The Knowledge Engineering Review, Vol. 27:4, 413–431. © Cambridge University Press, 2012
doi:10.1017/S026988891200029X

An overview of the phrase-based statistical machine translation techniques

MARTA RUIZ COSTA-JUSSÀ

Barcelona Media Innovation Center, Avenida Diagonal 177, 9th floor, 08018 Barcelona, Spain;
e-mail: marta.ruiz@barcelonamedia.org

Abstract

This work provides a general overview of the statistical machine translation (SMT) scientific field, which is a subfield of machine translation (MT). Specifically, this paper focuses on one of the most popular SMT approaches, that is, the phrase-based system.

The phrase-based translation units are typically extracted using statistical criteria, and they are weighted using different models. These models are log-linearly combined in the decoding, which is in charge of choosing the most probable translation. Significant quality improvements have been produced from original phrase-based SMT systems. Among others, the main challenges are reordering, domain adaptation and evaluation.

1 Introduction

The main goal of machine translation (MT) is to be able to translate from a source language *s* to a target language *t*. MT is a difficult task, mainly because natural languages are highly complex. Many words have more than one meaning and sentences may have various readings. Certain grammatical relations in one language might not exist in another language. Moreover, there are non-linguistic factors such as the problem that performing a translation might require world knowledge. Additional challenges arise when dealing with spoken language translation like confronting non-grammatical texts.

In order to face the MT challenge, many dependencies have to be taken into account. Often, these dependencies are weak and vague, which makes it rarely possible to describe simple and relevant rules that hold without exception for different language pairs. Specifically, statistical machine translation (SMT), which is a subfield of MT, treats MT as a decision problem. Given a source sentence and among all possible target sentences, the target sentence with the highest probability will be chosen according to a statistically learned model. SMT technology has received increasing interest leading to improved algorithms and it has been justified by various successful comparative evaluations since its revival by the work of the famous IBM research group more than 15 years ago. It has proved to be a competitive approach, which shows greater robustness than other methods for the translation of spontaneous speech. However, translations generated by SMT systems still have several significant challenges to pursue, like word reordering or word correspondences.

SMT systems have been implemented using different approaches. One of the most popular is the phrase-based approach and this paper focuses on the research around this approach. Other techniques, which are syntax-based, are just referenced.

This report is organized as follows. While Section 2 motivates the statistical approach, Section 3 reviews its formal description. Section 4 reports the main works in statistical word alignment that have been used in the phrase-based SMT systems. Then, Section 5 describes the phrase-based

https://doi.org/10.1017/S026988891200029X Published online by Cambridge University Press

translation model. Section 6 overviews the log-linear framework and the decoding research. Afterward, the following sections are dedicated to describe the main research lines from the last few years which have a high impact in the phrase-based translation: Section 7 reports the reordering work; Section 8 depicts the work related to the optimization procedure of the phrase-based systems; Section 9 reports the work related to rescoring and system combination; Sections 10 and 11 explain domain adaptation and source context information techniques, respectively. Finally, Section 12 describes the main evaluation procedures which have been used to test phrase-based SMT systems. Section 13 concludes with a summary of the possible future directions in the SMT research.

2 Motivation of the statistical approach

Increasing computational power was one of the reasons to raise the current interest in MT. The scientific community is quite involved in MT. There is an active participation in major events, such as evaluations (NIST[1],WMT[2], IWSLT[3]), and outstanding research groups dedicate great efforts to contribute to MT advances. The major approaches to MT are generally distinguished according to their core technology. Under this classification, there are the *rule-based* and the *corpus-based* approaches.

- In the *rule-based approach*, human experts specify a set of rules to describe the translation process. This approach requires an enormous amount of input from human experts (Dorr, 1994; Arnold and Balkan, 1995).
- Under the *corpus-based approach*, the knowledge is automatically extracted by analyzing translation examples from a parallel corpus (built by human experts). The advantage is that, once the required techniques have been developed for a given language pair, (in theory) MT systems can be very quickly developed for new language pairs using provided training data. Within the corpus-based approaches, they can be further distinguished between *example-based* and *statistical* MT (EBMT and SMT, respectively). The former makes use of previously seen examples in parallel corpora (bilingual aligned text at the level of sentence consisting in human translations) as its main knowledge base. It is essentially a translation by analogy. The latter applies statistical learning techniques to build a translation model given the training parallel corpora. Thus, it relies on statistical parameters and a set of translation and language models, among other data-driven features. This approach initially worked on a word-by-word basis. However, current systems attempt to introduce a certain degree of linguistic analysis into the SMT approach. SMT approaches are mainly distinguished depending on the translation model (Lopez (2007) reports a formal review of most SMT approaches):
  - The phrase-based system identifies any contiguous sequence of words (called *phrase*) as unit of translation. Each source phrase is non-empty and translates to exactly one non-empty target phrase. The phrase-based translation model is trained using relative frequencies.
  - The Ngram-based approach is a variation of the phrase-based approach. Differently to the phrase-based translation model, the Ngram-based translation model is trained on bilingual *n*-grams (Mariño *et al.*, 2006).
  - Formally syntax-based SMT learns a synchronous context-free grammar from a bitext without any syntactic information (Wu, 1996; Chiang, 2007).
  - Linguistically and formally syntax-based SMT models are tied to some linguistic representations of syntax. These SMT systems parse a source sentence as a target sentence when translating (Yamada and Knight, 2002).
  - Dependency treelet translations (Menezes *et al.*, 2006) align the parallel corpus, then project the source dependency parse onto the target sentence and extract dependency treelet

[1] http://www.nist.gov/speech/tests/mt/doc/mt06-evalplan.v4.pdf

[2] http://www.aclweb.org/anthology/W/W06/W06-15

[3] http://www.slc.atr.jp/IWSLT2006/

https://doi.org/10.1017/S026988891200029X Published online by Cambridge University Press

translation pairs. Then, they train a tree-based ordering model. One of the main advantages here is that reordering is based on syntax.

Comparing with the different MT approaches, the main advantages of SMT are cited as follows:

- SMT systems are not tailored to any specific pair of languages. To the translation system, any language is treated the same, and there is no manually created rule-set of grammar, metaphors and similar language considerations.
- There is a better use of resources. There is a great deal of natural language in machine-readable format. SMT relies on a large corpus of texts which are available in multiple languages.
- Generally, given that it is a corpus-based approach, SMT system is learning from existing human translations. Therefore, more natural translations have been achieved in international evaluations compared with other MT approaches.

However, notice that SMT requires resources such as parallel corpora and high computationally resources. Moreover, the main challenges in SMT are: syntactic such as word reordering, which may be quite difficult in cases of language pairs with different structures; morphological, specially when translating from a less inflected language into a more inflected language; and lexical such as out-of-vocabulary words and domain adaptation.

3 Statistical machine translation formal description

The main goal of SMT is the translation of a text given in some source language into a target language. A source string $s_1^J = s_1 \dots s_j \dots s_J$ has to be translated into a target string $t_1^I = t_1 \dots t_i \dots t_I$. Among all possible target strings, the string with the highest probability will be chosen:

$$\tilde{t}_1^I = \underset{t_1^I}{argmax}\ P(t_1^I|s_1^J) \tag{1}$$

The first SMT systems were reformulated using Bayes's rule. This approach, called the noisy channel, models the probability of a target language sentence $t_1^I = t_1 \dots t_I$ given a source language sentence $s_1^J = s_1 \dots s_J$ as follows:

$$\tilde{t}_1^I = \underset{t_1^I}{argmax}\ \frac{p(s_1^J|t_1^I)\,p(t_1^I)}{p(s_1^J)} \tag{2}$$

The denominator $p(s_1^J)$ inside the *argmax* function can be ignored since the best t_1^I sentence for a fixed sentence s_1^J is the objective function, and hence $p(s_1^J)$ is a constant:

$$\tilde{t}_1^I = \underset{t_1^I}{argmax}\ p(s_1^J|t_1^I)\,p(t_1^I) \tag{3}$$

Here $p(t_1^I)$ is the language model of the target language. A language model is usually formulated as a probability distribution over strings that attempts to reflect how likely a string occurs inside a language (Chen & Goodman, 1998). SMT systems make use of the same n-gram language models as do speech recognition and other applications. The language model component is monolingual, so acquiring training data is relatively easy. Then, $p(s_1^J|t_1^I)$ is the string translation model, which is the basis of the translation. The *argmax* operation denotes the search problem, that is, the generation of the output sentence in the target language.

In recent systems, such an approach has been expanded to a more general maximum entropy approach in which a log-linear combination of multiple feature functions (h_m) is implemented (Och, 2003). With respect to criterion shown in Equation (3), this approach leads to maximizing a linear combination of feature functions:

$$\tilde{t} = \underset{t}{argmax}\left\{\sum_{m=1}^{M} \lambda_m h_m(s_1^J, t_1^I)\right\} \tag{4}$$

https://doi.org/10.1017/S026988891200029X Published online by Cambridge University Press

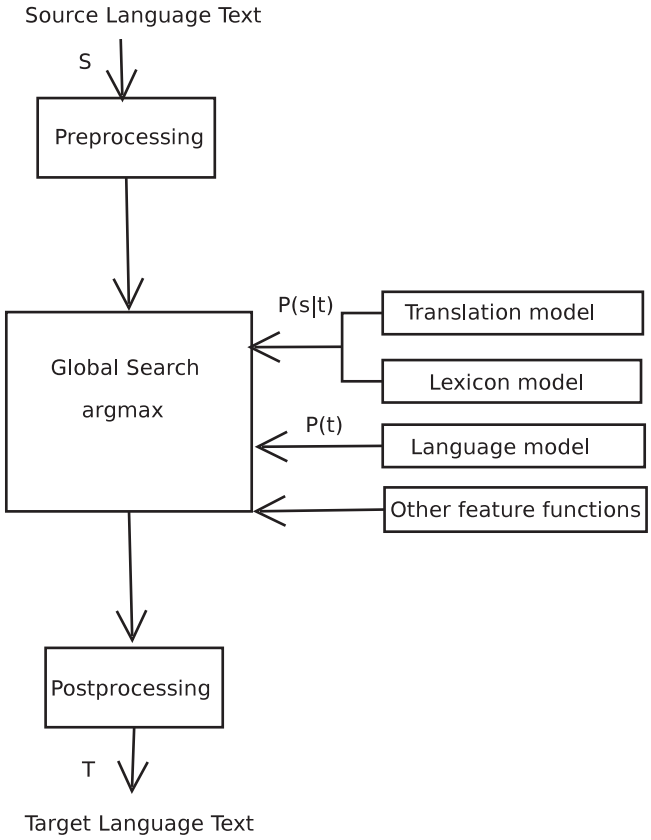


Figure 1 Architecture of the translation approach based on the log-linear framework approximation

where λ_m is the weight given to the $h_m(s_1^J, t_1^I)$ feature function and M is the total number of feature functions.

The overall architecture of this statistical translation approach is summarized in Figure 1.

4 Statistical word alignment

A key issue in modeling the string translation probability $p(s_1^J|t_1^I)$ is to define the correspondence between words of the target and source sentences. This section reports a brief description of the most popular word alignment approaches used to implement a phrase-based system.

In typical cases, it can be assumed a sort of pairwise dependence by considering all word pairs (s_j, t_i) for a given sentence pair (s_1^J, t_1^I). A word alignment is a mapping between the source words and the target words in a set of parallel sentences. Given parallel texts, the task of automatic word alignment focuses on detecting which tokens or set of tokens from each language are connected together in a given translation context or if any token has no alignment in which case is aligned to *NULL*.

Figure 2 shows a visualization of an alignment between the English sentence *For that is the decisive point here* and the Spanish sentence *Este es el punto decisivo*. Additionally, this figure shows what is called the word alignment matrix.

The alignment approach is an instance of unsupervised learning, where the system is not given examples of the kind of output desired, but it is instead trying to find values for the alignments which best explain the observed bitext.

In principle, there may be an arbitrary alignment relationships between the target and the source words. So-called statistical alignment models are decomposed into:

- *fertility model*, which accounts for the probability that a target word t_i generates ϕ_i words in the source sentence.

https://doi.org/10.1017/S026988891200029X Published online by Cambridge University Press

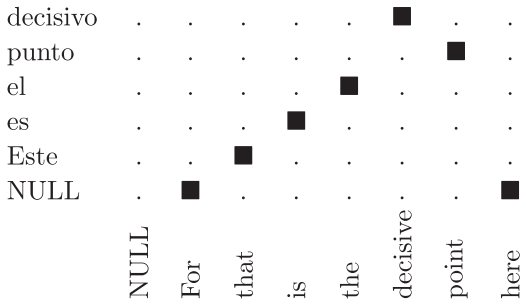


Figure 2 Word alignment example

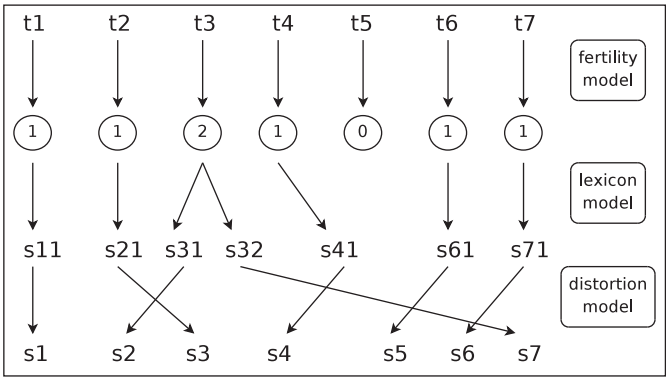


Figure 3 Illustration of the generative process underlying IBM models

- *lexicon model*, which models the probability to produce a source word s_j given a target word t_i.
- *distortion model*, which tries to explain the phenomenon of placing a source word in position j given that the target word is placed in position i in the target sentence (also used with inverted dependencies, and known as the alignment model).

The different combinations of these three models are commonly known in the literature as IBM models (Brown *et al.*, 1993). Currently, word alignments based on IBM and HMM models for which a systematic performance comparison can be found in Och (2003), are considered to be the state of the art. Typically, the implementation by Och and Ney (2000), which is freely available in the GIZA++ package, is used (Figure 3).

There are some current SMT systems that do not use IBM model parameters in their training schemes, but instead use only the *most probable* alignment (using a Viterbi search) given the estimated IBM models (typically by means of GIZA++). IBM models are asymmetric probabilistic models. Therefore, usually alignments are computed from source-to-target and target-to-source and, then, symmetrization strategies are applied. Several symmetrization algorithms have been proposed, the most widely known being the **union**, **intersection** and **refined** (Och & Ney, 2000) of source-to-target and target-to-source alignments, and the **grow-diag-final** (Koehn *et al.*, 2005), which employs the previous intersection and union alignments.

Lately, recent work has begun to explore supervised methods that rely on presenting the system with a (usually small) number of manually aligned sentences (Callison-Burch *et al.*, 2004). Besides the benefit of the additional information provided by supervision, these models are able to combine many features of the data, such as context, syntactic structure, part-of-speech or translation lexicon information, which are difficult to integrate into the generative statistical models traditionally used (Fraser & Marcu, 2006; Lambert, 2008). Additionally, discriminative alignment training allows to extend the IBM models with new (sub)models that leads to additional increases in word alignment accuracy.

https://doi.org/10.1017/S026988891200029X Published online by Cambridge University Press

In order to evaluate the quality of the word alignment task, the Alignment Error Rate (*AER*) measure as proposed in Och and Ney (2000) has been commonly used. Given a manual gold standard alignment with the criterion of Sure and Possible links, *Recall*, *Precision* and *AER* measures are defined as follows:

$$\begin{aligned} Recall &= \frac{|A \cap S|}{|S|}, \quad Precision = \frac{|A \cap P|}{|A|} \\ AER &= 1 - \frac{|A \cap S| + |A \cap P|}{|A| + |S|} \end{aligned} \tag{5}$$

where A is the hypothesis alignment and S is the set of Sure links in the gold standard reference, and P includes the set of Possible and Sure links in the gold standard reference.

In Fraser and Marcu (2006), the authors confirm experimentally that the AER does not correlate well with MT performance. On the other hand, they obtain strong correlations with the F-Measure:

$$\text{F-Measure with Sure and Possible}\,(A, P, S, \alpha) = \frac{1}{\frac{\alpha}{Precision(A,P)} + \frac{(1-\alpha)}{Recall(A,S)}} \tag{6}$$

The basic property of F-measure is that unbalanced *Precision* and *Recall* should be penalized.

Nowadays, the quality of the word alignment algorithms is highly dependent on the quality and quantity of parallel corpus. Additionally, the more monotonic are the language pairs, the better the word alignment.

5 Phrase-based statistical machine translation model

The job of the translation model, given a target sentence and a foreign sentence, is to assign a probability that t_1^I generates s_1^J. While these probabilities can be estimated by thinking about how each individual word is translated, modern SMT is based on the intuition that a better way to compute these probabilities is to consider the behavior of phrases (sequences of words). The intuition of phrase-based SMT is to use phrases as well as single words as the fundamental units of translation. Phrases are estimated from multiple segmentations of the word-aligned parallel corpora by using relative frequencies. The main difference between the phrase- and word-based models is that the former manages bilingual units of several words (i.e. *discurso extenso # long speech*) instead of only individual words themselves.

The basic idea of phrase-based translation is to segment the given source sentence into units (i.e. *phrases*), then translate each phrase and finally compose the target sentence from these phrase translations.

Basically, a bilingual phrase is a pair of m source words and n target words. For extraction from a bilingual word-aligned training corpus, two additional constraints are considered:

1. the words are consecutive and
2. they are consistent with the word alignment matrix, meaning that words inside the phrase are only aligned to words inside the phrase.

The phrase-based approach was first presented in Och (1999) and named *Alignment Templates*, consisting of pairs of generalized phrases which allow for word classes and include internal word alignments.

A simplification of this model is the so-called phrase-based SMT presented in Zens *et al.* (2002). This approach does not use word classes but instead uses bilingual phrases without internal alignment. The following criterion defines the set of bilingual phrases *BP* of the sentence pair $(s_1^J; t_1^I)$ that is consistent with the word alignment matrix A:

$$BP(s_1^J, t_1^I, A) = \{(s_j^{j+m}, t_i^{i+n}) : \forall (i', j')\, \varepsilon\, A : j \leqslant j' \leqslant j + m \leftrightarrow i \leqslant i' \leqslant i + n\} \tag{7}$$

https://doi.org/10.1017/S026988891200029X Published online by Cambridge University Press

For that ||| Este

For that is ||| Este es

is ||| es

is el ||| es the

el punto decisivo ||| the decisive point

punto decisivo ||| decisive point

punto decisivo ||| decisive point here

Figure 4 Phrase extraction example. Source and target are separated by |||

Figure 4 shows the phrases (maximum length of three words) extracted from the word alignment shown in Figure 2.

The extraction of bilingual phrases from a word alignment corpus can be done in a straightforward manner and pseudo-code is reported in Zens *et al.* (2002). To use the bilingual phrases in the translation model, the hidden variable B is introduced. This is a segmentation of the sentence pair $(s_1^J; t_1^I)$ into K phrases $(s_K^J; t_1^K)$. It is used a one-to-one phrase alignment, that is, one source phrase is translated by exactly one target phrase.

$$Pr(s_1^J|t_1^I) = \alpha(t_1^I) \times \sum_B Pr(\tilde{s}_k|\tilde{t}_k) \tag{8}$$

where the hidden variable B is the segmentation of the sentence pair in K bilingual phrases $(\tilde{s}_1^K, \tilde{t}_1^K)$, and $\alpha(t_1^I)$ assumes equal probability for all segmentations.

Given the collected phrase pairs, the phrase translation probability distribution is commonly estimated by relative frequency in both directions:

$$P(s|t) = \frac{N(s,t)}{N(t)} \tag{9}$$

$$P(t|s) = \frac{N(s,t)}{N(s)} \tag{10}$$

where $N(s, t)$ means the number of times the phrase s is translated by t. $N(s)$ and $N(t)$ are the number of times the source and the target phrase appear, respectively.

Recently, the phrase-based translation model has been further improved by using the factored Language Models (FLMs) technique. FLMs are a newer language modeling technique: more sophisticated than n-grams, but more widely applicable than the syntax-based models. FLMs do not reduce the complexity of an n-gram model, but they use a richer set of conditioning possibilities to improve performance. These models are capable of incorporating some amount of linguistic information, such as semantic classes or parts of speech. The appeal of FLMs for MT is this option to layer linguistic information in the corpus, when available, on top of the power of statistical n-gram language models (Axelrod, 2006).

6 Log-linear framework and decoding

Currently, most SMT systems consider a log-linear framework of probabilistic information (Och, 2002; Bertoldi *et al.*, 2006; Mariño *et al.*, 2006; Matusov *et al.*, 2006). This section describes the log-linear framework and names the main decoding used in a phrase-based system.

Following Figure 1, the simplest phrase-based approach (the noisy channel) would be the following combination of two feature functions (h):

$$h_{tm}(t_1^I, s_1^J) = log\, p(s_1^J|t_1^I); \;\; h_{lm}(t_1^I, s_1^J) = log\, p(t_1^I) \tag{11}$$

https://doi.org/10.1017/S026988891200029X Published online by Cambridge University Press

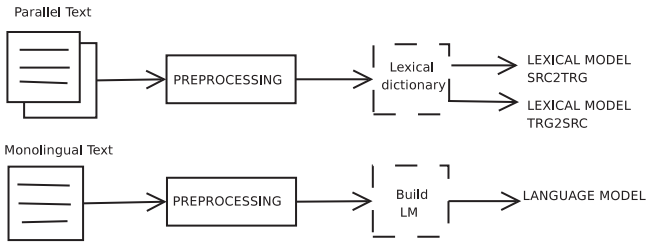


Figure 5 Training steps of the most common feature functions

Here, $p(t_1^I)$ denotes the trained language model and $p(s_1^J|t_1^I)$ denotes the trained translation model. Then, there can be obtained two maximum entropy model parameters λ_{lm} and λ_{tm} that can be trained using an optimization algorithm.

Additional feature functions widely used are cited as follows:

- A word bonus for each produced target word.

$$h_{wb}(t_1^I, s_1^J) = I \tag{12}$$

- Additional language models introducing linguistic knowledge as *Part of Speech* (POS) target language model (Costa-jussà *et al.*, 2006; Crego *et al.*, 2006).
- A distortion model. Distortion in SMT refers to a word having a different ('distorted') position in the target sentence than the corresponding word had in the source sentence. The feature functions related to distortion or reordering are reviewed and proposed in the next chapter.
- Lexical models (such as IBM model 1 from source-to-target and from target-to-source).

$$h_{ibm_1}(t_1^I, s_1^J) = log \frac{1}{(I+1)^J} \prod_{j=1}^{J} \sum_{i=0}^{I} p(t_j^n|s_i^n) \tag{13}$$

Figure 5 draws how to train three of the above mentioned feature functions.

All feature functions are combined in the decoder, which is in charge of solving the maximization problem of translation shown in Equation (3). Given that the decoding is an NP-hard problem, the phrase-based system uses a simplified search based on a beam search and its general framework is described in Koehn *et al.* (2007). The most relevant simplifications are the pruning away of low-scoring hypotheses and the reordering constraints in order to reduce the search space. With a well-tuned beam size, it is possible to gain large speedups with very little loss of accuracy (Tillmann & Ney, 2003; Lopez, 2007). *Moses*[4] is a freely available phrase-based decoder and it is widely used. When the feature functions do not have a feasible integration in the beam search procedure, feature functions are incorporated in rescoring (see Section 9).

7 Reordering approaches

This section describes several state-of-the-art reordering techniques employed in SMT systems. Reordering is understood as the word order redistribution of the translated words.

In initial SMT systems, this different order is only modeled within the limits of translation units. Relying only in the reordering provided by translation units may not be good enough in most language pairs, which might require longer reorderings. Therefore, additional techniques may be deployed to face the reordering challenge. Costa-jussà and Fonollosa (2009b) propose a classification of reordering techniques that tries to cover most recent reordering works. This classification can be summarized as follows:

- *Straight heuristic reordering search constraints*: These techniques are founded on the application of distance-based restrictions to the search space. The use of these constraints generates simple

[4] http://www.statmt.org/moses/

https://doi.org/10.1017/S026988891200029X Published online by Cambridge University Press

reordering models, which imply a necessary balance between translation accuracy and efficiency. One simple model is a '*weak*' distance-based distortion model that was initially used to penalize the longest reorderings. Other reordering constraints are: distortion (Koehn *et al*., 2003; Och & Ney, 2004); IBM (Berger *et al*., 1996; Tillmann & Ney, 2003); ITG (Wu, 1996); local (Kanthak *et al*., 2005) or maxjumps (Crego, 2008). In view of content independence of the distortion and flat reordering models, several researchers (Tillmann, 2004; Koehn *et al*., 2005) proposed a more powerful model called lexicalized reordering model that is phrase dependent. Lexicalized reordering model learns local orientations (monotone or non-monotone) with probabilities for each bilingual phrase from training data. This lexicalized reordering approach is implemented in the open source Moses toolkit[5].

- *Source reordering approaches*: Here the reordering rules are defined in the source language. The idea is to reorder the source language in a way that better matches the target language by using some of the following strategies:
  - *Deterministic Reordering Rules* (Popovic & Ney, 2006; Costa-jussà *et al*., 2011). The source corpus is reordered following a set of rules. These rules have been automatically learned using lexical and/or morphological information, that is, POS. The decoder search is monotonic.
  - *Clause Restructuring* (Xia & McCord, 2004; Collins *et al*., 2005; Wang *et al*., 2007). These methods, which are applied both in training and decoding steps, use syntactic information to reorder source words in SMT as a preprocessing step. This source reordering is complemented with a local reordering in search.
  - *Input Reordering Graph* (Kanthak *et al*., 2005; Mauser *et al*., 2006). The word alignment is then used as a function of source words to reorder the source corpus. Inspired by Knight and Al-Onaizan (1998), they permute the source sentence to provide a source input graph which extends the search graph. The reordering hypotheses of the source input graph are limited by several constraints, such as IBM or ITG. Similarly, in Crego and Mariño (2007) and Zhang *et al*. (2007), the reordering search problem is addressed through a source input graph. In this case, the reordering hypotheses are defined from a set of linguistically motivated rules (either using POS or chunks). In Costa-jussà and Fonollosa (2009a), the reordering hypotheses are automatically extracted by using an Ngram-based reordering approach, and in Khalilov *et al*. (2009), reordering hypotheses are derived through a syntactically augmented alignment of source and target texts.
- *Reordering in rescoring*: Typically the rescoring methods have generally provided small accuracy gains given the restriction of being applied to an *n*-best list. In these approaches, a baseline system is used to generate *n*-best translation hypotheses. Statistical or syntactic features are then used in a second model that re-ranks the *n*-best lists, in an attempt to improve over the baseline approach. Koehn and Knight (2003) apply a re-ranking approach to the sub-task of noun–phrase translation. Hassan *et al*. (2006) introduces super-tag information.
- *Reordering based on syntax structures*: This type of reordering is not carried out using standard phrases. In fact, it solves translation following some of the other SMT approaches, which were named in Section 2. Therefore, as this is out of this paper's scope, only some standard works are referenced. In Quirk *et al*. (2005) and Langlais and Gotti (2006), a dependency tree-based reordering model is inferred from aligned string-tree pairs. Parsing is performed on the source language and a corresponding dependency grammar is inferred on the aligned target side. Others use constituent trees (Chiang, 2007), in which context-free rules are inferred from string-to-string pairs (notice: no parsing is required). In this approach, phrases are reorganized into hierarchical by reducing subphrases to variables. This template-based scheme not only captures the reorderings of phrases, but also integrates some phrasal generalizations into the global model. This type of system has became more popular in recent International Evaluations, for example, see Vilar *et al*. (2008).

[5] http://www.statmt.org/moses/

https://doi.org/10.1017/S026988891200029X Published online by Cambridge University Press

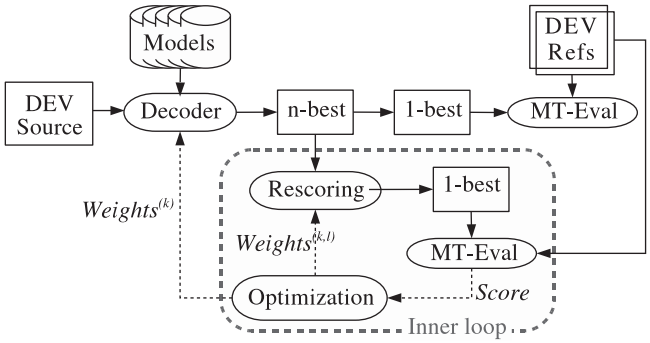


Figure 6 Double-loop optimization block diagram

8 Minimum error training

This section reports several optimization strategies that are used to train the model parameters λ_1^M of Equation (4).

An adequate algorithm for such a task is the *downhill simplex* algorithm (Nelder & Mead, 1965) or the SPSA (Lambert & Banchs, 2006). The method uses a geometrical figure called a simplex consisting in N dimensions of $N + 1$ points and all their interconnecting line segments, polygonal faces, etc. The starting point is a set of $N + 1$ points in parameter space, defining an initial simplex. At each step, the simplex performs geometrical operations (reflexions, contractions and expansions) until a local minimum is reached. Generally it adjusts the log-linear weights so as to maximize an objective function. Note that in this problem, only a local optimum is usually found. Tuning is performed according to MT quality measures and evaluated over development data.

Nowadays, the most widely used optimization scheme is an n-best list, which is produced by the decoder (see Figure 6). The optimization algorithm is used to minimize the translation error while rescoring this n-best list. With the optimal coefficients, a new decoding is performed so as to produce an updated n-best list (Bertoldi, 2006). This process converges after only 5–10 decodings. For each internal optimization, about 50 iterations are still required, but each iteration is much shorter since they only require to rescoring an n-best list. Most recent approaches propose to substitute the n-best lists to lattices (Macherey *et al.*, 2008).

9 Rescoring and system combination

This section reports an overview of the works, which have been developed in rescoring and system combination. Rescoring aims at integrating feature functions in the log-linear SMT framework, which cannot be easily incorporated into a dynamic programming search.

One straightforward solution to this problem is to use a SMT decoder that computes a list of n-best translations and then rescore the list with the feature functions that were not incorporated in decoding.

In Och *et al.* (2004), more than 450 different feature functions were used in order to improve the syntactic form of the MT output.

A widely known work is by Kumar and Byrne (2004), which presents the Minimum Bayes-Risk (MBR) approach in rescoring. It aims to minimize expected loss of translation errors under loss functions that measure translation performance. This algorithm was originally developed to work with n-best lists of translations, and recently has been extended to lattices and hypergraphs, which have the capacity of encoding many more hypotheses than typical n-best lists (Kumar *et al.*, 2009). DeNero *et al.* (2009) proposed a variant of the MBR procedure that applies efficiently to translation forests. These approaches can be interpreted as consensus decoding procedures because they choose a translation similar to other high posterior translations.

Later, some language modeling techniques have been used to rescore the n-best outputs. For example, clustered language models were implemented in Hasan and Ney (2005). They reported a

https://doi.org/10.1017/S026988891200029X Published online by Cambridge University Press

language model based on clusters obtained by applying regular expressions to the training data and thus discriminating several different sentence types, for example, interrogatives, imperatives or enumerations. The main motivation rested on the observation that different sentence types also have different syntactic structures, and thus yield a varying distribution of *n*-grams reflecting their word order. Also, techniques of neural language modeling (Schwenk *et al*., 2006) have been used for the purpose of rescoring. The basic idea of continuous space language models, also called neural network language models, is to project the word indices onto a continuous space and to use a probability estimator operating on this space. Since the resulting probability functions are smooth functions of the word representation, better generalization to unknown *n*-grams can be expected.

System combination in SMT has its inspiration from the fact that combining outputs from different systems was shown to be quite successful in automatic speech recognition (ASR). Voting schemes like the ROVER approach of Fiscus (1997) use edit distance alignment and time information to create confusion networks from the output of several ASR systems.

In MT, some approaches combine lattices or *n*-best lists from several different MT systems (Frederking & Nirenburg, 1994). To be successful, such approaches require compatible lattices and comparable scores of the (word) hypotheses in the lattices.

The most straightforward approach simply selects, for each sentence, one of the provided hypotheses. The selection is made based on the scores of translation, language and other models (Nomoto, 2004; Doi *et al*., 2005).

Bangalore *et al*. (2001) uses the edit distance alignment extended to multiple sequences to construct a confusion network from several translation hypotheses. This algorithm produces monotone alignments only, that is, allows insertion, deletion and substitution of words. Jayaraman and Lavie (2005) try to deal with translation hypotheses with significantly different word order. They introduce a method that allows non-monotone alignments of words in different translation hypotheses for the same sentence.

Confusion networks have been generated by choosing one hypothesis as the skeleton and other hypotheses are aligned against it. The skeleton defines the word order of the combination output. Previously mentioned MBR was used to choose the skeleton in Sim *et al*. (2007). The average translation edit rate (TER) score (Snover *et al*., 2006) was computed between each system's 1-best hypotheses in terms of TER. This work was extended by Rosti *et al*. (2007) by introducing system weights for word confidences.

More recently, experiments combining several kinds of MT systems have been presented in Matusov *et al*. (2008). They propose to use a consensus translation using an alignment procedure that explicitly models reordering of words in the hypotheses. In contrast to existing approaches, the context of the whole document, rather than a single sentence, is considered in this iterative, unsupervised procedure, yielding a more reliable alignment.

10 Domain adaptation

One of the main limitations of the phrase-based system is the out-domain generalization. This section reviews several techniques, which focus on solving this generalization challenge.

Training data for SMT is generally collected from wherever it is available. The application domain for a MT system has been investigated with and without in-domain available data. Initial adaptation efforts focused on adapting the target language model to the specific domain, such as it was done for speech recognition.

The main inconveniences of performing out-domain translations are the out-of-vocabulary words, the unknown expressions and the syntactic constructions. These challenges have mainly been addressed either by trying to incorporate source context information or, by trying to exploit different kinds of in-domain material to improve SMT. In-domain material can be mainly: parallel and/or monolingual. Additionally, works that focus on solving unknown vocabulary (Langlais & Patry, 2007) and adding morphological information (Nießen & Ney, 2001; Axelrod, 2006; de Gispert & Marino, 2008) are also dealing with out-domain challenges.

https://doi.org/10.1017/S026988891200029X Published online by Cambridge University Press

The exploitation of in-domain monolingual corpora have been mainly limited to the target language model (Eck *et al.*, 2004). Bulyko *et al.* (2007) explored discriminative estimation of language model weights by directly optimizing MT performances. Wu *et al.* (2008) investigated linear and log-linear interpolation of in-domain and out-domain language models. In Schwenk and Estève (2008), the authors present a promising technique of target language models linear interpolation.

More recently, Bertoldi and Federico (2009) have used the in-domain monolingual corpora to synthesize a bilingual corpus by translation of the monolingual adaptation data into the counterpart language. Translation, reordering and language models were estimated after translating in-domain texts with the baseline system.

To sum up, all the above works focus on introducing statistical techniques that can help to correctly translate in-domain expressions and vocabulary.

11 Source context information

The phrase-based translation model allows to introduce both source and target context information in comparison with the word-based translation model. However, the idea of introducing context information is simplified in the phrase-based systems given that all training sentences contribute equally to the final translation.

More complex works which introduce source context information can be found in the SMT literature. For example, Stroppa *et al.* (2007) and Haque *et al.* (2009) incorporate source language context using neighboring words, part-of-speech tags and/or supertags. They use a memory-based classification approach to obtain the probability for the given additional contexts with the source phrase. Works such as Carpuat and Wu (2007) are context-rich approaches from Word Sense Disambiguation methods. Other related works focus on extending the translation and target language model using neural networks (Schwenk *et al.*, 2007), which aim at smoothing both the translation and target language model in order to use the *n*-grams more adequate in the translated sentence.

To sum up, all these works focus on introducing semantic context in a standard phrase-based system.

Exploding in-domain parallel corpora have been recently addressed in the scientific community by the annual WMT international evaluations (Callison-Burch *et al.*, 2007, 2008, 2009).

The straightforward way to use in-domain parallel corpora is to simply concatenate the training corpora available and use the combined data for both translation model and language model training. However, in case the in-domain data is a much smaller set than the out-domain corpus, the gain expected through a simply concatenation is not much. The result can be even worse than using only the in-domain data when the in-domain data have a very particular style (Khalilov *et al.*, 2008).

Another proposal is to implement a TM interpolation strategy. Authors in Koehn and Schroeder (2007) show that the linear interpolation of translation models achieve gains in translation quality. Alternatively, a log-linear interpolation also improves translation (Khalilov *et al.*, 2008).

The technique of mixture models is exploited in works such as Foster and Kuhn (2007), Civera and Juan (2007) and Finch and Sumita (2008). Mixture modeling is a standard technique for density estimation that is capable of learning specific probability distributions that better fit subsets of the training data set.

In Rogati (2009), the author adds in-domain corpora by automatically weighting and combining multiple translation resources, according to several criteria, in order to better match a target corpus or a specific domain sample. The criteria include lexical-level domain match, translation quality estimates, size and taxonomy representation.

12 Evaluation measures

Automatic and human evaluation has been widely investigated by the scientific community. This section reports an overview of the most widely used SMT evaluation measures.

https://doi.org/10.1017/S026988891200029X Published online by Cambridge University Press

Having an automatic evaluation is a must in order to optimize a MT system. Interesting and widely used automatic measures are referenced as follows:

- Bilingual Evaluation Understudy (BLEU; Papineni *et al.*, 2002) has dominated most MT work. Essentially, it consists of an *n*-gram corpus-level measure and it is always referred to as a given *n*-gram order ($BLEU_n$, *n* usually being 4).
- NIST (Doddington, 2002) is an accuracy measure that calculates how informative a particular *n*-gram is, and the rarer a correct *n*-gram is, the more weight it will be given. Small variations in translation length do not impact much in the overall score.
- Word Error Rate (WER; McCowan *et al.*, 2004) is a standard speech recognition evaluation metric. One general difficulty of measuring performance lies in the fact that the translated word sequence can have a different length from the reference word sequence (supposedly the correct one). The WER is derived from the Levenshtein distance, working at the word level. Additionally for translation, its multiple-reference version (mWER) is computed on a sentence-by-sentence basis, so that the final measure for a given corpus is based on the cumulative WER for each sentence. Similar to WER, there is the Position-Independent Error Rate (mPER), which is again computed on a sentence-by-sentence basis. The main difference with WER is that it does not penalize the wrong order in the translation.
- Metric for Evaluation of Translation with Explicit ORdering (METEOR) evaluates a translation by computing a score based on explicit word-to-word matches between the translation and a reference translation. If more than one reference translation is available, the given translation is scored against each reference independently, and the best score is reported. It has been found (Lavie & Agarwal, 2007) that it correlates well with human adequacy.
- From a more intuitive point of view, in Snover *et al.* (2005), Translation Error Rate or TER is presented. This measures the amount of editing that a human would have to perform to change a system output so it exactly matches a reference translation. Its application in real-life situation is reported in Przybocki *et al.* (2006).

Actually, there are many interesting automatic measures, for example, the ones which have been presented and evaluated in the Annual Workshop of Machine Translation (WMT)[6]. Some measures include linguistic knowledge and do correlate with human criteria (Giménez & Màrquez, 2007; Popovic & Ney, 2009).

On the other hand, there are many schemas for human evaluation measures in MT. The most popular method is to obtain ratings from monolingual judges for segments of a document. Judges are presented with a segment, and are asked to rate it for two variables: adequacy and fluency. Adequacy is a rating of how much information is transferred between the original and the translation, and fluency is a rating of how good the English is. This technique is found to cover the relevant parts of the quality evaluation, while at the same time being easier to deploy, as it does not require expert judgment. Measuring systems based on adequacy and fluency, along with informativeness is now the standard methodology for the ARPA evaluation program.

Among the different challenges in human evaluation there are, in the first place, the consistency between the evaluator himself and the other evaluators. An evaluator normally evaluates from 1 to 5 in fluency and adequacy each. It is difficult to maintain consistency in the own evaluations and at the same time, it is even more difficult to keep pace with other evaluators. Generally, this challenge is overcome by calculating an average among the evaluators. And, in the second place, another challenge is to learn from evaluation. Evaluating different types of errors would allow to distinguish system which perform more or less the same.

Another trend is to manually post-edit the references with information from the test hypothesis translations, so that differences between translation and reference account only for errors and the final score is not influenced by the effects of synonymy. Actually, in the DARPA's Global Autonomous Language Exploitation (GALE) program (Olive, 2005), one effective way to evaluate was asking

[6] http://www.statmt.org/wmt[07-10]/

https://doi.org/10.1017/S026988891200029X Published online by Cambridge University Press

evaluators to edit the translation by means of HTER (Human targeted Translation Edit Rate). In that sense, the less number of edits, the better the translation. Authors in Callison-Burch *et al.* (2009) proposed to edit the translation output as fluent as possible, which reflects the annotators' understanding of the sentence.

Other proposals regarding evaluation classification schemas can be found in the literature. Vilar *et al.* (2006), for instance, proposed a 5-category schema that does not use linguistic criteria. The classification presented in the current paper offers more linguistic information about the type of error; for example, Vilar *et al.* (2006) use the concept of *incorrect words* that can be related to multiple linguistic levels: lexical, semantic and morphological. On the other hand, Flanagan classification (Flanagan, 1994) lists a series of errors that are pair language dependent. In the current paper, a similar list of subcategories for Catalan–Spanish is presented; however, these subcategories are included in a 5-category schema, which is language independent. Finally, Popovic (2009) presented a framework for automatic error analysis and categorization. The basic idea is to actually identify erroneous words using algorithms for the calculation of WER and PER. The extracted error details can be used in combination with different types of natural language knowledge (such as base forms, POS tags and others). The work focuses on the five error categories from Vilar *et al.* (2006), and the new measures correlate well with the results of human analysis when using the same categorization. Finally, in Farrús *et al.* (2010), they propose an evaluation method based on the assumption that all the errors can be classified into one of the following linguistic levels: orthographic, morphological, lexical, semantic and syntactic. Human evaluators use specific guidelines to classify and compute the number and type of errors encountered in the translations in order to analyze their linguistic quality.

13 Research directions

This paper overviewed the main phrase-based SMT research works: from the original approach (the Alignment Templates), to the nowadays SMT current research directions, which are mainly depicted in Sections 7–12. Actually, the principle SMT research objectives are to solve the following linguistic challenges:

- *Syntactic*, which includes word ordering: SVO or VSO languages; location modifiers and nouns.
- *Morphological*, which includes word correspondences: keeping number agreement among others.
- *Lexical*, which includes out-of-vocabulary words. The main reasons for out of vocabulary words are the limitation of training data, domain changes and morphology.

The new strategy that tries to face the above challenges altogether is to combine the basic ideas of different SMT approaches: for example, hierarchical and phrase-based or Ngram-based SMT systems. Works such as Lopez (2008) allow for introducing discontinuous phrases in a standard phrase-based system. Crego and Yvon (2009) allows for a translation of discontiguous tuples that is the translation unit of an Ngram-based SMT system.

Research in SMT may develop SMT systems of better quality for professional translators. Research fields such as domain adaptation should be really helpful in the recent initiative of Interactive Machine Translation (IMT), where an SMT system aids the translator by interactively making suggestions for completing the translation. The IMT idea have been developed in the TransType and TransType2 European projects (Barrachina *et al.*, 2009). Out of these projects, a recent online IMT (i.e. Caitra) system is now available[7].

Acknowledgments

The author would like to thank her PhD colleagues Patrik Lambert and Josep Maria Crego for their support and Simon Parsons for motivating this work. The author also wants to thank Barcelona Media Innovation Centre for its permission to publish this paper.

[7] http://www.caitra.org

https://doi.org/10.1017/S026988891200029X Published online by Cambridge University Press

This work has been partially funded by the Spanish Ministry of Economy and Competitiveness through the *Juan de la Cierva* fellowship program.

References

Arnold, D. & Balkan, L. 1995. Machine translation: an introductory guide. *Computational Linguistics* **210**(4), 577–578.

Axelrod, A. E. 2006. *Factored Language Models for Statistical Machine Translation*. Master Thesis, University of Edinburgh.

Bangalore, S., Bordel, G. & Riccardi, G. 2001. Computing consensus translation from multiple machine translation systems, In *IEEE Workshop on Automatic Speech Recognition and Understanding*, Madonna di Campiglio, Italy, 351–354.

Barrachina, S., Bender, O., Casacuberta, F., Civera, J., Cubel, E., Khadivi, S., Lagarda, A., Ney, H., Toms, J. & Vidal, E. 2009. Statistical approaches to computer-assisted translation. *Computational Linguistics* **350**(1), 3–28.

Berger, A., Della Pietra, S. & Della Pietra, V. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics* **220**(1), 39–72.

Bertoldi, N. 2006. *Minimum Error Training (Updates)*. Technical report, Slides of the JHU Summer Workshop.

Bertoldi, N. & Federico, M. 2009. Domain adaptation for statistical machine translation with monolingual resources. In *Proceedings of the 4th Workshop on Statistical Machine Translation*, Athens, Greece, 182–189. Association for Computational Linguistics. http://www.aclweb.org/anthology/W/W09/W09-0432.

Bertoldi, N., Cattoni, R., Cettolo, M., Chen, B. & Federico, M. 2006. ITC-irst at the 2006 TC-STAR SLT evaluation campaign. In *TC-STAR Workshop on Speech-to-Speech Translation*, Barcelona, Spain, 19–24.

Brown, P., Della Pietra, S., Della Pietra, V. & Mercer, R. 1993. The mathematics of statistical machine translation. *Computational Linguistics* **190**(2), 263–311.

Bulyko, I., Matsourkas, S., Schwartz, R., Nguyen, L. & Makhoul, J. 2007. Language model adaptation in machine translation from speech. In *Proceedings of the 32nd International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Honolulu, Hawai'i, 117–120.

Callison-Burch, C., Talbot, D. & Osborne, M. 2004. Statistical machine translation with word- and sentence-aligned parallel corpora. In *Proceedings of the 42th Annual Meeting of the Association for Computational Linguistics*, Barcelona, Spain, 175–182.

Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C. & Schroeder, J. 2007. (Meta-)evaluation of machine translation. In *Proceedings of the 2nd Workshop on Statistical Machine Translation*, Prague, Czech Republic. Association for Computational Linguistics, 136–158. http://www.aclweb.org/anthology/W/W07/W07-0218.

Callison-Burch, C., Koehn, P., Monz, C. & Schroeder, J. 2008. Further meta-evaluation of machine translation. In *Proceedings of the 3rd Workshop on Statistical Machine Translation*, Columbus, OH. Association for Computational Linguistics, 70–106. http://www.aclweb.org/anthology/W/W08/W08-0309.

Callison-Burch, C., Koehn, P., Monz, C. & Schroeder, J. 2009. Findings of the 2009 Workshop on Statistical Machine Translation. In *Proceedings of the 4th Workshop on Statistical Machine Translation*, Athens, Greece. Association for Computational Linguistics, 1–28. http://www.aclweb.org/anthology/W/W09/W09-0x01.

Carpuat, M. & Wu, D. 2007. Improving statistical machine translation using word sense disambiguation. In *Empirical Methods in Natural Language Processing (EMNLP)*, Prague, 61–72.

Chen, S. F. & Goodman, J. T. 1998. *An Empirical Study of Smoothing Techniques for Language Modeling*. Technical report, Harvard University.

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

Civera, J. & Juan, A. 2007. Domain adaptation in statistical machine translation with mixture modelling. In *Proceedings of the 2nd Workshop on Statistical Machine Translation*, Prague, Czech Republic, 177–180.

Collins, M., Koehn, P. & Kucerová, I. 2005. Clause restructuring for statistical machine translation. In *Proceedings of the 43th Annual Meeting of the Association for Computational Linguistics*, Michigan, 531–540.

Costa-jussà, M. R. & Fonollosa, J. A. R. 2009a. An Ngram-based reordering model. *Computer Speech & Language* **230**(3), 362–375.

Costa-jussà, M. R. & Fonollosa, J. A. R. 2009b. State-of-the-art word reordering approaches in statistical machine translation. *IEICE Transactions on Information and Systems* **920**(11), 2179–2185.

Costa-jussà, M. R., Crego, J. M., de Gispert, A., Lambert, P., Khalilov, M., Mariño, J. B., Fonollosa, J. A. R. & Banchs, R. 2006. TALP phrase-based statistical translation system for European language pairs. In *Human Language Technology Conference (HLT-NAACL'06): Proceedings of the Workshop on Statistical Machine Translation*, New York City, 142–145.

Costa-jussà, M. R., Fonollosa, J. A. R. & Monte, E. 2011. Recursive alignment block classification technique for word reordering in statistical machine translation. *Language Resources and Evaluation Journal* **450**(2), 165–179.

https://doi.org/10.1017/S026988891200029X Published online by Cambridge University Press

Crego, J. M. 2008. *Architecture and Modeling for N-gram-based Statistical Machine Translation*. PhD thesis, Department of Signal Theory and Communications, Universitat Politècnica de Catalunya (UPC).

Crego, J. M. & Mariño, J. B. 2007. Improving SMT by coupling reordering and decoding. *Machine Translation* **200**(3), 199–215.

Crego, J. M. & Yvon, F. 2009. Gappy translation units under left-to-right SMT decoding. In *Proceedings of the 13th Annual Conference of the European Association for Machine Translation (EAMT'09)*, Barcelona.

Crego, J. M., de Gispert, A., Lambert, P., Costa-jussà, M. R., Khalilov, M., Banchs, R., Mariño, J. B. & Fonollosa, J. A. R. 2006. N-gram-based SMT system enhanced with reordering patterns. In *Human Language Technology Conference (HLT-NAACL'06): Proceedings of the Workshop on Statistical Machine Translation*, New York City, 162–165.

de Gispert, A. & Marino, J. B. 2008. On the impact of morphology in English to Spanish statistical MT. *Speech Communication* **50**, 1034–1046.

DeNero, J., Chiang, D. & Knight, K. 2009. Fast consensus decoding over translation forests. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, Singapore, 567–575.

Doddington, G. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the Human Language Technology Conference, HLT-NAACL'02*, San Diego, 138–145.

Doi, T., Hwang, Y., Imamura, K., Okuma, H. & Sumita, E. 2005. Nobody is perfect: ATR's hybrid approach to spoken language translation. In *Proceedings of the International Workshop on Spoken Language Translation, IWSLT'04*, Pittsburgh, PA, USA, 55–62.

Dorr, B. J. 1994. Machine translation: a view from the lexicon. *Computational Linguistics* **200**(4), 670–676.

Eck, M., Vogel, S. & Waibel, A. 2004. Language model adaptation for statistical machine translation based on information retrieval. In *Proceedings of the LREC*, Lisbon, Portugal, 327–330.

Farrús, M., Costa-jussà, M. R., Mariño, J. B. & Fonollosa, J. A. R. 2010. Linguistic-based evaluation criteria to identify statistical machine translation errors. In *Proceedings of the 14th Annual Meeting of the EAMT: European Association for Machine Translation*, Saint Rapahel.

Finch, A. & Sumita, E. 2008. Dynamic model interpolation for statistical machine translation. In *Proceedings of the 3rd Workshop on Statistical Machine Translation*, Colombus, USA, 208–215.

Fiscus, G. 1997. A post-processing system to yield reduced word error rates: recognizer output voting error reduction (ROVER). In *IEEE Workshop on Automatic Speech Recognition and Understanding*, Santa Barbara, USA.

Flanagan, M. A. 1994. Error classification for MT evaluation. In *Proceedings of the AMTA*, Columbia, 65–72.

Foster, G. & Kuhn, R. 2007. Mixture-model adaptation for SMT. In *Proceedings of the 2nd Workshop on Statistical Machine Translation*, Prague, Czech Republic, 128–135.

Fraser, A. & Marcu, D. 2006. *Measuring Word Alignment Quality for Statistical Machine Translation*. Technical report, ISI/University of Southern California, California.

Frederking, R. & Nirenburg, S. 1994. Three heads are better than one. In *Proceedings of the 4th Conference on Applied Natural Language Processing*, Stuttgart, Germany.

Giménez, J. & Màrquez, L. 2007. Linguistic features for automatic evaluation of heterogenous MT systems. In *Proceedings of the 2nd Workshop on Statistical Machine Translation*, Prague, 256–264.

Haque, R., Kumar Naskar, S., Ma, Y. & Way, A. 2009. Using supertags as source language context in SMT. In *Proceedings of the 13th Annual Conference of the European Association for Machine Translation (EAMT)*, Barcelona, 234–241.

Hasan, S. & Ney, H. 2005. Clustered language models based on regular expressions for statistical machine translation. In *Proceedings of the 10th Annual Conference of The European Association for Machine Translation (EAMT)*, Budapest, Hungary, 119–125.

Hassan, H., Hearne, M., Way, A. & Sima'an, K. 2006. Syntactic phrase-based statistical machine translation. In *Proceedings of the 1st IEEE/ACL Workshop on Spoken Language Technology*, Aruba.

Jayaraman, S. & Lavie, A. 2005. Multi-engline machine translation guided by explicit word matching. In *Proceedings of the 10th Conference of the European Association for Machine Translation*, Budapest, Hungary, 143–152.

Kanthak, S., Vilar, D., Matusov, E., Zens, R. & Ney, H. 2005. Novel reordering approaches in phrase-based statistical machine translation. In *Annual Meeting of the Association for Computational Linguistics: Proceedings of the ACL Workshop on Building and Using Parallel Texts: Data-Driven Machine Translation and Beyond (WMT)*, Ann Arbor, MI, 167–174.

Khalilov, M., Costa-jussà, M. R., Henríquez, C. A., Fonollosa, J. A. R., Hernández, A., Mariño, J. B., Banchs, R. E., Chen, B., Zhang, M., Aw, A. & Li, H. 2008. The TALP & I2R SMT systems for IWSLT 2008. In *Proceedings of the International Workshop on Spoken Language Translation*, Hawaii, USA, 116–123.

Khalilov, M., Fonollosa, J. A. R. & Dras, M. 2009. A new subtree-transfer approach to syntax-based reordering for statistical machine translation. In *Proceedings of the 13th Annual Conference of the European Association for Machine Translation (EAMT'09)*, Barcelona, Spain, 198–204.

https://doi.org/10.1017/S026988891200029X Published online by Cambridge University Press

Knight, K. & Al-Onaizan, Y. 1998. Translation with finite-state devices. In *Proceedings of the 4th Conference of the Association for Machine Translation in the Americas, AMTA'02*, Langhorne, 421–437.

Koehn, K. & Knight, K. 2003. Empirical methods for compound splitting. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics*, Budapest, Hungary, 347–354.

Koehn, P. & Schroeder, J. 2007. Experiments in domain adaptation for statistical machine translation. In *Annual Meeting of the Association for Computational Linguistics: Proceedings of the 2nd Workshop on Statistical Machine Translation (WMT)*, Prague, 224–227.

Koehn, P., Och, F. J. & Marcu, D. 2003. Statistical phrase-based translation. In *Proceedings of the Human Language Technology Conference, HLT-NAACL'03*, Edmonton, Canada, 48–54.

Koehn, P., Amittai, A., Birch, A., Callison-Burch, C., Osborne, M., Talbot, D. & White, M. 2005. Edinburgh system description for the 2005 IWSLT speech translation evaluation. In *Proceedings of International Workshop on Spoken Languages Translation*, Pittsburgh.

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 *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, Prague, Czech Republic, 177–180.

Kumar, S. & Byrne, W. 2004. Minimum Bayes-risk decoding for statistical machine translation. In *Proceeding of the Human Language Technology Conference, HLT-NAACL'04*, Boston, MA, USA, 169–176.

Kumar, S., Macherey, W., Dyer, C. & Och, F. 2009. Efficient minimum error rate training and minimum Bayes-risk decoding for translation hypergraphs and lattices. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, Singapore, 163–171.

Lambert, P. 2008. *Exploiting Lexical Information and Discriminative Alignment Training in Statistical Machine Translation*. PhD thesis, Software Department, Universitat Politècnica de Catalunya (UPC).

Lambert, P. & Banchs, R. E. 2006. Tuning machine translation parameters with SPSA. In *Proceedings of the International Workshop on Spoken Language Translation*, Kyoto, Japan, 190–196.

Langlais, P. & Gotti, F. 2006. Phrase-based SMT with shallow tree-phrases. In *Proceedings of the Workshop on Statistical Machine Translation*, New York, USA, 39–46.

Langlais, P. & Patry, A. 2007. Translating unknown words by analogical learning. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, Prague, Czech Republic. Association for Computational Linguistics, 877–886. http://www.aclweb.org/anthology/D/D07/D07-1092.

Lavie, A. & Agarwal, A. 2007. METEOR: an automatic metric for MT evaluation with high levels of correlation with human judgments. In *Annual Meeting of the Association for Computational Linguistics: Proceedings of the 2nd Workshop on Statistical Machine Translation (WMT)*, Prague, Czech Republic, 228–231.

Lopez, A. 2007. *A Survey of Statistical Machine Translation*. Storming Media.

Lopez, A. 2008. *Machine Translation by Pattern Matching*. PhD thesis, University of Maryland.

Macherey, W., Och, F., Thayer, I. & Uszkoreit, J. 2008. Lattice-based minimum error rate training for statistical machine translation. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, Hawaii, 725–734.

Mariño, J. B., Banchs, R. E., Crego, J. M., de Gispert, A., Lambert, P., Fonollosa, J. A. R. & Costa-jussà, M. R. 2006. N-gram based machine translation. *Computational Linguistics* **320**(4), 527–549.

Matusov, E., Zens, R., Vilar, D., Mauser, A., Popovic, M., Hasan, S. & Ney, H. 2006. The RWTH machine translation system. In *TC-STAR Workshop on Speech-to-Speech Translation*, Barcelona, Spain, 31–36.

Matusov, E., Leusch, G., Banchs, R. E., Bertoldi, N., Dechelotte, D., Federico, M., Kolss, M., Lee, Y., Marino, J. B., Paulik, M., Roukos, S., Schwenk, H. & Ney, H. 2008. System combination for machine translation of spoken and written language. *IEEE Transactions on Audio, Speech and Language Processing* **160**(7), 1222–1237.

Mauser, A., Matusov, E. & Ney, H. 2006. Training a statistical machine translation system without GIZA++. In *Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC'06*, Genova, 715–720.

McCowan, I., Moore, D., Dines, J., Gatica-Perez, D., Flynn, M., Wellner, P. & Bourlard, H. 2004. On the use of information retrieval measures for speech recognition evaluation. In *Proceedings of the IDIAP-RR 73*, Martigny, Switzerland. IDIAP.

Menezes, A., Toutanova, K. & Quirk, C. 2006. Microsoft research treelet translation system: NAACL 2006 Europarl evaluation. In *Proceedings on the Workshop on Statistical Machine Translation*, New York City. Association for Computational Linguistics, 158–161.

Nelder, J. A. & Mead, R. 1965. A simplex method for function minimization. *The Computer Journal* **7**, 308–313.

Nießen, S. & Ney, H. 2001. Morpho-syntactic analysis for reordering in statistical machine translation. In *Proceedings of the MT-Summit VII*, Santiago de Compostela, Spain, 247–252.

https://doi.org/10.1017/S026988891200029X Published online by Cambridge University Press

Nomoto, T. 2004. Multi-engine machine translation with voted language model. In *Proceedings of the 42th Annual Meeting of the Association for Computational Linguistics*, Barcelona, Spain, 494–501.

Och, F. J. 1999. An efficient method for determining bilingual word classes. In *Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics*, Bergen, Norway, 71–76.

Och, F. J. 2002. *Statistical Machine Translation: From Single-Word Models to Alignment Templates*. PhD thesis, RWTH Aachen University, Aachen, Germany.

Och, F. J. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41th Annual Meeting of the Association for Computational Linguistics*, Sapporo, 160–167.

Och, F. J. & Ney, H. 2000. A comparison of alignment models for statistical machine translation. In *Proceedings of the 18th Conference on Computational Linguistics*, Morristown, NJ, USA, 1086–1090.

Och, F. J. & Ney, H. 2004. The alignment template approach to statistical machine translation. *Computational Linguistics* **300**(4), 417–449.

Och, F.-J., Gildea, D., Khudanpur, S., Sarkar, A., Yamada, K., Fraser, A., Kumar, S., Shen, L., Smith, D., Eng, K., Jain, V., Jin, Z. & Radev, D. 2004. A smorgasbord of features for statistical machine translation. In *Proceedings of the Human Language Technology Conference, HLT-NAACL'04*, 161–168.

Olive, J. 2005. Global autonomous language exploitation. *DARPA/IPTOProposer Information Pamphlet*.

Papineni, K., Roukos, S., Ward, T. & Zhu, W-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, PA, 311–318.

Popovic, M. 2009. *Machine Translation:Statistical Approach with Additional Linguistic Knowledge*. PhD thesis, RWTH University.

Popovic, M. & Ney, H. 2006. POS-based word reorderings for statistical machine translation. In *Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC'06*, Genoa, Genoa, Italy, 1278–1283.

Popovic, M. & Ney, H. 2009. Syntax-oriented evaluation measures for machine translation output. In *Proceedings of the 4th Workshop on Statistical Machine Translation*, Athens, 29–32.

Przybocki, M., Sanders, G. & Le, A. 2006. Edit distance: a metric for machine translation evaluation. In *Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC'06*, Genoa, Italy, 2038–2043.

Quirk, C., Menezes, A. & Cherry, C. 2005. Dependency treelet translation: syntactically informed phrasal SMT. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, Ann Arbor, MI, 271–279.

Rogati, M. 2009. *Domain Adaptation of Translation Models for Multilingual Applications*. PhD thesis, Carnegie Mellon University.

Rosti, A.-V.I., Ayan, N. F., Xiang, S. B., Schwartz Matsoukas, R. & Dorr, B. J. 2007. Combining outputs from multiple machine translation systems. In *Proceedings of the Human Language Technology Conference, HLT-NAACL'07*, Rocherster, USA, 228–235.

Schwenk, H. & Estève, Y. 2008. Data selection and smoothing in an open-source system for the 2008 NIST machine translation evaluation. In *Proceedings of the Interspeech'08*, Brisbane, Australia.

Schwenk, H., Costa-jussà, M. R. & Fonollosa, J. A. R. 2006. Continuous Space Language Models for the IWSLT 2006 Task. In *Proceedings of the International Workshop on Spoken Language Translation*, Kyoto, Japan, 166–173.

Schwenk, H., Costa-jussa, M. R. & Fonollosa, J. A. R. 2007. Smooth bilingual n-gram translation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, Prague, Czech Republic. Association for Computational Linguistics, 430–438. `http://www.aclweb.org/anthology/D/D07/D07-1045`.

Sim, K. C., Byrne, W. J., Gales, M. J. F., Sahbi, H. & Woodland, P. C. 2007. Consensus network decoding for statistical machine translation system combination. In *Proceedings of the ICASSP*, **4**, Rocherster, USA, 105–108.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L. & Makhoul, J. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the Association for Machine Translation in the Americas*, Sydney, Australia.

Snover, M., Dorr, B., Schwartz, R., Makhoul, J., Micciula, L. & Weischedel, R. 2005. *A Study of Translation Error Rate with Targeted Human Annotation*. Technical report LAMP-TR-126,CS-TR-4755,UMIACS-TR-2005-58, University of Maryland, College Park and BBN Technologies.

Stroppa, N., van de Bosch, A. & Way, A. 2007. Exploiting source similarity for SMT using context-informed features. In *Proceedings of the 11th Conference on Theoretical and Methodological Issues in Machine TRanslation (TMI)*, Skövde, 231–240.

Tillmann, C. 2004. A unigram orientation model for statistical machine translation. In *Proceedings of the Human Language Technology Conference, HLT-NAACL'04*, Boston, 101–104.

Tillmann, C. & Ney, H. 2003. Word reordering and a dynamic programming beam search algorithm for statistical machine translation. *Computational Linguistics* **290**(1), 97–133.

https://doi.org/10.1017/S026988891200029X Published online by Cambridge University Press

Vilar, D., Stein, D., Zhang, Y., Matusov, E., Mauser, A., Bender, O., Mansour, S. & Ney, H. 2008. The RWTH machine translation system for IWSLT 2008. In *Proceedings of the International Workshop on Spoken Language Translation*, Waikiki, Hawaii, 108–115.

Vilar, D., Xu, J., Fernando-D'Haro, L. & Ney, H. 2006. Error analysis of statistical machine translation output. In *Proceedings of the LREC*, Genoa, Italy.

Wang, C., Collins, M. & Koehn, P. 2007. Chinese syntactic reordering for statistical machine translation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, Prague, 737–745.

Wu, D. 1996. A polynomial-time algorithm for statistical machine translation. In *Annual Meeting of the Association for Computational Linguistics*, Santa Cruz.

Wu, H., Wang, H. & Zong, C. 2008. Domain adaptation for statistical machine translation with domain dictionary and monolingual corpora. In *Proceedings of the 22nd International Conference on Computational Linguistics*, Beijing, China, **1**, 993–1000.

Xia, F. & McCord, M. 2004. Improving a statistical mt system with automatically learned rewrite patterns. In *Proceedings of the 20th International Conference on Computational Linguistics*, Morristown, 508.

Yamada, K. & Knight, K. 2002. A decoder for syntax-based statistical MT. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 303–310.

Zens, R., Och, F. J. & Ney, H. 2002. Phrase-based statistical machine translation. In *KI-2002: Advances in Artificial Intelligence*, Jarke, M., Koehler, J. & Lakemeyer, G. (eds), Lecture Notes in Artificial Intelligence **2479**, 18–32. Springer Verlag.

Zhang, Y., Zens, R. & Ney, H. 2007. Chunk-level reordering of source language sentences with automatically learned rules for statistical machine translation. In *Proceedings of the Human Language Technology Conference (HLT-NAACL'06): Proceedings of the Workshop on Syntax and Structure in Statistical Translation (SSST)*, Rochester, 1–8.

https://doi.org/10.1017/S026988891200029X Published online by Cambridge University Press