Chapter 43
Statistical Machine Translation

Lucia Specia
University of Sheffield, UK

ABSTRACT

Statistical Machine Translation (SMT) is an approach to automatic text translation based on the use of statistical models and examples of translations. SMT is the current dominant research paradigm for machine translation and has been attracting significant commercial interest in recent years. In this chapter, the authors introduce the rationale behind SMT, describe the currently leading approach (phrase-based SMT), and present a number of emerging approaches (tree-based SMT, discriminative SMT). They also present popular metrics to evaluate the performance of SMT systems and discuss promising research directions in the field.

1. INTRODUCTION

Statistical Machine Translation (SMT) is an approach to automatically translate text based on the use of statistical models and examples of translations. Although Machine Translation (MT) systems developed according to rule-based approaches are still in use, SMT is the dominant research paradigm today and has recently been garnering significant commercial interest. The core of SMT research has developed over the last two decades, after the seminal paper by Brown et al. (1990). The field has progressed considerably since then, moving from word-to-word translation towards phrase-to-phrase translation and other more sophisticated models that take sentence structure into account. A trend observed in recent years is the shift from using pure statistical information extracted from large quantities of data to incorporating linguistic information about the source and/or the target language.

The idea of SMT is related to the late 1940's view of the translation task as a cryptography problem where a decoding process is needed to translate from a foreign "code" into the English language (Hutchins, 1997). This is the basis for the fundamental approach to SMT proposed in the early 1990s through the application of the *Noisy Channel Model* (Shannon, 1949) from the field of Information Theory. This model had proved to be successful in the area of Speech Recognition and was thus adapted to MT.

The use of the Noisy Channel Model for translation assumes that the original text has been accidentally scrambled or encrypted (using a different alphabet, for example) and the goal

DOI: 10.4018/978-1-4666-6042-7.ch043

is to find out the original text by "decoding" the encrypted/scrambled version, as depicted in Figure 1. According to this model, the message *I* is the input to the channel (text in a native language). *I* gets encrypted into *O* (text in a foreign language) using a certain coding scheme. The goal is to find a decoder that can reconstruct the input message as faithfully as possible into *I**.

In a probabilistic framework, finding *I**, i.e., the closest possible text to *I*, can be stated as finding the argument that maximizes the probability of recovering the original input given the noisy text, i.e.:

$$\underset{\text{noisy-free text}}{\operatorname{argmax}} \; P(\text{noise-free text} \mid \text{noisy text})$$

This problem is commonly exemplified as the task of translating from a *foreign* language sentence **f** into an *English* sentence **e**. Given **f**, we seek the translation **e** that maximizes $P(\mathbf{e}|\mathbf{f})$, i.e., the most likely translation:

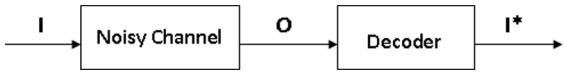
$$\underset{\mathbf{e}}{\operatorname{argmax}} \, P(\mathbf{e} \mid \mathbf{f})$$

This problem can be decomposed in smaller and simpler problems applying the Bayes Theorem:

$$\underset{\mathbf{e}}{\operatorname{argmax}} \; P(\mathbf{e} \mid \mathbf{f}) = \underset{\mathbf{e}}{\operatorname{argmax}} \frac{P(\mathbf{f} \mid \mathbf{e})P(\mathbf{e})}{P(\mathbf{f})}$$

Since the source text **f**, i.e., the input for the translation task is constant across all possible translations **e**, $P(\mathbf{f})$ can be disregarded:

Figure 1. The noisy channel model



$$\underset{\mathbf{e}}{\operatorname{argmax}} \; P(\mathbf{e} \mid \mathbf{f}) = \underset{\mathbf{e}}{\operatorname{argmax}} \, P(\mathbf{f} \mid \mathbf{e})P(\mathbf{e})$$

The process of decomposing the problem of translation into smaller problems and modeling each step individually is motivated by the fact that more reliable statistics can be collected for the smaller problems. The modeling of each smaller problem with probability distributions followed by their combination to find a model that best explains the data complies with a type of statistical learning called *generative modeling*. The generative models resulting from the decomposition of $P(\mathbf{e}|\mathbf{f})$ correspond to two of the fundamental components of a basic SMT system: the *translation model* $P(\mathbf{f}|\mathbf{e})$ and the *language model* $P(\mathbf{e})$. The translation model is used to search for the best translation given an input text. While mathematically it represents the inverse translation probability, i.e., the probability of the source text given the target text, in practice the initial direction, i.e., the probability of the target text given the source text, can be estimated in the very same way using the same data, as we will discuss later. In most SMT systems, both probability directions are used. The decomposition is relevant mostly to isolate the language model component from the translation model. The language model searches for the best translation regardless of the input text.

The third fundamental component of an SMT system, the *decoder*, is a module that performs the search for the best translation **e** given the space of all possible translations (or a subset of it) based on the probability estimates $P(\mathbf{e})$ and $P(\mathbf{f}|\mathbf{e})$. These are major components of the basic word-to-word SMT approaches proposed in the early 1990s. The state-of-the art SMT approaches, however, use extra components to provide additional information to translate phrases or hierarchical structures, and discriminative methods to combine such models by weighting them according to their relevance to discriminate between good and bad translations.