CMPE 409 Machine Translation

Example-Based MT

Murat ORHUN
Adopted from original slides of Josef van Genabith, CNGL, Dublin City University Khalil Sima'an, University of Amsterdam

Istanbul Bilgi University

June 1, 2022

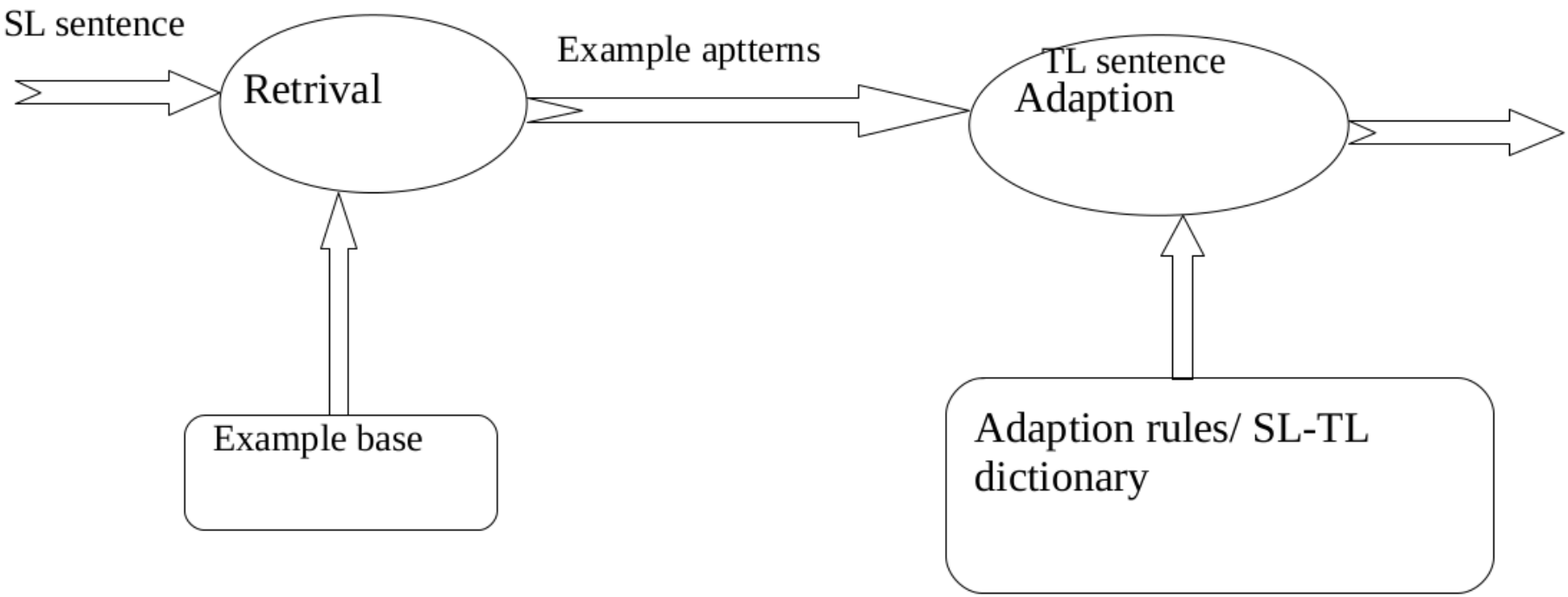
Example-Based Machine Translation (EBMT)

- The EBMT system uses past translation examples to generate translation for a given SL text.
- EBMT systems maintains an example-base consisting of translation examples between source and target languages.
- When a SL sentence is given to the system, the system retrieves a similar SL sentence from the example-base and its translation.

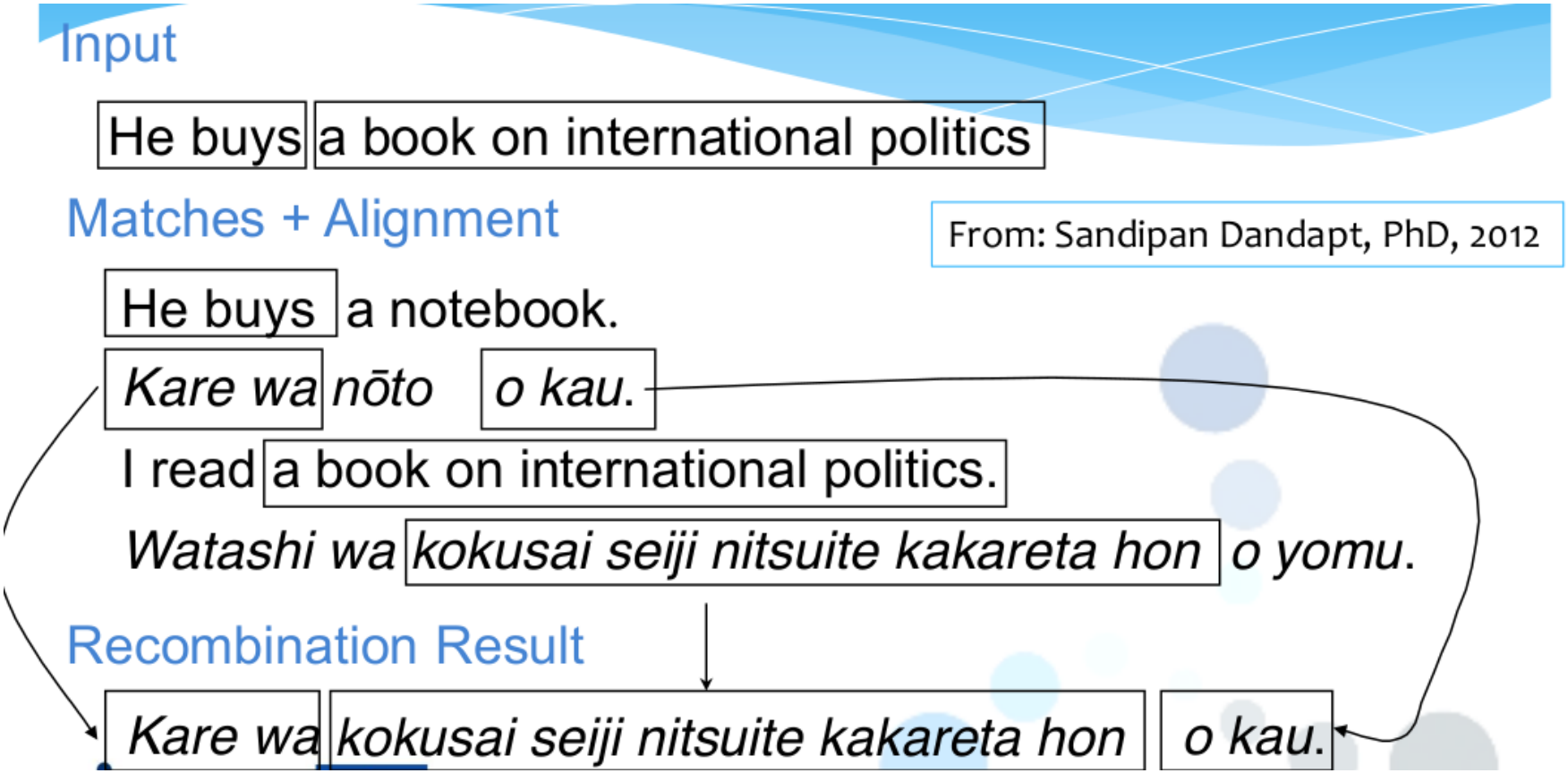
Example-Based Machine Translation (EBMT)

- Then it adapts the example to generate the TL sentence of the input sentence.
- The EBMT system rest on the idea that smilar sentence will be have smilar translations.
- EBMT systems are corpus-based approaches to MT.

EBMT


SL sentence
Retrival
Example aptterns
TL sentence
Adaption
Example base
Adaption rules/ SL-TL
dictionary

EBMT


Input
He buys a book on international politics
Matches + Alignment
From: Sandipan Dandapt, PhD, 2012
He buys a notebook.
Kare wa nōto o kau.
I read a book on international politics.
Watashi wa kokusai seiji nitsuite kakareta hon o yomu.
Recombination Result
Kare wa kokusai seiji nitsuite kakareta hon o kau.

EBMT

- EBMT generally uses a sentence-aligned parallel text (TM) as the primary source of data.
- EBMT systems search the source side of the example-base for close matches to the input sentences
- Obtain corresponding target segments at runtime
- EBMT is a fully automatic translation

EBMT

- EBMT: supposed to be good on limited amounts of data and homogeneous data (lots of repetition)
- EBMT systems produce a good translation while SMT systems fail and vice versa

EBMT and SMT

- phrase-based SMT approach has proven to be the most successful MT approach in MT competitions e.g. NIST, WMT, IWSLT etc.
- SMT systems discard the actual training data once the translation model and language model have been estimated
- cannot always guarantee good quality translations for sentences which closely match those in the training corpora

EBMT Approaches

1. Runtime using proportional analogy
2. Compile time using generalized translation template-based EBMT model

EBMT Background

- Rule-based or data-driven MT
- Data driven MT: EBMT and SMT
- Corpus-based data driven approaches derive knowledge from parallel corpora to translate new input
- Mostly SMT today
- A few EBMT (hybrid) systems include CMU-EBMT (Brown, 2011) and Cunei (Phillips, 2011)

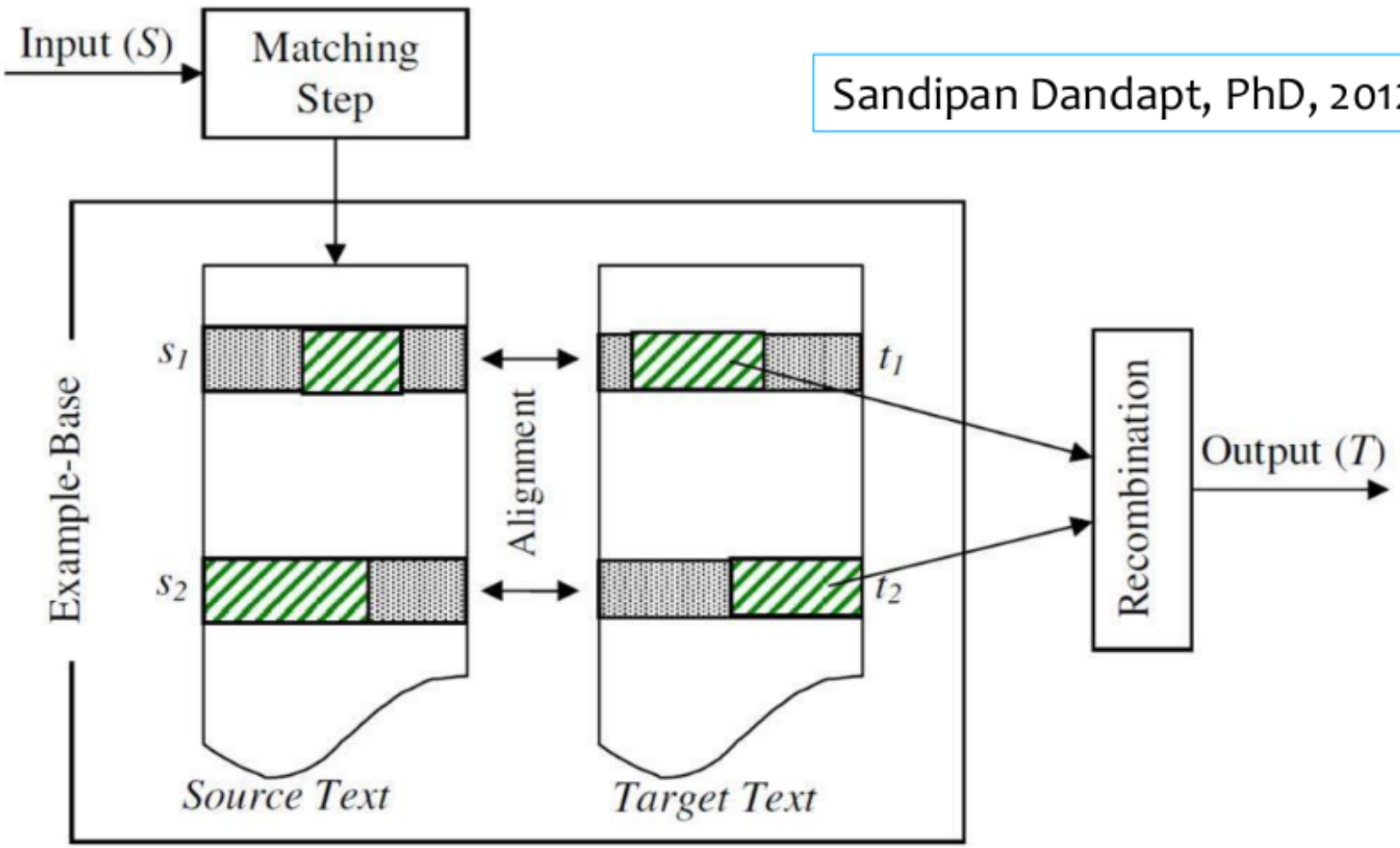
Where did is start?

- Nagao (1984)
- "MT by analogy principle"
- "Man does not translate a simple sentence by doing deep linguistic analysis, rather, man does translation, first, by properly decomposing an input sentence into certain fragmental phrases, ... then by translating these phrases into other language phrases, and finally by properly composing these fragmental translations into one long sentence. The translation of each fragmental phrase will be done by the analogy translation principle with proper examples as its reference." (Nagao, 1984, p.178)

EBMT Core steps

- Translation in three steps: matching, alignment and recombination
  - **Matching:** finds the example or set of examples from the bitext which most closely match the source-language string to be translated.
  - **Alignment**: extracts the source–target translation equivalents from the retrieved examples of the matching step.
  - **Recombination**: produces the final translation by combining the target translations of the relevant subsentential fragments.

EBMT Core steps



Informal Example

- *Where can I find tourist information*
- Where can I find ladies dresses
  (tr) bayan kıyafetlerini nereden bulabilirim
- just in front of the tourist information
  (tr)turist bilgilerini hemen önünde

Informal Example

- Where can I find = nereden bulabilirim
- tourist information = turist bilgilerini
- Where can I find tourist information = turist bilgilerini nereden bulabilirim

Varieties of EBMT

- EBMT systems differ widely in their matching stages
- involve a distance or similarity measure of some kind (e.g. edit distance)

Character-Based Matching

- dynamic programming technique, e.g. Levenshtein distance
  1. The President **agrees** with the decision.
  2. The President **disagrees** with the decision.
  3. The President **concurs** with the decision.

  System will chose (2) given (1)

Varieties of EBMT

Word-Based Matching

- Nagao (1984)
- uses dictionaries and thesauri to determine the relative word distance in terms of meaning
  1. The President **agrees** with the decision.
  2. The President **disagrees** with the decision.
  3. The President **concurs** with the decision.

System will chose (3) given (1)

Varieties of EBMT

Pattern-Based Matching

- similar examples can be used to abstract "generalised" translation templates
- Brown (1999):
- NE equivalence classes, such as person, date and city
- some linguistic information, such as gender and number

a. John Miller flew to Frankfurt on December 3rd.
b. ⟨FIRSTNAME-M⟩ ⟨LASTNAME⟩ flew to ⟨CITY⟩ on ⟨MONTH⟩ ⟨ORDINAL⟩.
c. ⟨PERSON-M⟩ flew to ⟨CITY⟩ on ⟨DATE⟩.

Varieties of EBMT

- Syntax-Based Matching
- Marker-Based Matching

Approaches to EBMT

- first introduced as an analogy-based approach to MT
- "case-based", "memory-based" and "experience-guided" MT
- Many, many varieties ...
- Two main approaches
  1. With or without preprocessing/training stage
  2. Pure/runtime EBMT vs. compiled EBMT

Approaches to EBMT

Pure/runtime EBMT:

- (e.g. Lepage and Denoual, 2005b, see further reading)
- No time consumed for training/preprocessing
- But: runtime/translation complexity very considerable ...

Approaches to EBMT

Compiled approaches:

- (e.g. Al-Adhaileh and Tang,1999; Cicekli and Guvenir, 2001, (see further reading))
- Pre-compute units below sentence level before predictiontranslation time

Analogical Reasoning

Compiled approaches:

- A : B :: C : D
- "A is to B as C is to D"
- A global relationship between 4 objects
- "Analogical equation"
- A : B :: C : D?

Analogies

Compiled approaches:

- lungs are to humans as gills are to fish
- cat : kitten :: dog : puppy
- speak : spoken :: break : broken

Analogies

- lungs are to humans as gills are to X? X = fish
- cat : kitten :: dog : X? X = puppy
- speak : spoken :: break : X? X = broken

Pure EBMT

- "The 'purest' EBMT system ever built: no variables, no templates, no training, examples, just examples, only examples"
- see further reading

Further Reading

- Lepage, Y. and Denoual, E. (2005c). The 'purest' EBMT System Ever Built: No Variables, No Templates, No Training, Examples, Just examples, Only Examples. In Proceedings of the 2nd Workshop on Example-based Machine Translation, a Workshop at the MT Summit X, page 81–90, Phuket, Thailand.
- Nagao, M. (1984). A Framework of a Mechanical Translation between Japanese and English by Analogy Principle. In Elithorn, A. and Banerji, R., editors, Artificial and Human Intelligence, page 173–180. North-Holland, Amsterdam.
- Sandipan Dandapat "Mitigating the Problems of SMT using EBMT" PhD Thesis, DCU, 2012

GIZA++ demo

- Explain
- installation
- test a demo
- explain results

Moses

- Explain
- installation
- test a demo
- explain results

https://www.youtube.com/watch?v=88UXtB9tnbc&t=9s
https://www.youtube.com/watch?v=7QA9quB3IZA
http://www.statmt.org/moses/?n=Development.GetStarted
http://www.statmt.org/moses_steps.html