Part I
Foundations

Chapter 1
Introduction

This chapter provides a gentle introduction to the field of statistical machine translation. We review the history of the field and highlight current applications of machine translation technology. Some pointers to available resources are also given.

But first let us take a broad look at the contents of this book. We have broken up the book into three main parts: **Foundations** (Chapters 1–3), which provides essential background for novice readers, **Core Methods** (Chapters 4–8), which covers the principles of statistical machine translation in detail, and **Advanced Topics** (Chapters 9–11), which discusses more recent advances.

Chapter 1: Introduction
Chapter 2: Words, Sentences, Corpora
Chapter 3: Probability Theory
Chapter 4*: Word-Based Models
Chapter 5*: Phrase-Based Models
Chapter 6*: Decoding
Chapter 7: Language Models
Chapter 8: Evaluation
Chapter 9: Discriminative Training
Chapter 10: Integrating Linguistic Information
Chapter 11: Tree-Based Models

This book may be used in many ways for a course on machine translation or data-driven natural language processing. The three chapters flagged with a star (*) should be part of any such course. See

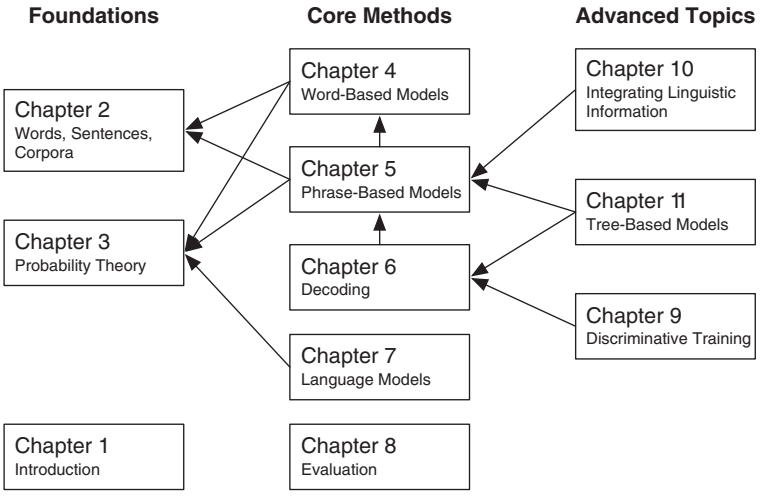

Figure 1.1 Dependencies between the chapters of this book.

also Figure 1.1 for the dependencies between the chapters. Some suggestions for courses that last one semester or term:

- Undergraduate machine translation course: Lectures on Chapters 1–6, and optionally Chapters 7 and 8.
- Graduate machine translation course: Lectures on Chapter 1, give Chapters 2–3 as reading assignment, lectures on Chapters 4–8, and optionally Chapters 9, 10, or 11.
- Graduate data-driven NLP course: Chapters 2–7 (with stronger emphasis on Chapters 2, 3, and 7), and optionally Chapters 9 or 11.

Chapter 2 may be skipped if students have taken undergraduate classes in linguistics. Chapter 3 may be skipped if students have taken undergraduate classes in mathematics or computer science.

The web site for this book[1] provides pointers to slides that were prepared from this book, as well as a list of university classes that have used it.

1.1 Overview

In the following, we give an extended overview of the book that touches on major concepts, chapter by chapter.

1.1.1 Chapter 1: Introduction

The history of machine translation goes back over 60 years, almost immediately after the first computers had been used to break encryption

[1] Hosted at `http://www.statmt.org/book/`

codes in the war, which seemed an apt metaphor for translation: what is a foreign language but encrypted English? Methods rooted in more linguistic principles were also investigated. The field flourished, until the publication of a report by ALPAC, whose negative assessment stalled many efforts. In the 1970s the foundations for the first commercial systems were laid, and with the advent of the personal computer and the move towards translation memory tools for translators, machine translation as a practical application was meant to stay. The most recent trend is towards data-driven methods, especially statistical methods, which is the topic of this book.

Work in machine translation research is not limited to the grand goal of fully automatic, high-quality (publishable) translation. Often, rough translations are sufficient for browsing foreign material. Recent trends are also to build limited applications in combination with speech recognition, especially for hand-held devices. Machine translation may serve as a basis for post-editing, but translators are generally better served with tools such as translation memories that make use of machine translation technology, but leave them in control.

Many resources are freely available for statistical machine translation research. For many of the methods described in this book, tools are available as reference implementation. Also, many translated text collections are available either by simple web download or through the Linguistic Data Consortium. Other sources are ongoing evaluation campaigns that provide standard training and test sets, and benchmark performance.

1.1.2 Chapter 2: Word, Sentences, Corpora

Many statistical machine translation methods still take a very simple view of language: It is broken up into sentences, which are strings of words (plus punctuation which is separated out by a tokenization step). Words have a very skewed distribution in language: Some are very frequent, while many rarely occur. Words may be categorized by their part-of-speech (noun, verb, etc.) or by their meaning. Some languages exhibit rich inflectional morphology, which results in a large vocabulary. Moreover, the definition of what words are is less clear, if writing systems do not separate them by spaces (e.g. Chinese).

Moving beyond the simple notion of sentences as strings of words, one discovers a hierarchical structure of clauses, phrases and dependencies between distant words that may be best represented graphically in tree structures. One striking property of language is recursion, which

enables arbitrarily deep nested constructions. Several modern linguistic theories of grammar provide formalism for representing the syntax of a sentence. Theories of discourse address the relationships between sentences.

Text collections are called corpora, and for statistical machine translation we are especially interested in parallel corpora, which are texts, paired with a translation into another language. Texts differ in style and topic, for instance transcripts of parliamentary speeches versus news wire reports. Preparing parallel texts for the purpose of statistical machine translation may require crawling the web, extracting the text from formats such as HTML, as well as document and sentence alignment.

1.1.3 Chapter 3: Probability Theory

Probabilities are used when we have to deal with events with uncertain outcomes, such as a foreign word that may translate into one of many possible English words. Mathematically, a probability distribution is a function that maps possible outcomes to values between 0 and 1. By analyzing an event, we may find that a standard distribution (such as uniform, binomial, or normal distributions) can be used to model it. Alternatively, we may collect statistics about the event and estimate the probability distributions by maximum likelihood estimation.

We typically deal with multiple uncertain events, each with a different probability distribution, for instance the translation of multiple words in a sentence. The mathematics of probability theory provides a toolset of methods that allow us to calculate more complex distributions, such as joint or conditional distributions for related events. Rules such as the chain rule or the Bayes rule allow us to reformulate distributions. Interpolation allows us to compensate for poorly estimated distributions due to sparse data.

Our methods of dealing with probabilistic events are often motivated by properties of probability distributions, such as the mean and variance in outcomes. A powerful concept is entropy, the degree of uncertainty, which guides many machine learning techniques for probabilistic models.

1.1.4 Chapter 4: Word-Based Models

The initial statistical models for machine translation are based on words as atomic units that may be translated, inserted, dropped, and reordered.

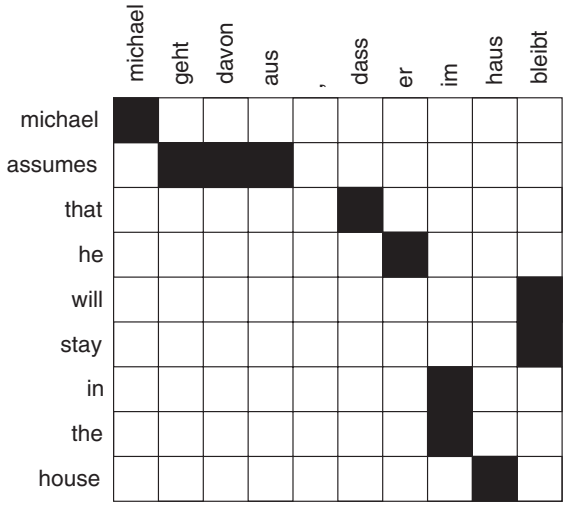

Figure 1.2 Aligning the words in a sentence pair is the first step of many statistical machine translation methods (Chapter 4).

Viewing the translation between a sentence pair as a mapping of the words on either side motivates the notion of word alignment (see Figure 1.2 for an illustration), which may be modeled with an alignment function.

Since parallel corpora provide us with sentences and their translation, but not with word alignments, we are confronted with the problem of learning from incomplete data. One way to address this is the expectation maximization algorithm, that alternately computes the probability of possible word alignments and collects counts, and builds an improved model of these alignments.

In statistical machine translation, we use both a translation model and a language model, which ensures fluent output. The combination of these models is mathematically justified by the noisy-channel model.

The original work at IBM on statistical machine translation uses word-based translation models of increasing complexity, that not only take lexical translation into account, but also model reordering as well as insertion, deletion, and duplication of words. The expectation maximization algorithm gets more computationally complex with the increasingly sophisticated models. So we have to sample the alignment space instead of being able to exhaustively consider all possible alignments.

The task of word alignment is an artifact of word-based translation models, but it is a necessary step for many other multilingual applications, and forms a research problem of its own. Automatically obtained word alignments may be evaluated by how closely they match word alignments created by humans. Recent advances in word alignment are

symmetrization methods to overcome the one-to-many problem of the IBM models, integrate additional linguistic constraints, or directly align groups of words.

1.1.5 Chapter 5: Phrase-Based Models

The currently most successful approach to machine translation uses the translation of phrases as atomic units. See Figure 1.3 for an illustration. These phrases are any contiguous sequences of words, not necessarily linguistic entities. In this approach, the input sentence is broken up into a sequence of phrases; these phrases are mapped one-to-one to output phrases, which may be reordered.

Commonly, phrase models are estimated from parallel corpora that were annotated with word alignments by methods discussed in Chapter 4. All phrase pairs that are consistent with the word alignment are extracted. Probabilistic scores are assigned based on relative counts or by backing off to lexical translation probabilities.

Phrase-based models typically do not strictly follow the noisy-channel approach proposed for word-based models, but use a log-linear framework. Components such as language model, phrase translation model, lexical translation model, or reordering model are used as feature functions with appropriate weights. This framework allows the straightforward integration of additional features such as number of words created or number of phrase translations used.

Reordering in phrase models is typically modeled by a distance-based reordering cost that discourages reordering in general. Reordering is often limited to movement over a maximum number of words. The lexicalized reordering model learns different reordering behavior for each specific phrase pair (for instance in French–English, adjectives like *bleue* are more likely to be reordered).

Alternatively to learning phrase-based models from a word-aligned parallel corpus, we may also use the expectation maximization algorithm to directly find phrase alignments for sentence pairs.

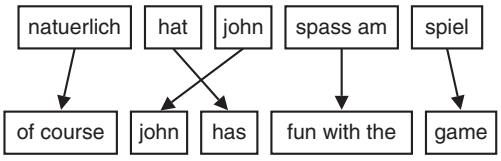

Figure 1.3 Phrase-based machine translation. The input is segmented into phrases (not necessarily linguistically motivated), translated one-to-one into phrases in English and possibly reordered (Chapter 5).

1.1.6 Chapter 6: Decoding

Probabilistic models in statistical machine translation assign a score to every possible translation of a foreign input sentence. The goal of decoding is to find the translation with the best score. In the decoding process, we construct the translation word by word, from start to finish. The word-based and phrase-based models are well suited for this, since they allow the computation of scores for partial translations.

Before translating a foreign input sentence, we first consult the translation table and look up the applicable translation options. During decoding, we store partial translations in a data structure called a hypothesis. Decoding takes the form of expanding these hypotheses by deciding on the next phrase translation, as illustrated in Figure 1.4. Due to the computational complexity of decoding (NP-complete), we need to restrict the search space. We do this by recombination, a dynamic programming technique to discard hypotheses that are not possibly part of the best translation, and by organizing the hypotheses into stacks to prune out bad ones early on. Limits on reordering also drastically reduce the search space.

When comparing hypotheses for the purpose of pruning, we also have to take the translation cost of the remaining untranslated words into account. Such future costs may be efficiently computed up front before decoding.

Variations and alternatives to the beam search stack decoding algorithm have been proposed, such as A* search, a standard search technique in artificial intelligence. Greedy hill-climbing decoding first creates a rough translation and then optimizes it by applying changes. Machine translation decoding may also be fully implemented with finite state transducer toolkits.

1.1.7 Chapter 7: Language Models

Language models measure the fluency of the output and are an essential part of statistical machine translation. They influence word choice,

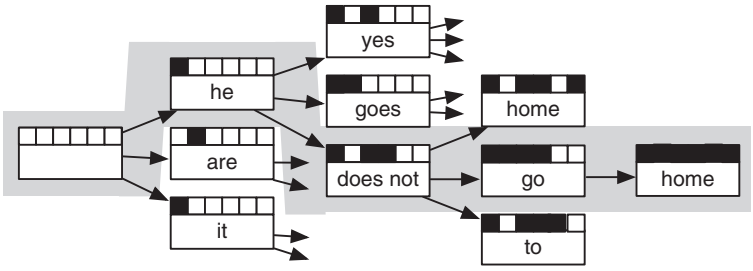

Figure 1.4 The translation process (decoding): The translation is built from left to right, and the space of possible extensions of partial translation is explored (Chapter 6).

reordering and other decisions. Mathematically, they assign each sentence a probability that indicates how likely that sentence is to occur in a text. N-gram language models use the Markov assumption to break the probability of a sentence into the product of the probability of each word, given the (limited) history of preceding words. Language models are optimized on perplexity, a measure related to how much probability is given to a set of actual English text.

The fundamental challenge in language models is handling sparse data. Just because something has not been seen in the training text does not mean that it is impossible. Methods such as add-one smoothing, deleted estimation or Good–Turing smoothing take probability mass from evident events and assign it to unseen events.

Another angle on addressing sparse data in n-gram language models is interpolation and back-off. Interpolation means that n-gram models with various orders (i.e., length of history) are combined. Back-off uses the highest order n-gram model, if the predicted word has been seen given the history, or otherwise resorts to lower-order n-gram models with shorter histories. Various methods exist to determine the back-off costs and adapt the elementary n-gram probability models. Kneser–Ney smoothing takes both the diversity of predicted words and histories into account.

Especially for English there is no lack of training data for language models. Billions, if not trillions, of words of text can be acquired. Larger language models typically lead to better results. Handling such large models, however, is a computational challenge. While the models may be trained on disk, for machine translation decoding they need to be accessed quickly, which means that they have to be stored in RAM. Efficient data structures help, and we may reduce vocabulary size and n-grams to what is needed for the translation of a specific input set. Computer clusters have also been used for very large models.

1.1.8 Chapter 8: Evaluation

A hotly debated topic in machine translation is evaluation, since there are many valid translations for each input sentence (see Figure 1.5 for an illustration). At some point, we need some quantitative way to assess the quality of machine translation systems, or at least a way to be able to tell if one system is better than another or if a change in the system led to an improvement. One way is to ask human judges to assess the adequacy (preservation of meaning) and fluency of machine translation output, or to rank different translations of an individual sentence. Other criteria, such as speed, are also relevant in practical deployments.