Machine Translation Systems

Abdullah Alrajeh
asar1a10@ecs.soton.ac.uk

ISIS Postgraduate Conference
School of Electronics and Computer Science
The University of Southampton

27 April 2011
Outline

1. Introduction
2. Rule-based Systems
3. Example-based Systems
4. Statistical Systems
5. Advanced Systems
6. MT Evaluation
7. Resources
Translation involves 2 stages

- First: decoding the meaning of the source language
- Second: re-encoding the meaning into the target language
1) Introduction

- Behind this process set of complex operations
  - full knowledge of the source language which include
    - morphology
    - syntax
    - semantics
  - the context features of the text
  - the same in-depth knowledge to re-encode the meaning in the target language

- Different MT systems have been proposed:
  - Rule-based Systems
  - Example-based Systems
  - Statistical Systems (popular)
  - Advanced Systems
2) Rule-based Systems

- They are meaning-oriented MT systems (1980s-1990s)

- Intermediate representations: morphology, syntax and semantics patterns
  → collected and revised manually by linguistics experts

- Rule-based systems are effective because much of the linguistics knowledge is static

- Two major limitations:
  ▶ construction of linguistics rules (time, cost, ambiguity[idioms])
  ▶ processing efficiency (solving conflicts)

- Commercial Systems: SYSTRAN, LOGOS and EUROTRA
3) Example-based Systems

- Started early 1990s and were major turning point for MT

- Sometimes called corpus-based, analogy-based, memory-based or experience-guided

- Three main tasks
  - phrase matching
  - phrase alignment
  - phrase recombination

- They are faster than rule-based but do not guarantee better translations

- Example-based MT and Statistical MT are nearly similar
4) Statistical Systems

- In 1988, Brown et al. from IBM present the mathematics of statistical MT and suggest to find the best translation $e_{\text{best}}$ with the highest probability based on a parallel corpus

$$e_{\text{best}} = \arg\max_e p(e|f)$$

- Bayes rule, note the denominator $p(f)$ is independent of $e$ then it can be discarded

$$\arg\max_e p(e|f) = \arg\max_e \frac{p(f|e)p(e)}{p(f)}$$

- $p(f|e)$ (translation model) and $p(e)$ (language model) simulate human translators by first understanding the text and then expressing it
5) Advanced Systems

- Syntactic tree prediction with the perceptron algorithm
- Syntactic tree prediction with the boosting technique
- Kernel-based methods for MT
- They explore extensive context and syntax information using Machine Learning technologies but complicate the MT (memory and power)

Open questions:
  - which Machine Learning methodologies are appropriate for MT?
  - how to make use of these methods?

- Yizhao Ni has worked in this area and I will to continue his research
6) MT Evaluation

- How good is a given machine translation system?

- Hard problem, since many different translations acceptable
  \(\rightarrow\) semantic equivalence / similarity

- Adequacy and fluency
  - Does the output convey the same meaning as the input sentence?
  - Is the output good fluent English?

- Evaluation metrics
  - Subjective judgments by human evaluators
  - Automatic evaluation metrics
    - Precision and Recall of Words
    - WER (Word Error Rate)
    - BLEU (Bilingual Evaluation Understudy) - popular
6) MT Evaluation

- Chinese example from the 2001 NIST evaluation set
  这个 机场 的 安全 工作 由 以色列 方面 负责.
6) MT Evaluation

- Chinese example from the 2001 NIST evaluation set

这个 机场 的 安全 工作 由 以色列 方面 负责．

1- Israeli officials are responsible for airport security.
2- Israel is in charge of the security at this airport.
3- Israeli side was in charge of the security of this airport.
4- Israel is responsible for the airport’s security.
5- Israel is responsible for safety work at this airport.
6- Israel presides over the security of the airport.
7- Israel took charge of the airport security.
8- The safety of this airport is taken charge of by Israel.
9- This airport’s security is the responsibility of the Israeli security officials.
10- The security work for this airport is the responsibility of the Israel government.

Statistical Machine Translation, Philipp Koehn, textbook, Cambridge University Press, January 2010

http://www.statmt.org

http://www.mt-archive.info
Thank you!