itself in a number of other forms, with an extremely large number of rare words existing due to morphological productivity, word compounding, numbers, and proper names. A myriad of methods have been proposed to handle each of these phenomena individually in the context of MT, including morphological analysis, stemming, compound breaking, number regularization, word segmentation optimization, and transliteration, which are outlined in more detail in Sect. 2.

These difficulties stem from the basic premise that we are translating sequences of *words* as our basic unit. On the other hand, Vilar et al. (2007) examine the possibilities of eschewing the concept of words, treating each sentence as sequences of *characters* to be translated. This method is attractive, as it is theoretically able to handle almost all sparsity phenomena in a single unified framework, but has only proven feasible between similar language pairs such as Spanish–Catalan (Vilar et al. 2007), Swedish–Norwegian (Tiedemann 2009), and Thai–Lao (Sornlertlamvanich et al. 2008), which have a large number of cognates and a strong co-occurrence between single characters. As Xu et al. (2004) and Vilar et al. (2007) state and we further confirm here, accurate translations cannot be achieved when simply applying the traditional MT pipeline to character-based translation for less similar language pairs.

This paper is an extension of our work presented in Neubig et al. (2012), supplemented with a more complete description of the proposed alignment technique, additional experimental results investigating the effect of varying reordering limits or using character strings on only one side of the translation, and a subjective analysis of what type of alignments benefit or suffer when using character strings. We propose improvements to character-based translation, and demonstrate that it is, in fact, possible to achieve competitive translation accuracy for distant language pairs using only character strings. In particular, we focus on the bitext alignment process, and demonstrate that poor alignments achieved by more traditional alignment methods are one of the major reasons for character-based alignment failing to generalize to distant language pairs in previous work. We then propose an improved alignment strategy for character-based translation, which is made possible through recent advances in many-to-many word alignment, which we overview in Sect. 3. In comparison with the popular one-to-many IBM alignment models (Brown et al. 1993; Och and Ney 2003) used in previous work on character-based translation, many-to-many alignment can choose to align arbitrary substrings, which may consist of characters, morphemes, words, or multi-word phrases, automatically adjusting the granularity of alignment based on the paucity or abundance of data available.

One barrier to applying many-to-many alignment models to character strings is training cost. In the inversion transduction grammar (ITG) framework (Wu 1997), used in many previous works in many-to-many alignment, search is cumbersome for longer sentences, a problem that is further exacerbated when using characters instead of words as the basic unit. Even with more efficient search techniques for phrasal ITGs such as those proposed by Saers et al. (2009) or Blunsom and Cohn (2010), most previous research has limited the number of words in a sentence to at most 40. In order to overcome this computational burden and make character-based alignment feasible, we propose two improvements to the alignment model. The first proposed improvement, described in Sect. 4.3, increases the efficiency of the beam-search technique of Saers et al. (2009) by augmenting it with look-ahead probabilities in the spirit of A* search. As a heuristic function, we consider the monolingual cost of covering the strings in the source and target languages independently, which can be calculated efficiently but provides a reasonable estimate of the bilingual alignment cost.

A second problem with existing many-to-many alignment models in the context of character-based translation lies in the fact that they use one-to-many alignment models to seed the many-to-many alignment models in the form of a prior probability over the phrase pair distribution. While this has proven critical for accuracy in many-to-many systems for word-based translation (DeNero et al. 2008), in the character-based context one-to-many probabilities are not reliable. The second proposed improvement, described in Sect. 5, seeds the search process using counts of all substring pairs in the corpus to bias the phrase alignment model. We present an efficient method to calculate these substring pairs using enhanced suffix arrays (Abouelhoda et al. 2004) and sparse matrix operations. After these statistics have been collected, we transform them into prior probabilities and use them to seed the less efficient, but more accurate Bayesian ITG-based many-to-many alignment model.

Finally, to evaluate the effectiveness of the method, we perform end-to-end MT experiments on four language pairs with differing morphological properties. The evaluation results presented in Sect. 6 show that for distant language pairs, character-based SMT can achieve translation accuracy that is comparable to word-based systems. In addition, ablation studies show that the use of our proposed look-ahead parsing technique as well as substring-based priors both significantly help accuracy, and the look-ahead parsing method doubles the speed of alignment. Finally, we perform a qualitative analysis of the translation results that shows that the character-based method is not only able to translate unsegmented text, conjugated words, and proper names in a unified framework, but also uses a larger fraction of locally correct translation rules than word-based translation.

2 Related work on lexical processing in SMT

As traditional SMT systems treat all words as single tokens without considering their internal structure, major problems of data sparsity occur for less frequent tokens. In fact, it has been shown that there is a direct negative correlation between vocabulary size (and thus sparsity) of a language and translation accuracy (Koehn 2005). Rare words causes trouble for alignment models, both in the form of incorrect alignments, and in the form of garbage collection, where rare words in one language are incorrectly aligned to large segments of the sentence in the other language (Och and Ney 2003). Unknown words are also a problem during the translation process, and the default approach is to map them ‘as is’ into the translated sentence.

This is a major problem in morphologically rich languages such as Finnish and Korean, as well as highly compounding languages such as Dutch and German. Many previous works have attempted to handle morphology, decompounding and regularization through lemmatization, morphological analysis, or unsupervised techniques (Nießen and Ney 2000; Brown 2002; Lee 2004; Goldwater and McClosky 2005; Talbot and Osborne (2006); Macherey et al. (2011)). Other research has noted that it is more difficult to translate into morphologically rich languages with word-based systems, and methods for modeling target-side morphology have attracted interest in recent years (Bojar 2007; Subotin 2011). It is also notable that morphology and compounding remain problematic regardless of the size of the training data, with systems trained on hundreds of millions of words still seeing significant gains in accuracy due to lexical processing (Macherey et al. 2011).

Another major source of rare words in all languages is proper names, which have been handled by using cognates or transliteration to improve translation (Knight and Graehl 1998; Kondrak et al. 2003; Li et al. 2004; Finch and Sumita 2007). More sophisticated methods for named entity translation that combine translation and transliteration have also been proposed (Al-Onaizan and Knight 2002). In addition, while transliteration uses the underlying phonetic similarity of proper names to translate between writing systems, there has also recently been work on direct phoneme-to-word speech translation with the motivation of improving robustness to speech recognition errors (Jiang et al. 2011).

Choosing word units is also essential for creating good translation results for languages that do not explicitly mark word boundaries, such as Chinese, Japanese, and Thai. A number of works have addressed this word segmentation problem in translation, mainly focusing on translation of unsegmented languages such as Chinese or Japanese (Bai et al. 2008; Chang et al. 2008; Zhang et al. 2008b; Chung and Gildea 2009; Nguyen et al. 2010; Wang et al. 2010; Chu et al. 2012). However, these works generally assume that a word segmentation exists in one language (e.g. English) and attempt to optimize the word segmentation in the other language (e.g. Chinese). There have also been a number of works which propose evaluation measures for these languages that consider matches over characters instead of words (Denoual and Lepage 2005; Li et al. 2011; Liu and Ng 2012).

This enumeration of related work demonstrates the range of problems caused by the concept of ‘words’ in MT, and the large number of solutions proposed to address these problems. Character-based translation has the potential to handle all of the phenomena in the previously mentioned research in a single unified framework, while at the same time requiring no language-specific tools such as morphological analyzers or word segmenters. However, while the approach is conceptually attractive, previous research has only been shown to be effective for closely related language pairs (Vilar et al. 2007; Sornlertlamvanich et al. 2008; Tiedemann 2009, 2012), or when word- and character-based alignment is combined (Nakov and Tiedemann 2012). This work proposes effective alignment and decoding techniques that allow character-based translation to achieve accurate results for both close and distant language pairs. It should also be noted that there are other many-to-many alignment methods that have been used for simultaneously discovering morphological boundaries over multiple languages (Snyder and Barzilay 2008; Naradowsky and Toutanova 2011), but these have generally been applied to single words or short phrases, and it is not immediately clear that they will scale to aligning full sentences.

3 Alignment methods

Statistical machine translation systems are generally constructed from a parallel corpus consisting of target-language sentences $\mathcal{E}$ and source-language sentences $\mathcal{F}$. The first step of training is to find alignments $\mathcal{A}$, which indicate which parts of the target sentence align to which parts of the source sentence.

Here, we will represent our target and source sentences as $\boldsymbol{e}_1^I$ and $\boldsymbol{f}_1^J$. e_i and f_j represent single elements of the target and source sentences respectively, and I and J indicate the number of elements in the target and source sentences. Each element may be a word in word-based alignment models or a single character in character-based alignment models.[1] We define our alignment as $\boldsymbol{a}_1^K$, where each element is a span $a_k = \langle s, t, u, v \rangle$ indicating that the target string $e_s, \ldots, e_t$ and source string $f_u, \ldots, f_v$ are alignments of each other.[2]

3.1 One-to-many alignment

The most well-known and widely-used models for bitext alignment are for one-to-many alignment, including the IBM models (Brown et al. 1993) and HMM alignment model (Vogel et al. 1996). These models are by nature directional, attempting to find the alignments that maximize the conditional probability of the target sentence $P(\boldsymbol{e}_1^I | \boldsymbol{f}_1^J, \boldsymbol{a}_1^K)$. For computational reasons, the IBM models are restricted to aligning each word on the target side to a single word on the source side. In the formalism presented above, this means that each e_i must be included in at most one span, and for each span $u = v$. Traditionally, these models are run in both directions and combined using heuristics to create many-to-many alignments (Koehn et al. 2003).

However, in order for one-to-many alignment methods to be effective, each element f_j must contain enough information to allow for effective alignment with its

[1] Some previous work has also performed alignment using morphological analyzers to normalize or split the sentence into morpheme streams (Corston-Oliver and Gamon 2004).

[2] Null alignments can be represented implicitly with no span in $\boldsymbol{a}_1^K$ covering the unaligned words.

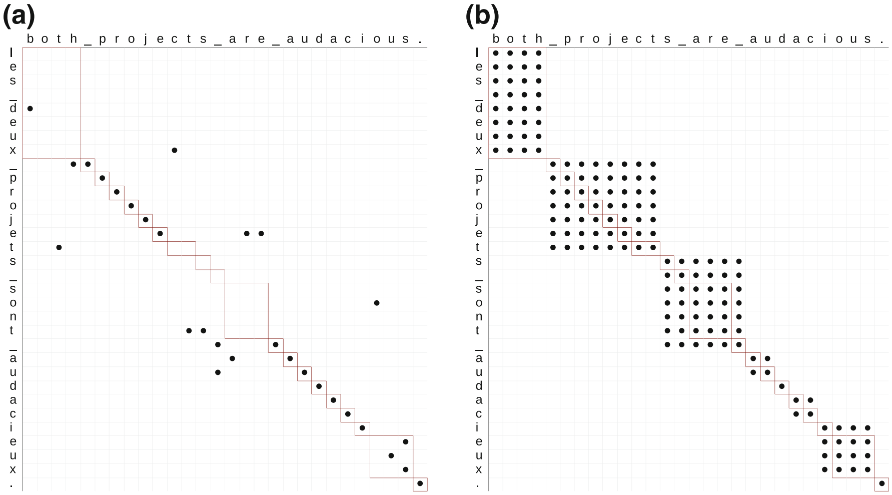


Fig. 1 Character-based alignments using **a** one-to-many models and **b** many-to-many models

corresponding elements in $\boldsymbol{e}_1^I$. While this is often the case in word-based models, for character-based models this assumption breaks down, as there is generally no clear correspondence between single characters.

An example of the alignments that result when applying one-to-many alignment to character strings is shown in Fig. 1a. It can be seen that in general, the alignments are less than desirable, with only cognates with similar spellings (e.g. "projects" and "projets") being aligned properly. The remaining words are not aligned properly, and a number of spurious alignment links are introduced, preventing even some of the properly aligned cognates from being extracted correctly.

3.2 Many-to-many alignment

On the other hand, in recent years, there have been advances in many-to-many alignment techniques that are able to align multi-element chunks on both source and target sides (Marcu and Wong 2002; DeNero et al. 2008; Blunsom et al. 2009; Neubig et al. 2011; Levenberg et al. 2012). Many-to-many methods can be expected to achieve superior results when applied to character-based alignment, as the aligner can use information about substrings, which may correspond to single characters, morphemes, words, or short phrases. An example of alignments acquired using the many-to-many models described in later sections is shown in Fig. 1b. From this example it can be confirmed that the alignments are not only of higher quality than those obtained by one-to-many alignment models, but also in units that correspond to human intuition: words ("project"/"projet"), phrases ("both"/"les deux"), sub-words ("ious"/"ieux"). The above example also shows the somewhat surprising alignment "s are"/"s sont", in which the plural suffix of the noun and the plural form of the copula are combined into a single phrase, capturing agreement between the two words.

Given our objective of finding multi-character alignments over character strings, there are a number of requirements for the alignment model that can be used.

1. **Efficiency:** The number of characters in a sentence is greater than the number of words in a sentence, so an alignment model that can handle longer sentences becomes a greater concern.
2. **Automatic Granularity Adjustment:** Given that it is possible to find alignments of any number of granularities, we must be able to choose an appropriate size based on the amount of data at our disposal. If we have more data, we can use longer units to achieve more accurate alignments, and if we have less data, we can fall back to sub-words or characters to maintain robustness.
3. **Compact Translation Model:** When performing word-based translation, phrases are generally restricted to a maximum size of around seven words. However, for character-based translation, seven characters is not enough to achieve reasonable accuracy, so we would like a model that can utilize longer phrases without creating enormous and unwieldy translation models.

Out of the many-to-many alignment methods proposed in the literature, the model we introduced in Neubig et al. (2011) satisfies most of these desiderata. In order to achieve efficient many-to-many alignment, it formulates the alignment process using ITGs (Wu 1997), which allow for many-to-many alignment through biparsing (described in the following section) in polynomial time. In order to automatically adjust the granularity of alignment, alignment model probabilities are calculated according to non-parametric Bayesian statistics, which allows for a balance between complex, expressive models that memorize long segments, and small but less expressive models that use shorter segments. Finally, the method reduces the translation model size by not using all phrases licensed by alignments as is typically done in traditional translation systems (Koehn et al. 2003), but only those licensed by the ITG tree.

This model is trained using Gibbs sampling in a multi-step process that can be very simply outlined below (readers may refer to Neubig et al. (2011) for more details):

1. **Calculate prior probabilities** with a less accurate but highly efficient alignment model (such as IBM Model 1 (Brown et al. 1993)).
2. **Sample** alignments A for each sentence:
   (a) **Remove** statistics from the model for the current sentence.
   (b) **Biparse** the two sentences according to the current ITG statistics.
   (c) **Sample** a new alignment using the information from the parse and **add** the statistics back into the model.

While previous work has shown this model to be effective for word-based alignment, in this paper we examine its effectiveness with regards to character-based alignment, and propose two improvements that are described in detail in the following sections. In particular, Sect. 4.3 describes an improvement to the *biparsing* step that improves the efficiency and accuracy for long sentences, while Sect. 5 describes improvements to the step of *calculating prior probabilities* using substring co-occurrence statistics.

4 Efficient sampling of ITG-based many-to-many alignments

In this section we briefly explain the process of alignment in the ITG framework, describe the process of biparsing that is used to find these alignments, and finally touch upon our proposed method to improve the efficiency of biparsing in many-to-many alignment models through the use of look-ahead probabilities.

4.1 Inversion transduction grammars (ITGs)

Inversion transduction grammars are generative models that were designed to simultaneously describe the generative process of equivalent strings of tokens $\boldsymbol{e}$ and $\boldsymbol{f}$ in two different languages. They are a limited form of synchronous context-free grammar (SCFG) in Chomsky normal form (Chomsky 1956), where "synchronous" indicates that the grammar is defined over two languages instead of one. Figure 2a shows an example of the word-based ITG derivation that has generated two phrases "to admit it" and "de le admettre" in English and French, which we will use to demonstrate how ITGs work. The ITG describes how these two equivalent sentences were created through a recursive process that passes through two phases.

The first phase consists of generating the sentence structure, which in the case of ITGs is particularly important for specifying the reordering that occurs between the sentences in the two languages. It can be seen from the reordering matrix in Fig. 2b that for some phrase pairs the word order is the same in both languages ("to" precedes "admit it" and "de" precedes "le admettre"). On the other hand, there are also some places where the order is inverted ("admit" precedes "it" while "admettre" follows "le"). ITGs represent this reordering structure as a binary tree, with each internal node labeled as *straight* (STR) or *inverted* (INV), where each of these node types represents the case where the order is the same or inverted in both languages, respectively.[3] Much like standard CFGs, each leaf node is labeled with the *pre-terminal* (TERM) to indicate that we have finished the first step of generating the sentence structure.

The second phase takes place after generating the pre-terminal symbol, and consists of generating short parallel phrases. These phrases can be one-to-one alignments as shown in the above example, but can just as easily be one-to-many or many-to-many alignments without a significant increase in the time required for alignment.

In addition, by assigning a probability to each of the ITG productions, it is possible to create a generative model for parallel phrase pairs. The ITG generative probability can be characterized by $P_x(x)$, which is a distribution over non- and pre-terminals, and $P_t(\langle \boldsymbol{e}, \boldsymbol{f} \rangle)$, which is a distribution over parallel phrase pairs. In this work, we follow the model of Neubig et al. (2011) which defines the probability through a hierarchical backoff scheme that attempts to generate parallel phrase pairs from $P_t(\langle \boldsymbol{e}, \boldsymbol{f} \rangle)$, but smooths the probability of longer phrase pairs by combining shorter phrase pairs in the order specified by the non-terminals generated by $P_x(x)$.

Inversion transduction grammar-based models can be used to find alignments for words in parallel sentences through the process of biparsing (Wu 1997). Within the ITG framework, a sentence pair $\langle \boldsymbol{e}_1^I, \boldsymbol{f}_1^J \rangle$ can be defined as the phrase pair that is generated by the node at the top of the derivation tree. Biparsing for ITGs finds the

[3] Here we are specifically referring to a special case of ITGs with only a single symbol each for straight and inverted productions, which is also known as the *bracketing* ITG. ITGs with multiple straight and inverted terminals are also conceivable, but are generally not used in alignment as they significantly increase the computational burden of learning the ITG.

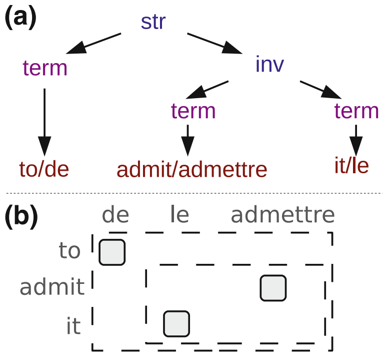


Fig. 2 An example of **a** an inversion transduction grammar (ITG) derivation tree, **b** its corresponding alignment matrix

most likely derivation for this sentence pair given the ITG probabilities. Once we have this most likely derivation, we treat all phrase pairs that were generated from the same terminal symbols as aligned (for example, in Fig. 2: “to/de,” “admit/admettre,” and “it/le”).

4.2 Biparsing and beam biparsing

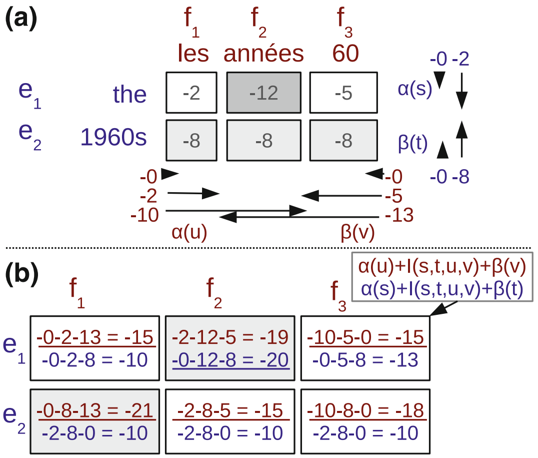
Biparsing ITGs is quite similar to standard chart parsing algorithms for monolingual PCFGs, efficiently calculating marginal probabilities for each span using bottom-up dynamic programming. We define the chart as a data structure with a single cell for each alignment $a_{s,t,u,v}$ spanning $\boldsymbol{e}_s^t$ and $\boldsymbol{f}_u^v$. Each cell has an accompanying “inside” probability $I(a_{s,t,u,v})$. This probability is the combination of the generative probability of each phrase pair $P_t(\boldsymbol{e}_s^t, \boldsymbol{f}_u^v)$ and the sum of the probabilities over all shorter spans in straight and inverted order, as in (1):

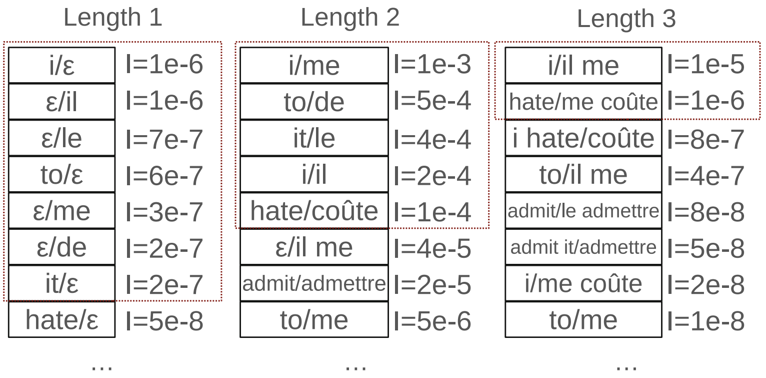
$$\begin{aligned} I(a_{s,t,u,v}) = P_t(e_s^t, f_u^v) & \\ + \sum_{s \le S \le t} \sum_{u \le U \le v} & P_x(x = \textsc{str}) I(a_{s,S,u,U}) I(a_{S,t,U,v}) \\ + \sum_{s \le S \le t} \sum_{u \le U \le v} & P_x(x = \textsc{inv}) I(a_{s,S,U,v}) I(a_{S,t,u,U}) \end{aligned} \tag{1}$$

where $P_x(x = \textsc{str})$ and $P_x(x = \textsc{inv})$ are the probability of straight and inverted ITG productions, respectively. An example of part of the chart used in this bottom-up parsing can be found in Fig. 3a, where we show the cells that have one-to-one alignments.

The exact calculation of these probabilities can be performed in $O(n^6)$ time, where $n = \max(I, J)$ is the length of the longer of $\boldsymbol{e}_1^I$ and $\boldsymbol{f}_1^J$ (Wu 1997). This calculation is performed using a dynamic programming algorithm that separates each of the spans into queues based on their length $l = t - s + u - v$, and queues are processed in ascending order of l. An example of the queues for the first three lengths is shown in Fig. 4.

Fig. 3 **a** A chart with inside log probabilities $I(s, t, u, v)$ in *boxes* and forward/backward look-ahead log probabilities marking surrounding *arrows*. **b** Spans with corresponding look-ahead probabilities added, and the minimum probability *underlined*. *Light* and *dark shaded* spans will be trimmed when the beam is $\log(P) \geq -3$ and $\log(P) \geq -6$ respectively





Length 1		Length 2		Length 3	
i/ε	I=1e-6	i/me	I=1e-3	i/il me	I=1e-5
ε/il	I=1e-6	to/de	I=5e-4	hate/me coûte	I=1e-6
ε/le	I=7e-7	it/le	I=4e-4	i hate/coûte	I=8e-7
to/ε	I=6e-7	i/il	I=2e-4	to/il me	I=4e-7
ε/me	I=3e-7	hate/coûte	I=1e-4	admit/le admettre	I=8e-8
ε/de	I=2e-7	ε/il me	I=4e-5	admit it/admettre	I=5e-8
it/ε	I=2e-7	admit/admettre	I=2e-5	i/me coûte	I=2e-8
hate/ε	I=5e-8	to/me	I=5e-6	to/me	I=1e-8
…		…		…	

Fig. 4 An example of the first three queues used in ITG parsing along with their inside probabilities. The hypotheses that would be processed if the beam is set to $c = 1e - 1$ are surrounded by *boxes*

The motivation behind this algorithm is that when calculating a particular span's inside probability $I(a_{s,t,u,v})$ according to Eq. (1), all of the other inside spans that we reference on the right-hand side of the equation are shorter than $a_{s,t,u,v}$ itself. Thus, if we process all spans in ascending order of length, it is simple to calculate these sums for every span in the chart. The computational complexity of the algorithm is $O(n^6)$ because Eq. (1) must be calculated for all of the $O(n^4)$ spans in the sentence, and there are $O(n^2)$ elements in each calculation of the sum.

However, exact computation of these probabilities in $O(n^6)$ time is impractical for all but the shortest sentences. Saers et al. (2009) note that in order to increase the efficiency of processing, queues can be trimmed based on a fixed histogram beam, only processing the b hypotheses with the highest probability for each queue. Here, we instead utilize a probability beam, expanding only hypotheses that are more than c times as likely as the best hypothesis $\hat{a}$. In other words, we have a queue discipline based on the inside probability, and all spans a_k where $I(a_k) < cI(\hat{a})$ are pruned. c is a constant between 0 and 1 describing the width of the beam, and a smaller constant probability will indicate a wider beam. Figure 4 shows an example of this, with boxes surrounding part of each queue showing the hypotheses that fall within the beam when $c = 10^{-1}$.

It should also be noted that slice sampling has been proposed as a way to improve efficiency in the learning of Bayesian ITGs (Blunsom and Cohn 2010). Comparing these two methods, slice sampling has the ability to derive exact samples from the true probability distribution for biparse trees, and the cost of faster parsing time is reflected in a larger number of samples required to converge to a high-probability section of the probability space. In contrast, beam search is by nature an approximate search method, and removes the guarantees of selecting from the true probability distribution. However, beam search is also comparatively simple and conducive to introduction of the look-ahead probabilities that we introduce in the following section, so we opt to use it instead.[4]

4.3 Look-ahead biparsing

While this pruning significantly increases the speed of biparsing, this method is insensitive to the existence of competing hypotheses when performing pruning. Figure 3a provides an example of what a competing hypothesis is, and why it is unwise to ignore them. Particularly, the alignments “les/1960s” and “les/the” both share the word “les,” and thus cannot both exist in a single derivation according to the ITG framework. We will call hypotheses that are mutually exclusive in this manner *competing* hypotheses. As the probability of “les/1960s” is much lower than its competing hypothesis “les/the,” it is intuitively unlikely to be chosen, and thus a good candidate for pruning. However, its inside probability is the same as that of “années/1960s,” which has no competing hypotheses and thus should not be removed from consideration. This section proposes the use of a look-ahead probability to increase the efficiency of this chart parsing by considering competing hypotheses.

In order to take into account competing hypotheses, we can use for our queue discipline not only the inside probability $I(a_k)$, but also the outside probability $O(a_k)$, the probability of generating all spans other than a_k, as in A* search for CFGs (Klein and Manning 2003), and tic-tac-toe pruning for word-based ITGs (Zhang and Gildea 2005). As the calculation of the true outside probability $O(a_k)$ is just as expensive as parsing itself, it is necessary to approximate this with heuristic function O^* that can be calculated efficiently.

This section proposes a heuristic function that is designed specifically for phrasal ITGs and is computable with worst-case complexity of n^2, compared with the n^3 amortized time of the tic-tac-toe pruning algorithm described by Zhang et al. (2008a). During the calculation of the phrase generation probabilities P_t, we save the best probability O^* for each monolingual span.

[4] It is also likely that the look-ahead probabilities could be integrated into the auxiliary variable sampling function for slice sampling to improve efficiency while maintaining correctness guarantees, an interesting challenge that we will leave to future work.

$$O_e^*(s,t) = \max_{\{\tilde{a}=\langle\tilde{s},\tilde{t},\tilde{u},\tilde{v}\rangle;\tilde{s}=s,\tilde{t}=t\}} P_t(\tilde{a}) \tag{2}$$

$$O_f^*(u,v) = \max_{\{\tilde{a}=\langle\tilde{s},\tilde{t},\tilde{u},\tilde{v}\rangle;\tilde{u}=u,\tilde{v}=v\}} P_t(\tilde{a}) \tag{3}$$

For each language independently, we calculate forward probabilities α and backward probabilities β. For example, $\alpha_e(s)$ is the maximum probability of the span $(0, s)$ of e that can be created by concatenating together consecutive values of O_e^*, as in (4):

$$\alpha_e(s) = \max_{\{S_1,\ldots,S_x\}} O_e^*(0, S_1) O_e^*(S_1, S_2) \ldots O_e^*(S_x, s). \tag{4}$$

Backwards probabilities and probabilities over f can be defined similarly. These probabilities are calculated for e and f independently, and can be calculated in n^2 time by processing each α in ascending order, and each β in descending order in a fashion similar to that of the forward-backward algorithm. Finally, for any span, we define the outside heuristic as the minimum of the two independent look-ahead probabilities over each language, as in (5):

$$O^*(a_{s,t,u,v}) = \min(\alpha_e(s) * \beta_e(t), \alpha_f(u) * \beta_f(v)). \tag{5}$$

It should be noted that both of the monolingual probabilities are optimistic estimates of the one-best outside probability $O(a_k)$ (in a manner similar to the heuristic function in A* search). Thus, taking the minimum of the two is motivated by the fact that we would like to choose the less optimistic of the two as a more accurate estimate of the true one-best probability.

Taking a look again at the example in Fig. 3b, it can be seen that when using these look-ahead probabilities, the relative probability difference between the highest probability span “les/the” and the spans “années/1960s” and “60/1960s” decreases, allowing for tighter beam pruning without losing these good hypotheses. In contrast, the relative probability of “les/1960s” remains low, as it is in conflict with a high-probability alignment, allowing it to be discarded.

5 Prior probabilities

One of the most critical elements to achieving accurate alignments in the probabilistic ITG is the accuracy of the phrase distribution P_t. Previous work on many-to-many alignment (DeNero et al. 2008; Neubig et al. 2011) helps achieve more accurate translations through the definition of a phrase pair prior probability $P_{base}(\boldsymbol{e}_s^t, \boldsymbol{f}_u^v)$, also referred to as the “base measure”. This can help efficiently seed the search process with a bias towards phrase pairs that satisfy certain properties. In particular, there are three pieces of prior knowledge that we would like to provide through the base measure. First, we would like to minimize the number of phrases that are not aligned to any phrase in the other language, as we can assume that most of the phrases will have some corresponding translation. Second, we would like to bias against overly long phrases, as these are likely to cause sparsity and hurt generalization performance when the model is tested on new data. Finally, to the best extent possible, we would like to provide information about whether phrase pairs are likely potential alignments. This can be done by using a simpler alignment model that is more efficient but less accurate than the ITG-based many-to-many alignment model.

5.1 One-to-many prior probabilities

First, we describe a formulation of the base probability similar to that of DeNero et al. 2008, which uses efficiently calculable IBM Model 1 probabilities to seed the ITG translation model. P_{base} is first calculated by choosing whether to generate an unaligned phrase pair (where $|\boldsymbol{e}| = 0$ or $|\boldsymbol{f}| = 0$) according to a fixed probability p_u. p_u should generally be a small value (10^{-2} in our experiments) to minimize the number of unaligned phrases. Based on this choice, we next generate an aligned phrase pair from P_{ba}, or an unaligned phrase pair from P_{bu}

For P_{ba}, we follow DeNero et al. (2008) in using the geometric mean of unidirectional IBM Model 1 probabilities, defined according to the probabilities in (6) and (7):

$$P_{ba}(\langle \boldsymbol{e},\boldsymbol{f}\rangle) = M_0(\langle \boldsymbol{e},\boldsymbol{f}\rangle)P_{pois}(|\boldsymbol{e}|;\lambda)P_{pois}(|\boldsymbol{f}|;\lambda) \tag{6}$$

$$M_0(\langle \boldsymbol{e},\boldsymbol{f}\rangle) = (P_{m1}(\boldsymbol{f}|\boldsymbol{e})P_{uni}(\boldsymbol{e})P_{m1}(\boldsymbol{e}|\boldsymbol{f})P_{uni}(\boldsymbol{f}))^{\frac{1}{2}}. \tag{7}$$

P_{pois} is the Poisson distribution with the average length parameter λ, where k represents the phrase length $|\boldsymbol{f}|$ or $|\boldsymbol{e}|$, as in (8):

$$P_{pois}(k|\lambda) = \frac{(\lambda-1)^{k-1}}{(k-1)!}e^{-(\lambda-1)}. \tag{8}$$

λ was set to a relatively small value (10^{-2} in our experiments), which allows us to bias against overly long phrases.

P_{uni} is the unigram probability of a particular phrase, and P_{m1} is the word-based Model 1 (Brown et al. 1993) probability of one phrase given the other. Model 1 probabilities are word-based translation probabilities that help to indicate whether the words in each phrase are good translations of each other. The phrase-based Model 1 probability is calculated as in (9):

$$P_{m1}(\boldsymbol{e}|\boldsymbol{f}) = \prod_{i=1}^{|\boldsymbol{e}|}\frac{1}{|\boldsymbol{f}|+1}\sum_{j=0}^{|\boldsymbol{f}|}P_{m1}(e_i|f_j) \tag{9}$$

where e_i and f_j are the i-th and j-th words in phrases $\boldsymbol{e}$ and $\boldsymbol{f}$ respectively, and f_0 is a token for null alignments. The word-based probabilities $P_{m1}(e_i|f_j)$ and $P_{m1}(f_j|e_i)$ are parameters of the model, and can be calculated efficiently using the expectation maximization algorithm (Brown et al. 1993) before starting phrase alignment. Following Liang et al. (2006), we combine the Model 1 probabilities in both directions using the geometric mean, which assigns a high probability to spans where both models agree, and lower probability to any span where any one of the models assigns a low probability.

For P_{bu}, in the case of $|\boldsymbol{f}| = 0$, we calculate the probability as in (10):

$$P_{bu}(\langle \boldsymbol{e}, \boldsymbol{f} \rangle) = P_{uni}(\boldsymbol{e}) P_{pois}(|\boldsymbol{e}|; \lambda)/2. \tag{10}$$

The probability can be calculated similarly when $|\boldsymbol{e}| = 0$. Note that P_{bu} is divided by 2 as the probability is considering null alignments in both directions.

5.2 Substring-based prior probabilities

While the method suggested in the previous section is effective, it is also highly dependent on the quality of the IBM Model 1 probabilities. However, for reasons previously stated, these methods are less satisfactory when performing character-based alignment, as the amount of information contained in a character does not allow for proper alignment. In this section, we propose a novel method for using raw substring co-occurrence statistics to bias alignments towards substrings that often co-occur in the entire training corpus. This is similar to the method of Cromières (2006), but instead of using these co-occurrence statistics as a heuristic alignment criterion, they are incorporated as a prior probability in a statistical model that can take into account mutual exclusivity of overlapping substrings in a sentence.

We define this prior probability using three counts over substrings $c_{\boldsymbol{e}}$, $c_{\boldsymbol{f}}$, and $c_{\langle \boldsymbol{e}, \boldsymbol{f} \rangle}$. $c_{\boldsymbol{e}}$ and $c_{\boldsymbol{f}}$ count the total number of sentences in which the substrings $\boldsymbol{e}$ and $\boldsymbol{f}$ occur, respectively. $c_{\langle \boldsymbol{e}, \boldsymbol{f} \rangle}$ is a count of the total number of sentences in which the substring $\boldsymbol{e}$ occurs on the target side, and $\boldsymbol{f}$ occurs on the source side.[5] We can perform the calculation of these statistics using enhanced suffix arrays, a data structure that can efficiently calculate all substrings in a corpus (Abouelhoda et al. 2004).[6]

While suffix arrays allow for efficient calculation of these statistics, storing all co-occurrence counts $c_{\langle \boldsymbol{e}, \boldsymbol{f} \rangle}$ is an unrealistic memory burden for larger corpora. In order to reduce the amount of memory used, each count is discounted fixed value d, which is set to 5. This has a dual effect of reducing the amount of memory needed to hold co-occurrence counts by removing values for which $c_{\langle \boldsymbol{e}, \boldsymbol{f} \rangle} < d$, as well as helping to prevent overfitting the training data. In addition, we can heuristically prune values for which the conditional probabilities $P(\boldsymbol{e}|\boldsymbol{f})$ or $P(\boldsymbol{f}|\boldsymbol{e})$ are less than some fixed value, which is set to 0.1 for the reported experiments.

In preliminary experiments designed to determine how to combine $c_{\boldsymbol{e}}$, $c_{\boldsymbol{f}}$, and $c_{\langle \boldsymbol{e}, \boldsymbol{f} \rangle}$ into prior probabilities we tested a number of methods proposed by previous research

[5] It should be noted that we are not counting duplicate occurrences of substrings in a single sentence. This was a design choice to prevent the over-counting of one-character or very short strings that tend to occur many times in a single sentence.

[6] Using the open-source implementation esaxx http://code.google.com/p/esaxx/.

including plain co-occurrence counts, the Dice coefficient, and Chi-squared statistics (Cromières 2006). As a result of these experiments, we found the most effective to be a new method of defining substring pair probabilities to be proportional to bidirectional conditional probabilities; as in (11) and (12):

$$P_{cooc}(\boldsymbol{e},\boldsymbol{f}) = P_{cooc}(\boldsymbol{e}|\boldsymbol{f})P_{cooc}(\boldsymbol{f}|\boldsymbol{e})/Z \tag{11}$$

$$= \left(\frac{c_{\langle \boldsymbol{e},\boldsymbol{f}\rangle} - d}{c_{\boldsymbol{f}} - d}\right)\left(\frac{c_{\langle \boldsymbol{e},\boldsymbol{f}\rangle} - d}{c_{\boldsymbol{e}} - d}\right)/Z \tag{12}$$

for all substring pairs where $c_{\langle \boldsymbol{e},\boldsymbol{f}\rangle} > d$ and where Z is a normalization term equal to that in (13):

$$Z = \sum_{\{\boldsymbol{e},\boldsymbol{f};c_{\langle \boldsymbol{e},\boldsymbol{f}\rangle}>d\}} P_{cooc}(\boldsymbol{e}|\boldsymbol{f})P_{cooc}(\boldsymbol{f}|\boldsymbol{e}). \tag{13}$$

The motivation for combining the probabilities in this fashion is similar to that of the base measure in Eq. (7), finding highly reliable alignments that are supported by both models. The preliminary experiments showed that the bidirectional conditional probability method gave significantly better results than all other methods, so this method will be adopted for the remainder of the experiments.

It should be noted that as we are using discounting, many substring pairs will be given zero probability according to P_{cooc}. As the prior is only supposed to bias the model towards good solutions and not explicitly rule out any possibilities, we can instead linearly interpolate the co-occurrence probability with the one-to-many Model 1 probability, which will give at least some probability mass to all substring pairs, as in (14):

$$P_{base}(\boldsymbol{e},\boldsymbol{f}) = \lambda P_{cooc}(\boldsymbol{e},\boldsymbol{f}) + (1-\lambda)P_{m1}(\boldsymbol{e},\boldsymbol{f}). \tag{14}$$

In order to find an appropriate value, we put a Beta prior ($\alpha = 1$, $\beta = 1$) on the interpolation coefficient λ and learn it during training.

6 Experiments

This section describes experiments over a variety of language pairs designed to test the effectiveness of the proposed substring-based translation method.

6.1 Experimental setup

Evaluation was performed on a combination of four languages with English, using freely available data. The first three language pairs, French–English, German–English, and Finnish–English, used data from EuroParl (Koehn 2005), with development and test sets designated for the 2005 ACL shared task on machine translation.[7] Experiments

[7] http://www.statmt.org/wpt05/mt-shared-task/.

Table 1 The number of sentences and words in each corpus for TM and LM training, tuning, and testing

Type	de-en		fi-en		fr-en		ja-en	
	Sent	Word	Sent	Word	Sent	Word	Sent	Word
TM (en)	457k	2.80M	467k	3.10M	457k	2.77M	286k	2.13M
TM (other)		2.56M		2.23M		3.05M		2.34M
LM (en)	751k	16.0M	717k	15.5M	688k	13.8M	440k	11.5M
LM (other)		15.3M		11.3M		15.6M		11.9M
Tune (en)	2.00k	58.7k	2.00k	58.7k	2.00k	58.7k	1.24k	30.8k
Tune (other)		55.1k		42.0k		67.3k		34.4k
Test (en)	2.00k	58.0k	2.00k	58.0k	2.00k	58.0k	2.00k	26.6k
Test (other)		54.3k		41.4k		66.2k		28.5k

were also performed with Japanese–English Wikipedia articles from the Kyoto Free Translation Task (Neubig 2011) using the designated training and tuning sets, and reporting results on the test set. These languages were chosen as they have a variety of interesting characteristics. French has some level of inflection, but among the test languages has the strongest one-to-one correspondence with English, and is generally considered to be easy to translate. German has many compound words, which must be broken apart in order to translate properly into English. Finnish is an agglutinative language with extremely rich morphology, resulting in long words and the largest vocabulary of the languages in EuroParl. Japanese does not have any clear word boundaries, and uses logographic characters, which contain more information than phonetic characters.

With regards to data preparation, the EuroParl data was pre-tokenized, so the experiments simply used the tokenized data ‘as is’ for the training and evaluation of all models. For word-based translation in the Kyoto task, training was performed using the tokenization scripts provided. For character-based translation, no tokenization was performed, using the original text for both training and decoding. For both tasks, all sentences for which both source and target were 100 characters or less were selected as training data, the total size of which is shown in Table 1.[8] In character-based translation, white spaces between words were treated as any other character and not given any special treatment. Evaluation was performed on tokenized and lower-cased data.

For alignment, GIZA++ (Och and Ney 2003) was used as an implementation of one-to-many alignment, with pialign used as an implementation of the ITG models[9] modified with the proposed improvements. For GIZA++, the default settings were used for word-based alignment, but for character-based alignment the training process was stopped at the HMM model, omitting IBM Models 3 and above, as the more advanced

[8] The 100-character limit results in the use of somewhat shorter sentences than when using limits based on words. For example, using a more traditional limit of a maximum of 40 words on both sides for Japanese-English results in a total of 5.91M words of English, 2.7 times greater than when a 100-character limit is used. The 100-character limit was mainly for efficient experimentation in the character-based models, and we describe possible directions for raising this limit in the future work section.

[9] http://phontron.com/pialign/.

models caused training to fail for longer sentences. For pialign, default settings were used except for character-based ITG alignment, which used a probability beam of 10^{-4} instead 10^{-10}. Decoding was performed with the Moses decoder (Koehn et al. 2007), with the default settings except for the stack size, which was set to 1000 instead of 200. Minimum error rate training (Och 2003) was performed to maximize word-based BLEU score for all systems.[10] For language models, word-based translation used a word 5-g model, and character-based translation used a character 12-g model, both smoothed using interpolated Kneser–Ney smoothing (Kneser and Ney 1995).[11]

6.2 Quantitative evaluation

This section presents a quantitative analysis of the translation results for each of the proposed methods. As previous research has shown that it is more difficult to translate into morphologically rich languages than into English (Koehn 2005), experiments are performed to test the accuracy translating in both directions for all language pairs. Translation quality is evaluated using BLEU score (Papineni et al. 2002), both on the word and character level, as well as METEOR (Denkowski and Lavie 2011) on the word level. For METEOR, we used the language-independent setting for Japanese and Finnish, and the language-dependent settings for the remaining languages.

Table 2 shows the results of the evaluation. It can be seen that in general, character-based translation with all of the proposed alignment improvements greatly exceeds character-based translation using the IBM models, confirming the hypothesis that substring-based information is necessary for accurate alignments. In general, the accuracy of character-based translation is comparable or slightly inferior to that of word-based translation. The evaluation of character-based BLEU shows that character-based translation is superior, comparable, or inferior depending on the language pair, word-based METEOR shows that character-based translation is comparable or inferior, and word-based BLEU shows that character-based translation is inferior.

For translation into English, character-based translation achieves higher relative accuracy compared to word-based translation on Japanese and Finnish input, followed by German, and finally French. This is notable in that it confirms the fact that character-based translation is performing well on languages that have long words or ambiguous boundaries, and less well on language pairs with a relatively strong one-to-one correspondence between words in both languages.

[10] This setup was chosen to minimize the effect of the tuning criterion on the comparison between the baseline and the proposed system, although it does imply that we must have access to tokenized data for the development set.

[11] We also performed experiments in which we incorporated a word-based language model in character-based translation, but found that this consistently gave neutral to negative results, a similar finding to that of Vilar et al. (2007). We suspect that this is due to the fact that word-based language models assign a sudden, large penalty when a word completes, hurting decoding. In addition, the modeling of unknown words is not trivial, and while we provided a fixed penalty for each unknown word (tuned using MERT), a more sophisticated unknown word model is probably necessary.

Table 2 Translation results for word-based BLEU (wBLEU), character-based BLEU (cBLEU), and METEOR for the GIZA++ and ITG models for word and character-based translation, with bold numbers indicating a statistically insignificant difference from the best system according to the bootstrap resampling method at $p = 0.05$

	wBLEU	cBLEU	METEOR	wBLEU	cBLEU	METEOR
	de-en			en-de		
GIZA-word	**24.58**	64.28	30.43	**17.94**	62.71	**37.88**
ITG-word	23.87	**64.89**	**30.71**	17.47	**63.18**	**37.79**
GIZA-char	08.05	45.01	15.35	06.17	41.04	19.90
ITG-char	21.79	64.47	30.12	15.35	61.95	35.45
	fi-en			en-fi		
GIZA-word	20.41	60.01	27.89	**13.22**	58.50	**27.03**
ITG-word	**20.83**	61.04	28.46	**13.12**	**59.27**	**27.09**
GIZA-char	06.91	41.62	14.39	04.58	35.09	11.76
ITG-char	18.38	**62.44**	**28.94**	12.14	**59.02**	25.31
	fr-en			en-fr		
GIZA-word	**30.23**	**68.79**	34.20	**32.19**	69.20	**52.39**
ITG-word	**29.92**	68.64	**34.29**	31.66	**69.61**	51.98
GIZA-char	11.05	48.23	17.80	10.31	42.84	25.06
ITG-char	26.70	66.76	32.47	27.74	67.44	48.56
	ja-en			en-ja		
GIZA-word	**17.95**	56.47	**24.70**	**20.79**	27.01	**38.41**
ITG-word	17.14	56.60	**24.89**	**20.26**	**28.34**	**38.34**
GIZA-char	09.46	49.02	18.34	01.48	00.72	06.67
ITG-char	15.84	**58.41**	24.58	17.90	**28.46**	35.71

Table 3 METEOR scores for alignment with and without look-ahead and co-occurrence priors, bold numbers indicate a statistically insignificant difference from the best system according to the bootstrap resampling method at $p = 0.05$

	fi-en	en-fi	ja-en	en-ja
ITG +cooc +look	**28.94**	**25.31**	**24.58**	**35.71**
ITG +cooc −look	28.51	24.24	**24.32**	**35.74**
ITG −cooc +look	**28.65**	24.49	**24.36**	35.05
ITG −ooc −look	27.45	23.30	23.57	34.50

6.3 Effect of improvements to the alignment method

This section compares the translation accuracy for character-based translation using the ITG model with and without the proposed improvements of substring co-occurrence priors and look-ahead parsing as described in Sects. 4 and 5.

METEOR scores for experiments translating Japanese and Finnish are shown in Table 3.[12] It can be seen that the co-occurrence prior probability provides gains in all

[12] Character-based BLEU and word-based BLEU showed similar relative gains.

Table 4 An adequacy evaluation of word- and character-based MT (0–5 scale)

	fi-en	ja-en
ITG-word	2.851	2.085
ITG-char	2.826	2.154

Table 5 The major gains of character-based translation, source-side unknown words (Src), target-side unknown words (Trg), and rare words (Rare)

Type	#	Reference	ITG-word	ITG-char
Src	13	directive on equality	tasa-arvodirektiivi	equality directive
Trg	5	yoshiwara-juku station	yoshiwara no eki	yoshiwara-juku station
Rare	5	world health organisation	world health	world health organisation

# Indicates the total number of occurrences of each gain

cases, indicating that the using substring statistics over the whole corpus are providing effective prior knowledge to the ITG aligner. The introduced look-ahead probabilities improve accuracy significantly when substring co-occurrence counts are not used, but only slightly when co-occurrence counts are used. More importantly, they allow for more aggressive beam pruning, increasing sampling speed from 1.3 sent/s to 2.5 sent/s on the Finnish task, and 6.8 sent/s to 11.6 sent/s on the Japanese task.

6.4 Qualitative evaluation

This section presents the results of a subjective evaluation of Japanese–English and Finnish–English translations. In the evaluation, two raters evaluated 100 sentences each, assigning an adequacy score of 0–5 based on how well the translation conveys the information contained in the reference translation. The raters were asked to rate on shorter sentences of 8–16 English words to ease rating and interpretation. The results of this evaluation are shown in Table 4. It can be seen that the results are comparable, with no significant difference in average scores for either language pair.

A breakdown of the types of sentences for which character-based translation was given a score of at least two points more than word-based is shown in Table 5. It can be seen that character-based translation is, in fact, properly handling a number of sparsity phenomena. On the other hand, word-based translation was generally stronger with reordering and lexical choice of more common words.

6.5 Phrases used in translation

This section presents an analysis of the phrases used in the translation of 50 sentences using word- and character-based ITG alignment for the Finnish–English and Japanese–English tasks. First, Table 6 shows the number of phrases where the phrase used by one of the two systems was subjectively better than the phrase used by the other system. It can be seen that there are a greater number of accurate translations at the phrase level for the character-based system than for the word-based system across both languages.

Table 6 The number of phrases that were the same, different but of equal quality, or subjectively better translations in one of the two models

	fi-en	ja-en
Same phrase	220	215
Equal quality	209	217
ITG-char better	67	96
ITG-word better	35	69

Table 7 A phrase-by-phrase analysis of gains and losses for Finnish–English translation categorized by errors due to misalignment (Mis), conjugation (Conj.), deletion of a word (Del.), insertion of a word (Ins.), compound words (Comp.), and lexical choice (Lex.)

#	Type	Source	ITG-char	ITG-word
ITG-char better				
19	Mis.	itsenäisille *independent* (*pos./all.*)	independent for	economic reform
18	Conj.	perustuslaillisempi *constitution* (*comp.*)	constitution more	perustuslaillisempi
12	Del.	kuuluisi *belong* (*cond.*)	would include	would
12	Comp.	yleismietintöä *the general report*	the general report,	yleismietintöä
10	Ins.	myös *also*	also	will also be
8	Lex.	pelkojen *fears/emotions*	fears	emotions
ITG-word better				
19	Del.	itsellemme *ourselves* (*all.*)	*space*	ourselves
15	Ins.	haluan *i would like to*	i would like to make	i would like to
6	Lex.	jo *in/already*	already	in
5	Mis.	on vastattava *is answer*	must be	is answer
4	Conj.	vertailuanalyysiä *benchmarking*	comparative analysis	benchmarking

# The total number of instances of each class

In order to examine the types of phrases where one of the two systems is more accurate than the other, Tables 7 and 8 provide more detailed break-downs by type of the mistranslated phrases used by each of the models for Finnish–English and Japanese–English translation, respectively. It can be seen that character-based translation naturally handles a number of phenomena due to unknown words that are not handled by word-based systems, such as those requiring transliteration, decompounding, and division of morphological components. It should also be noted that this process is not perfect; there are a number of cases where character-based translation splits or transliterates words that would be more accurately translated as a whole, although the total number of correctly translated compounds and inflected words is more than twice the number of incorrectly translated ones.

With regards to Finnish–English, it is interesting to note that character-based translation also succeeded in discovering a number of inflectional suffixes that have a clear grammatical function in the language (Karlsson 1999). Examples of the most common sub-word units used in translation are shown in Table 9. It can be seen that all but one

Table 8 A phrase-by-phrase analysis of gains and losses for Japanese–English translation categorized by errors due to transliteration (Tran.), insertion of a word (Ins.), deletion of a word (Del.), insertion of a word (Ins.), misalignment (Mis.), lexical choice (Lex.), compound words (Comp.), or partial transliteration (Part.)

#	Type	Source	ITG-char	ITG-word
ITG-char better				
38	Tran.	希玄 *kigen*	kigen	希玄
19	Ins.	半年間 *six months*	half year	six months between
19	Del.	病のため *due to illness*	due to his illness	illness
17	Mis.	を求めて *seeking*	seeking	the
2	Lex.	俗 *lay/commonly called*	lay	commonly called
2	Comp.	顔洗い *face washing*	face washing	顔洗い
ITG-word better				
28	Del.	も用いた。 *was also used.*	.	was also used.
11	Ins.	招請 *invited*	invited cont	invited
11	Tran.	無常 *vanity*	mujo	vanity
10	Mis.	書かれた *written*	the	written
5	Part.	大佛 *osaragi*	os	osaragi
2	Lex.	で *in/and*	and	in

# The total number of instances of each class

Table 9 Common Finnish sub-word phrases along with their grammatical function

String	#	Grammatical function
n	564	Genitive (“of X”)
a	467	Partitive (“some X”)
i	307	Plural, non-nominative (“Xs”)
t	241	Plural, nominative (“Xs”)
sta	235	Elative (“out of X”)
e	156	Similar to “e” in “play**e**d”
lle	134	Allative (“onto X”)
s	133	–
ä	121	Partitive (“some X”)
in	114	Plural, genitive (“of Xs”)
ssa	94	Inessive (“in X”)

have a clear grammatical function in Finnish. The only exception “s” is used in the transliteration of unknown words, as well as part of some morphological paradigms (similarly to “e”). This demonstrates that despite using no sort of explicit morphological knowledge, character-based translation is able to handle, to some extent, the more common morphological paradigms in morphologically rich languages.

One significant area for improvement in the character-based model is that it has a tendency to create alignments of actual content words on the source side to the white space character on the target side, effectively deleting content words. While deleted words are a problem in the word-based model as well, the problem is more prevalent in the character-based model, so it will be worth examining the possibilities of giving space characters a special status in the translation model in the future.

Finally, we note that the character-based model helps not only with unknown words, but also words that do exist, but are misaligned by the word-based model because they are rare, or do not have a consistent translation. In fact, this was the single most common error category for Finnish–English, and a significant portion of the Japanese–English errors. This indicates that simply applying character-based methods to process unknown words will not be sufficient to overcome the sparsity issues of the word-based model.

6.6 Character-to-word and word-to-character translation

Up until now, we have mainly considered the traditional combination of translation from words to words, and translation from characters to characters. However, it is easy to imagine the translation from word strings on the source side to character strings on the target side, or vice-versa. In order to examine the effect of character-to-word or word-to-character translation, we performed additional experiments where only the source or target side was divided into characters, and the other parts of the text were left as words.

The results of these experiments for Finnish–English and Japanese–English translation are shown in Table 10.[13] From this table, it can be seen that there is no clear strategy for obtaining the highest accuracy across all language pairs. In general it can be seen that the largest positive effect of character-based translation can be obtained by dividing Japanese or Finnish on the source side. This is a reasonable result given the features of the languages, as well as our previous analysis, which showed that the largest number of gains from character-based translation were for unknown words on the source side.

6.7 Effect of the reordering limit

Finally we examine the effect of the reordering limit on word- and character-based translation for Finnish and Japanese to and from English. The reordering limit is a hard constraint on the length of reorderings allowed by phrase-based translation that is

[13] These numbers were produced with a different version of Moses than the numbers in previous sections, so should not be directly compared.

Table 10 Translation results in word-based BLEU (wBLEU), character-based BLEU (cBLEU), and METEOR for the ITG model for word- or character-based input and output

	wBLEU	cBLEU	METEOR	wBLEU	cBLEU	METEOR
	fi-en			en-fi		
Word → word	**20.58**	60.86	28.47	**13.06**	**60.14**	**27.20**
Char → char	17.97	61.30	28.71	11.92	58.86	25.09
Word → char	16.48	57.23	25.73	11.18	58.41	24.70
Char → word	19.44	**61.94**	**29.09**	10.78	56.49	24.85
	ja-en			en-ja		
Word → word	**17.07**	56.52	**24.73**	**19.96**	28.15	**38.26**
Char → char	15.41	**58.06**	**24.57**	17.59	**28.51**	35.74
Word → char	14.98	55.64	23.37	19.08	27.35	**37.08**
Char → word	15.65	56.62	24.42	17.60	**28.72**	36.04

Bold indicates a statistically insignificant difference from the best system $p = 0.05$

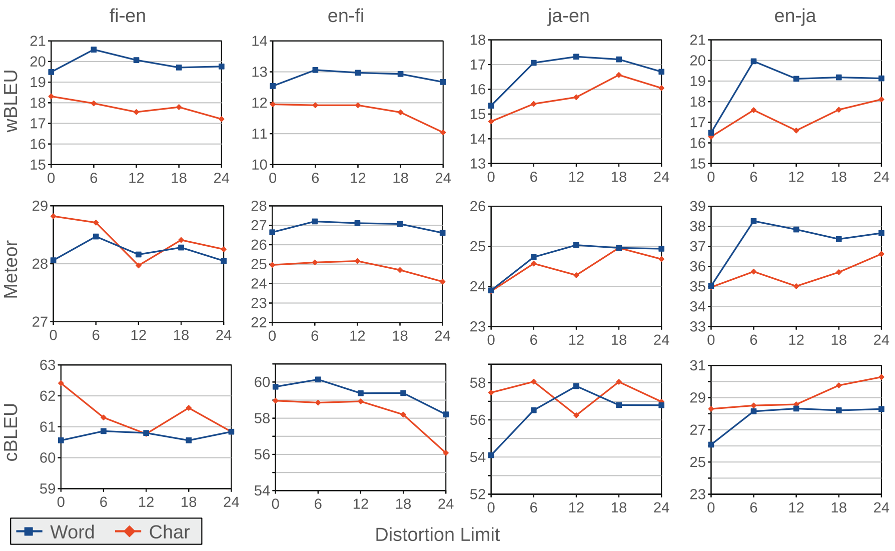


Fig. 5 Accuracy for four language pairs at each reordering limit

essential for producing translations efficiently (Koehn et al. 2005). All previous experiments used Moses' default reordering limit of 6 elements, which is often a reasonable limit for word-based translation, especially for similar language pairs. However, for character-based translation, a limit of 6 characters will often only translate into the reordering of a single word (or less). Thus, it could be expected that the effect of different reordering limits will have different effects on word- and character-based translation.

In order to examine the effect of the reordering limit, in Fig. 5 we show results for ITG-word and ITG-char over four language pairs with reordering limits of 0, 6, 12, 18, and 24. First, examining the results for word-based translation, we can see that a reordering limit of 6 is ideal for all language pairs except Japanese–English, which achieves the highest accuracy with a reordering limit of 12.

On the other hand, the results are much less consistent for character-based translation. For translation to or from Japanese, a large reordering limit generally helps character-based translation, with limits of 18 and 24 achieving ideal results for Japanese–English and English–Japanese, respectively. This is a somewhat expected result, as we require a larger reordering limit to handle the same types of reordering that may be covered by the character-based model. However, for English–Finnish translation there is no clear improvement by increasing the reordering limit, and for Finnish–English translation accuracy actually decreases significantly for any limit over 0. This indicates that the search space for character-based translation is too large, and the lexicalized phrase-based reordering models are too weak to find an appropriate reordering within this search space. Given this, it is likely that improved methods of decoding or constraining the search space within the character-based translation framework could further improve translation accuracy quite significantly.

7 Conclusion

This paper demonstrated that given improvements to alignment, character-based translation is able to act as a unified framework for handling a number of difficult problems in translation: morphology, compound words, transliteration, and word segmentation. It also presented two advances to many-to-many alignment methods that allow them to be run on much longer sentences, and improve accuracy through substring-based prior probabilities. However, while this is a first step towards moving beyond the concept of words in MT, there are still a number of remaining challenges.

One of the major challenges for the future is the development of efficient decoding methods for character-based translation models. As shown in the analysis of phrase quality in the system, the character-based model is able to produce better translations on the phrase level, but nevertheless achieves results that are approximately equal to those of the word-based systems. The major reason for this gap is that the word-based model tends to be better at reordering, as it is able to treat whole words as single units, which gives it both more freedom to handle reorderings over long distances and a more constrained search space that only considers more reasonable reorderings. Given more effective and efficient decoding methods, it is likely that we will be able to further close this gap in reordering quality, resulting in a clear advantage of the character-based models over word-based models.

In addition, there are still significant improvements that could be made to alignment speed to scale to longer sentences. This can probably be achieved through methods such as the heuristic span pruning of Haghighi et al. (2009) or sentence splitting of Vilar et al. (2007).

Finally, an interesting future direction is the consideration of discontiguous spans in character-based alignment. As noted in Fig. 1, the proposed model was able to capture a rudimentary concept agreement by learning phrases that combine the plural suffix of nouns with the plural form of a verb. Learning discontiguous spans (possibly with a method similar to that of Levenberg et al. (2012)) could further allow for the entirely unsupervised learning of morphological agreement, even when there are words intervening between the words that must agree.

References

Abouelhoda MI, Kurtz S, Ohlebusch E (2004) Replacing suffix trees with enhanced suffix arrays. J Discret Algorithms 2(1):53–86

Al-Onaizan Y, Knight K (2002) Translating named entities using monolingual and bilingual resources. In: 40th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, Philadelphia, pp 400–408

Bai MH, Chen KJ, Chang JS (2008) Improving word alignment by adjusting Chinese word segmentation. In: IJCNLP 2008, Proceedings of the 3rd International Joint Conference on Natural Language Processing. Hyderabad, pp 249–256

Blunsom P, Cohn T (2010) Inducing synchronous grammars with slice sampling. In: Human Language Technologies: The 2010Annual Conference of the North American Chapter of the Association for Computational Linguistics. Proceedings of the Main Conference, Los Angeles, pp 238–241

Blunsom P, Cohn T, Dyer C, Osborne M (2009) A Gibbs sampler for phrasal synchronous grammar induction. In: ACL-IJCNLP 2009, Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and 4th International Joint Conference on Natural Language Processing of the AFNLP. Proceedings of the Conference, Suntec, pp 782–790

Bojar O (2007) English-to-Czech factored machine translation. In: ACL2007: Proceedings of the Second Workshop on Statistical Machine Translation. Czech Republic, Prague, pp 232–239

Brown PF, Della-Pietra VJ, Della-Pietra SA, Mercer RL (1993) The mathematics of statistical machine translation: parameter estimation. Comput Linguist 19:263–312

Brown RD (2002) Corpus-driven splitting of compound words. In: TMI-2002 Conference: Proceedings of the 9th International Conference on Theoretical and Methodological issues in Machine Translation, Keihanna, pp 12–21

Chang PC, Galley M, Manning CD (2008) Optimizing Chinese word segmentation for machine translation performance. In: ACL-08: HLT: Third Workshop on Statistical Machine Translation. Proceedings of the Workshop, Columbus, pp 224–232

Chomsky N (1956) Three models for the description of language. IRE Trans Inf Theory 2(3):113–124

Chu C, Nakazawa T, Kawahara D, Kurohashi S (2012) Exploiting shared Chinese characters in Chinese word segmentation optimization for Chinese–Japanese machine translation. In: EAMT 2012, Proceedings of the 16th Annual Conference of the European Association for Machine Translation. Trento, pp 35–42

Chung T, Gildea D (2009) Unsupervised tokenization for machine translation. In: EMNLP 2009: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, Singapore, pp 718–726

Corston-Oliver S, Gamon M (2004) Normalizing German and English inflectional morphology to improve statistical word alignment. In: Machine Translation: from Real Users to Research, 6th Conference of the Association for Machine Translation in the Americas, AMTA (2004) Washington, DC, pp 48–57

Cromières F (2006) Sub-sentential alignment using substring co-occurrence counts. In: COLING—ACL2006, 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics. Proceedings of the Student Research Workshop, Sydney, pp 13–18

DeNero J, Bouchard-Côté A, Klein D (2008) Sampling alignment structure under a Bayesian translation model. In: EMNLP 2008: 2008 Conference on Empirical Methods in Natural Language Processing. Proceedings of the Conference, Honolulu, pp 314–323

Denkowski M, Lavie A (2011) Meteor 1.3: automatic metric for reliable optimization and evaluation of machine translation systems. In: Proceedings of the 6th Workshop on Statistical Machine Translation (WMT), Edinburgh, pp 85–91

Denoual E, Lepage Y (2005) BLEU in characters: towards automatic MT evaluation in languages without word delimiters. In: Proceedings of the 2nd International Joint Conference on Natural Language Processing, IJCNLP-05, Jeju Island, pp 81–86

Finch A, Sumita E (2007) Phrase-based machine transliteration. In: Proceedings of the Workshop on Technologies and Corpora for Asia-Pacific Speech Translation (TCAST), Hyderabad, pp 13–18

Goldwater S, McClosky D (2005) Improving statistical MT through morphological analysis. In: HLT/EMNLP 2005: Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing. Proceedings of the Conference, Vancouver, British Columbia, pp 676–683

Haghighi A, Blitzer J, DeNero J, Klein D, (2009) Better word alignments with supervised ITG models. In: ACL-IJCNLP 2009, Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and 4th International Joint Conference on Natural Language Processing of the AFNLP. Proceedings of the Conference, Suntec, pp 923–931

Jiang J, Ahmed Z, Carson-Berndsen J, Cahill P, Way A (2011) Phonetic representation-based speech translation. In: Proceedings of Machine Translation Summit XIII, Xiamen, pp 81–88

Karlsson F (1999) Finnish: an essential grammar. Routledge, London

Klein D, Manning CD (2003) A* parsing: fast exact Viterbi parse selection. In: HLT-NAACL 2003: Conference combining Human Language Technology conference series and the North American Chapter of the Association for Computational Linguistics conference series. Edmonton, pp 40–47

Kneser R, Ney H (1995) Improved backing-off for M-gram language modelling. In: International Conference on Acoustics, Speech, and Signal Processing (ICASSP), vol. 1, Detroit, pp 181–184

Knight K, Graehl J (1998) Machine transliteration. Comput Linguist 24(4):599–612

Koehn P (2005) Europarl: a parallel corpus for statistical machine translation. In: MT Summit X: The Tenth Machine Translation Summit, Phuket, pp 79–86

Koehn P, Axelrod A, Mayne AB, Callison-Burch C, Osborne M, Talbot D (2005) Edinburgh system description for the 2005 IWSLT speech translation evaluation. In: International Workshop on Spoken Language Translation: Evaluation Campaign on Spoken Language Translation [IWSLT 2005], Pittsburgh, 8pp [no page numbers]

Koehn P, Hoang H, Birch A, Callison-Burch C, Federico M, Bertoldi N, Cowan B, Shen W, Moran C, Zens R, Dyer C, Bojar O, Constantin A, Herbst E, (2007) Moses: open source toolkit for statistical machine translation. In: ACL 2007: proceedings of demo and poster sessions. Czech Republic, Prague, pp 177–180

Koehn P, Och FJ, Marcu D (2003) Statistical phrase-based translation. In: HLT-NAACL, (2003)conference combining Human Language Technology conference series and the North American Chapter of the Association for Computational Linguistics conference series. Edmonton, pp 48–54

Kondrak G, Marcu D, Knight K (2003) Cognates can improve statistical translation models. In: HLT-NAACL 2003:conference combining Human Language Technology conference series and the North American Chapter of the Association for Computational Linguistics conference series. Edmonton, pp 46–48

Lee YS (2004) Morphological analysis for statistical machine translation. In: HLT-NAACL, 2004: Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics. Proceedings of the Main Conference, Boston, Massachusetts, pp 57–60

Levenberg A, Dyer C, Blunsom P (2012) A Bayesian model for learning SCFGs with discontiguous rules. In: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL). Jeju Island, pp 223–232

Li H, Zhang M, Su J (2004) A joint source-channel model for machine transliteration. In: Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL), Barcelona, pp 159–166

Li M, Zong C, Ng HT (2011)Automatic evaluation of Chinese translation output: word-level or character-level? In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL). Portland, pp 159–164

Liang P, Taskar B, Klein D (2006) Alignment by agreement. In: Proceedings of the 2006Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL). Montreal, pp 104–111

Liu C, Ng HT (2012) Character-level machine translation evaluation for languages with ambiguous word boundaries. In: [ACL, 2012] Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics. Republic of Korea, pp 921–929

Macherey K, Dai A, Talbot D, Popat A, Och F (2011) Language-independent compound splitting with morphological operations. In: ACL-HLT, 2011: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics. Portland, pp 1395–1404

Marcu D, Wong W (2002) A phrase-based, joint probability model for statistical machine translation. In: Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing. Philadelphia, pp 133–139

Nakov P, Tiedemann J (2012) Combining word-level and character-level models for machine translation between closely-related languages. In: [ACL, 2012] Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics. Republic of Korea, Jeju, pp 301–305

Naradowsky J, Toutanova K (2011) Unsupervised bilingual morpheme segmentation and alignment with context-rich hidden semi-Markov models. In: ACL-HLT, 2011: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics. Portland, pp 895–904

Neubig G (2011) The Kyoto free translation task. http://www.phontron.com/kftt. Accessed 16 May 2011

Neubig G, Watanabe T, Mori S, Kawahara T (2012) Machine translation without words through substring alignment. In: : [ACL, 2012] Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics. Republic of Korea, pp 165–174

Neubig G, Watanabe T, Sumita E, Mori S, Kawahara T (2011) An unsupervised model for joint phrase alignment and extraction. In: ACL-HLT, 2011: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics. Portland, pp 632–641

Nguyen T, Vogel S, Smith NA, (2010) Nonparametric word segmentation for machine translation. In: Coling 2010, 23rd International Conference on Computational Linguistics. Proceedings of the Conference, Beijing, pp 815–823

Nießen S, Ney H (2000) Improving SMT quality with morpho-syntactic analysis. In: The 18th International Conference on Computational Linguistics, COLING 2000 in Europe. Proceedings of the Conference, Saarbrücken, pp 1081–1085

Och, FJ (2003) Minimum error rate training in statistical machine translation. In: ACL-2003: 41st Annual meeting of the Association for Computational Linguistics, Sapporo, pp 160–167

Och FJ, Ney H (2003) A systematic comparison of various statistical alignment models. Comput Linguist 29(1):19–51

Papineni K, Roukos S, Ward T, Zhu WJ (2002) BLEU: a method for automatic evaluation of machine translation. In: 40th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, Philadelphia, pp 311–318

Saers M, Nivre J, Wu D (2009) Learning stochastic bracketing inversion transduction grammars with a cubic time biparsing algorithm. In: IWPT-09: Proceedings of the 11th International Conference on Parsing Technologies, Paris, pp 29–32

Snyder B, Barzilay R (2008) Unsupervised multilingual learning for morphological segmentation. In: ACL-08: HLT, 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, Columbus, pp 737–745

Sornlertlamvanich V, Mokarat C, Isahara H (2008) Thai-lao machine translation based on phoneme transfer. In: Proceedings of the 14th Annual Meeting of the Association for Natural Language Processing, Tokyo, pp 65–68

Subotin M (2011) An exponential translation model for target language morphology. In: ACL-HLT,2011: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics. Portland, Oregon, pp 230–238

Talbot D, Osborne M (2006) Modelling lexical redundancy for machine translation. In: COLING—ACL, 2006, 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics. Proceedings of the Conference, Sydney, pp 969–976

Tiedemann J (2009) Character-based PSMT for closely related languages. In: EAMT-2009: Proceedings of the 13th Annual Conference of the European Association for Machine Translation, Barcelona, pp 12–19

Tiedemann J (2012) Character-based pivot translation for under-resourced languages and domains. In: [EACL2012] Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics. Avignon, pp 141–151

Vilar D, Peter JT, Ney H (2007)Can we translate letters? In: ACL 2007: Proceedings of the Second Workshop on Statistical Machine Translation. Czech Republic, Prague, pp 33–39

Vogel S, Ney H, Tillmann C (1996) HMM-based word alignment in statistical translation. In: COLING-96: The 16th International Conference on Computational Linguistics, Proceedings, Copenhagen, pp 836–841

Wang Y, Uchimoto K, Kazama J, Kruengkrai C, Torisawa K (2010) Adapting Chinese word segmentation for machine translation based on short units. In: LREC 2010: proceedings of the seventh international conference on Language Resources and Evaluation. La Valetta, Malta, pp 1758–1764

Wu D (1997) Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. Comput Linguist 23(3):377–403

Xu J, Zens R, Ney H (2004)Do we need Chinese word segmentation for statistical machine translation? In: Proceedings of the 3rd SIGHAN workshop on Chinese language processing. Barcelona, pp 122–128

Zhang H, Gildea D (2005) Stochastic lexicalized inversion transduction grammar for alignment. In: ACL-05: 43rd Annual Meeting of the Association for Computational Linguistics Ann Arbor, Michigan, pp 475–482

Zhang H, Quirk C, Moore RC, Gildea D (2008a) Bayesian learning of non-compositional phrases with synchronous parsing. In: ACL-08: HLT, 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, Columbus, pp 97–105

Zhang R, Yasuda K, Sumita E (2008b) Improved statistical machine translation by multiple Chinese word segmentation. In: ACL-08: HLT: Third Workshop on Statistical Machine Translation, Proceedings of the Workshop, Columbus, pp 216–223