2 Statistical MT Preliminaries

First, before talking about any specific models, this chapter briefly lays out statistical machine translation [2] more formally.

First, we define our task of machine translation as translating a J-word source sentence $F = f_1, \ldots, f_J = f_1^J$ into an I-word target sentence $E = e_1, \ldots, e_I = e_1^I$.[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 and finding the target sentence that maximizes this probability:

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

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 (details about data can be found in Section 9). 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 $P(E|F; \theta)$ will look like? What parameters will it have, and how will the parameters specify a probability distribution?

Learning: Next, we need a method to go from parallel training data to appropriate parameters θ,

Search: Finally, the problem of finding the most probable sentence (solving "argmax") requires searching for the best hypothesis and is often called **decoding**.[2]

The remainder of this course will focus on solving these problems.

Chapter 2 References

[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.

---

[1] Note for the time being, we are assuming that we translate each sentence independently, although we will discuss document-level translation in Section 22.

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