WSTA 20: Machine Translation

- Introduction
  - examples
  - applications
- Why is MT hard?
- Symbolic Approaches to MT
- Statistical Machine Translation
  - Bitexts
- Computer Aided Translation

Slides adapted from: Steven Bird
Machine Translation Uses

- **Fully automated translation**
  - Informal translation, *gisting*
    - Google & Bing translate
    - Cross-language information retrieval
  - Translating technical writing, literature
    - Manuals
    - Proceedings
  - Speech-to-speech translation

- **Computer aided translation**
Introduction

- Classic “hard-AI” challenge, natural language understanding
- Goal: Automate of some or all of the task of translation.
  - Fully-Automated Translation
  - Computer Aided Translation
- What is "translation"?
  - Transformation of utterances from one language to another that preserves "meaning".
- What is "meaning"?
  - Depends on how we intend to use the text.
Why is MT hard: Lexical and Syntactic Difficulties

- One word can have multiple translations
  - know: Fr: savoir or connaitre
- Complex word overlap
- Words with many senses, no translation, idioms
- Complex word forms
  - e.g., noun compounds, ‘Kraftfahrzeug’ = power + drive + machinery
- Syntactic structures differ between languages
  - SVO, SOV, VSO, OVS, OSV, VOS (V=verb, S=subject, O=object)
  - Free word order languages
- Syntactic ambiguity
  - resolve in order to do correct translation
### Why is MT hard: Grammatical Difficulties

- **E.g. Fijian Pronoun System**
  - INCL = includes hearer, EXCL = excludes hearer

<table>
<thead>
<tr>
<th></th>
<th>SNG</th>
<th>DUAL</th>
<th>TRIAL</th>
<th>PLURAL</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1P EXCL</strong></td>
<td>au</td>
<td>keirau</td>
<td>keitou</td>
<td>keimami</td>
</tr>
<tr>
<td><strong>1P INCL</strong></td>
<td></td>
<td>kedaru</td>
<td>kedatou</td>
<td>keda</td>
</tr>
<tr>
<td><strong>2P</strong></td>
<td>iko</td>
<td>kemudrau</td>
<td>kemudou</td>
<td>kemunii</td>
</tr>
<tr>
<td><strong>3P</strong></td>
<td>koya</td>
<td>irau</td>
<td>iratou</td>
<td>ira</td>
</tr>
</tbody>
</table>

**cf English:**

- I
- we
- you
- you
- he, she, it
- they
Why is MT hard: Semantic and Pragmatic Difficulties

- Literal translation does not produce fluent speech:
  - Ich esse gern: *I eat readily*.
  - La botella entro a la cueva flotando: *The bottle entered the cave floating*.

- Literal translation does not preserve semantic information
  - e.g., "I am full" translates to "I am pregnant" in French.
  - literal translation of slang, idioms

- Literal translation does not preserve pragmatic information.
  - e.g., focus, sarcasm
Symbolic Approaches to MT

1. Direct Translation
   - English (word string) → French (word string)

2. Syntactic Transfer
   - English (syntactic parse) → French (syntactic parse)

3. Semantic Transfer
   - English (semantic representation) → French (semantic representation)

4. Knowledge-based Transfer
   - Interlingua (knowledge representation) → English (semantic representation)
   - Interlingua (knowledge representation) → French (semantic representation)
   - Interlingua (knowledge representation) → English (syntactic parse)
   - Interlingua (knowledge representation) → French (syntactic parse)
Difficulties for symbolic approaches

- **Machine translation should be robust**
  - Always produce a sensible output
  - even if input is anomalous

- **Ways to achieve robustness:**
  - Use robust components (robust parsers, etc.)
  - Use fallback mechanisms (e.g., to word-for-word translation)
  - Use statistical techniques to find the translation that is *most likely* to be correct.

- **Fallen out of use...**
  - symbolic MT efforts largely dead (except SYSTRANS)
  - from 2000s, field has moved to statistical methods
Statistical MT

- Noisy Channel Model
  - *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)
    - Assume that we *started* with an English sentence.
    - The sentence was then corrupted by translation into French.
    - We want to recover the original.
  - Use Bayes' Rule:
    \[
    P(e|f) = \frac{P(e)P(f|e)}{P(f)}
    \]
    \[
    \hat{e} = \arg\max_e P(e)P(f|e)
    \]
Statistical MT (cont)

\[ \hat{e} = \arg\max_e P(e)P(f|e) \]

- **Two components:**
  - \( P(e) \): Language Model
  - \( P(f|e) \): Translation Model

- **Task:**
  - \( P(f|e) \) rewards good translations
    - but permissive of disfluent \( e \)
  - \( P(e) \) rewards \( e \) which look like fluent English
    - helps put words in the correct order

- **Estimate** \( P(f|e) \) **using a parallel corpus**
  - \( e = e_1 \ldots e_p, f = f_1 \ldots f_m \)
  - alignment: \( f_j \) is the translation of which \( e_i \)?
  - content: which word is selected for \( f_j \)?
Noisy Channel example

Bilingual Corpora
French/English

Monolingual Corpora
English

Statistical Translation table

Statistical Language Model

French

I don't want to work
I no will work
...

I not work
I do not work

Je ne veux pas travailler

English

Slide from Phil Blunsom
Benefits of Statistical MT

• **Data-driven**
  - Learns the model directly from data
  - More data = better model

• **Language independent (largely)**
  - No need for expert linguists to craft the system
  - Only requirement is parallel text

• **Quick and cheap to get running**

See GIZA++ and Moses toolkits, http://www.statmt.org/moses/
Parallel Corpora: Bitexts and Alignment

- **Parallel texts (or bitexts)**
  - one text in multiple languages
  - Produced by human translation; readily available on web
    - news, legal transcripts, literature, subtitles, bible, ...

- **Sentence alignment:**
  - translators don't translate each sentence separately
    - 90% of cases are 1:1, but also get 1:2, 2:1, 1:3, 3:1
  - Which sentences in one language correspond with which sentences in another?

- **Algorithms:**
  - Dictionary-based
  - Length-based (Church and Gale, 1993)
Representing Alignment

- **Representation:**

  \[ e = e_1 \ldots e_l = \text{And the program has been implemented} \]
  \[ f = f_1 \ldots f_m = \text{Le programme a été mis en application} \]
  \[ a = a_1 \ldots a_m = 2,3,4,5,6,6,6 \]

Figure from Brown, Della Pietra, Mercer, 1993
Estimating $P(f|e)$

- If we know the alignments this can be easy
  - assume translations are *independent*
  - assume word-alignments are *observed* (given)
- Simply count frequencies:
  - e.g., $p(\text{programme} | \text{program}) = \frac{c(\text{programme}, \text{program})}{c(\text{program})}$
  - aggregating over all aligned word pairs in the corpus
- However, word-alignments are rarely observed
  - have to infer the alignments
  - define probabilistic model and use the Expectation-Maximisation algo
  - akin to unsupervised training in HMMs
Assume simple model, aka ‘IBM model 1’

\[ p(f, a | e) = \frac{\epsilon}{(l + 1)^m} \prod_{j=1}^{m} t(f_j | e \alpha_j) \]

- length of result independent of length of source, \( \epsilon \)
- alignment probabilities depend only on length of target, \( l \)
- each word translated from aligned word

Learning problem: estimate ‘t’ table of translations from

- instance of “expectation maximization” (EM) algorithm
  1. make initial guess of ‘t’ parameters, e.g., uniform
  2. estimate alignments of corpus \( p(a | f, e) \)
  3. learn new t values, using corpus frequency estimates
  4. repeat from step 2
Modelling problems

- Problems with this model:
  - ignores the positions of words in both strings (solution: HMM)
  - need to develop a model of alignment probabilities
  - tendency for proximity across the strings, and for movements to apply to whole blocks

- More general issues:
  - not building phrase structure, not even a model of source language P(f)
  - idioms, non-local dependencies
  - sparse data (solution: using large corpora)

Figure from Brown, Della Pietra\(^2\), Mercer, 1993
Word- and Phrase-based MT

- Typically use different models for *alignment* and *translation*
  - word based translation can be used to solve for best translation
  - overly simplistic model, makes unwarranted assumptions
  - often words translated and move in ‘blocks’
- Phrase based MT
  - treats n-grams as translation units, referred to as ‘phrases’ (not linguistic phrases though)
  - phrase-pairs memorise:
    - common translation fragments
    - common reordering patterns
  - architecture underlying Google & Bing online translation tools
Decoding

\[ e^* = \arg \max_e f(e, f) \]

- **Objective**

- **Where model, \( f \), incorporates**
  - translation probability, \( P(f|e) \)
  - language model probability, \( P(e) \)
  - distortion cost based on word reordering (translations are largely left-to-right, penalise big ‘jumps’)
  - ...

- **Search problem**
  - find the translation with the best overall score
Score the translations based on translation probabilities (step 2), reordering (step 3) and language model scores (steps 2 & 3).
Search problem

• **Given options**

<table>
<thead>
<tr>
<th><strong>er</strong></th>
<th><strong>geht</strong></th>
<th><strong>ja</strong></th>
<th><strong>nicht</strong></th>
<th><strong>nach</strong></th>
<th><strong>hause</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>he</td>
<td>is</td>
<td>yes</td>
<td>not</td>
<td>after</td>
<td>house</td>
</tr>
<tr>
<td>it</td>
<td>are</td>
<td>is</td>
<td>do not</td>
<td>to</td>
<td>home</td>
</tr>
<tr>
<td>, it</td>
<td>goes</td>
<td>, of course</td>
<td>does not</td>
<td>according to</td>
<td>chamber</td>
</tr>
<tr>
<td>, he</td>
<td>go</td>
<td></td>
<td>is not</td>
<td>in</td>
<td>at home</td>
</tr>
</tbody>
</table>

• **1000s of possible output strings**
  - he does not go home
  - it is not in house
  - yes he goes not to home …

• **Millions of possible translations for this short example…**

Figure from Koehn, 2009
Search insight

• Consider the sorted list of all derivations
  – he does not go after home
  – he does not go after house
  – he does not go home
  – he does not go to home
  – he does not go to house
  – he does not goes home

• Many similar derivations
  – can we avoid redundant calculations?
Dynamic Programming Solution

• **Instance of Viterbi algorithm**
  - factor out repeated computation (like Viterbi for HMMs, chart used in parsing)
  - efficiently solve the maximisation problem

• **What are the key components for “sharing”?**
  - don’t have to be exactly identical; need same:
    - set of translated words
    - righter-most output words
    - last translated input word location
Phrase-based Decoding

Start with empty state

Figure from Koehn, 2009
Phrase-based Decoding

Expand by choosing input span and generating translation

Figure from Koehn, 2009
Phrase-based Decoding

Consider all possible options to start the translation

Figure from Koehn, 2009
Phrase-based Decoding

Continue to expand states, visiting uncovered words. Generating outputs left to right.
Phrase-based Decoding

Read off translation from best complete derivation by backtracking

he goes home
are does not go home
it to
Complexity

• Search process is intractable
  – word-based and phrase-based decoding is NP complete (Knight 99)

• Complexity arises from
  – reordering model allowing all permutations
    solution: allow no more than 6 uncovered words
  – many translation options
    solution: no more than 20 translations per phrase
  – coverage constraints, i.e., all words to be translated once
MT Evaluation

- **Human evaluation of MT**
  - quantifying fluency and faithfulness
  - expensive and very slow (takes months)
  - but MT developers need to re-evaluate daily
  - thus evaluation is a bottleneck for innovation

- **BLEU: bilingual evaluation understudy**
  - data: corpus of reference translations
    - there are many good ways to translate the same sentence
  - translation closeness metric
    - weighted average of variable length phrase matches between the MT output and a set of professional translations
  - correlates highly with human judgements
MT Evaluation Example

- Two candidate translations from a Chinese source:
  - It is a guide to action which ensures that the military always obeys the commands of the party.
  - It is to insure the troops forever hearing the activity guidebook that party direct.

- Three reference translations
  - It is a guide to action that ensures that the military will forever heed Party commands.
  - It is the guiding principle which guarantees the military forces always being under the command of the Party.
  - It is the practical guide for the army always to heed the directions of the party.

- The BLEU metric has had a huge impact on MT
  - e.g. NIST Scores: Arabic->English 51% (2002), 89% (2003)
Summary

- Applications
- Why MT is hard
- Early symbolic motivations
- Statistical approaches
  - alignment
  - decoding
- Evaluation
- Reading
  - Either JM #25 or MS #13