It's funny because it's true. Literally?
CS460/626: Natural Language Processing/Speech, NLP and the Web (Lecture 9–Machine Translation)

Somya Gupta

Under the guidance of Prof. Pushpak Bhattacharyya

CSE Dept., IIT Bombay
19th Jan, 2012
I’ll let you in on a little secret

study of linguistics makes me fret

Back to English?

I studied linguistics

a little secret makes teasing me
Excerpt from an English Novel:

Ms. X: God is omnipresent
Mr. Y: I don’t believe it exists
Ms. X: it?

Hindi Translation:

Ram: उसके मामा के चाचा की मौसी और नानी दिल्ली में रहते हैं

John: Excuse me?

English Translation:

Of Nephews, nieces, cousins, aunts and uncles!
Machine Translation

• Machine translation (MT) is the use of computers to automate the production of translations from one natural language into another, with or without human assistance (Hutchins and Somers, 92)

• Translated text should have two desired properties:
  – **Adequacy**: Meaning should be conveyed
  – **Fluency**: Text should be fluent in the target language
Motivation

According to SIL International, There are 6,909 spoken languages, as catalogued and described in the book Languages of the World.

Uses:  - Judicial Domain – work going on at IITB
    - Internet mostly in English, reach out to masses?
    - Lots of monolingual data, sharing of information through translation
    - Business channels
    - Other Experiences?

Translation – A difficult task
Translators – Expensive, Take time
Contents

• Why is MT difficult?
• Taxonomy of MT Systems
• MT Approaches – Vauquois Triangle
• POS Tagging for MT
• References and further reading
Why is MT difficult? - I

- Lexical Ambiguity - *Bank*
- Differing Word Orders – *SVO, SOV*
- Syntactic Structure is not Preserved Across Translations
- Syntactic Ambiguity Causes Problems – *I saw the boy with a telescope*
- Pronoun Resolution – *The King appointed Birbal. He was a genius.*
- Boundary Friction
Why is MT difficult? - II

• Some of these are difficult problems also for human translators.

• Many require real-world knowledge, intuitions about the meaning of the text, etc. to get a good translation.

• Existing MT systems opt for a strategy of structure-preservation where possible, and do what they can to get lexical choices right.
Ram saw Shyam

- Word to word translation?
- Add a reordering rule? SVO -> SOV
- Identify Subject, Object – Case markers
- Morphology?
- Exhaustive and consistent set of rules possible? Language Divergence!
- Data to the rescue?
Taxonomy

Machine Translation (MT)

Direct MT  Rule-Based MT  Data-Driven MT

Transfer  Interlingua  EBMT  SMT
MT Approaches – Vauquois Triangle

The Vauquois triangle.
Direct Approach

- English Sentence: I will go to the market
- Morphological Analysis: I go FUTURE to market
- Constituent Identification: <I> <go FUTURE> <to market>
- Reorder: <I> <to market> <go FUTURE>
- Dictionary Lookup: <मैं> <बाज़ार> <जाना FUTURE>
- Inflect: मैं बाज़ार जाऊँगा

- Too focused on individual words
- Only morphological analysis (GNPTAM) is done
- Problem dealing with real examples requires phrasal and structural knowledge
Transfer Approach

- Involves three stages:
  - Source representation
  - Transfer
  - Target representation

There was a book on the table.

There was a book on the table.

मेज़ पर एक किताब थी
Interlingua Based Approach
Dave, Parikh and Bhattacharyya, 2002

- Interlingua based approach eliminates transfer altogether by creating a language independent canonical form known as an *interlingua*.

- Represent all sentences that mean the same thing in the same way independent of language.

- Avoids explicit descriptions of the relationship between source and target language; rather it uses abstract elements, like Agent, Event, Tense, etc.
Example Based MT

- First developed in contrast with rule-based MT
- Idea of translation by analogy (Nagao 1984)
  - Translate by adapting previously seen examples rather than by linguistic rule
  - Translate a sentence by using the closest match in parallel data

**Input:** He buys a book on international politics

**Matches**

- **He buys** mangoes from the market
- **This is a matter of** international politics
- **They read** a book

**Result:** वह अंतरराष्ट्रीय राजनीति पर एक किताब खरीदता है
Statistical MT

“Every time I fire a linguist, my system’s performance improves”

(Brown et al. 1988)

• Goal is to find out the English sentence $e$ given foreign language sentence $f$ whose $p(e|f)$ is maximum.

$$
\hat{e} = \arg\max_{e \in e^*} p(e|f) = \arg\max_{e \in e^*} p(f|e)p(e)
$$

• Translations are generated on the basis of statistical model
• Parameters are estimated using bilingual parallel corpora
MT Approaches

• Advantages of techniques discussed?
• Limitations of each?

• Empiricist vs. Rationalist ideology?

• What will work? Hybrid?
Some **POS Tagging** could really help in clearing doubts here!
Power of POS Tagging - I

• For machine translation, the accuracy of POS tagging is a crucial pre-processing step for a high-quality translation

• We cannot move up the Vauquois Triangle without POS Tagging

• Word sense disambiguation
• Parsing – Syntax
• Semantics
Power of POS Tagging - II

• An example of a sentence where POS tags disambiguate the meaning:

   Time flies like an arrow but fruit flies like a banana

   Noun Verb Prep Art Noun Conj Adj Noun Verb Art Noun

• Distinguishing proper nouns is especially important for information extraction and machine translation.
Statistical MT and POS Tagging

Factored PB-SMT
I (Phrase-based) SMT: no linguistic analysis
Factored Phrase Based SMT
Koehn 2007

• No Linguistic analysis in Phrase Based SMT
  – e.g., no notion of morphology
    • treat “look”, “looks”, “looked” as completely different words!
    • in learning: knowing “look” doesn’t help to translate “looks”
    • works fine for English (and reasonable amount of data)
  – problem: morphologically rich languages!

• Idea:
  – include morphological analysis (lemma + morph. features)
  – translate lemmas, translate features
  – I generate target surface forms
In this model the translation process is broken up into the following three mapping steps:

1. **Translate input lemmas into output lemmas**
   \{ house, home, building, shell \}

2. **Translate morphological and POS factors**
   \{ house\_NN\_plural, home\_NN\_plural, building\_NN\_plural, shell\_NN\_plural, house\_NN\_singular \}

3. **Generate surface forms given the lemma and linguistic factors**
   \{ houses\_house\_NN\_plural, homes\_home\_NN\_plural, buildings\_building\_NN\_plural, shells\_shell\_NN\_plural \}

---

<table>
<thead>
<tr>
<th>Model</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>best published result</td>
<td>18.15%</td>
</tr>
<tr>
<td>baseline (surface)</td>
<td>18.04%</td>
</tr>
<tr>
<td>surface + POS</td>
<td>18.15%</td>
</tr>
<tr>
<td>surface + POS + morph</td>
<td>18.22%</td>
</tr>
</tbody>
</table>
Syntax, Semantics
Discourse, Pragmatics
Alignment, Multi-Words, Punctuation

Sarcasm, Intention
Poetry, Convention
Phrase Sense, Word Sense Disambiguation

Language Divergence,
Its core essence,
Are all a nuisance for Machine Translation

No matter which you try unriddling
You would still be lagging
If you forget or underrate

*The Power of POS Tagging!*

-Somya
NLP Trinity and Machine Translation?

“Unfortunately, no one can be told what the Matrix is. You have to see it for yourself.”

A Machine Translation Fix? 😊
References / Further Reading


