Language and Computers

Machine Translation

L245
(Based on Dickinson, Brew, & Meurers (2013))

Spring 2017

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

What is MT good for?

- When you need the gist of something and there are no human translators around:
  - translating e-mails & webpages
  - obtaining information from sources in multiple languages (e.g., search engines)
- If you have a limited vocabulary and a small range of sentence types:
  - translating weather reports
  - translating technical manuals
  - translating terms in scientific meetings
- If you want your human translators to focus on interesting/difficult sentences while avoiding lookup of unknown words and translation of mundane sentences.

Why 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
- European Union has 24 official languages (as of 2013)
  - All federal laws and other documents have to be translated into all languages.
  - Also: 5 semi-official 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$_{1st,sg}$ Spanish
'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$_{2nd,sg}$ Spanish
'Do you speak Spanish?'

b. Hablas español?
Speak$_{2nd,sg}$ Spanish
'Do you speak Spanish?'

Expressions can also vary:

(3) Me llamo Markus
(To) me/myself I call Markus
'My name is Markus (I call myself Markus)'

Q: What have you seen in languages you've studied?

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, we can ask:

- What makes translation difficult?
- What else might we need to handle a wider range of translations?

What makes MT hard?

MT is a difficult task because two languages can be 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.

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*.

Semantic relationships

Often we find (rough) **synonyms** between two languages:

- English *book* = Russian *kniga*
- English *music* = Spanish *música*

But words don't always line up exactly between languages.

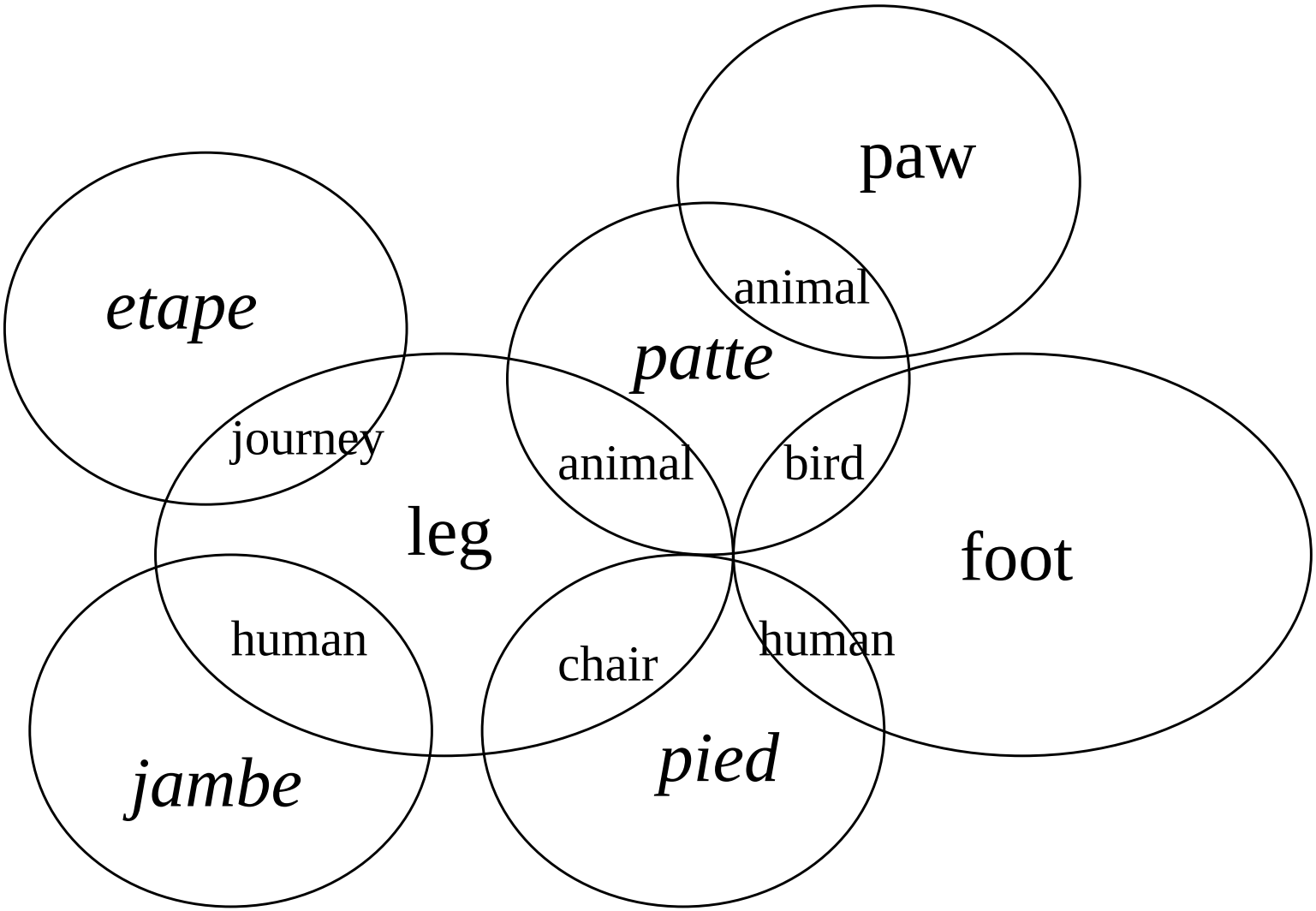
- English **hypernyms** = words that are more general in English than in their counterparts in other languages
  - English *know* is rendered by the French *savoir* ('to know a fact') and *connaitre* ('to know a thing')
  - English *library* is German *Bücherei* if it is open to the public, but *Bibliothek* if it is intended for scholarly work.
- English **hyponyms** = words that are more specific in English than in their foreign language counterparts.
  - The German word *berg* can mean either *hill* or *mountain* in English.
  - The Russian word *ruka* can mean either *hand* or *arm*.

Semantic overlap

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)

Venn diagram of semantic overlap



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*

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')

Different approaches to MT

Despite all the difficulties, MT can be feasible & practical in many contexts

We'll look at some basic approaches to MT:

- Systems based on linguistic knowledge (Rule-Based MT (RBMT))
  - Direct transfer systems
- Machine learning approaches, i.e., statistical machine translation (SMT)
  - SMT is the most popular form of MT

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
- Can contain 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:

(4) Londres plaît à 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 plaît 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		English UR
X plaire Y	↔	Eng(Y) like Eng(X)

   (where Eng(X) means the English translation of X)

   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

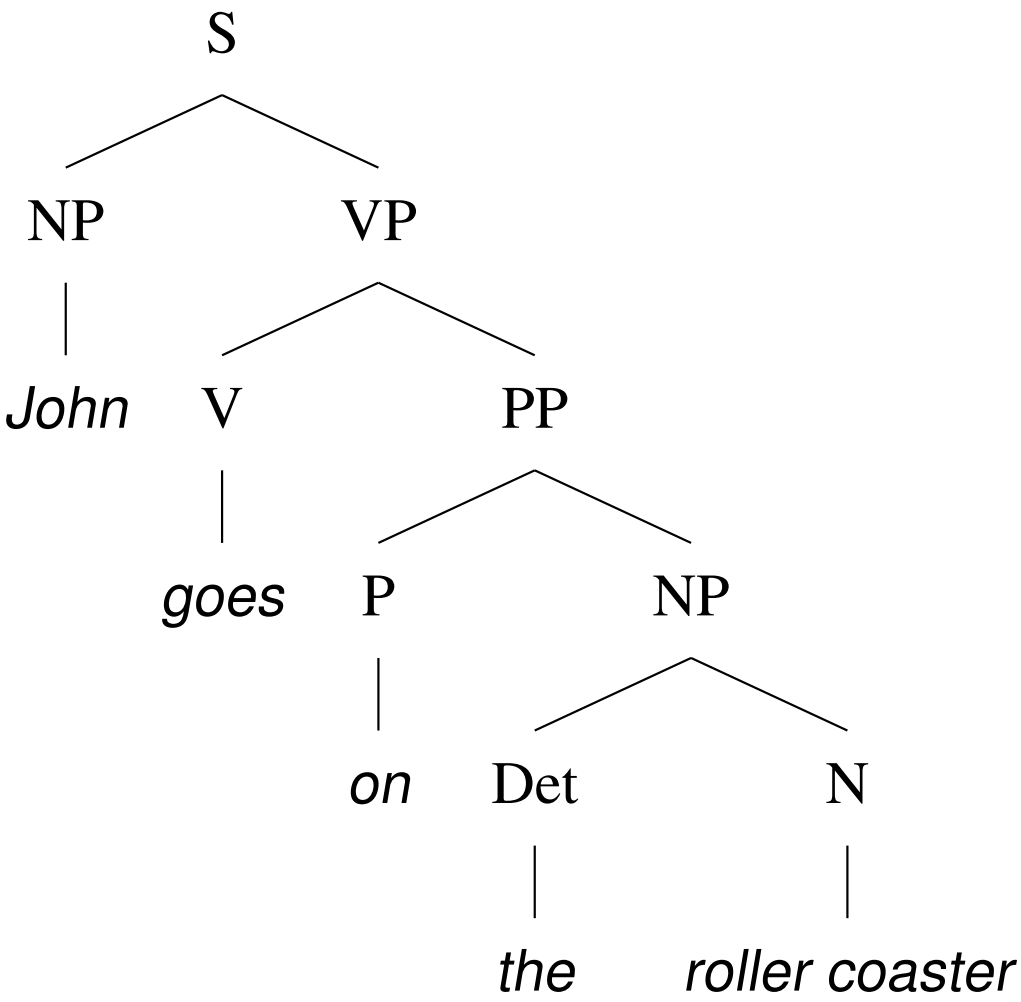
- The transfer mechanism is in theory reversible; e.g., the *plaire* rule works in both directions
  - Not clear if this is desirable: e.g., Dutch *aanvangen* should be translated into English as *begin*, but *begin* should be translated as *beginnen*.
- Because we have a separate target language grammar, 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

Levels of abstraction

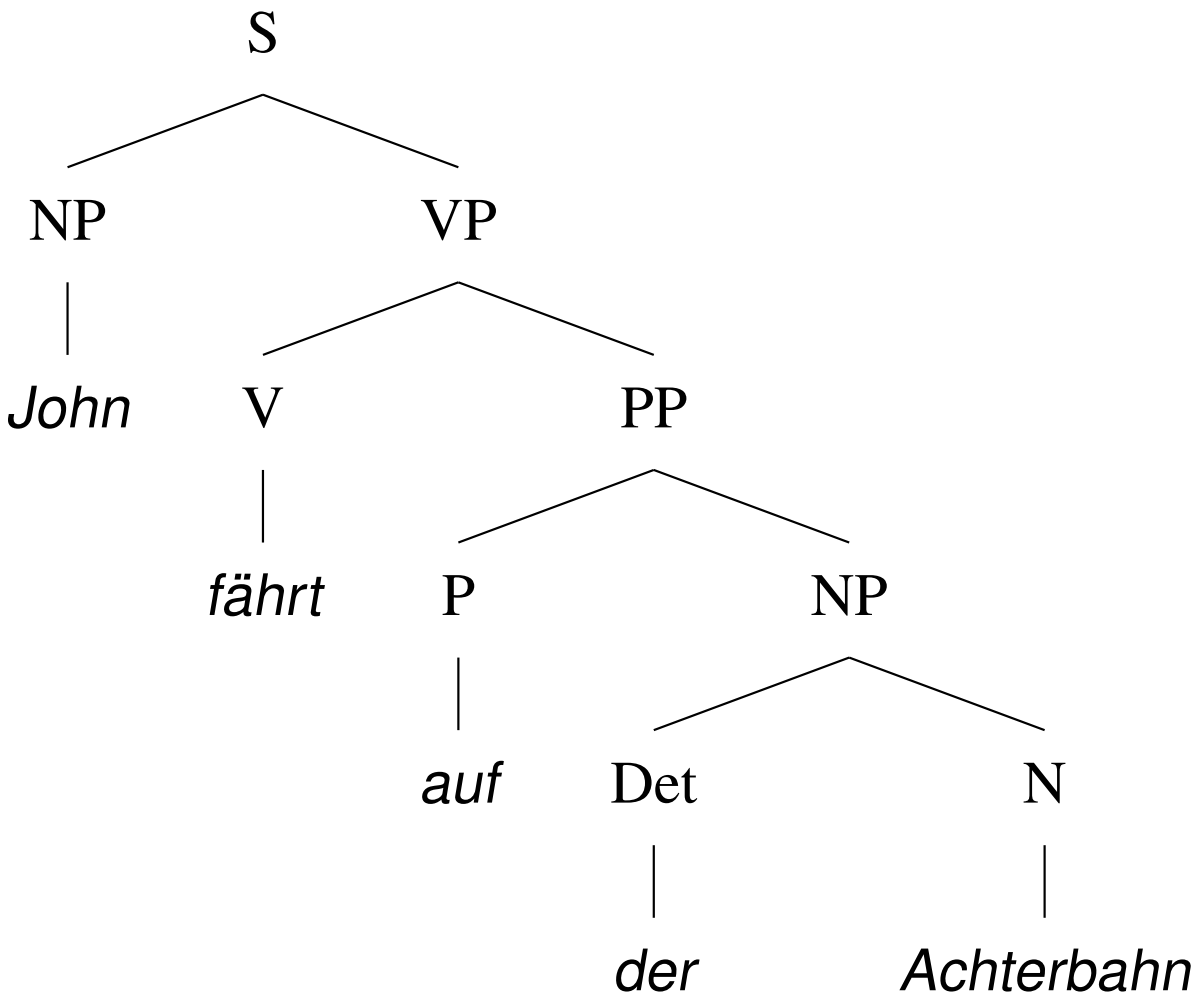
- There are differing levels of abstraction at which transfer can take place. So far we have looked at URs that represent only word information.
- We can do a full syntactic analysis, which helps us to know how the words in a sentence relate.
- Or we can do only a partial syntactic analysis, such as representing the dependencies between words.

Direct transfer & syntactic similarity

This method works best for structurally-similar languages



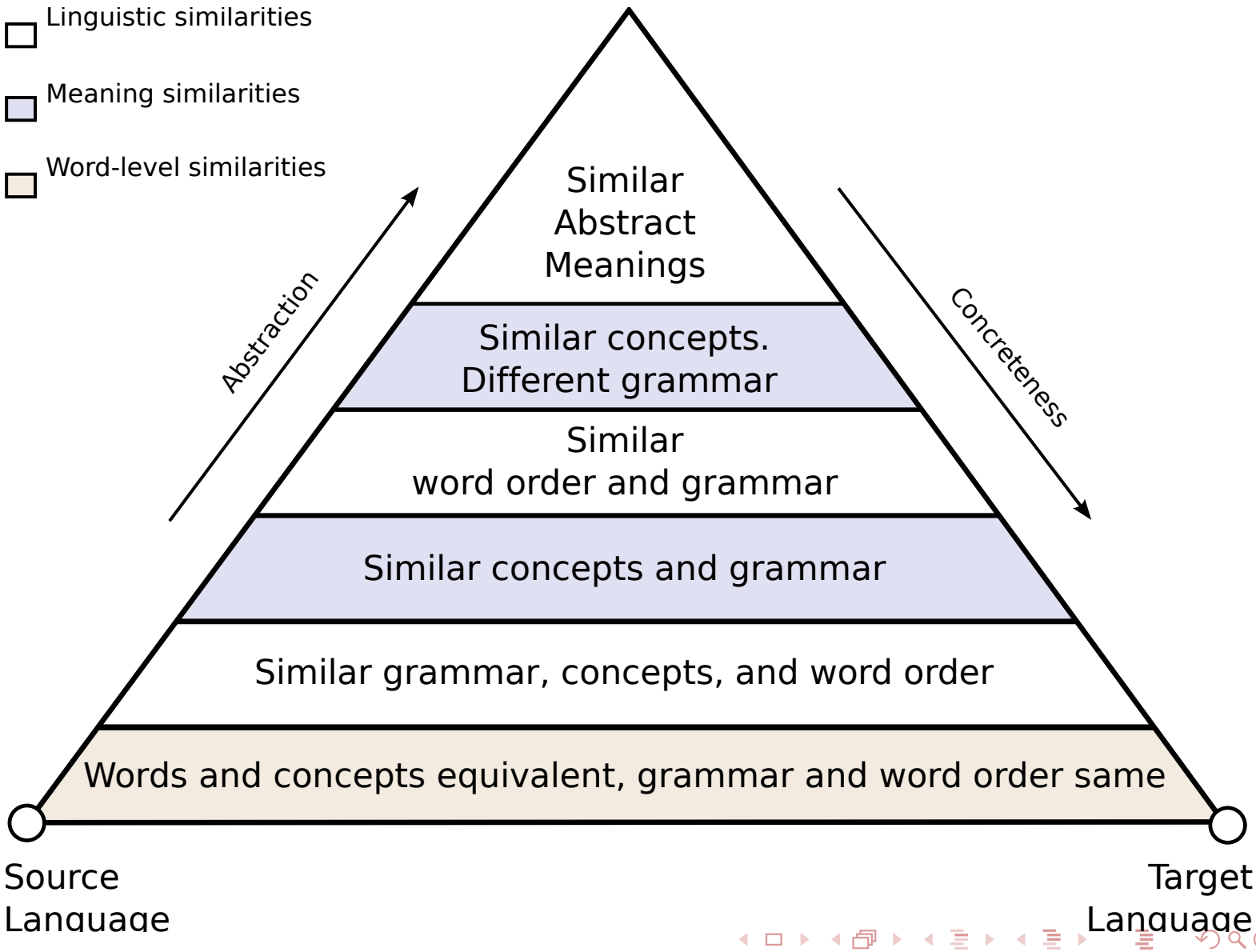
Direct transfer & syntactic similarity (2)



Interlinguas

- 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.

The translation triangle



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'.

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
- We'll examine alignments:
  1. Done by hand: this will give us a good idea about calculating a word's translation equivalent
  2. Done automatically: bag of words method

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')

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)

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

- English *He speaks Russian well.*
- Russian *On khorosho govorit po-russki.*

Eng	Rus	Eng	Rus
He	On	speaks	On
He	khorosho	speaks	khorosho
He	govorit	...	...
He	po-russki	well	po-russki

The idea is that, over thousands, or even millions, of sentences, *He* will tend to appear more often with *On*, *speaks* will appear with *govorit*, and so on.

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%.

Example for bag of words method

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 times 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:

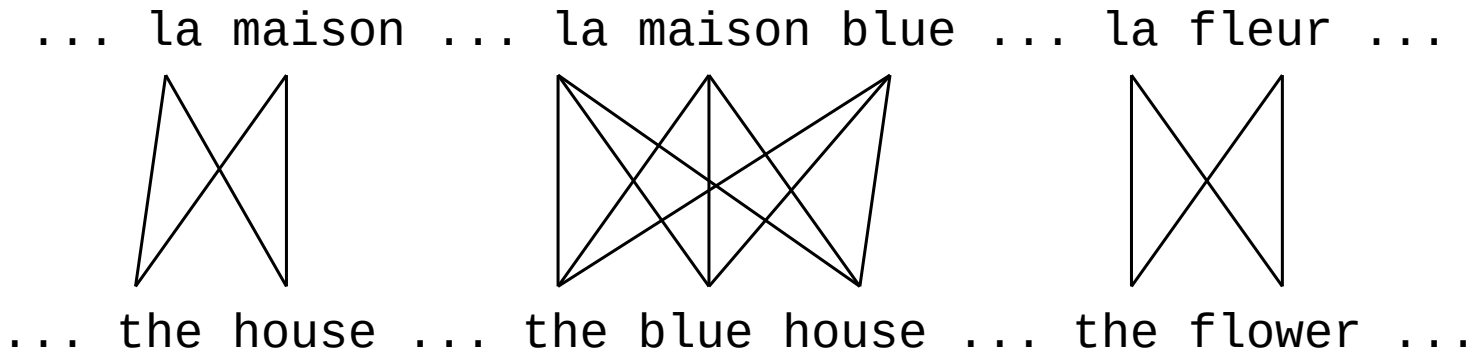
- $n(\#|word)$ = probability of the number of words in the target language that the source word generates
- *p-null* = probability of a null word appearing
- $t(tword|sword)$ = probability of a target word, given the source word (i.e., what we've just discussed)
- $d(tposition|sposition)$ = probability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

But we need alignments to estimate these parameters.

Generative story (IBM models)

Source: Introduction to Statistical Machine Translation, Chris Callison-Burch and Philipp Koehn, http://www.iccs.inf.ed.ac.uk/~pkoehn/publications/esslli-slides-day3.pdf

Beyond bags of words



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.

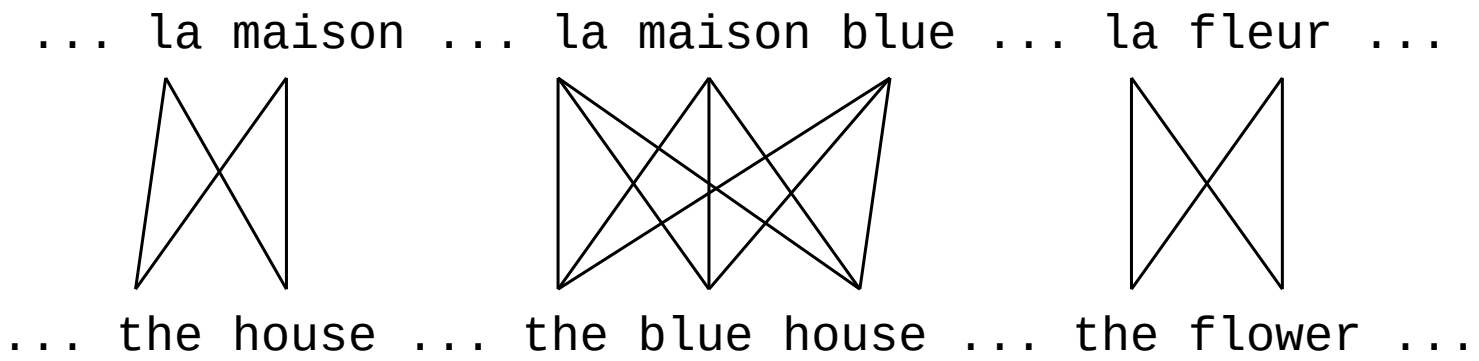
Expectation Maximization algorithm

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

EM in a nutshell

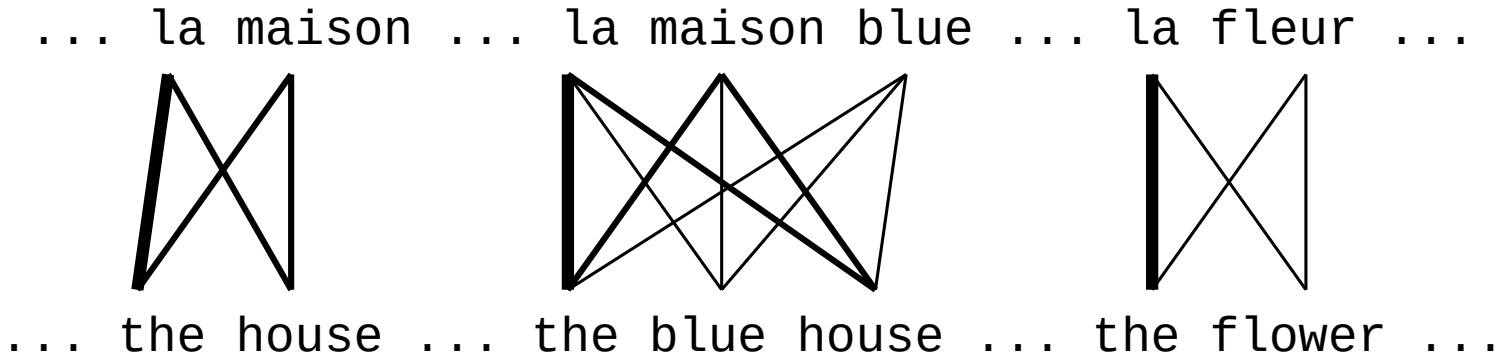
1. initialize model parameters (e.g. uniform)
2. (re-)assign probabilities to the missing data
3. (re-)estimate model parameters from completed data (**weighted counts**)
4. iterate, i.e., repeat steps 2&3 until you hit some stopping point

Initial step



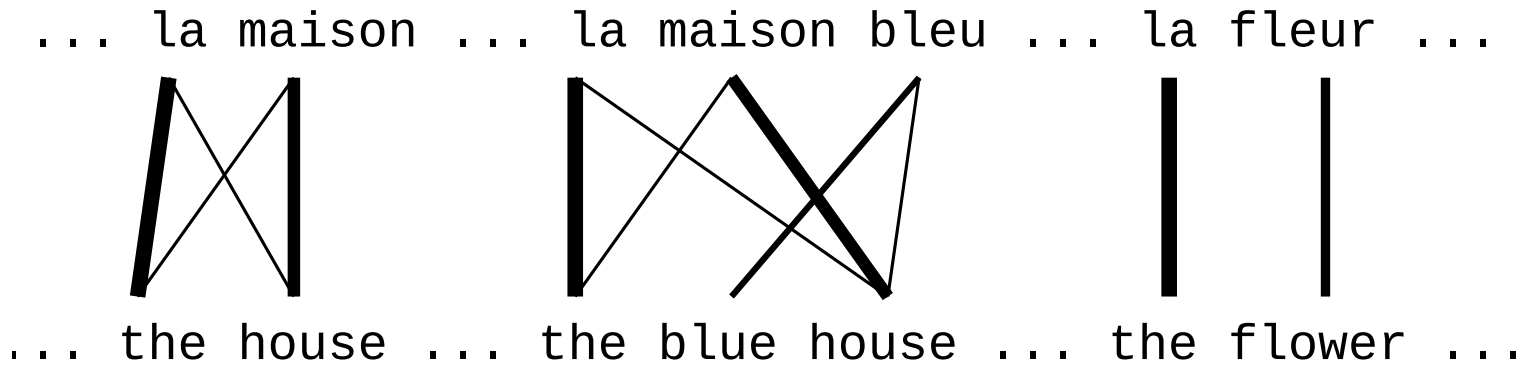
- All connections equally likely.

After 1st iteration



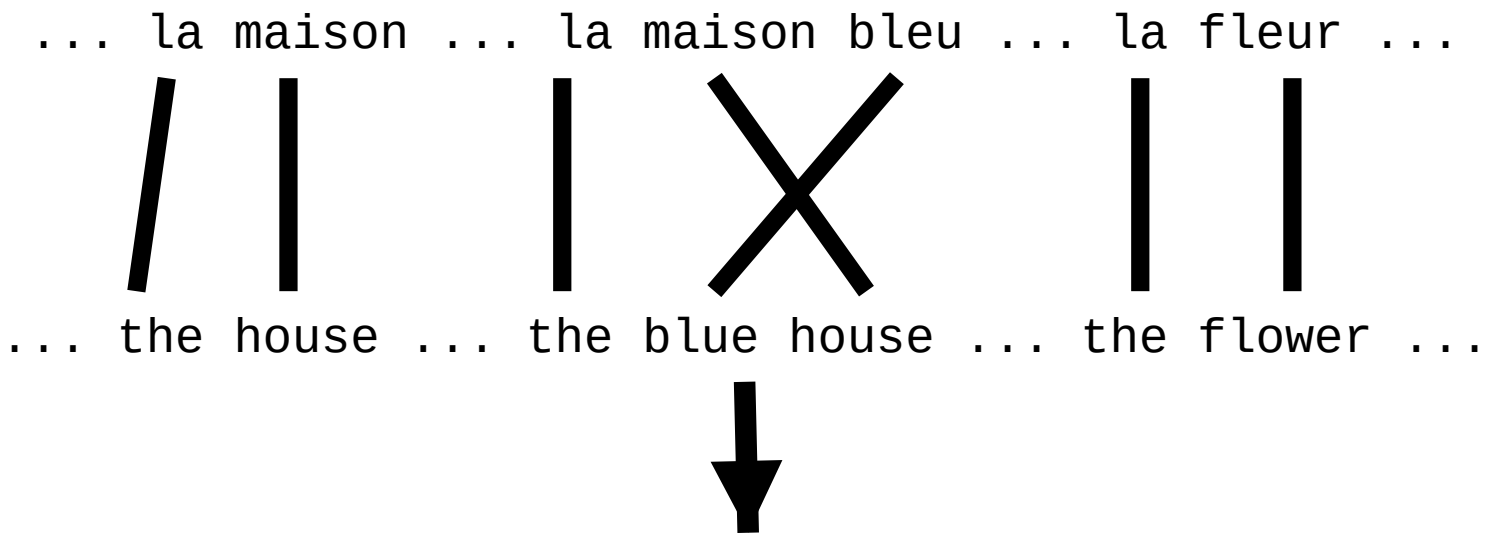
- Connections between e.g. *la* and *the* are more likely.

After another iteration



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

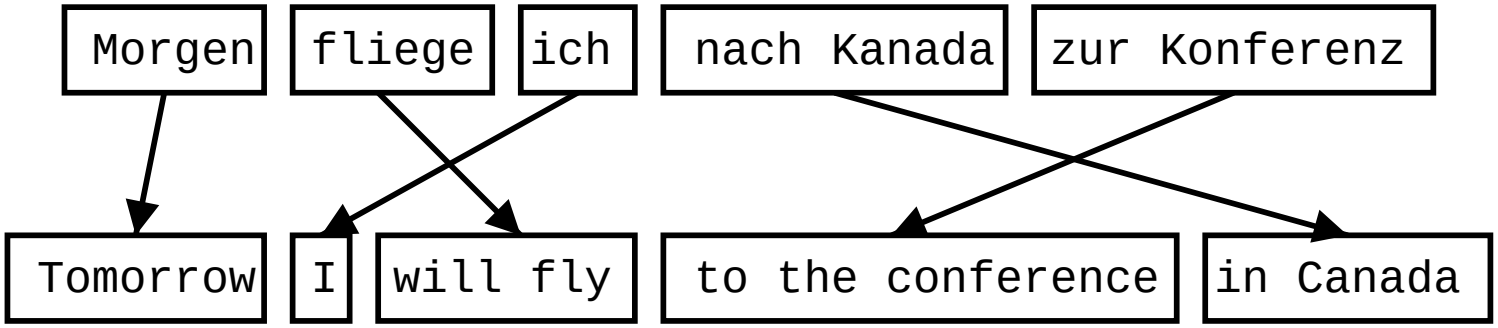
Convergence



Phrase-based translation

Overview

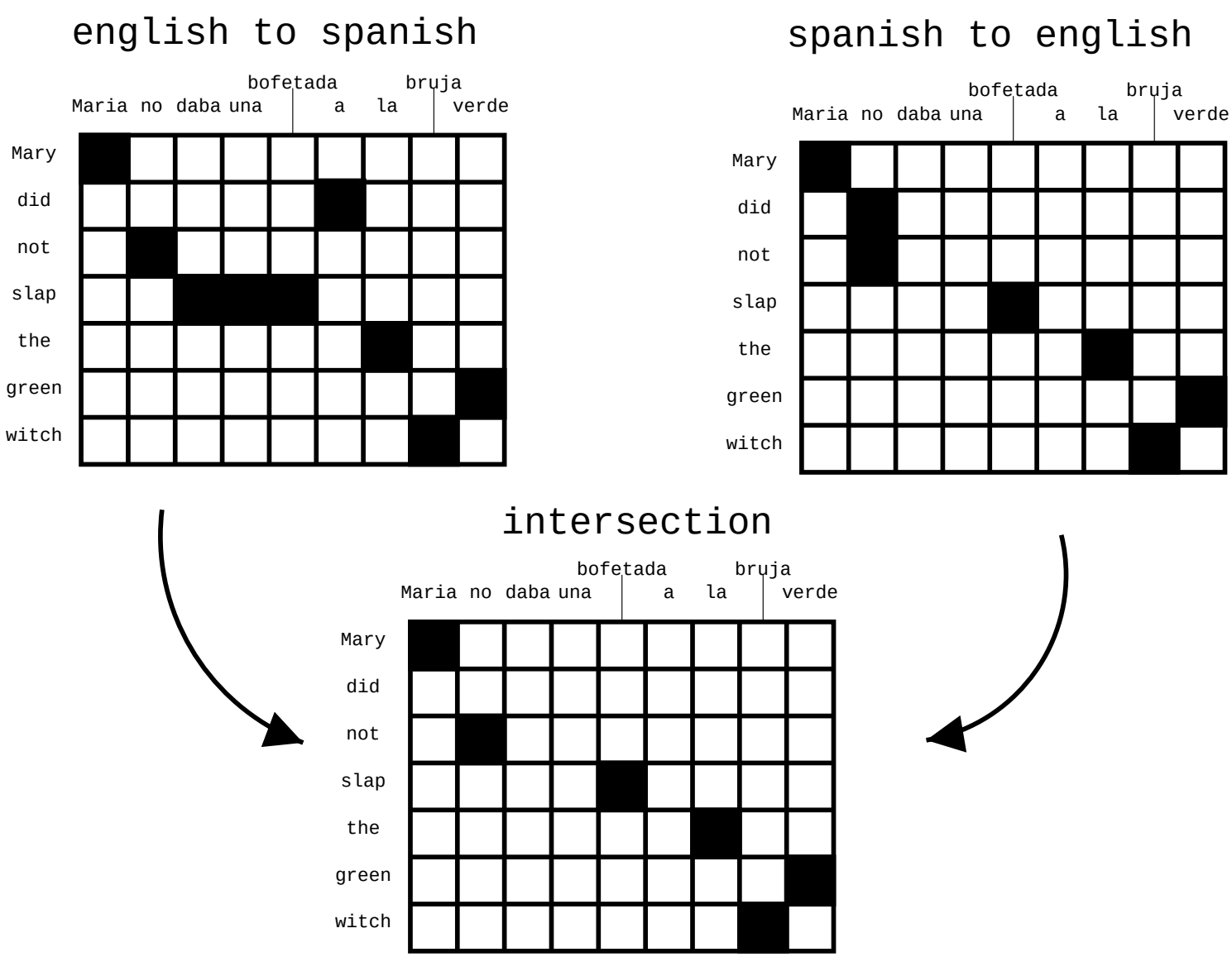
But this word-based translation doesn't account for many-to-many mappings between languages



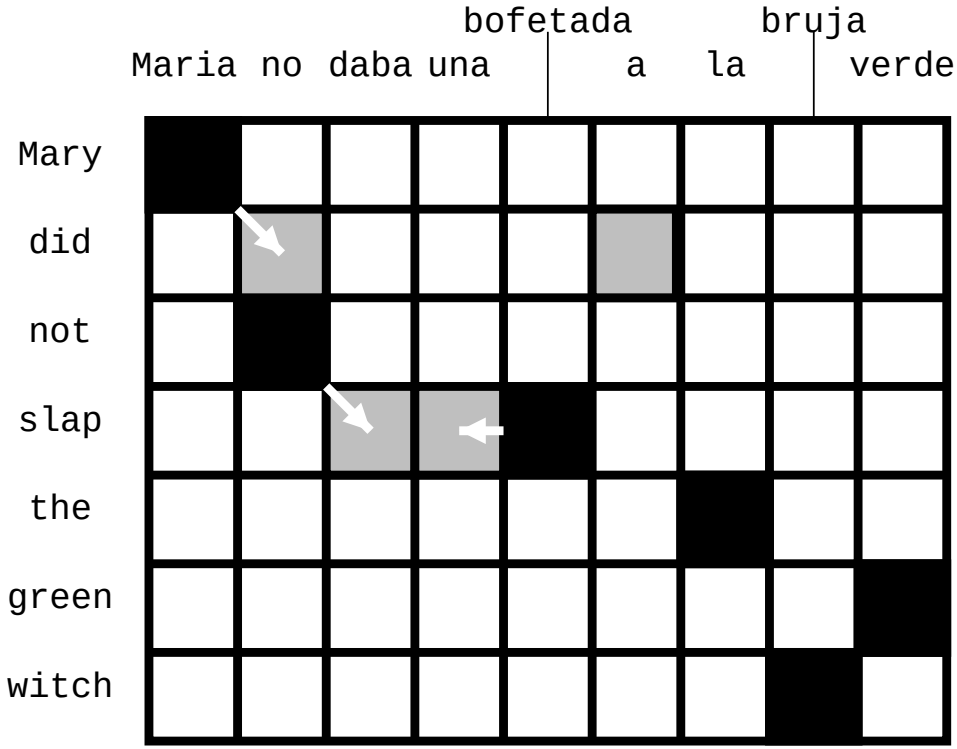
- Foreign "phrases" are translated into English.
- Phrases may be reordered.

Current models allow for many-to-one mappings → we can use those to induce many-to-many mappings

Intersecting alignments

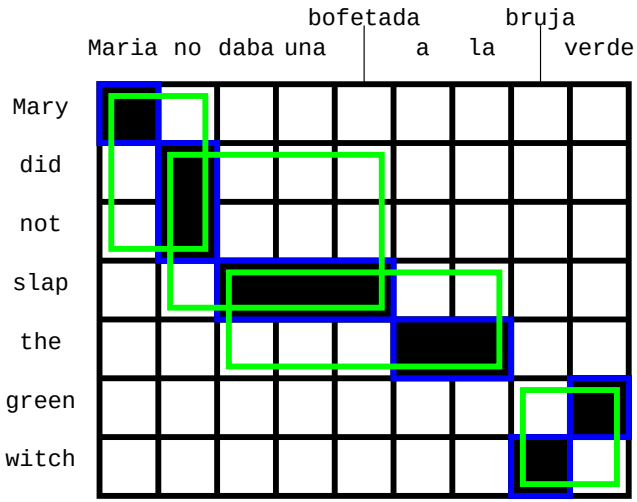


Growing alignments



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

Induced phrases



(Maria, Mary), (no, did not), (slap, daba una bofetada), (a la, the), (bruja, witch), (verde, green), (Maria no, Mary did not), (no daba una bofetada, did not slap), (daba una bofetada a la, slap the), (bruja verde, green witch)

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

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.

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

Further reading

Some of the examples are adapted from the following books:

- Doug J. Arnold, Lorna Balkan, Siety Meijer, R. Lee Humphreys and Louisa Sadler (1994). *Machine Translation: an Introductory Guide*. Blackwells-NCC, London. 1994. Available from http://www.essex.ac.uk/linguistics/clmt/MTbook/
- Jurafsky, Daniel, and James H. Martin (2000). *Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics*. Prentice-Hall. More info at http://www.cs.colorado.edu/~martin/slp.html.