2 Machine Learning for MT Preliminaries

First, before talking about any specific models, this chapter describes the overall framework of models that use data to *learn* how to perform MT. There are a number of (non-mutually-exclusive) concepts that fall under this overall umbrella, including **example-based machine translation** (EBMT; [3]), **statistical machine translation** (SMT; [2]), and **neural machine translation** (NMT; [1]) more formally.

First, we define our task of machine translation as translating a source sentence $F = f_1, \ldots, f_J = f_1^{|F|}$ into a target sentence $E = e_1, \ldots, e_I = e_1^{|E|}$.[1] Thus, any type of translation system can be defined as a function

$$\hat{E} = \text{mt}(F), \tag{1}$$

which returns a translation hypothesis $\hat{E}$ given a source sentence F as input.

Statistical machine translation systems are systems that perform translation by creating a probabilistic model for the probability of E given F, $P(E \mid F; \theta)$, and finding the target sentence that maximizes this probability:

$$\hat{E} = \underset{E}{\text{argmax}}\ P(E \mid F; \theta), \tag{2}$$

where θ are the parameters of the model specifying the probability distribution. The parameters θ are learned from data consisting of aligned sentences in the source and target languages, which are called **parallel corpora** in technical terminology.[2] Within this framework, there are three major problems that we need to handle appropriately in order to create a good translation system:

Modeling: First, we need to decide what our model $P(E \mid F; \theta)$ will look like. What parameters will it have, and how will the parameters specify a probability distribution?

Learning: Next, we need a method to learn appropriate values for parameters θ from training data.

Search: Finally, we need to solve the problem of finding the most probable sentence (solving "argmax"). This process of searching for the best hypothesis and is often called **decoding**.[3]

The remainder of the material here will focus on solving these problems.

References

[1] Robert B Allen. Several studies on natural language and back-propagation. In *Proceedings of the IEEE First International Conference on Neural Networks*, volume 2, page 341. IEEE Piscataway, NJ, 1987.

[1] Note for the time being, we are assuming that we translate each sentence independently, although we will discuss document-level translation in Section 24.

[2] Details about data can be found in Section 10.

[3] This is based on the famous quote from Warren Weaver, likening the process of machine translation to decoding an encoded cipher.

[2] Peter F. Brown, Vincent J.Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19:263–312, 1993.

[3] Makoto Nagao. A framework of a mechanical translation between Japanese and English by analogy principle. In *Proc. International NATO Symposium on Artificial and Human Intelligence*, pages 173–180, 1984.