# Scenarios for Customizing an SMT Engine Based on Availability of Data

**Kirti Vashee**
Asia Online
18 Soi Sukhumvit 41 (Pirom), Sukhumvit Road
Klongton-Nua, Wattana, Bangkok 10110, Thailand
kirti.vashee@asiaonline.net

**Rustin Gibbs**
Moravia IT, Inc.
199 E Thousand Oaks Blvd., Ste 203
Thousand Oaks, CA 91360
rusting@moraviaworldwide.com

## Abstract

Although still in a nascent state as a professional translation tool, customized SMT engines already have multiple applications, each of which require clear definitions about quality and productivity. Three engine-training scenarios have emerged which are representative of real-world applications for the development and use of a customized SMT engines based on the availability of data. In the case that limited or no bilingual training data is available, a unique development process can be used to harvest and translate n-grams directly. Using this approach Asia Online and Moravia IT have successfully customized SMT engines for use in various domains. A partnership between an MT engine provider and a qualified LSP is essential to deliver quality results using this approach.

## 1 Introduction

SMT engine customization is the process of training an engine with domain-specific terminology and data to narrow the range of possible candidate translations used during the translation process. Narrowing the scope of translation to very specific matching patterns has the effect of greatly increasing the quality of the translation and, thus, greatly improves the productivity of the translation process. Using this approach, Asia Online and Moravia IT have been able to increase quality while decreasing translation errors and turnaround time.

Traditionally, the SMT engine customization process has been dependent on the use of large corpora of aligned bilingual data from dictionaries, glossaries, translation memories, and aligned bilingual documents. In practice, however, this data is not readily available in most languages or domains. In order to offer the benefits of SMT to customers with large volumes of monolingual data without corresponding bilingual data, Asia Online and Moravia IT developed a unique process to harvest and translate n-grams directly. This approach has proven successful in real-world scenarios.

## 2 SMT Engine Customization Process

The process to customize and train an SMT engine will vary depending to the scenario. Applications can be grouped into scenarios based on a number of factors, including the amount of training material available, the ratio of monolingual training material to bilingual training material, the complexity of the terminology, the expected volume of translated output, and the desired level of quality of the translated output. The traditional process to customize and train and SMT engine is based on an optimal scenario where a large corpora of bilingual data is available and the quality expectation is moderately low.

### 2.1 Traditional Process

The traditional process to customize and train an SMT engine has four phases: Harvesting Data, MT Engine Customization, Analysis and Evaluation, and Post-editing translations. The process assumes

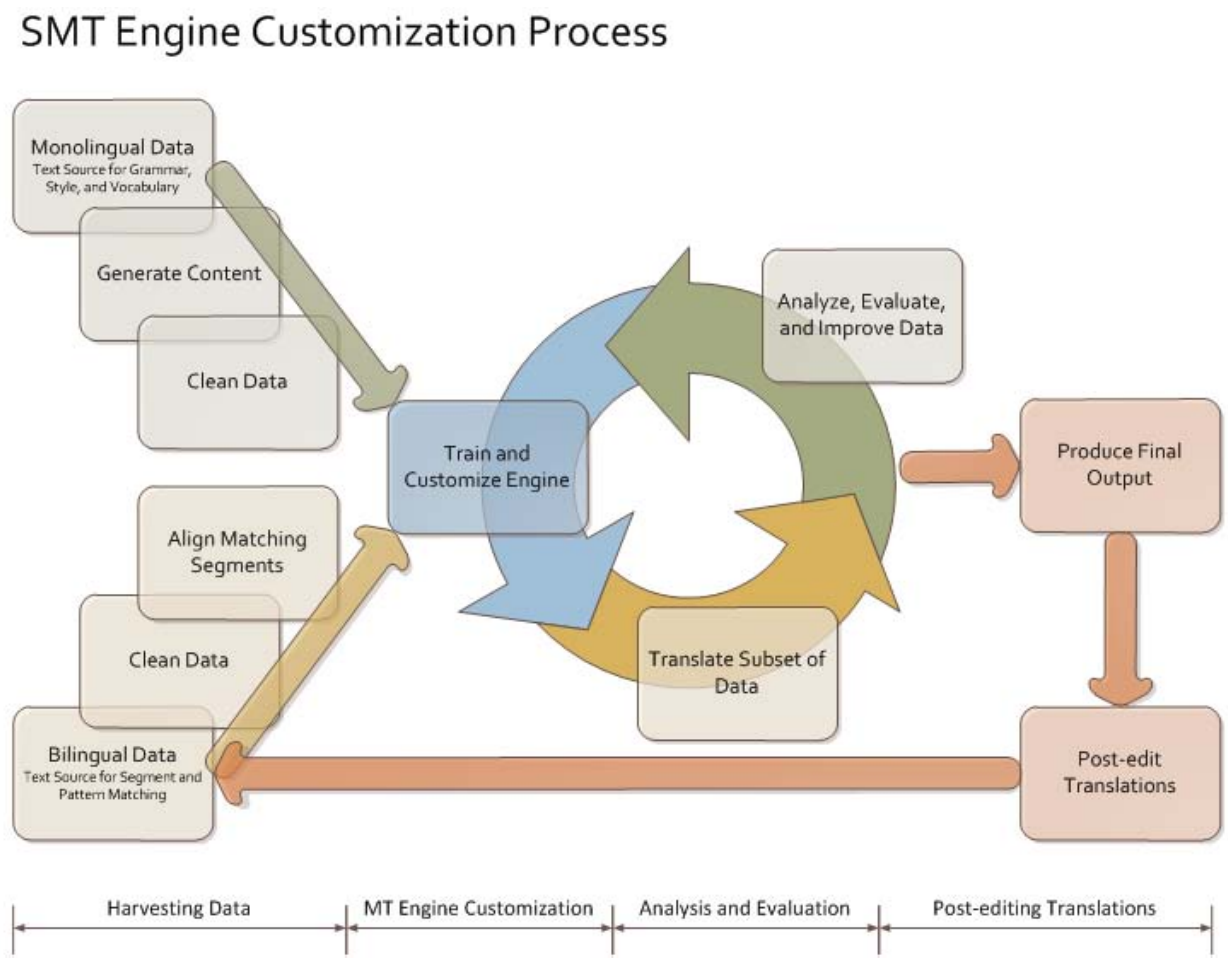


Figure 1: Traditional Process to Customize an SMT Engine

a large corpus of aligned bilingual training material exists. The process, illustrated in Figure 1, has the unique feature of allowing incremental training by feeding post-edited translations back into the engine as approved terminology for new iterations of training. The chief benefit of this approach is fine-grain control over the quality. The limitation of this approach, however, is the dependency on the availability of the training data.

### 2.2 n-gram Process

It is often the case that aligned bilingual data does not exist for training an SMT engine. In this case, the data must be created using direct manipulation of n-grams. The n-gram process is a modified version of the traditional customization process. Many of the steps are the same. The key difference is the collection and treatment of the n-grams. N-grams are the statistically weighted terminology sequences gleaned from the training material and are used as the building blocks of the SMT output. In the absence of aligned bilingual n-grams, n-grams can be translated in isolation and inserted back into the process, as illustrated in Figure 2.

### 2.3 Gleaning n-grams from Source Material

In the traditional process, translated n-grams are gleaned from bilingual data translated into the target language. In cases where this data does not exist, n-grams can still be created by harvesting the n-grams from the training data in the source language. By creating a list of significant, high frequency terminology in the source language, it is possible to filter, select, and translate the n-grams that will have the greatest impact on the output. As not all n-gram combinations are useful according to the selection criteria, this process requires qualified translation partners who understand the n-gram process sufficiently to translate the n-gram list with accuracy and efficiency.

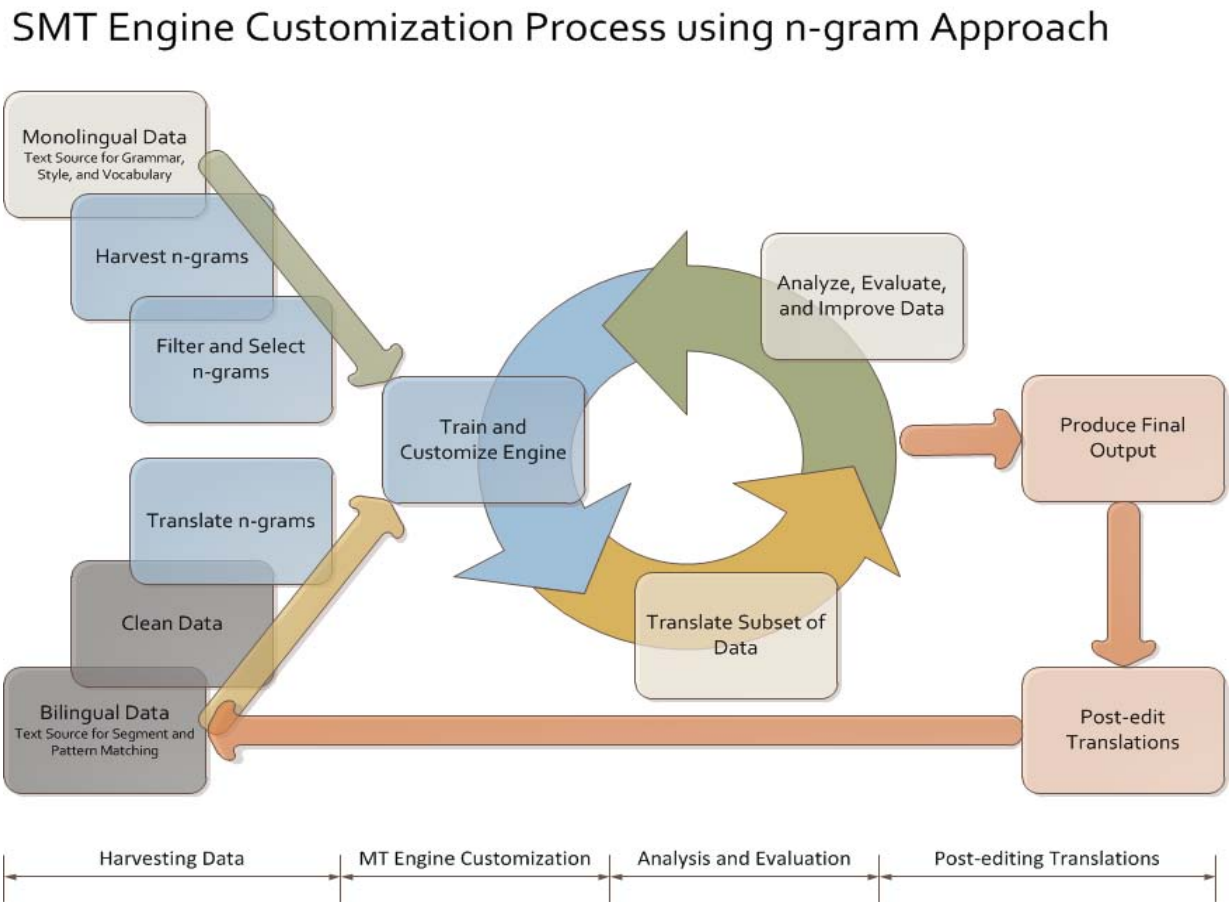


Figure 2: Modified Approach Using n-grams

## 3 Scenarios Based on Availability of Data

The n-gram approach to SMT engine customization opens the way to new applications for SMT as scenarios with limited or no data availability are excluded from the traditional SMT customization and training process.

In our experience, three scenarios based on availability of data have emerged as representative examples of the application of SMT, each of which require a slightly different development process to accommodate the different requirements and parameters. As previously mentioned, the scenarios can be categorized according to the amount of training data, the complexity of the terminology, and the desired quality of the output. Based on availability of training data, the three categories are Extensive, Moderate, and Limited. Asia Online and Moravia IT have successfully customized, trained, and implemented SMT engines in each of these scenarios. Some real-world applications are provided as examples.

### 3.1 Extensive

The traditional scenario is also described as Extensive in the sense that bilingual training data is broadly available. In this scenario, terminology is usually complex and well-established. Volume of both input and output is high, but expected translation quality is moderate to low.

As a leading provider of information and business solutions to professionals in a variety of industries, including legal, risk solutions, corporate, government, law enforcement, accounting and academic, LexisNexis is a good example of a client that utilizes SMT in the Extensive data scenario. LexisNexis had a need to provide discoverability and search ability in Japanese for a large library of English-language patent documents. The requirements are summarized as follows:

- Large amount of training data available
- Highly complex and extensive terminology
- Large volume of new data to be translated
- Moderate quality expectation

Compared to free RbMT alternatives, Asia Online implemented the best patent-focused SMT solution in Japan for LexisNexis, making it possible to translate millions of words a day using material that was previously impossible to translate. The custom solution for LexixNexis provides superior quality due to extensive terminology research and management in the training of the engine.

### 3.2 Moderate

The moderate scenario is representative of a typical localization project in the IT domain. The volume of training material is moderate. The extent and complexity of the terminology is moderate to low, but new terminology is frequently coined in this domain. The expected translation output is moderate to high, so a post-editing step is required following SMT to raise the quality to the level of human translation. The requirements for the moderate scenario are:

- Moderate amount of training data available
- Blended data from multiple sources
- Same content into a few languages
- n-gram translation as needed based on gap analysis of a diagnostic engine
- Intensive and comprehensive post-editing

Moravia IT successfully implemented English-to-Polish and English-to-Swedish SMT engines based on this approach. Upon deployment of the Polish engine, Moravia saw an immediate increase in productivity of 15%. As this is a general purpose engine for the IT domain, global optimization issues are a factor when trying to build a single engine to service multiple clients. The n-gram process is not critical as in the next scenario, but n-gram selection and translation is still important to cover potential gaps in the input bilingual training data and to improve engine performance prior to production post-editing.

### 3.3 Limited

The limited scenario is challenging to implement in the traditional process and requires a unique approach to manipulate n-grams directly to make up for the lack sufficient training data. This application works best when terminology is narrow, focused, and well-established. Translation volume can be moderate to high. Translation quality can be moderate to high provided there is an essential post-editing step to ensure quality.

- Small amount of training data available
- Training data created via n-gram process
- Same content into many languages
- Large number of named entities
- High quality expectation
- Intensive post-editing phase

Asia Online successfully implemented this application with a major, online travel site for 20% of the cost of a full human process due to savings in start-up costs. This approach requires a partnership with a qualified LSP with the technical skills to understand and implement the n-gram translation process.

## 4 MT Partnerships with LSPs

In partnership, Asia Online and Moravia IT have successfully implemented several SMT applications according to the three scenarios based on availability of data. Each scenario is unique and requires a different process to configure and implement the SMT engine successfully. In the case of the moderate and limited data scenarios, a post-editing translation step is essential to meeting expected levels of quality to satisfy project requirements.

As SMT technology continues to mature, SMT applications will multiply and become broadly available in the marketplace. Successful SMT projects will require an incrementally customizable engine configured and deployed according to the scenarios described herein. In addition to versatile and powerful SMT technology, successful SMT projects will require qualified linguistic partners to perform post-editing to ensure that SMT output meets expectations.

Having successfully implemented the n-gram translation approach for SMT, we are optimistic that new SMT applications will come online that previously were inaccessible due to the limitations of the traditional SMT customization process.

## Acknowledgments

David Filip, Ph.D., Director, Moravia Research, Moravia IT a.s.