CS5740: Natural Language Processing

Machine Translation

Instructor: Yoav Artzi

Some slides adapted from Michael Collins

Overview

- Challenges in machine translation (MT)
- Classical MT
- Statistical MT (very briefly)
- MT evaluation

Challenges: Lexical Ambiguity

Book the flight → reservar

Read the **book** → libro

Kill a man → matar

Kill a process → acabar

Examples from Dorr et al. 1999

Challenges: Differing Word Order

- English: subject-verb-object
- Japanese: subject-object-verb

English: IBM bought Lotus
“Japanese”: IBM Lotus bought

English: Sources said that IBM bought Lotus yesterday
“Japanese”: Sources yesterday IBM Lotus bought that said

Syntactic Structure is not Always Preserved

The bottle floated into the cave

La botella entro a la cuerva flotando
(the bottle entered the cave floating)

Examples from Dorr et al. 1999

Syntactic Ambiguity Causes Problems

John hit the dog with the stick

John golpeo el perro [con palo / que tenia el palo]

Examples from Dorr et al. 1999

Pronoun Resolution

The computer outputs the data; it is fast.

La computadora imprime los datos; **es** rapida.

The computer outputs the data; it is stored in ascii.

La computadora imprime los datos; **estan** almacendos en ascii.

Classical I: Direct MT

- Translation is word-by-word
- Very little analysis of source text – no syntax, no semantics
- Relies on large bilingual dictionary:
  - For each word in the source language, specifies a set of translation rules
- After words are translated, simple re-ordering rules are applied
  - Example: move adjectives after nouns when translating from English to French

Classical I: Direct MT

- Rules for translating *much* or *many* into Russian:

if preceding word is *how* **return** *skol'ko*
else if preceding word is *as* **return** *stol'ko zhe*
else if word is *much*
    if preceding word is *very* **return** nil
    else if following word is a noun **return** *mnogo*
else (word is many)
    if preceding word is a preposition and following word is noun **return** *mnogii*
    else return *mnogo*

(From Jurafsky and Martin, edition 2, chapter 25. Originally from a system from Panov 1960)

Classical I: Direct MT

- Lack of analysis of source language causes problems:
  - Difficult to capture long-range orderings

English:	Sources said that IBM bought Lotus yesterday
Japanese:	Sources yesterday IBM Lotus bought that said

  - Words are translated without disambiguation of their syntactic role

e.g., *that* can be a complementizer or determiner, and will often be translated differently for these two cases

They said that ...
They like that ice-cream

Classical II: Transfer-based Approaches

- Three phases in translation:
  - Analysis of the source language sentence
    - Example: build a syntactic analysis of the source language sentence
  - Transfer (convert) the source-language parse tree to a target-language parse tree
  - Generation: Convert the target-language parse tree to an output sentence

Classical III: Interlingua-based Translation

- Two phases:
  - Analysis of the source language sentence into a (language-independent!) representation of its meaning
  - Generation of the output sentence from the meaning representation

Classical III: Interlingua-based Translation

- Advantage: if we need to translate between *n* languages, need only *n* analysis and generation systems.
  - In transfer systems, would need n^2
- Disadvantage: what would a language-independent representation look like?

Classical III: Interlingua-based Translation

- How to represent different concepts in an interlingua?
- Different languages break down concepts in quite different ways:
  - German has two words for wall: one for an internal wall, one for a wall that is outside
  - Japanese has two words for brother: one for an elder brother, one for a younger brother
  - Spanish has two words for leg: pierna for a human's leg, pata for an animal's leg, or the leg of a table
- A simple intersection of these different ways of breaking down concepts is not satisfactory
  - And very hard to design

Data

- Parallel corpora are available in multiple language pairs
- Basic idea: use a parallel corpus as a training set of translation examples
- Classic example: IBM work on French-English translation using Candian Hansards (1.7M pairs)
- Idea goes back to Warren Weaver's (1949) suggestion to use cryptanalytic techniques

... one naturally wonders if the problem of translation could conceivably be treated as a problem in cryptography. When I look at an article in Russian, I say: “This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode.”

Warren Weaver, 1949,

in a letter to Norbert Wiener

The Noisy Channel Model

- Goal: translate from French to English
- Have a model $p(e|f)$ to estimate the probability of an English sentence e given a French sentence f
- Estimate the parameters from training corpus
- A noisy channel model has two components:

  $p(e)$ the language model

  $p(f|e)$ the translation model
- Giving:

$$p(e|f) = \frac{p(e,f)}{p(f)} = \frac{p(e)p(f|e)}{\sum_e p(e)p(f|e)}$$

and

$$\arg\max_e p(e|f) = \arg\max_e p(e)p(f|e)$$

Example

- Translating from Spanish to English

Que hombre tengo yo

What hunger have	$p(s\|e) = 0.000014$
Hungry I am so	$p(s\|e) = 0.000001$
I am so hungry	$p(s\|e) = 0.0000015$
Have I that hunger	$p(s\|e) = 0.000020$

(From Koehn and Knight tutorial)

Example

- Translating from Spanish to English

Que hombre tengo yo

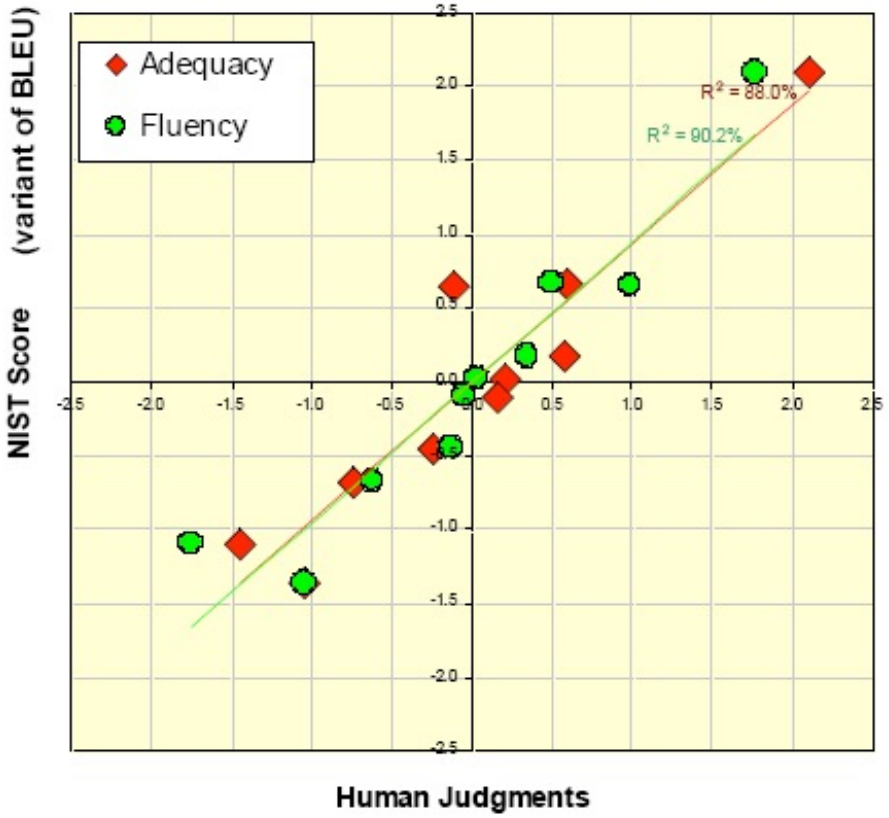
What hunger have	$p(s\|e)p(e) = 0.000014 \; x \; 0.000001$
Hungry I am so	$p(s\|e)p(e) = 0.000001 \; x \; 0.0000014$
I am so hungry	$p(s\|e)p(e) = 0.0000015 \; x \; 0.0001$
Have I that hunger	$p(s\|e)p(e) = 0.000020 \; x \; 0.00000098$

(From Koehn and Knight tutorial)

Automatic Evaluation

- Human evaluations: subjective measures, fluency/adequacy
- Automatic measures: n-gram match to references
  - NIST measure: n-gram recall (worked poorly)
  - BLEU: n-gram precision (no one really likes it, but everyone uses it)
- BLEU:
  - P1 = unigram precision
  - P2, P3, P4 = bi-, tri-, 4-gram precision
  - Weighted geometric mean of P1-4
  - Brevity penalty (why?)
  - Somewhat hard to game…

Correlation with Human Evaluataion



slide from G. Doddington (NIST)