Outline

1. Course Information

2. Introduction to Machine Translation
   - A few things that make MT difficult
   - Approaches to Machine Translation
Course Information
General information

• Formally:
  – Lecture (Vorlesung): Wednesday 14:15 – 15:45
  – Exercise (Übung): Tuesday 16:15 – 17:45

• However, I will be mostly doing lectures earlier in the semester on *BOTH DAYS*

• Schedule, lecture slides, videos posted on web page (see my homepage, Google: fraser CIS)

• NOTE: to receive a grade you must be registered for *both* Vorlesung and Übung, two of you have forgotten the Übung.
This course will look at machine translation:

- **Primarily from a computational side**
  - Understanding the challenges of modeling translation computationally
  - Basic understanding of rule-based machine translation
  - In-depth understanding of statistical machine translation
  - In-depth understanding of deep learning and neural machine translation
    (several guest lectures)

- **But also a little bit from a linguistic side**
  - Understanding the linguistic challenges of translation
  - Thinking about the implications of working with different language pairs
The statistical machine translation (SMT) material is mainly based on the book:

Koehn, Philipp (2009):
*Statistical Machine Translation*

There is unfortunately no good textbook for neural machine translation (NMT) yet, but we will point you to interesting blog posts.
Course Requirements: IN SS 2020

- Due to the current situation, group projects will not be possible.
- To pass this course ...
  - Exercises and assignments (bonus points for exam)
  - Written exam
Any questions about logistics, etc., before I briefly introduce machine translation?
1. Course Information

2. Introduction to Machine Translation
   
   A few things that make MT difficult
   
   Approaches to Machine Translation
Acknowledgements

The content of this lecture is based on a previous lecture by Chris Callison-Burch (probably with some new errors introduced by me).
What is machine translation?

- Automatic translation of text in one language to another language.
- Examples: Google Translate, Bing Translator, DeepL, many more
A few things that make MT difficult

Not an exhaustive list:

- POS ambiguity
- Word sense
- Word order
- Pronouns
- Tense
- Idioms
- etc…
In many languages, the POS of a word is ambiguous

- Consider translation of the word “fire” to German
- “The fire was large.”?
- “I will fire them.”?
Word sense ambiguity is a big problem for many NLP systems:

- “Bank” as in river
  “Bank” as in financial institution
- “Plant” as in a tree
  “Plant” as in a factory
- Different word senses often translate into different words in another language
Differing word orders

- English word order is: subject - verb - object
- Japanese order is: subject - object - verb
- English: IBM bought Lotus
- Japanese: IBM Lotus bought
- English: Reporters said IBM bought Lotus
- Japanese: Reporters IBM Lotus bought said
Problem of pronouns

Pronouns can be a big difficulty in translation:

- Some languages like Spanish can drop subject pronouns
- In Spanish the verbal inflection often indicates which pronoun should be restored
  -o = I
  -as = you
  -a = he / she / it
  -amos = we
  -an = they
- When should we use ‘she’ or ‘he’ or ‘it’?
- Think about translating “it” from English to German.
Different tenses

- Spanish has two versions of the past tense: one for a definite time in the past, and one for an unknown time in the past.
- When translating from English to Spanish we need to choose which version of the past tense to use.
Idioms

- "to kick the bucket" means "to die"
- "a bone of contention" does not have anything to do with skeletons
- "a lame duck", "tongue in cheek", "to cave in"
- etc...
• Word-for-word translation
• Syntactic transfer
• Semantic transfer
• Interlingual approaches
• Controlled language

• Example-based translation
• Statistical machine translation
• Neural machine translation
Word-for-word translation

- Use a machine-readable bilingual dictionary to translate each work in a text
- Advantages: Easy to implement, results give a rough idea about what the text is about
- Disadvantages: Problems with word order (and word sense) means that this results in low-quality translation
Syntactic transfer

- Parse the sentence
- Rearrange constituents
- Then translate the words
Syntactic Transfer

- Advantages:
  - Deals with the word-order problem
  - Components are reusable - can use English parser developed for English to French for a subsequent English to German system

- Disadvantages:
  - Must construct grammars for each language that you deal with
  - Sometimes there is a syntactic mismatch between languages
  - Example 1:
    English: The bottle floated into the cave
    Spanish: La botella entró a la cueva flotando
    = The bottle entered the cave floating
  - Example 2:
    Peter likes to swim — > Peter schwimmt gerne
Semantic Transfer

- The sentence is first converted into a source-language-specific logical form
  
  \[\text{John likes to swim} \Rightarrow \text{LIKE(SWIM(JOHN))}\]

- This logical form is translated to a target-language-specific logical form
  
  \[\text{LIKE(SWIM(JOHN))} \Rightarrow \text{GERNE(SCHWIMMEN(JOHN))}\]

- Then from the target-language logical form, we generate the text
  
  \[\text{GERNE(SCHWIMMEN(JOHN))} \Rightarrow \text{John schwimmt gerne}\]

- In these approaches, it is typical to use a syntactic analysis of the source sentence (parse tree) as an intermediate step.
- Sometimes a syntactic parse tree for the target is also created.
Assign a logical form to input sentences

John must not go = OBLIGATORY(NOT(GO(JOHN)))

John may not go = NOT(PERMITTED(GO(JOHN)))

Use this logical form to generate a sentence in another language
Interlingua

- **Advantages:**
  Single logical form means that we can translate between all languages and only write a parser/generator for each language once

- **Disadvantages:**
  Difficult to define a single logical form that covers all situations in all languages. English words in all capital letters probably won't cut it.
Controlled language

- Define a subset of a language which can be used to compose text to be translated
- Issue editorial guidelines that limit each word to only one word sense, and which forbid certain difficult constructions
- Apply syntactic transfer or interlingual approaches
- Famous example: Weather Reports
Controlled language

- Advantages: Results in more reliable, higher quality translation for subset of language that it deals with
- Disadvantages: Does not cover all language use, so can only be applied in limited settings
Example-based MT

- Uses a translation memory or parallel corpus as a starting point
- When a human translator types a sentence that is similar to one in the memory, it is retrieved
- Some rules/heuristics to change the sentence to match the new sentence
### Source

<table>
<thead>
<tr>
<th>A-t-on acheté les actions ou les biens des entreprises nationalisées?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quel était le genre de travaux exécutés aux termes de ces contrats?</td>
</tr>
<tr>
<td>Le recours est rejeté comme manifestement irrecevable</td>
</tr>
<tr>
<td>Les propositions ne seront pas mises en application maintenant.</td>
</tr>
<tr>
<td>La République française supportera ses propres dépens</td>
</tr>
<tr>
<td>Production domestique exprimée en pourcentage de l'utilisation domestique</td>
</tr>
<tr>
<td>La séance est ouverte à 2 heures.</td>
</tr>
</tbody>
</table>

### Translation

<table>
<thead>
<tr>
<th>Have the shares or properties of nationalized companies been purchased?</th>
</tr>
</thead>
<tbody>
<tr>
<td>What was the nature of the work performed under these contracts?</td>
</tr>
<tr>
<td>The action is dismissed as manifestly inadmissible</td>
</tr>
<tr>
<td>The proposal will not now be implemented.</td>
</tr>
<tr>
<td>France was ordered to bear its own costs</td>
</tr>
<tr>
<td>Domestic output as a % of domestic use</td>
</tr>
<tr>
<td>The House met at 2 p.m.</td>
</tr>
</tbody>
</table>
Example-based MT

- Advantages: Uses human translations which are higher quality than machine translations
- Disadvantages: May have limited coverage depending on the size of the translation memory, and flexibility of heuristics
Statistical machine translation

- Find most probable English sentence given a French sentence
- Probabilities are determined automatically by training a statistical model using a parallel corpus
Advantages:
- Has a way of dealing with lexical ambiguity
- Can deal with idioms that occur in the training data
- Requires minimal human effort
- Can be created for any language pair that has enough training data

Disadvantages:
- Requires plentiful parallel training data
- Does not explicitly deal with syntax (but later work on this)
- Complex pipeline, can be computationally expensive to translate new sentences
- Can be difficult to understand decision process
Neural machine translation

- Find most probable English sentence given a French sentence
- Probabilities are determined automatically by training a statistical model using a parallel corpus
- Model is implemented using a neural network
Neural machine translation

Neural machine translation is a new form of statistical machine translation, relying on neural networks, but for convenience we tend to refer to the two as distinct.

- Advantages:
  - Has a better way of dealing with lexical ambiguity
  - Can deal with idioms that occur in the training data
  - Requires minimal human effort
  - Can be created for any language pair that has enough training data
  - Simple pipeline
  - Seems to work better than previous statistical machine translation approaches

- Disadvantages:
  - Requires plentiful parallel training data
  - Expensive to train (requires heavy computing resources and/or specialized processors)
  - Very very difficult (but probably not impossible?) to understand decision process
Conclusion

I hope to have convinced you that Machine Translation is an interesting problem!

In this introduction I presented:

- Some basic linguistic problems in machine translation
- An overview of previous approaches to machine translation

In future lectures:

- We will see a little bit more about linguistic problems and previous approaches to machine translation
- We will go into much more detail in terms of statistical and neural machine translation
Thank you for your attention.