Language and Computers

Machine Translation

Based on Dickinson, Brew, & Meurers (2013)
Indiana University
Spring 2015

What is Machine Translation?

Translation is the process of:
- moving texts from one (human) language (source language) to another (target language),
- in a way that preserves meaning.

Machine translation (MT) automates (part of) the process:
- Fully automatic translation
- Computer-aided (human) translation

Is MT needed?

- Translation is of immediate importance for multilingual countries (Canada, India, Switzerland, . . . ), international institutions (United Nations, International Monetary Fund, World Trade Organization, . . . ), multinational or exporting companies.
- The European Union has 24 official languages (as of 2013). All federal laws and other documents have to be translated into all languages.

Example translations

The simple case

It will help to look at a few examples of real translation before talking about how a machine does it.
Take the simple Spanish sentence and its English translation below:

(1) (Yo) hablo español.  
I speak Spanish.

- Words in this example pretty much translate one-for-one
- But we have to make sure hablo matches with Yo, i.e., that the subject agrees with the form of the verb.

Example translations

A slightly more complex case

The order and number of words can differ:

2. a. Tu hablas español?  
You speak Spanish?
   ‘Do you speak Spanish?’

   b. Hablas español?  
Speak español?  
   ‘Do you speak Spanish?’
What goes into a translation

Some things to note about these examples and thus what we might need to know to translate:

- Words have to be translated → dictionary
- Words are grouped into meaningful units & word order can differ across languages → syntax
- The forms of words within a sentence are systematic, e.g., verbs have to be conjugated, etc. → morphology

As we move beyond simple examples, what else might we need for translation?

Dictionaries

An MT dictionary differs from a “paper” dictionary:

- must be computer-usable (electronic form, indexed)
- needs to be able to handle various word inflections
- can contain (syntactic and semantic) restrictions that a word places on other words
  - e.g., subcategorization information: give needs a giver, a person given to, and an object that is given
  - e.g., selectional restrictions: if X eats, X must be animate
- contains frequency information
  - for SMT, may be the only piece of additional information

Direct transfer systems

A direct transfer systems consists of:

- A source language grammar
- A target language grammar
- Rules relating source language underlying representation (UR) to target language UR
  - A direct transfer system has a transfer component which relates a source language representation with a target language representation.
  - This can also be called a comparative grammar.

We’ll walk through the following French to English example:

(3) Londres plaire Sam

London is pleasing to Sam

'Sam likes London.'

Steps in a transfer system

1. Source language grammar analyzes the input and puts it into an underlying representation (UR).
   - Londres plait à Sam → Londres plaire Sam (source UR)
2. The transfer component relates this source language UR (French UR) to a target language UR (English UR).
   - French UR: Londres plaire Sam (source UR) → Sam like London (target UR)
3. Target language grammar translates the target language UR into an actual target language sentence.
   - Sam like London → Sam likes London

Notes on transfer systems

- The transfer mechanism is in theory reversible; e.g., the plaire rule works in both directions
  - Not clear if this is desirable: e.g., Dutch aanvangen should be translated into English as begin, but begin should be translated as beginnen.
  - Because we have a separate target language grammar, we are able to ensure that the rules of English apply: like → likes.
  - RBMT systems are still in use today, especially for more exotic language pairs.
There are differing levels of abstraction at which transfer can take place. So far we have looked at URs that represent only word information. We can do a full syntactic analysis, which helps us to know how the words in a sentence relate. Or we can do only a partial syntactic analysis, such as representing the dependencies between words.

A fine-grained interlingua can require extra work: representing the dependencies between words. Language and the interlingua, and then you can put it into the interlingua. If the dependency is not quite right, you can disambiguate English older brother, so we have to disambiguate older brother in French.

Ideally, we could use an interlingua = a language-independent representation of meaning. Benefit: To add new languages to your MT system, you merely have to provide mapping rules between your language and the interlingua, and then you can translate into any other language in your system.

What exactly should be represented in the interlingua? e.g., English corner = Spanish rincón = ‘inside corner’ or esquina = ‘outside corner’.

A fine-grained interlingua can require extra (unnecessary) work: e.g., Japanese distinguishes older brother from younger brother, so we have to disambiguate English brother to put it into the interlingua. Then, if we translate into French, we have to ignore the disambiguation and simply translate it as frère, which simply means ‘brother’.

### Direct transfer & syntactic similarity

This method works best for structurally-similar languages. What exactly should be represented in the interlingua?

- Similar word order and grammar
- Similar concepts and grammar
- Similar concepts, different grammar
- Similar abstract meanings
- Words and concepts equivalent, grammar and word order same

### Interlingual problems

What exactly should be represented in the interlingua?

- e.g., Japanese distinguishes older brother from younger brother, so we have to disambiguate English brother to put it into the interlingua.
- Then, if we translate into French, we have to ignore the disambiguation and simply translate it as frère, which simply means ‘brother’.

### The translation triangle

- Linguistic similarities
- Lexical similarities
- Word-level similarities
- Meanings similarities
- Similarity concepts, different grammar
- Similar abstract meanings
- Words and concepts equivalent, grammar and word order same

### Direct transfer & syntactic similarity (2)

#### John goes on the roller coaster

- S
- NP
- VP
- John
- V
- PP
- goes
- on
- Det
- N
- der
- Achterbahn

---

**Direct transfer & syntactic similarity**

This method works best for structurally-similar languages.

**Interlingual problems**

What exactly should be represented in the interlingua?

- e.g., English corner = Spanish rincón = ‘inside corner’ or esquina = ‘outside corner’

A fine-grained interlingua can require extra (unnecessary) work:

- e.g., Japanese distinguishes older brother from younger brother, so we have to disambiguate English brother to put it into the interlingua.

Then, if we translate into French, we have to ignore the disambiguation and simply translate it as frère, which simply means ‘brother’.

**Interlingua-based systems**

- Direct transfer systems
- Phrase-based systems
- Statistical modeling
- Alignment

**Language and computers**

Machine Translation

Introduction

Examples for Translations

Background: Dictionaries

Linguistic knowledge-based systems

Machine learning-based systems

What makes MT hard?

Evaluating MT systems

References
**Machine learning**

Instead of trying to tell the MT system how we're going to translate, we might try a **machine learning** approach

- We can look at how often a source language word is translated as a target language word, i.e., the **frequency** of a given translation, and choose the most frequent translation.
- But how can we tell what a word is being translated as?
  - Two different scenarios to consider:
    - We are told what each word is translated as: **text alignment**
    - We are not told what each word is translated as: use a **bag of words**

We can also attempt to learn alignments, as a part of the process, as we will see.

---

**Word alignment**

- **word alignment** = determine which source language words align with which target language ones
  - Much harder than sentence alignment to do automatically.
  - But if it has already been done for us, it gives us good information about a word's translation equivalent.

---

**Calculating probabilities**

- With word alignments, it is relatively easy to calculate probabilities.
  - e.g., What is the probability that *run* translates as *correr* in Spanish?
    1. Count up how many times *run* appears in the English part of your bi-text. e.g., 500 times
    2. Out of all those times, count up how many times it was translated as (i.e., aligns with) *correr*. e.g., 275 (out of 500) times.
    3. Divide to get a probability: 275/500 = 0.55, or 55%
- Word alignment gives us some frequency numbers, which we can use to align new cases, using other information, too (e.g., contextual information)

---

**Sentence alignment**

- **sentence alignment** = determine which source language sentences align with which target language ones (what we assumed in the bag of words example).
- Intuitively easy, but can be difficult in practice since different languages have different punctuation conventions.

---

**Different word alignments**

- One word can map to one word or to multiple words. Likewise, sometimes it is best for multiple words to align with multiple words.
- **English-Russian examples:**
  - one-to-one: *khorosho* = *well*
  - one-to-many: *kniga* = *the book*
  - many-to-one: *to take a walk* = *gulyat’*
  - many-to-many: *at least* = *khotya* by (*although if/would*)

---

**The bag of words method**

- What if we're not given word alignments?
- How can we tell which English words are translated as which German words if we are only given an English text and a corresponding German text?
  - We can treat each sentence as a **bag of words** = unordered collection of words.
  - If word A appears in a sentence, then we will record all of the words in the corresponding sentence in the other language as appearing with it.
Example for bag of words method

### Calculating probabilities: sentence 1

So, for *He* in *He speaks Russian well*/*On khorosho govorit po-russki*, we do the following:

1. Count up the number of Russian words: 4.
2. Assign each word equal probability of translation: $1/4 = 0.25$, or 25%.

### Calculating probabilities: sentence 2

If we also have *He is nice*./*On simpatich'nyi*, then for *He*, we do the following:

1. Count up the number of possible translation words: 4 from the first sentence, 2 from the second = 6 total.
   - Note that we are NOT counting the number of English words: we count the number of possible translations.
2. Count up the number of times *On* is the translation = 2 out of 6 = $1/3 = 0.33$, or 33%.

All other words have the probability $1/6 = 0.17$, or 17%, so *On* is the best translation for *He*.

### Probabilities used in IBM models

Probabilistic models are generally more sophisticated, treating the problem as the source language generating the target and taking into account probabilities such as:

- $n(\#|word)$ = probability of the number of words in the target language that the source word generates
- $p-null$ = probability of a null word appearing
- $t(\text{word}|\text{source})$ = probability of a target word, given the source word (i.e., what we’ve just discussed)
- $d(position|position)$ = probability of a target word appearing in position $tposition$, given the source position $sposition$

But we need alignments to estimate these parameters.

### A Generative Story (IBM Models)

Example: Mary did not slap the green witch

- $n(3|\text{slap})$
- $p-null$
- $t(\text{la}|\text{the})$
- $d(4|4)$


### Beyond Bags of Words

- *la maison* ... *la maison blue* ... *la fleur* ...
- *the house* ... *the blue house* ... *the flower* ...

### A chicken-and-egg problem

- If we had the word alignments, we could estimate the parameters of our generative story.
- If we had the parameters, we could estimate the alignments.
Expectation Maximization Algorithm

The Expectation Maximization (EM) algorithm works forwards and backwards to estimate the probabilities:

**EM in a nutshell**
1. initialize model parameters (e.g. uniform)
2. (re-)assign probabilities to the missing data
3. (re-)estimate model parameters from completed data (weighted counts)
4. iterate, i.e., repeat steps 2&3 until you hit some stopping point

**Initial Step**

... la maison ... la maison bleu ... la fleur ...
... the house ... the blue house ... the flower ...

- connections between e.g. la and the are more likely.

**After 1st Iteration**

... la maison ... la maison bleu ... la fleur ...  
... the house ... the blue house ... the flower ...

**After Another Iteration**

... la maison ... la maison bleu ... la fleur ...  
... the house ... the blue house ... the flower ...

- connections between e.g. fleur and flower are more likely (pigeon hole principle).

**Convergence**

... la maison ... la maison bleu ... la fleur ...  
... the house ... the blue house ... the flower ...

\[
\begin{align*}
p(a | the) &= 0.453 \\
p(le | the) &= 0.334 \\
p(maison | house) &= 0.876 \\
p(blue | blue) &= 0.563 \\
\end{align*}
\]

**Phrase-Based Translation Overview**

But this word-based translation doesn’t account for many-to-many mappings between languages

- Foreign “phrases” are translated into English.
- Phrases may be reordered.

Current models allow for many-to-one mappings → we can use those to induce many-to-many mappings
Flaws of Word-Based MT

Multiple alignment points

Heuristically add alignments along the diagonal (Och & Ney, Computational Linguistics, 2003)

Advantages of Phrase-Based Translation

- Many-to-many translation can handle non-compositional phrases.
- Use of local context.
- The more data, the longer the phrases that can be learned.

Lexical ambiguity

Words can be **lexically ambiguous** = have multiple meanings.
- *bank* can be a financial institution or a place along a river.
- *can* can be a cylindrical object, as well as the act of putting something into that cylinder (e.g., *John cans tuna*), as well as being a word like *must, might, or should*.

We can now use these phrase pairs as the units of our probability model.

What makes MT hard?

We've seen how MT systems can work, but MT is a very difficult task because languages are vastly different.

Languages differ:
- Lexically: In the words they use
- Syntactically: In the constructions they allow
- Semantically: In the way meanings work
- Pragmatically: In what readers take from a sentence.

In addition, there is a good deal of real-world knowledge that goes into a translation.
A common (though problematic) evaluation metric is the BLEU metric, based on n-gram comparisons. But we might want to translate it as mourir ('die') and we want to treat it differently than kick the table.

Examples for Translations

Venn diagram of semantic overlap

- hemp = pied (human), patte (bird)
- paw = patte (animal)

The situation can be fuzzy, as in the following English and French correspondences (Jurafsky & Martin 2000, Figure 21.2)

- leg = etape (journey), jambe (human), pied (chair), patte (animal)
- foot = pied (human), patte (bird)
- paw = patte (animal)

Semantic non-compositionality

Some verbs carry little meaning, so-called light verbs

- French faire une promenade is literally 'make a walk,' but it has the meaning of the English take a walk
- Dutch een poging doen 'do an attempt' means the same as the English make an attempt

And we often face idioms = expressions whose meaning is not made up of the meanings of the individual words.

- e.g., English kick the bucket
  - approximately equivalent to the French casser sa pipe ('break his/her pipe')
  - but we might want to translate it as mourir ('die')
  - and we want to treat it differently than kick the table

Evaluating quality

Two main components in evaluating quality:

- Intelligibility = how understandable the output is
- Accuracy = how faithful the output is to the input

A common (though problematic) evaluation metric is the BLEU metric, based on n-gram comparisons

Idiosyncratic differences

Some words do not exist in a language and have to be translated with a more complex phrase: lexical gap or lexical hole.

- French gratiner means something like 'to cook with a cheese coating'
- Hebrew stam means something like 'I'm just kidding' or 'Nothing special.'

There are also idiosyncratic collocations among languages, e.g.:

- English heavy smoker
- French grand fumeur ('large smoker')
- German starker Raucher ('strong smoker')
Further reading

Some of the examples are adapted from the following books: