Linguistics 384: Language and Computers

Topic 5: Machine Translation

Scott Martin[*]

Dept. of Linguistics, OSU
Winter 2008

[*] The course was created together with Chris Brew, Markus Dickinson and Detmar Meurers.

Outline

Introduction

Background: Dictionaries

Transformer approaches

Linguistic knowledge based systems

Machine learning based systems

What makes MT hard?

Evaluating MT systems

References

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
  - determining if certain words or ideas appear in suspected terrorist documents → help pin down which documents need to be looked at closely
- If you want your human translators to focus on interesting/difficult sentences while avoiding lookup of unknown words and translation of mundane sentences.

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 used to have 11 official languages, since May 1, 2004 it has 20. All federal laws and other documents have to be translated into all languages.

What is MT not good for?

- Things that require subtle knowledge of the world and/or a high degree of (literary) skill:
  - translating Shakespeare into Navaho
  - diplomatic negotiations
  - court proceedings
  - . . .
- Things that may be a life or death situation:
  - Pharmaceutical business
  - Automatically translating frantic 911 calls for a dispatcher who speaks only Spanish

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?'

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. → dictionaries
- Words are grouped into meaningful units (cf. our discussion of syntax for grammar checkers).
- Word order can differ from language to languge.
- The forms of words within a sentence are systematic, e.g., verbs have to be conjugated, etc.

Different approaches to MT

- Transformer systems
- Systems based on linguistic knowledge
  - Direct transfer systems
  - Interlinguas
- Machine learning approaches

Most of these use dictionaries in one form or another, so we will start by looking at dictionaries.

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: *have* is the dictionary entry, but we want the entry to specify how to conjugate this verb.

Language and Computers
Topic 5: Machine Translation
Introduction
Examples for Translations
Background: Dictionaries
Transformer approaches
Linguistic knowledge based systems
Direct transfer systems
Interlingua-based systems
Machine learning based systems
Alignment
What makes MT hard?
Evaluating MT systems
References

Dictionaries (cont.)

- contains (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 is *eating*, then X must be animate
- may also contain frequency information
- can be hierarchically organized, e.g.:
  - all nouns have person, number, and gender
  - verbs (unless irregular) conjugate in the past tense by adding *ed*.

Language and Computers
Topic 5: Machine Translation
Introduction
Examples for Translations
Background: Dictionaries
Transformer approaches
Linguistic knowledge based systems
Direct transfer systems
Interlingua-based systems
Machine learning based systems
Alignment
What makes MT hard?
Evaluating MT systems
References

What dictionary entries might look like

- WORD: *button*
  PART OF SPEECH: noun
  HUMAN: no
  CONCRETE: yes
  GERMAN: Knopf
- WORD: *knowledge*
  PART OF SPEECH: noun
  HUMAN: no
  CONCRETE: no
  GERMAN: Wissen, Kenntnisse
  - There can be extra rules which tell you whether to choose *Wissen* or *Kenntnisse*.

Language and Computers
Topic 5: Machine Translation
Introduction
Examples for Translations
Background: Dictionaries
Transformer approaches
Linguistic knowledge based systems
Direct transfer systems
Interlingua-based systems
Machine learning based systems
Alignment
What makes MT hard?
Evaluating MT systems
References

A dictionary entry with frequency

- WORD: *knowledge*
  PART OF SPEECH: noun
  HUMAN: no
  CONCRETE: no
  GERMAN: Wissen: 80%, Kenntnisse: 20%
- Probabilities can be derived from various machine learning techniques → to be discussed later.

Language and Computers
Topic 5: Machine Translation
Introduction
Examples for Translations
Background: Dictionaries
Transformer approaches
Linguistic knowledge based systems
Direct transfer systems
Interlingua-based systems
Machine learning based systems
Alignment
What makes MT hard?
Evaluating MT systems
References

Transformer approaches

- **Transformer** architectures transform example sentences from one language into another.
- They consist of
  - a grammar for the source/input language
  - a source-to-target language dictionary
  - source-to-target language rules
- Note that there is no grammar for the target language, only mappings from the source language.

Language and Computers
Topic 5: Machine Translation
Introduction
Examples for Translations
Background: Dictionaries
Transformer approaches
Linguistic knowledge based systems
Direct transfer systems
Interlingua-based systems
Machine learning based systems
Alignment
What makes MT hard?
Evaluating MT systems
References

An example for the transformer appraoch

We'll work through a German-to-English example.

(3) a. Drehen Sie den Knopf eine Position zurück.
    b. Turn the button back one position.

1. Using the grammar, assign parts-of-speech:

   (4) Drehen Sie den Knopf eine Position zurück.
       verb pron. article noun article noun prep.

2. Using the grammar, give the sentence a (basic) structure

   (5) Drehen Sie [den Knopf] [eine Position] zurück.

Language and Computers
Topic 5: Machine Translation
Introduction
Examples for Translations
Background: Dictionaries
Transformer approaches
Linguistic knowledge based systems
Direct transfer systems
Interlingua-based systems
Machine learning based systems
Alignment
What makes MT hard?
Evaluating MT systems
References

An example (cont.)

3. Using the dictionary, find the target language words

   (6) Drehen Sie [den Knopf] [eine Position] zurück.
       turn you the button one position back

4. Using the source-to-target rules, reorder, combine, eliminate, or add target language words, e.g.,
   - 'turn' and 'back' form one unit.
   - because 'Drehen . . . zurück' is a command, in English it is expressed without 'you'.

⇒ End result: *Turn back the button one position.*

Language and Computers
Topic 5: Machine Translation
Introduction
Examples for Translations
Background: Dictionaries
Transformer approaches
Linguistic knowledge based systems
Direct transfer systems
Interlingua-based systems
Machine learning based systems
Alignment
What makes MT hard?
Evaluating MT systems
References

Transformers: Less than meets the eye

- By their very nature, transformer systems are non-**reversible** because they lack a target language grammar.
  If we have a German to English translation system, for example, we are incapable of translating from English to German.
- However, as these systems do not require sophisticated knowledge of the target language, they are usually very **robust** = they will return a result for nearly any input sentence.

Language and Computers
Topic 5: Machine Translation
Introduction
Examples for Translations
Background: Dictionaries
Transformer approaches
Linguistic knowledge based systems
Direct transfer systems
Interlingua-based systems
Machine learning based systems
Alignment
What makes MT hard?
Evaluating MT systems
References

Linguistic knowledge-based systems

- Linguistic knowledge-based systems include knowledge of both the source and the target languages.
- We will look at direct transfer systems and then the more specific instance of interlinguas.
  - Direct transfer systems
  - Interlinguas

Direct transfer systems

A direct transfer systems consists of:

- A source language grammar
- A target language grammar
- Rules relating source language underlying representation to target language underlying representation

Language and Computers
Topic 5: Machine Translation
Introduction
Examples for Translations
Background: Dictionaries
Transformer approaches
Linguistic knowledge based systems
Direct transfer systems
Interlingua-based systems
Machine learning based systems
Alignment
What makes MT hard?
Evaluating MT systems
References

Direct transfer systems (cont.)

- 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:

(7) Londres plaît à Sam.
London is pleasing to Sam
'Sam likes London.'

Language and Computers
Topic 5: Machine Translation
Introduction
Examples for Translations
Background: Dictionaries
Transformer approaches
Linguistic knowledge based systems
Direct transfer systems
Interlingua-based systems
Machine learning based systems
Alignment
What makes MT hard?
Evaluating MT systems
References

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.

Language and Computers
Topic 5: Machine Translation
Introduction
Examples for Translations
Background: Dictionaries
Transformer approaches
Linguistic knowledge based systems
Direct transfer systems
Interlingua-based systems
Machine learning based systems
Alignment
What makes MT hard?
Evaluating MT systems
References

Things to note about transfer systems

- The transfer mechanism is essentially reversible; e.g., the *plaire* rule works in both directions (at least in theory)
- Because we have a separate target language grammar, we are able to ensure that the rules of English apply; *like* → *likes*.
- Word order is handled differently than with transformers: the URs are essentially unordered.
- The underlying representation can be of various levels of abstraction – words, syntactic trees, meaning representations, etc.; we will talk about this with the **translation triangle**.

Language and Computers
Topic 5: Machine Translation
Introduction
Examples for Translations
Background: Dictionaries
Transformer approaches
Linguistic knowledge based systems
Direct transfer systems
Interlingua-based systems
Machine learning based systems
Alignment
What makes MT hard?
Evaluating MT systems
References

Caveat about reversibility

- It seems like reversible rules are highly desirable—and in general they are—but we may not always want reversible rules.
  - e.g., Dutch *aanvangen* should be translated into English as *begin*, but English *begin* should be translated into Dutch as *beginnen*.

Language and Computers
Topic 5: Machine Translation
Introduction
Examples for Translations
Background: Dictionaries
Transformer approaches
Linguistic knowledge based systems
Direct transfer systems
Interlingua-based systems
Machine learning based systems
Alignment
What makes MT hard?
Evaluating MT systems
References

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.

Language and Computers
Topic 5: Machine Translation
Introduction
Examples for Translations
Background: Dictionaries
Transformer approaches
Linguistic knowledge based systems
Direct transfer systems
Interlingua-based systems
Machine learning based systems
Alignment
What makes MT hard?
Evaluating MT systems
References

Czech-English example

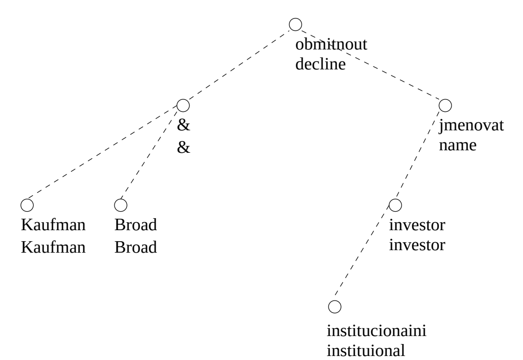
(8) Kaufman & Broad odmítla institucionální investory jmenovat.
Kaufman & Broad declined institutional investors to name/identify
'Kaufman & Broad refused to name the institutional investors.'

Example taken from Čmejrek, Cuřín, and Havelka (2003).

- They find the base forms of words (e.g., *obmidout* 'to decline' instead of *odmítla* 'declined')
- They find which words depend on which other words and represent this in a tree (e.g., the noun *investory* depends on the verb *jmenovat*)
- This dependency tree is then converted to English (comparative grammar) and re-ordered as appropriate.

Language and Computers
Topic 5: Machine Translation
Introduction
Examples for Translations
Background: Dictionaries
Transformer approaches
Linguistic knowledge based systems
Direct transfer systems
Interlingua-based systems
Machine learning based systems
Alignment
What makes MT hard?
Evaluating MT systems
References

Dependency tree for Czech-English example



Language and Computers
Topic 5: Machine Translation
Introduction
Examples for Translations
Background: Dictionaries
Transformer approaches
Linguistic knowledge based systems
Direct transfer systems
Interlingua-based systems
Machine learning based systems
Alignment
What makes MT hard?
Evaluating MT systems
References

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.
- What your interlingua looks like depends on your goals; an example for *I shot the sheriff.* is shown on the following slide.

Language and Computers
Topic 5: Machine Translation
Introduction
Examples for Translations
Background: Dictionaries
Transformer approaches
Linguistic knowledge based systems
Direct transfer systems
Interlingua-based systems
Machine learning based systems
Alignment
What makes MT hard?
Evaluating MT systems
References

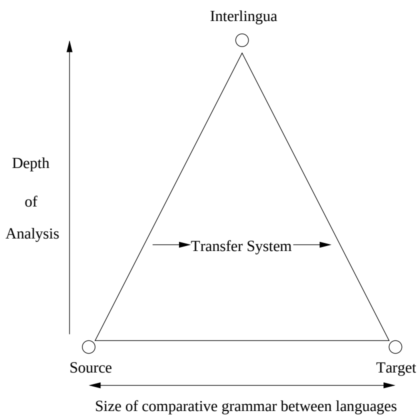
Interlingua example

- ACTION
  - *wound*
  - MEANS: *gun*
  - TENSE: *past*
  - KILL: *maybe*
  - WOUNDER
    - *speaker*
    - PERSON: *first*
    - NUMBER: *sg*
    - GENDER: *?*
  - WOUNDEE
    - *sheriff*
    - DEFINITE: *yes*
    - PERSON: *third*
    - NUMBER: *singular*
    - GENDER: *?*
    - HUMAN: *yes*
    - ANIMATE: *yes*
    - NOUN-TYPE: *kind of job*
    - IS-A-KIND-OF: *officer*

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'.

The translation triangle



Machine learning

- Instead of trying to tell the MT system how we're going to translate, we might try a **machine learning** approach = the computer will learn how to translate based on example translations.
- For this, we need
  - examples of translations as **training data**, and
  - a way of learning from that data.

Using frequency (statistical methods)

- 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? There are two different cases:
  - 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**

Text alignment

Sometimes humans have provided informative training data:

- sentence alignment
- word alignment

The process of text alignment can also be automated and then used to train an MT system.

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.

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 what a word's translation equivalent is.

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 difficulties

- Knowing how words align in the training data will not tell us how to handle the new data we see.
  - we may have many cases where *fool* is aligned with the Spanish *engañar* = 'to fool'
  - but we may then encounter *a fool*, where the translation should be *tonto* (male) or *tonta* (female)
- So, word alignment only helps us get some frequency numbers; we still have to do something intelligent with them.

Word alignment difficulties (cont.)

- Sometimes it is not even clear that word alignment is possible.

(9) Ivan aspirant.
Ivan graduate student
'Ivan is a graduate student.'

- What does *is* align with?
- In cases like this, a word can be mapped to a "null" element in the other language.

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.
2. Count up the number of times *On* is the translation = 2 times out of 6 = 1/3 = 0.33, or 33%.

Every other word has the probability 1/6 = 0.17, or 17%, so *On* is clearly the best translation for *He*.

What makes MT hard?

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

⇒ We have to know which meaning before we translate.

Language and Computers
Topic 5: Machine Translation
Introduction
Examples for Translations
Background: Dictionaries
Transformer approaches
Linguistic knowledge based systems
Direct transfer systems
Interlingua-based systems
Machine learning based systems
Alignment
What makes MT hard?
Evaluating MT systems
References

How words divide up the world (lexical issues)

Words don't line up exactly between languages.
Within a language, we have synonyms, hyponyms, and hypernyms.

- *sofa* and *couch* are synonyms (mean the same thing)
- *sofa* is a hyponym (more specific term) of *furniture*
- *furniture* is a hypernym (more general term) of *sofa*

Language and Computers
Topic 5: Machine Translation
Introduction
Examples for Translations
Background: Dictionaries
Transformer approaches
Linguistic knowledge based systems
Direct transfer systems
Interlingua-based systems
Machine learning based systems
Alignment
What makes MT hard?
Evaluating MT systems
References

Synonyms

Often we find **synonyms** between two languages (as much as there are synonyms within a language):

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

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

Language and Computers
Topic 5: Machine Translation
Introduction
Examples for Translations
Background: Dictionaries
Transformer approaches
Linguistic knowledge based systems
Direct transfer systems
Interlingua-based systems
Machine learning based systems
Alignment
What makes MT hard?
Evaluating MT systems
References

Hypernyms and Hyponyms

- 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*.

Language and Computers
Topic 5: Machine Translation
Introduction
Examples for Translations
Background: Dictionaries
Transformer approaches
Linguistic knowledge based systems
Direct transfer systems
Interlingua-based systems
Machine learning based systems
Alignment
What makes MT hard?
Evaluating MT systems
References

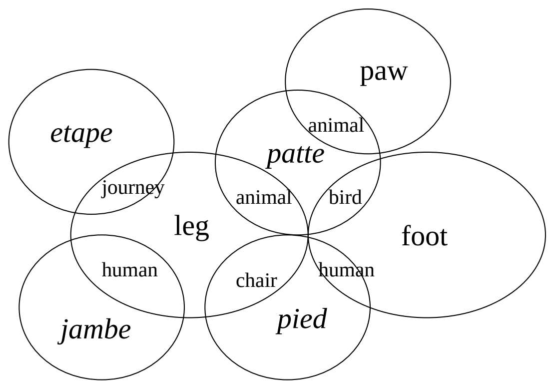
Semantic overlap

And then there's just fuzziness, as in the following English and French correspondences

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

Language and Computers
Topic 5: Machine Translation
Introduction
Examples for Translations
Background: Dictionaries
Transformer approaches
Linguistic knowledge based systems
Direct transfer systems
Interlingua-based systems
Machine learning based systems
Alignment
What makes MT hard?
Evaluating MT systems
References

Venn diagram of semantic overlap



Language and Computers
Topic 5: Machine Translation
Introduction
Examples for Translations
Background: Dictionaries
Transformer approaches
Linguistic knowledge based systems
Direct transfer systems
Interlingua-based systems
Machine learning based systems
Alignment
What makes MT hard?
Evaluating MT systems
References

Lexical gaps

Sometimes there is no simple equivalent for a word in a language, and the word has to be translated with a more complex phrase. We call this a **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.'

Language and Computers
Topic 5: Machine Translation
Introduction
Examples for Translations
Background: Dictionaries
Transformer approaches
Linguistic knowledge based systems
Direct transfer systems
Interlingua-based systems
Machine learning based systems
Alignment
What makes MT hard?
Evaluating MT systems
References

Light verbs

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*

Language and Computers
Topic 5: Machine Translation
Introduction
Examples for Translations
Background: Dictionaries
Transformer approaches
Linguistic knowledge based systems
Direct transfer systems
Interlingua-based systems
Machine learning based systems
Alignment
What makes MT hard?
Evaluating MT systems
References

Idioms

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*

Language and Computers
Topic 5: Machine Translation
Introduction
Examples for Translations
Background: Dictionaries
Transformer approaches
Linguistic knowledge based systems
Direct transfer systems
Interlingua-based systems
Machine learning based systems
Alignment
What makes MT hard?
Evaluating MT systems
References

Idiosyncracies

There are idiosyncratic choices among languages, e.g.:

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

Taboo words

There are **taboo words** = words which are "forbidden" in some way or in some circumstances (i.e., swear/curse words)

- You of course know several English examples. Note that the literal meanings of these words lack the emotive impact of the actual words.
- Other languages/cultures have different taboos: often revolving around death, body parts, bodily functions, disease, and religion.
  - e.g., The word 'skin' is taboo in a Western Australian (Aboriginal) language (http://www.aija.org.au/online/ICABenchbook/BenchbookChapter5.pdf)
  - Imagine encountering the word 'skin' in English and translating it without knowing this.

Structure and word order differences

- Word order (and syntactic structure) differs across languages.
- E.g., in English, we have what is called a subject-verb-object (SVO) order, as in (10).

(10) John punched Bill.
SUBJECT VERB OBJECT

- In contrast, Japanese is SOV. Arabic is VSO. Dyirbal (Australian aboriginal language) has free word order.
- MT systems have to account for these differences.

More on word order differences

- Sometimes things are conceptualized differently in different languages, e.g.:

(11) a. His name is Jerome.
b. Er heißt Jerome. (German)
He goes-by-name-of Jerome
c. Il s' appelle Jerome. (French)
He himself call Jerome.

- Words don't really align here.

How syntactic grouping and meaning relate (Syntax/Semantics)

Even within a language, there are syntactic complications. We can have **structural ambiguities** = sentences where there are multiple ways of interpreting it.

(12) John saw the boy (with the binoculars).

with the binoculars can refer to either *the boy* or to how John saw the boy.

- This difference in structure corresponds to a difference in what we think the sentence means, i.e., meaning is derived from the words and how they are grouped.
- Do we attempt to translate only one interpretation? Or do we try to preserve the ambiguity in the target language?

How language is used (Pragmatics)

Translation becomes even more difficult when we try to translate something in context.

- *Thank you* is usually translated as *merci* in French, but it is translated as *s'il vous plaît* 'please' when responding to an offer.
- *Can you drive a stick-shift?* could be a request for you to drive my manual transmission automobile, or it could simply be a request for information about your driving abilities.

Real-world knowledge

- Sometimes we have to use **real-world knowledge** to figure out what a sentence means.

(13) Put the paper in the printer. Then switch **it** on.

- We know what *it* refers to only because we know that printers, not paper, can be switched on.

Ambiguity resolution

- If the source language involves ambiguous words/phrases, but the target language does not have the same ambiguity, we have to resolve ambiguity before translation.
  e.g., the hyponyms/hypernyms we saw before.
- But sometimes we might want to preserve the ambiguity, or note that there was ambiguity or that there are a whole range of meanings available.
  ⇒ In the Bible, the Greek word *hyper* is used in 1 Corinthians 15:29; it can mean 'over', 'for', 'on behalf of', and so on. How you treat it affects how you treat the theological issue of salvation of the already dead. i.e., people care deeply about how you translate this word, yet it is not entirely clear what English meaning it has.

Evaluating MT systems

- We've seen some translation systems and we know that translation is hard.
- The question now is: How do we evaluate MT systems, in particular for use in large corporations as likely users?
  - How much change in the current setup will the MT system force?
  - How will it fit in with word processors and other software?
  - Will the company selling the MT system be around in the next few years for support and updates?
  - How fast is the MT system?
  - How good is the MT system (quality)?

Evaluating quality

- **Intelligibility** = how understandable the output is
- **Accuracy** = how faithful the output is to the input
- **Error analysis** = how many errors we have to sort through (and how do the errors affect intelligibility & accuracy)
- **Test suite** = a set of sentences that our system should be able to handle

Introduction
Examples for Translations
Background: Dictionaries
Transformer approaches
Linguistic knowledge based systems
Direct transfer systems
Interlingua-based systems
Machine learning based systems
Alignment
What makes MT hard?
Evaluating MT systems
References

Intelligibility

Intelligibility Scale (from Arnold et al., 1994)

1. The sentence is perfectly clear and intelligible. It is grammatical and reads like ordinary text.
2. The sentence is generally clear and intelligible. Despite some inaccuracies or infelicities of the sentence, one can understand (almost) immediately what it means.
3. The general idea of the sentence is intelligible only after considerable study. The sentence contains grammatical errors and/or poor word choices.
4. The sentence is unintelligible. Studying the meaning of the sentence is hopeless; even allowing for context, one feels that guessing would be too unreliable.

Language and Computers
Topic 5: Machine Translation
Introduction
Examples for Translations
Background: Dictionaries
Transformer approaches
Linguistic knowledge based systems
Direct transfer systems
Interlingua-based systems
Machine learning based systems
Alignment
What makes MT hard?
Evaluating MT systems
References

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.

Language and Computers
Topic 5: Machine Translation
Introduction
Examples for Translations
Background: Dictionaries
Transformer approaches
Linguistic knowledge based systems
Direct transfer systems
Interlingua-based systems
Machine learning based systems
Alignment
What makes MT hard?
Evaluating MT systems
References