Orientation and Introduction to Machine Translation

Alexander Fraser
fraser@cis.uni-muenchen.de

CIS, Ludwig-Maximilians-Universität München

Erweiterungsmodul: Machine Translation
SoSe 2017
2017-04-26

Outline

1. Course Information

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

Course Information

General information

- Lecture (Vorlesung): Wednesay 14:15 – 15:45 here
- Exercise (Übung): Tuesday 16:15 – 17:45 in room C003 (sometimes in computer pool)
- There will probably not be a strict separation of lectures and exercises
- Schedule and lecture slides posted on web page (see my homepage, Google: fraser CIS)

Course information

Contents and goals of this course

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
  - Introduction to deep learning and neural machine translation
- **But also somewhat from a linguistic side**
  - Understanding the linguistic challenges of translation
  - Thinking about the implications of working with different language pairs

Who is who

- **Alexander Fraser**
  - Dr. Fraser is a permanent staff member at CIS (and coordinator of the Masters program), leads three large research projects on machine translation
  - Working in all areas of (mostly machine learning based) machine translation, also on other structured prediction problems
- **Fabienne Braune and Matthias Huck**
  - Dr. Fabienne Braune: word embeddings, deep learning, recurrent neural networks
  - Matthias Huck: both statistical and neural machine translation, morphology in neural MT, other topics
- **Tutor: Costanza Conforti**
  - Costanza Conforti will be the tutor for this course

Course material

The course material is mainly based on the book:

- **Theoretical background**

  Koehn, Philipp (2009):
  Statistical Machine Translation

We will also look at the open source toolkit Moses in the exercises

Course Requirements

- To pass this course ...
  - Exercises and assignments
  - Regular attendance
  - Course project: implementation of a small project including extensive documentation; presentation
    - Roughly last 5-6 weeks of semester
    - Programming and/or data analysis intensive
    - Short presentation

Questions?

Any questions about logistics, etc., before I briefly introduce machine translation?

Outline

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 yours truly)

What is machine translation?

- Automatic translation of text in one language to another language.
- Examples: Systran Babelfish, Google Translate, Bing Translator, many more

A few things that make MT difficult

Not an exhaustive list:

- POS ambiguity
- Word sense
- Word order
- Pronouns
- Tense
- Idioms
- etc...

POS ambiguity

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

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

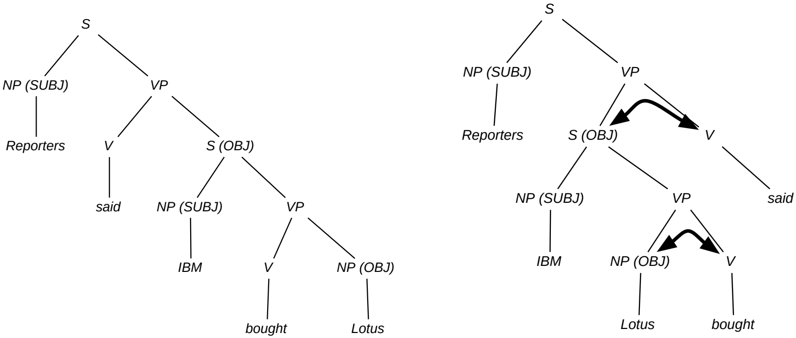
- "to kick the bucket” means "to die”
- "a bone of contention" does not have anything to do with skeletons
- "a lame duck", "tongue in cheek", "to cave in"
- etc...

- Word-for-word translation
- Syntactic transfer
- 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:

    English: The bottle floated into the cave

    Spanish: La botella entró a la cueva flotando

    = The bottle entered the cave floating

Interlingua

- 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

Parallel corpus

Source	Translation
A-t-on acheté les actions ou les biens des entreprises nationalisées?	Have the shares or properties of nationalized companies been purchased?
Quel était le genre de travaux exécutés aux termes de ces contrats?	What was the nature of the work performed under these contracts?
Le recours est rejeté comme manifestement irrecevable	The action is dismissed as manifestly inadmissible
Les propositions ne seront pas mises en application maintenant.	The proposal will not now be implemented.
La République française supportera ses propres dépens	France was ordered to bear its own costs
Production domestique exprimée en pourcentage de l'utilisation domestique	Domestic output as a % of domestic use
La séance est ouverte à 2 heures.	The House met at 2 p.m.
. . .	. . .

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

Statistical machine translation

- 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.