HOW MACHINE TRANSLATION OF WEB PAGES WORKS?

E-BUSINESS TECHNOLOGIES, BCM1, WS 2011

Student: Lei Sio Meng
Professor: Dr. Eduard Heindl

Introduction

• Machine Translation (MT)
  • Translate one natural language to other language by the computational power of computer

• When computer was developed, MT has brought up

• A demonstration to solve the MT as early as 1950’s
Introduction

• Processing was slow before 1980’s
  • the result is poor
  • computation ability not enough to solve MT problem

• After 1980s’, people pay attention back to MT
  • the growth of computation

• Nowadays, different approaches can be implement
  • rule-based machine translation (RBMT)
  • statistical machine translation (SMT)...

Introduction

• Web make a new stage of MT after 1990s’
  • The application of MT in Global, also e-Business becomes more important

• 260 countries are connected by internet, over 26 major languages.

• Non-English language speakers take about 43 percent of online population.
Introduction

• Key issue of web page in globalization:
  • How user gets Multi-language information
  • How to present and translate information of company

• Solution: Web based Machine Translation services
  • General
  • Fast
  • Instant
  • Effective
  • Low-cost

• For accurate result, still need human post-edit
• MT can speed-up the process of traditional translation

Web-MT Service

• Various translation service
• For instance: Google translate

Online translator

Document/ web page translation
Web-MT Service

Basic Work flow

- Build a *client/server* translation service by various architectures
  - Provide translation service by connect web server
  - MT application still relies on tradition MT approach

- Different translation language rules and approaches will be developed
  - as a series of modules
  - in the back-end server side
- Front-end client interface accept the translate requirement, send to server side
- Result send back to client after processing
Basic Work flow

E.g. Use Google translator to translate English to Germany.

English = Source language (SL)
Germany = Target language (TL)

Basic Work flow

- Different MT approaches
  - rule-based (RBMT)
  - statistical (SMT),
  - example-based
  - hybrid (RBMT + SMT)...  

- Different Modules
  - HTML fetching
  - Word segmentation
  - Part of Speech tagging...
  - Depend on MT approaches

- Different Resource:
  - Probabilistic data
  - Lexicon
  - Grammar rule
Translation approaches

1st Generation (1960s - 1980s)
• Direct approach

2nd Generation (1980s - )
• Rule-Based approach
• Transfer
• Interlingua

3rd Generation (1990s - )
• Corpus-Based
• Example
• Statistical

1st Generation is simple
2nd Generation is Linguistic analysis approaches,
3rd Generation is using corpus to train a statistical data for obtain result.

Direct approach

• Earliest, basic approach
• Dictionary Approach
• Linguist Model is not involved
• Translate words by words
• Result is poor
Transfer approach

- 3 Steps:
  1. Analysis: Parses input into Abstract Source Representation (SL Intermediate)
  2. Transfer: Translate Intermediate into Abstract Target Representation (TL Intermediate)
  3. Generation: Map TL Intermediate into output

Interlingua approaches

- Similar as Transfer approach

- 2 steps:
  1. Analysis: input is converted to one Interlingua representation
     - A summarized, abstract meaning, Neutral Universal Language
  2. General: transfer Interlingua to target text
Statistical Machine approaches

- The most *widely-use* approach

- Training a large Parallel Bilingual Corpus
  - Bilingual corpus is a set of documentations with SL, TL and translated relationship

- Easier to build Multilingual MT
  - No matter the closely or un-closely language-pair

---

Statistical Machine approaches

1. Input text is segmented into phrases and strings
2. Translate word segments by probability theory and training data
3. Arranging, combining the Segments by probability theory and training data

![Diagram of Statistical Machine approaches]
Statistical Machine approaches

- Statistical approach is implemented by Bayes’ rule

\[ P(T|S) = P(T)P(S|T) \]

- Translation Model
  - calculating probabilities of matching the source segments to target segment by a bilingual corpus

- Language Model
  - calculating best sequences from target segments and combine them as a final output

Example-Based approaches

- Uses the *aligned bilingual corpus* and *TL model*

1. Input is decomposed into a set of segments
2. Translated to target segments
   - Find a *closely translation-pair example* from Examples in Aligned bilingual corpus.
3. Target segments recombined together to be a target output text.
Example-Based approaches

- MT system is “Imitating” the translation of similar segment in corpus

![Diagram showing the decomposition and recombination of a sentence in English and its translation in Indonesian.]

Hybrid approach

- Build rule-based MT is expensive
  - Add linguistic rule, the result will be inconsistent
  - Add new language is different

- Statistical approach can’t reach the quality for people to fully understand
  - Especially when translating long sentences

- New idea: **Build Hybrid approach**
  - E.g. SYSTRAN: Hybrid Rule-based + Statistical approach in 2010
Modules and Resources

- Varies modules and resources are used between different MT applications
- Case: Multi-languages MT system based on the Statistical approaches

![Diagram of modules and resources]

Process and architectures of Web MTs

- Varies architectures are used between MT applications
- Case 1: Translate English to Bangla and Punjabi to Hindi

1. Parsing HTML Source Code: Use a HTML Parser, omits HTML tags, obtain content texts, combined as a SL input text
2. Translate input text to TL Text
3. Modifying original HTML code: Replace SL content by the TL text. The modified HTML code redirected to client
Process and architectures of Web MTs

- Case 2: Web MT translates Arabic, Chinese, Spanish to English by statistical approach

![Diagram of Web MT architectures]

**Process and architectures of Web MTs**

- **2-Level Layer architecture**
  - Web site *user-interface*: user input SL text and choose language-pair
  - Send request by *HTML form*
  - A CGI (Common Gateway Interface): Communicate between web site and machine translator in server side
  - MT front-end: Forward translation requests to appropriate languages wrapper.
Process and architectures of Web MTs

- Different Language-pair wrappers (as Chinese to English, Spanish to English)
- Include kernel of MT system
- Also Pre-processing module for SL
- The Translation implement MT program, result send back to client by opposition direction.

Process and architectures of Web MTs

- Case 3: Use Moses toolkit to build 3 different web MT systems
  - Moses is a open-source development software
  - Design 1:
Process and architectures of Web MTs

- **Apache Web Server and Tomcat Server:** Communicate between kernel MT system (Moses Toolkit) and Clients

- Moses Toolkit only *handles 1 request*

- When multi-client send requests, the Tomcat Server queues them (FIFO)

- Moses Toolkit *activate* and *re-load* related language-pair MT system every time

- Simple, not efficiency

---

Process and architectures of Web MTs

- **Design 2:**

  - Client

  - Server

  - *Apache Web Server and Tomcat Server*

  - Moses Server

  - Modules A

  - Modules B

  - Languages Pair 1

  - Languages Pair 2

  - Languages Pair 3

  - Languages Pair 4
Process and architectures of Web MTs

- Moses Server is used instead of Moses Toolkit
- Moses server: load multiple language-pair translation modules, handle multiple requests at a time
- Haven’t re-load language-pair MT system
- Resources are shared.
  - E.g. if translate same target language continuously
- Reduce the server’s memory.

Process and architectures of Web MTs

- **Design 3:**

  ![Diagram of Web MT architectures]

  - Client A
  - Client B
  - Client C
  - Apache Web Server
  - Tomcat Server
  - Moses Server 1
  - Moses Server 2
  - Moses Server 3
  - Moses Server 4
  - Modules A
  - Modules B
  - Language Pair 1
  - Language Pair 2
  - Language Pair 3
  - Language Pair 4
Process and architectures of Web MTs

- Each language-pair, system creates a separate Moses server
- Keeps translation resources, such as CPU, memory, independent of each other
- Work-Load of resource is reduced.
- Compare the result of the second design, this design is better

Conclusion

- Numerous web MT application is existing in market
  - Various approaches, resources, architecture
  - Yahoo! Babe Fish : Rule-based approaches
  - Google translator/ Bing translator : Statistics approaches
- Give a fast, instant translation to user
- Although can’t provide the high quality output
- It is satisfied for general purpose
Conclusion

• Various ways to use web-based MT service
  • Basic Client-Server MTs
  • Automatic webpage translation
  • Toolkit integration
  • Browser integration.
  • Mobile version web translation application
  • Real-time Speech translation

• By the powerful computation and resource on Web
• Web machine translation become convenient and easier accessible