Natural Language Processing
Lecture 08

Qun Liu, Valentin Malykh
Huawei Noah's Ark Lab

Spring 2022
A course delivered at KFU, Kazan

Content

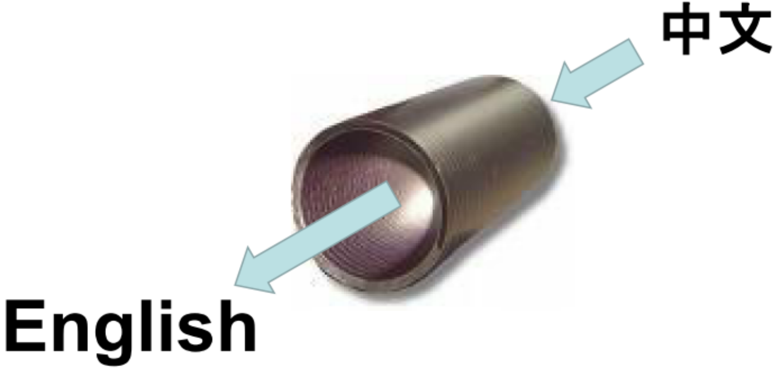
1. Machine Translation (MT)
2. Machine translation evaluation
3. Statistical machine translation (SMT)
4. Neural machine translation (NMT) based on RNNs
5. Neural machine translation (NMT) with attentions

Content

Holy grails of NLP (Recap)

- **Accurate machine translation between human languages**
- Free conversation between humans and computers

Accurate machine translation (Recap)



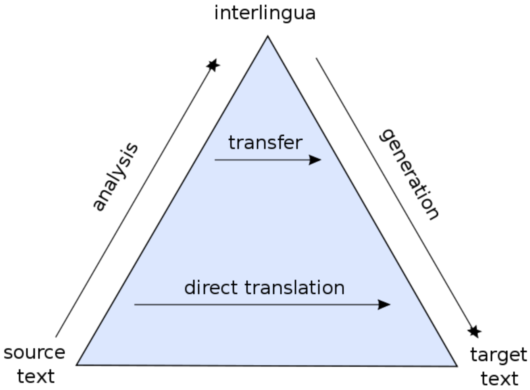
A brief history of MT

- 1940s-1950s: Early systems
- 1960s: ALPEC report (1966)
- 1970s: Operational systems: Systran, METEO
- 1980s: Eurotra project (EC-funded)
- Late 1980s–late1990s: the dawn of SMT (IBM), EBMT
- 2000s-early2010s: dominance of SMT
  - free on-line translation
- 2014-present: Neural MT

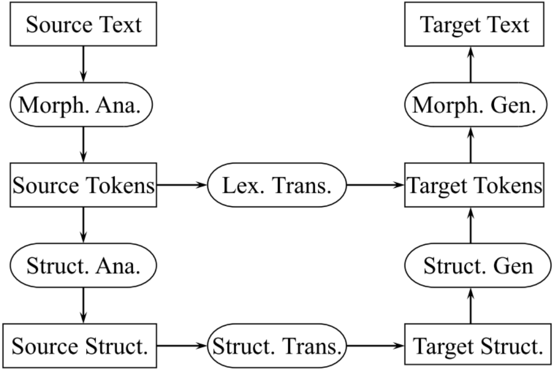
Different Approaches to MT

- Direct Translation
- Rule-based Machine Translation (RBMT)
- Memory-based Translation
- Example-based Machine Translation (EBMT)
- Statistical Machine Translation (SMT)
- Neural Machine Translation (NMT)

The Vauquois Pyramid

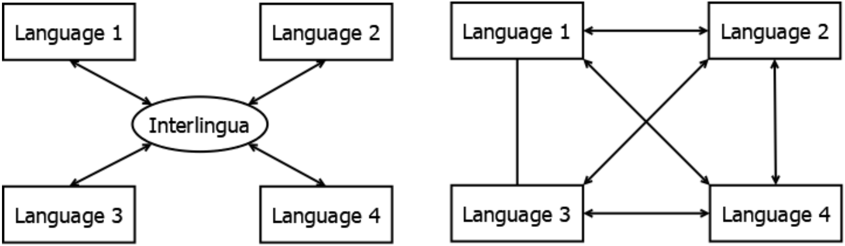


Transfer-based MT



Transfer-based MT

Interlingua approach vs. transfer approach



Interlingua-based Approach

Transfer-based Approach

Multilingual MT

- Theoretically, interlingua-based approach is ideal because it requires much less components compared with transfer-based approach.
- However, using a human-defined interlingua (i.e. a specific knowledge expression) is practically too difficult.
- In business, a feasible way is to use a natural language (mostly English) as the interlingua, which is usually called a pivot language in this case.

Multilingual MT

- Theoretically, interlingua-based approach is ideal because it requires much less components compared with transfer-based approach.
- However, using a human-defined interlingua (i.e. a specific knowledge expression) is practically too difficult.
- In business, a feasible way is to use a natural language (mostly English) as the interlingua, which is usually called a pivot language in this case.

Multilingual MT

Makoto Nagao (Kyoto University):
When the pivot language (i.e. interlingua) is used, the results of the analytic stage must be in a form which can be utilized by all of the different languages into which translation is to take place. This level of subtlety is a practical impossibility. (Machine Translation, Oxford, 1989)

- In business, a feasible way is to use a natural language (mostly English) as the interlingua, which is usually called a pivot language in this case.

Multilingual MT

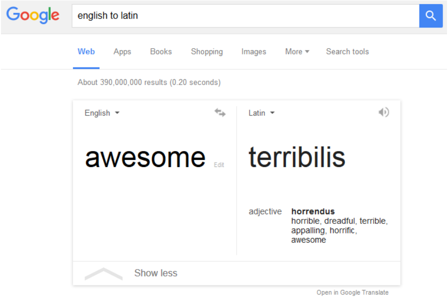
Patel-Schneider (METAL system):
METAL employs a modified transfer approach rather than an interlingua. If a meta-language (an interlingua) were to be used for translation purposes, it would need to incorporate all possible features of many languages. That would not only be an endless task but probably a fruitless one as well. Such a system would soon become unmanageable and perhaps collapse under its own weight. (A four-valued semantics for terminological reasoning, Artificial Intelligence, 38, 1989)

Multilingual MT

- Theoretically, interlingua-based approach is ideal because it requires much less components compared with transfer-based approach.
- However, using a human-defined interlingua (i.e. a specific knowledge expression) is practically too difficult.
- In business, a feasible way is to use a natural language (mostly English) as the interlingua, which is usually called a pivot language in this case.

Applications of MT

Since neural machine translation become popular, the quality of MT systems is greatly improved, and MT is used in more and more scenarios.



Web Translator

Applications of MT

Since neural machine translation become popular, the quality of MT systems is greatly improved, and MT is used in more and more scenarios.

Photo Translator

Applications of MT

Since neural machine translation become popular, the quality of MT systems is greatly improved, and MT is used in more and more scenarios.

Voice Translator

Applications of MT

Since neural machine translation become popular, the quality of MT systems is greatly improved, and MT is used in more and more scenarios.

Real-time Conference Translator

Content

Machine translation evaluation is not easy

- Machine translation evaluation is not easy because corrent translations for the same text may differ a lot literally.
- Human evaluation:
  - Reliable
  - Inconsistent
  - Expensive
  - Slow
- Automatic evaluation:
  - Unreliable
  - Consistent
  - Cheap
  - Fast

Goals for Automatic MT Evaluation

- Meaningful: score should give intuitive interpretation of translation quality.
- Consistent: repeated use of metric should give same results.
- Correct: metric must rank better systems higher.
- Low cost: reduce time and money spent on carrying out evaluation.
- Tunable: automatically optimise system performance towards metric.

Automatic MT Evaluation Metrics

- Input:
  - output of machine translation systems.
  - reference translations give by human translators.
  - source text (optional).
- Output:
  - a score which represents the similarity between the MT output and the human reference given the source sentence.

Automatic MT Evaluation Metrics

- Plenty of automatic MT evaluation metrics are proposed by researchers, among which the most popular metrics include:
  - WER (Word Error Rate)
  - BLEU (BiLingual Evaluation Understudy)
  - TER (Translation Error Rate)
  - METEOR (Metric for Evaluation of Translation with Explicit ORdering)

BLEU: Motivation

- N-gram precision between the machine translation and the reference
- Compute n-gram precisions for 1 to n (typically $n = 4$)
- A harmony average of the n-gram precisions is calculated over different n's
- Compute on the entire test set, rather than single sentences, to avoid 0 value for higher n-gram precisions

BLEU: Definition

$$\mathrm{BLEU} = \underbrace{\min(1, \frac{length\_of\_MT}{length\_of\_reference})}_{\text{Brevity Penalty}} \times \underbrace{(\prod_{i=1}^{n} \mathrm{Precision}_i)^{\frac{1}{n}}}_{\text{N-gram Precision}}$$

$$\mathrm{Precision}_i = \frac{clipped\_number\_of\_matched\_\{i\}grams\_in\_MT}{number\_of\_total\_\{i\}grams\_in\_MT}$$

- Brevity Penalty is used to avoid that a very short MT get a high BLEU score.
- The number of matched $\{i\}$grams is *clipped* to avoid that a single word in the reference is matched multiple times in MT.

Multiple References

To avoid translation variability, multiple references can by used:

- n-grams may match in any of the references
- closest reference length is used for brevity penalty

BLEU: An example

- **Candidate 1: the book is on the desk**

This example is extracted from Dr. Ming Zhou's slides on MT Evaluation

BLEU: An example

• **Candidate 1: the book is on the desk**	• **Reference 1: there is a book on the desk**
	• **Reference 2: the book is on the table**

This example is extracted from Dr. Ming Zhou's slides on MT Evaluation

BLEU: An example

- **Candidate 1: the book is on the desk**

- **Reference 1: there is a book on the desk**
- **Reference 2: the book is on the table**

unigram:
$Count_{clip}(the) = 2$
$Count_{clip}(book) = 1$
$Count_{clip}(is) = 1$
$Count_{clip}(on) = 1$
$Count_{clip}(desk) = 1$

This example is extracted from Dr. Ming Zhou's slides on MT Evaluation

BLEU: An example

- **Candidate 1: the book is on the desk**

- **Reference 1: there is a book on the desk**
- **Reference 2: the book is on the table**

unigram:

$Count_{clip}(the) = 2$

$Count_{clip}(book) = 1$

$Count_{clip}(is) = 1$

$Count_{clip}(on) = 1$

$Count_{clip}(desk) = 1$

$\sum_{unigram \in C} Count(unigram) = 6$

This example is extracted from Dr. Ming Zhou's slides on MT Evaluation

BLEU: An example

- **Candidate 1: the book is on the desk**

- **Reference 1: there is a book on the desk**
- **Reference 2: the book is on the table**

unigram:
$Count_{clip}(the) = 2$
$Count_{clip}(book) = 1$
$Count_{clip}(is) = 1$
$Count_{clip}(on) = 1$
$Count_{clip}(desk) = 1$
$\sum_{unigram \in C} Count(unigram) = 6$
$p_1 = 1$

This example is extracted from Dr. Ming Zhou's slides on MT Evaluation

BLEU: An example

- **Candidate 1: the book is on the desk**

- **Reference 1: there is a book on the desk**
- **Reference 2: the book is on the table**

unigram:
$Count_{clip}(the) = 2$
$Count_{clip}(book) = 1$
$Count_{clip}(is) = 1$
$Count_{clip}(on) = 1$
$Count_{clip}(desk) = 1$
$\sum_{unigram \in C} Count(unigram) = 6$
$p_1 = 1$

bigram:
$Count_{clip}(the, book) = 1$
$Count_{clip}(book, is) = 1$
$Count_{clip}(is, on) = 1$
$Count_{clip}(on, the) = 1$
$Count_{clip}(the, desk) = 1$

This example is extracted from Dr. Ming Zhou's slides on MT Evaluation

BLEU: An example

- **Candidate 1: the book is on the desk**

- **Reference 1: there is a book on the desk**
- **Reference 2: the book is on the table**

unigram:
$Count_{clip}(the) = 2$
$Count_{clip}(book) = 1$
$Count_{clip}(is) = 1$
$Count_{clip}(on) = 1$
$Count_{clip}(desk) = 1$
$\sum_{unigram \in C} Count(unigram) = 6$
$p_1 = 1$

bigram:
$Count_{clip}(the, book) = 1$
$Count_{clip}(book, is) = 1$
$Count_{clip}(is, on) = 1$
$Count_{clip}(on, the) = 1$
$Count_{clip}(the, desk) = 1$
$\sum_{bigram \in C} Count(bigram) = 5$

This example is extracted from Dr. Ming Zhou's slides on MT Evaluation

BLEU: An example

- **Candidate 1: the book is on the desk**

- **Reference 1: there is a book on the desk**
- **Reference 2: the book is on the table**

unigram:
$Count_{clip}(the) = 2$
$Count_{clip}(book) = 1$
$Count_{clip}(is) = 1$
$Count_{clip}(on) = 1$
$Count_{clip}(desk) = 1$
$\sum_{unigram \in C} Count(unigram) = 6$
$p_1 = 1$

bigram:
$Count_{clip}(the, book) = 1$
$Count_{clip}(book, is) = 1$
$Count_{clip}(is, on) = 1$
$Count_{clip}(on, the) = 1$
$Count_{clip}(the, desk) = 1$
$\sum_{bigram \in C} Count(bigram) = 5$
$p_2 = 1$

This example is extracted from Dr. Ming Zhou's slides on MT Evaluation

BLEU: An example

- **Candidate 1: the book is on the desk**

- **Reference 1: there is a book on the desk**
- **Reference 2: the book is on the table**

unigram:	**bigram:**	**trigram:**
$Count_{clip}(the) = 2$	$Count_{clip}(the, book) = 1$	$Count_{clip}(the, book, is) = 1$
$Count_{clip}(book) = 1$	$Count_{clip}(book, is) = 1$	$Count_{clip}(book, is, on) = 1$
$Count_{clip}(is) = 1$	$Count_{clip}(is, on) = 1$	$Count_{clip}(is, on, the) = 1$
$Count_{clip}(on) = 1$	$Count_{clip}(on, the) = 1$	$Count_{clip}(on, the, desk) = 1$
$Count_{clip}(desk) = 1$	$Count_{clip}(the, desk) = 1$	
$\sum_{unigram \in C} Count(unigram) = 6$	$\sum_{bigram \in C} Count(bigram) = 5$	
$p_1 = 1$	$p_2 = 1$	

This example is extracted from Dr. Ming Zhou's slides on MT Evaluation

BLEU: An example

- **Candidate 1: the book is on the desk**

- **Reference 1: there is a book on the desk**
- **Reference 2: the book is on the table**

unigram:	**bigram:**	**trigram:**
$Count_{clip}(the) = 2$	$Count_{clip}(the, book) = 1$	$Count_{clip}(the, book, is) = 1$
$Count_{clip}(book) = 1$	$Count_{clip}(book, is) = 1$	$Count_{clip}(book, is, on) = 1$
$Count_{clip}(is) = 1$	$Count_{clip}(is, on) = 1$	$Count_{clip}(is, on, the) = 1$
$Count_{clip}(on) = 1$	$Count_{clip}(on, the) = 1$	$Count_{clip}(on, the, desk) = 1$
$Count_{clip}(desk) = 1$	$Count_{clip}(the, desk) = 1$	
$\sum_{unigram \in C} Count(unigram) = 6$	$\sum_{bigram \in C} Count(bigram) = 5$	$\sum_{trigram \in C} Count(trigram) = 4$
$p_1 = 1$	$p_2 = 1$	

This example is extracted from Dr. Ming Zhou's slides on MT Evaluation

BLEU: An example

- **Candidate 1: the book is on the desk**

- **Reference 1: there is a book on the desk**
- **Reference 2: the book is on the table**

unigram:	**bigram:**	**trigram:**
$Count_{clip}(the) = 2$	$Count_{clip}(the, book) = 1$	$Count_{clip}(the, book, is) = 1$
$Count_{clip}(book) = 1$	$Count_{clip}(book, is) = 1$	$Count_{clip}(book, is, on) = 1$
$Count_{clip}(is) = 1$	$Count_{clip}(is, on) = 1$	$Count_{clip}(is, on, the) = 1$
$Count_{clip}(on) = 1$	$Count_{clip}(on, the) = 1$	$Count_{clip}(on, the, desk) = 1$
$Count_{clip}(desk) = 1$	$Count_{clip}(the, desk) = 1$	
$\sum_{unigram \in C} Count(unigram) = 6$	$\sum_{bigram \in C} Count(bigram) = 5$	$\sum_{trigram \in C} Count(trigram) = 4$
$p_1 = 1$	$p_2 = 1$	$p_3 = 1$

This example is extracted from Dr. Ming Zhou's slides on MT Evaluation

BLEU: An example

- **Candidate 1: the book is on the desk**

- **Reference 1: there is a book on the desk**
- **Reference 2: the book is on the table**

unigram:	**bigram:**	**trigram:**
$Count_{clip}(the) = 2$	$Count_{clip}(the, book) = 1$	$Count_{clip}(the, book, is) = 1$
$Count_{clip}(book) = 1$	$Count_{clip}(book, is) = 1$	$Count_{clip}(book, is, on) = 1$
$Count_{clip}(is) = 1$	$Count_{clip}(is, on) = 1$	$Count_{clip}(is, on, the) = 1$
$Count_{clip}(on) = 1$	$Count_{clip}(on, the) = 1$	$Count_{clip}(on, the, desk) = 1$
$Count_{clip}(desk) = 1$	$Count_{clip}(the, desk) = 1$	
$\sum_{unigram \in C} Count(unigram) = 6$	$\sum_{bigram \in C} Count(bigram) = 5$	$\sum_{trigram \in C} Count(trigram) = 4$
$p_1 = 1$	$p_2 = 1$	$p_3 = 1$

$$\left.\begin{array}{l} c = 6 \\ r = 6 \end{array}\right\} = e^{1-\frac{r}{c}} = e^0 = 1 = BP$$

This example is extracted from Dr. Ming Zhou's slides on MT Evaluation

BLEU: An example

- **Candidate 1: the book is on the desk**

- **Reference 1: there is a book on the desk**
- **Reference 2: the book is on the table**

unigram:	**bigram:**	**trigram:**
$Count_{clip}(the) = 2$	$Count_{clip}(the, book) = 1$	$Count_{clip}(the, book, is) = 1$
$Count_{clip}(book) = 1$	$Count_{clip}(book, is) = 1$	$Count_{clip}(book, is, on) = 1$
$Count_{clip}(is) = 1$	$Count_{clip}(is, on) = 1$	$Count_{clip}(is, on, the) = 1$
$Count_{clip}(on) = 1$	$Count_{clip}(on, the) = 1$	$Count_{clip}(on, the, desk) = 1$
$Count_{clip}(desk) = 1$	$Count_{clip}(the, desk) = 1$	
$\sum_{unigram \in C} Count(unigram) = 6$	$\sum_{bigram \in C} Count(bigram) = 5$	$\sum_{trigram \in C} Count(trigram) = 4$
$p_1 = 1$	$p_2 = 1$	$p_3 = 1$

$$\left.\begin{array}{l} c = 6 \\ r = 6 \end{array}\right\} = e^{1-\frac{r}{c}} = e^0 = 1 = BP$$

$$\begin{aligned} \text{BLEU} &= BP \times (p_1 \times p_2 \times p_3)^{\frac{1}{3}} \\ &= 1 \times (1 \times 1 \times 1)^{\frac{1}{3}} = 1 \end{aligned}$$

This example is extracted from Dr. Ming Zhou's slides on MT Evaluation

Content

Content

Motivation

- In the time of rule-based MT, the translation knowledge was explicitly encoded by linguists, which is labor intensive, time consuming and inefficient.
- The basic idea of statistical machine translation is to learn a machine translation system directly from parallel translation data, without involving linguistics to encode translation knowledge.

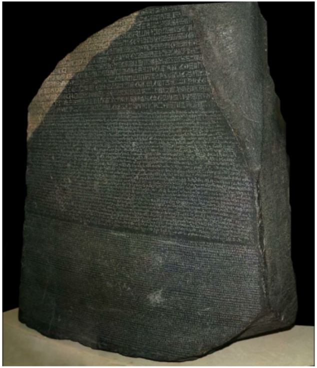
A brief history of SMT

- The first statistical machine translation was conducted by IBM researchers in late 1980s to early 1990s.
- The research of SMT was not very active in 1990s because the computing resources used by IBM researchers were higher than most researchers could have.
- In 1999, a group of researchers gathered in the JHU summer workshop and repeated IBM's early work successfully and release GIZA++, the implementation of the training algorithm of IBM models.
- After 2000 SMT became a dominate MT paradigm until it was replaced by NMT in recent years.

Parallel corpus

- A parallel corpus, or a bitext, is a collection of texts in one language and its corrending translation in another language.
- Parallel corpara used for MT research are normally aligned at sentence level.
- Document-level MT needs parallel corpara aligned in both word level and document level.
- Many parallel corpara are aligned only in document level in their original form, so a technique called sentence alignement is developed to align the sentences in a document aligned corpus.

Rosette Stone



	The Rosetta Stone
Material	Granodiorite
Size	114.4 × 72.3 × 27.93 cm (45 x 28.5 x 11 in)
Writing	Ancient Egyptian hieroglyphs, Demotic script, and Greek script
Created	196 BC
Discovered	1799
Present	British Museum

Rosette stone is a good example that it is feasible to learn translation knowledge from a parallel corpus.

Statistical Translation Model

A statistical translation model is the probability that a target text e is the translation of a given source text f:

$$p(e|f), \text{where} \sum_{e} p(e|f) = 1$$

Statistical Machine Translation

Given a statistical translation $p(e|f)$, the machine translation problem can be transfered to a search problem: to search a target sentence e which has the highist translation probability given f:

$$\hat{e} = \operatorname*{argmax}_{e} p(e|f)$$

Noisy Channel Model

- Statistical machine translation with a single translation model does not work well.
- IBM researchers proposed a Noisy Channel Model for statistical machine translation, which includes an additional language model.

Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, Paul S. Roossin, A Statistical Approach to Machine Translation, Computational Linguistics,1990

Noisy Channel Framework