Cambridge University Press & Assessment
978-1-108-49732-9 — Neural Machine Translation
Philipp Koehn
Excerpt
More Information

Part I
Introduction

Chapter 1
The Translation Problem

Imagine that you are a translator. You are asked to translate from German to English and you come across the word *Sitzpinkler*. Its literal meaning is *someone who pees sitting down*, but its intended meaning is *wimp*. The implication is that a man who sits down to pee is not a real man.

But there is more going on here. This word was popularized on a comedy show that coined several other terms in this fashion. One is *Warmduscher*, *someone who takes a warm shower,* or even *Frauenversteher*, *someone who understands women.* In fact, a whole fad emerged to come up with new terms like this. All these terms are used as insults, but not as real serious insults. They are used very much in jest, a slight mocking.

These terms are also firmly a reflection of the current zeitgeist, when the expectations of what it means to be a man are changing. Using such terms is a light-hearted commentary on this change. It is not really unmanly to sit down to pee, although it is something that women do and hence a man who wants to be a traditional "real" man loses some of his identity this way. As you can see, there is a lot going on here.

So, what is a translator going to do? Probably use *wimp* and move on. This example demonstrates that translation is basically impossible. The meaning of words in a language are tied to their prior use in a specific culture. *Four score and seven years* is not just any way to say *87 years*. And *I have a dream* implies much more than just announcing a vision of the future. Words carry not only an explicit meaning but also an undercurrent of implications that often does not have any equivalent in another language and another culture.

Figure 1.1 Ten translators translate the same short French sentence—*Sans se démonter, il s'est montré concis et précis.*—in 10 different ways. Human evaluators also disagree for each translation if it is correct or wrong.

Assessment Correct/Wrong	Translation
1/3	*Without fail, he has been concise and accurate.*
4/0	*Without getting flustered, he showed himself to be concise and precise.*
4/0	*Without falling apart, he has shown himself to be concise and accurate.*
1/3	*Unswayable, he has shown himself to be concise and to the point.*
0/4	*Without showing off, he showed himself to be concise and precise.*
1/3	*Without dismantling himself, he presented himself consistent and precise.*
2/2	*He showed himself concise and precise.*
3/1	*Nothing daunted, he has been concise and accurate.*
3/1	*Without losing face, he remained focused and specific.*
3/1	*Without becoming flustered, he showed himself concise and precise.*

1.1 Goals of Translation

goals of translation

There are many different ways to translate a sentence. See Figure 1.1 for an example (from a study on a computer aided translation tool). Ten translators translated the same short French sentence—*Sans se démonter, il s'est montré concis et précis.*—in 10 different ways. There is the challenge of the French phrase *Sans se démonter*, which does not seem to have a nice equivalent, so translators make choices from very literal translations that are awkward English (say, *Without dismantling himself*) to fairly free translations (*Unswayable*), to just dropping this phrase. But there is also a lot of variance for the rest of the sentence. In fact, no two translations are the same. And this is by far the most typical outcome when several translators translate the same sentence. In this study, the translations were also evaluated by four human assessors each as either correct and wrong. For most translations, there is disagreement.

adequacy **fluency**

Translation is always an approximation. Translators have to make choices, and different translators make different choices. The main competing goals are **adequacy** and **fluency**. Adequacy means retaining the meaning of the original text. Fluency requires producing output text that reads just like any well-written text in the target language.

Often, these two goals are in conflict. To closely maintain the meaning of the original sentence may make a translation clumsy. Different genres of text make different trade-offs here. Translations of literature are more concerned with style, that text flows well, so it may completely change some of the meaning to maintain the overall spirit of a text. Think about the translation of song lyrics. It is more important that the translated song sounds right and carries across the same emotion.

However, when translating an operations manual or a legal text, concerns about fluency are secondary. It is fine to produce wooden and awkward phrases when this is the only way to express the same facts.

Consider an example that may show up in a newspaper article: the phrase *about the same population as Nebraska*. Let's say you want to translate this into Chinese. Very few people in China will have any idea of how many people live in Nebraska. So, you may want to change *Nebraska* to the name of a Chinese city or province that the reader will be familiar with. This was the whole intention of the author—to provide a concrete example that is meaningful to the reader.

A more subtle example is a foreign phrase that literally translates to *the American newspaper the New York Times*. For any American reader this would come across at least as odd. It is well known that the *New York Times* is an American newspaper, so what is the reason to point this out? It is likely the original phrase did not intend to place special emphasis on the American nature of the paper. It is just there to inform the readers who may not know the paper. Consider the converse. A literal translation from German may be *Der Spiegel reported*, which leaves most American readers unsure about the reliability of the source. So, a professional translator may decide to render this as *the popular German news weekly Der Spiegel reported*.

A goal of translation is to be invisible. At no point should a reader think *This is translated really well/badly* or even worse *What did this say in the original?* Readers should not notice any artifacts of translation and should be given the illusion that the text was originally written in their own language.

1.2 Ambiguity

ambiguity

If there is one word that encapsulates the challenge of natural language processing with computers, it is **ambiguity**. Natural language is ambiguous on every level: word meaning, morphology, syntactic properties and roles, and relationships between different parts of a text. Humans are able to deal with this ambiguity somewhat by taking in the broader context and background knowledge, but even among humans there is a lot of misunderstanding. Sometimes the speaker is purposely ambiguous to not make a firm commitment to a particular interpretation. In that case, the translation has to retain that ambiguity.

1.2.1 Word Translation Problems

word translation problems

The first obvious example of ambiguity is that some words have strikingly different meanings. Consider the example sentences:

- *He deposited money in a* **bank** *account with a high* **interest** *rate.*
- *Sitting on the* **bank** *of the Mississippi, a passing ship piqued his* **interest**.

The words *bank* and *interest* have different meanings in these two sentences. A *bank* may be the shore of a river or a financial institution, while *interest* may mean curiosity or have the financial meaning of a fee charged for a loan.

How could computers ever know the difference? Well, how do humans know the difference? We consider the surrounding words and the overall meaning of the sentence. In the examples, the word *rate* following *interest* is already a very strong indicator. Computers have to take this context into account as well.

1.2.2 Phrase Translation Problems

phrase translation problems

The next challenge is that meaning is not always compositional. This prevents us from cleanly breaking up the translation problem into small subproblems. The clearest examples for this are idiomatic phrases such as *It's raining cats and dogs*. This will not translate well word for word into any other language. A good German translation may be *es regnet Bindfäden*, which translates literally to English as *it rains strings of yarn* (the rain droplets are so close that they string together).

You may sometimes be able to track down an idiom through its origin story or the metaphor it builds on, but in practice human users of language just memorize these and do not think too much about them.

1.2.3 Syntactic Translation Problems

syntactic translation problems

The classic example for syntactic ambiguity is prepositional phrase attachment. There is a difference between *eating steak with ketchup* and *eating steak with a knife*, in the first case the noun in the prepositional phrase is connected to the object *steak* while in the second case it is connected to the verb *eating*. However, this problem often does not matter much for translation, since the target language may allow for the same ambiguous structure, so there is no need to resolve it.

However, languages often differ in their sentence structure in ways that matter for translation. One of the main distinctions between languages is if they use word order or morphology to mark the relationships between words. English mostly relies on word order, the standard sentence structure is subject–verb–object. Other languages, like German, allow the subject or object at the beginning of the sentence, and they use morphology, typically changes to word endings, to make the distinction clear.

Consider the following short German sentence, with possible translations for each word below it.

das	*behaupten*	*sie*	*wenigstens*
that	*claim*	*they*	*at least*
the		*she*	

There is a lot going on here.

- The first word *das* could mean *that* or *the*, but since it is not followed by a noun, the translation *that* is more likely.

- The third word *sie* could mean *she* or *they*.
- The verb *behaupten* means *claim,* but it is also morphologically inflected for plural. The only possible plural subject in the sentence is *sie* in the interpretation of *they*.

So, the closest English translation *they claim that at least* requires the reordering from object–verb–subject word order to subject–verb–object word order. Google Translate translates this sentence as *at least, that's what they say*, which avoids some of the reordering (*that* is still in front of the verb). This is also a common choice of human translators who would like to retain the emphasis on *that* by placing it early in the English sentence.

1.2.4 Semantic Translation Problems

semantic translation problems

Translation becomes especially tricky when meaning is expressed differently in different languages or, even worse, requires some inference over several distant literal items or may even be just implied.

pronominal anaphora

Consider the problem of **pronominal anaphora**. Pronouns are used to refer to other mentions, typically prior to the occurrence of the pronoun but not always. Here is one example:

I saw the movie, and **it** *is good.*

This is straightforward example where *it* refers to *movie*. When translating this sentence into languages such as German or French, we also have to find a pronoun for the translation of *it*. However, German and French have gendered nouns. Not all things are of neutral gender as in English, they may be masculine, feminine, or neutral, with apparently arbitrary assignment (*moon* is male in German but female in French, *sun* is female in German but male in French). In our example, a good translation for movie is *Film* in German, which has masculine gender. Hence the pronoun *it* has to be rendered as the masculine pronoun *er* and not the feminine *sie* or the neutral *es*.

So there is quite a lot of inference required: the co-reference between the English pronoun *it* and the English noun *movie*, the decision of translating *movie* into *Film*, the acquisition of the knowledge that *Film* is a masculine noun, and the use of all this information when translating *it* into *er*. So, a lot of information needs to tracked, and the hard problem of co-reference resolution (detecting which entities in a text refer to the same thing) has to be solved.

Let us consider an even more difficult example that involves co-reference resolution.

Whenever I visit my uncle and his daughters, I can't decide who is my favorite **cousin**.

The English word *cousin* is gender neutral, but there is no gender neutral translation of the word into German. Compare that to the strong preference in English for the gendered nouns *brother* and *sister* opposed to the gender neutral *sibling* which is very unusual in certain circumstances (*I'll visit my sibling this weekend* sounds rather odd).

In this case, there is even more complex inference required to detect that the cousin is female—because it is the daughter of my uncle. This requires **world knowledge** about facts of family relationships, in addition to the need for co-reference resolution (*cousin* and *daughters* are connected) and knowledge of grammatical gender of German nouns.

world knowledge

Finally, let us look at problems posed by **discourse** relationships. Consider the two examples:

discourse

Since *you suggested it, I now have to deal with it.*

Since *you suggested it, we have been working on it.*

Here, the English discourse connective *since* has two different senses. In the first example, it is equivalent to *because*, marking a **causal** relationship between the two clauses. In the second example, it has a **temporal** sense. The word will be translated differently for these different senses into most languages. However, detecting the right sense requires information about how the two clauses relate to each other. Analyzing the **discourse structure** of a document, i.e., how all the sentences hang together, is an open and very hard research problem in natural language processing.

causal relationship
temporal relationship
discourse structure

Moreover, discourse relationships may not even be marked by discourse connectives like *since*, *but*, or *for example*. Instead, they may be revealed through the choice of grammatical sentence structure. To give one example:

Having said that, I see the point.

The first clause here has a grammatical form that is used to mark a **concession**. We could also use the word *although* there. When translating this into other languages, this implicit encoding of the **concession** relationship may need to be made explicit with a discourse connective.

concession

1.3 The Linguistic View

linguistics

The examples in the previous section suggest that the problem of translation requires not only several levels of abstractions over natural language but also ultimately commonsense reasoning informed by knowledge about the world, making machine translation an **AI hard** problem. In other words, solving machine translation ultimately requires

AI hard

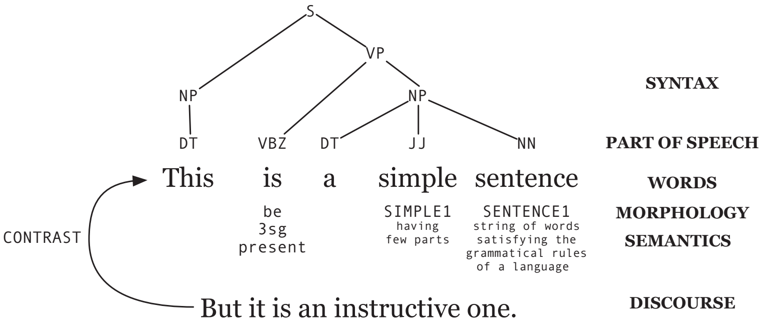

Figure 1.2 Levels of abstraction used in natural language processing.

solving the core problem of **artificial intelligence**. Translating speech acts ultimately requires understanding what these speech acts mean in the world.

artificial intelligence

Let us be a more explicit about the types of abstraction that have been developed over the decades in natural language processing research. See Figure 1.2, which shows various types of linguistic annotation for the sentence *This is a simple sentence*.

Words: While breaking up speech acts into sentences and words seems uncontroversial, it is actually not totally obvious. Consider the case of languages that do not separate words by spaces (such as Chinese), where breaking up a sentence into words requires linguistic tools.

word

Parts of speech: We like to distinguish between nouns, verbs, determiners, etc. Parts of speech fall into two main classes: content words (also called open class words), which describe objects, actions, and properties of the world, and function words, which provide the glue to make the relationships between these words clear. Languages differ quite a bit in the type of open class words that exist (for instance, Chinese does not have determiners, which are admittedly kind of useless).

part of-speech

Morphology: The endings of words may be changed to clarify some of their syntactic or semantic properties. We distinguish between inflectional morphology (e.g., *dog* and *dogs*, *eats* and *eating*), which accounts for count, gender, case, tense, etc., and derivational morphology, which changes the part of speech of a word (*eat*, *eater*, *eatery*). For the task of translation it is sometimes useful to break up words into **stems** (which carry the dictionary meaning) and **morphemes** (which carry inflectional or derivational information), for example, *eats* → *eat* + *s*.

morphology

stem

morpheme

Syntax: We can understand the meaning of a sentence by understanding the connections between its words. Sentences may have multiple clauses (such as the main clause and a relative clause), each clause has at its center a verb, which requires arguments such as subjects and objects, and additional adjuncts such as adverbs (say, *quickly*) temporal phrases (say, *for five minutes*). Subjects and objects are typically noun phrases that break up into

syntax

syntax tree
dependency structure

the main noun, which may be further refined by adjectives and determiners but also relative clauses. A core property of natural language is its recursive structure, so a good way to represent this structure is a **syntax tree**, as shown in Figure 1.2. Another way to represent syntax is by **dependency structure**, where each word has a link to its parent (e.g., the object noun *sentence* to the verb *is*, in our example).

semantics
lexical semantics

Semantics: There are several levels of semantics that could be considered. At the most basic level, **lexical semantics** addresses the different senses of a word. In our example, the meaning of *sentence* is detected as SENTENCE1, which has the definition *string of words satisfying the grammatical rules of a language*, opposed to, say, a prison sentence. But we may also describe the meaning of the entire sentence. One formalism to do this is **abstract meaning representation (AMR)**. For our example sentence, this looks like this:

AMR
abstract meaning representation

```
(b / be
 :arg0 (t / this)
 :arg1 (s / sentence
         :mod (s2 / simple)))
```

Compared to syntax structure, it contains mostly only content words and pronouns, and defines their relationships in form of semantic roles (such as actor, patient, temporal modifier, quantity, etc.). There is much disagreement about the correct formalisms to use for higher-level semantics, and even AMR is a work in progress.

discourse

Discourse: Finally, discourse deals with the relationship between clauses (or elementary discourse units) in a text. It attempts to define the structure of a text, for instance to aid applications such as summarization. There is not much consensus about the right formalisms here and even trained human annotators cannot agree very well on which discourse relationships to assign to a given text.

interlingua

One vision for machine translation is shown in Figure 1.3, initially proposed by Vauquois (1968). The ultimate goal is to analyze a source sentence into its meaning, hopefully in a language-independent meaning representation called **interlingua**, and then to generate the target sentence from that interlingua representation. The research strategy toward this goal is to start with simple lexical transfer models and then move on to more complex intermediate representations at the level of syntax and language-dependent semantics.

Before the advent of neural machine translation, the field of statistical machine translation made great strides along this path. The best performing systems for language pairs such as Chinese–English and German–English were syntax-based systems that generated

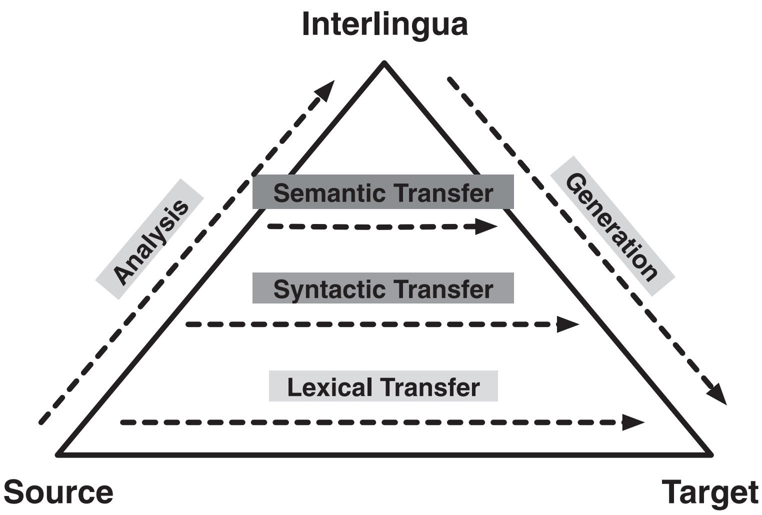

Figure 1.3 The Vauquois triangle. The linguistic vision to analyze the meaning of a source sentence into a language-independent meaning representation and then the generation of the target sentence.

syntax structures during the translation process. With neural machine translation, we are currently back to the level of lexical transfer, but there is a plausible argument to be made that once we mastered that level, we can make another climb up the Vauquois triangle.

1.4 The Data View

data

During the twenty-first century, machine translation research has been firmly grounded in the paradigm that it is futile to write down all the necessary dictionaries and rules that govern language and translation. Instead, all information should be automatically acquired from large amounts of translation examples.

corpus

There are two main types of text **corpora** (a corpus is a collection of text): monolingual and parallel. If we acquire large amounts of text in a single language, we can learn a lot from it, i.e., the words used in the language, how these words are used, the structure of sentences, and so on. There is even the dream to learn how to translate purely from large amounts of monolingual text, called **unsupervised machine translation**. But better resources to learn how to translate are parallel corpora, also called bi-texts, that typically come in the form of sentence pairs, a source sentence and its translation.

unsupervised machine translation

1.4.1 Adequacy

adequacy

Let us take a look at how data will help us solve translation problems, beginning with adequacy, i.e., matching the meaning of the source sentence. To start, take the German word *Sicherheit*, which has three main