WIKIPEDIA

Rule-based machine translation

Rule-based machine translation (**RBMT**; "Classical Approach" of MT) is machine translation systems based on **linguistic information** about source and target languages basically retrieved from (unilingual, bilingual or multilingual) **dictionaries** and **grammars** covering the main semantic, morphological, and syntactic regularities of each language respectively. Having input sentences (in some source language), an RBMT system generates them to output sentences (in some target language) on the basis of morphological, syntactic, and semantic analysis of both the source and the target languages involved in a concrete translation task.

Contents

History

The first RBMT systems were developed in the early 1970s. The most important steps of this evolution were the emergence of the following RBMT systems:

- Systran (http://www.systran.de/)
- Japanese MT systems (http://aamt.info/english/mtsys.htm, http://www.wtec.org/loyola/ar93_94/mt.htm)
- EUROTRA (Eurotra)

Today, other common RBMT systems include:

- Apertium
- GramTrans

Types of RBMT

There are three different types of rule-based machine translation systems:

1. **Direct Systems** (Dictionary Based Machine Translation) map input to output with basic rules.
2. **Transfer RBMT Systems** (Transfer Based Machine Translation) employ morphological and syntactical analysis.
3. **Interlingual RBMT Systems** (Interlingua) use an abstract meaning.[1][2]

RBMT systems can also be characterized as the systems opposite to Example-based Systems of Machine Translation (Example Based Machine Translation), whereas Hybrid Machine Translations Systems make use of many principles derived from RBMT.

Basic principles

The main approach of RBMT systems is based on linking the structure of the given input sentence with the structure of the demanded output sentence, necessarily preserving their unique meaning. The following example can illustrate the general frame of RBMT:

A girl eats an apple. Source Language = English; Demanded Target Language = German

Minimally, to get a German translation of this English sentence one needs:

1. A dictionary that will map each English word to an appropriate German word.
2. Rules representing regular English sentence structure.
3. Rules representing regular German sentence structure.

And finally, we need rules according to which one can relate these two structures together.

Accordingly, we can state the following **stages of translation**:

1st: getting basic part-of-speech information of each source word:

a = indef.article; girl = noun; eats = verb; an = indef.article; apple = noun

2nd: getting syntactic information about the verb "to eat":

NP-eat-NP; here: eat – Present Simple, 3rd Person Singular, Active Voice

3rd: parsing the source sentence:

(NP an apple) = the object of eat

Often only partial parsing is sufficient to get to the syntactic structure of the source sentence and to map it onto the structure of the target sentence.

4th: translate English words into German

a (category = indef.article) => ein (category = indef.article)

girl (category = noun) => Mädchen (category = noun)

eat (category = verb) => essen (category = verb)

an (category = indef. article) => ein (category = indef.article)

apple (category = noun) => Apfel (category = noun)

5th: Mapping dictionary entries into appropriate inflected forms (final **generation**):

A girl eats an apple. => Ein Mädchen isst einen Apfel.

Components

The RBMT system contains:

from nltk.translate import AlignedSent, Alignment

- a **SL morphological analyser** - analyses a source language word and provides the morphological information;
- a **SL parser** - is a syntax analyser which analyses source language sentences;
- a **translator** - used to translate a source language word into the target language;
- a **TL morphological generator** - works as a generator of appropriate target language words for the given grammatica information;
- a **TL parser** - works as a composer of suitable target language sentences;
- **Several dictionaries** - more specifically a minimum of three dictionaries:

> a *SL dictionary* - needed by the source language morphological analyser for morphological analysis,
>
> a *bilingual dictionary* - used by the translator to translate source language words into target language words,
>
> a *TL dictionary* - needed by the target language morphological generator to generate target language words.[3]

The RBMT system makes use of the following:

- a **Source Grammar** for the input language which builds syntactic constructions from input sentences;
- a **Source Lexicon** which captures all of the allowable vocabulary in the domain;
- **Source Mapping Rules** which indicate how syntactic heads and grammatical functions in the source language are mapped onto domain concepts and semantic roles in the interlingua;
- a **Domain Model**/**Ontology** which defines the classes of domain concepts and restricts the fillers of semantic roles for each class;
- **Target Mapping Rules** which indicate how domain concepts and semantic roles in the interlingua are mapped onto syntactic heads and grammatical functions in the target language;
- a **Target Lexicon** which contains appropriate target lexemes for each domain concept;
- a **Target Grammar** for the target language which realizes target syntactic constructions as linearized output sentences.[4]

Advantages

- No bilingual texts are required. This makes it possible to create translation systems for languages that have no texts in common, or even no digitized data whatsoever.
- Domain independent. Rules are usually written in a domain independent manner, so the vast majority of rules will "just work" in every domain, and only a few specific cases per domain may need rules written for them.
- No quality ceiling. Every error can be corrected with a targeted rule, even if the trigger case is extremely rare. This is in contrast to statistical systems where infrequent forms will be washed away by default.
- Total control. Because all rules are hand-written, you can easily debug a rule based system to see exactly where a given error enters the system, and why.
- Reusability. Because RBMT systems are generally built from a strong source language analysis that is fed to a transfer step and target language generator, the source language analysis and target language generation parts can be shared between multiple translation systems, requiring only the transfer step to be specialized. Additionally, source language analysis for one language can be reused to bootstrap a closely related language analysis.

Shortcomings

- Insufficient amount of really good dictionaries. Building new dictionaries is expensive.
- Some linguistic information still needs to be set manually.
- It is hard to deal with rule interactions in big systems, ambiguity, and idiomatic expressions.

- Failure to adapt to new domains. Although RBMT systems usually provide a mechanism to create new rules and extend and adapt the lexicon, changes are usually very costly and the results, frequently, do not pay off.[5]

References

1. Koehn, Philipp (2010). *Statistical Machine Translation*. Cambridge: Cambridge University Press. p. 15.

2. Nirenburg, Sergei (1989). "Knowledge-Based Machine Translation". *Machine Trandation 4 (1989), 5 - 24.* Kluwer Academic Publishers. JSTOR 10.2307/40008396 (https://www.jstor.org/stable/10.2307/40008396). Missing or empty |url= (help)

3. Hettige, B.; Karunananda, A.S. (2011). "Computational Model of Grammar for English to Sinhala Machine Translation" (http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6075022). *The International Conference on Advances in ICT for Emerging Regions - ICTer20 11 : 026-031.* Retrieved 20 June 2012.

4. Lonsdale, Deryle; Mitamura, Teruko; Nyberg, Eric (1995). "Acquisition of Large Lexicons for Practical Knowledge-Based MT" (http://www.springerlink.com/content/v51w13n70m145786/fulltext.pdf) (PDF). *Machine Translation 9: 251-283.* Kluwer Academic Publishers. Retrieved 20 June 2012.

5. Lagarda, A.-L.; Alabau, V.; Casacuberta, F.; Silva, R.; Díaz-de-Liaño, E. (2009). "Statistical Post-Editing of a Rule-Based Machine Translation System" (http://www.aclweb.org/anthology/N/N09/N09-2055.pdf) (PDF). *Proceedings of NAACL HLT 2009: Short Papers, pages 217–220, Boulder, Colorado.* Association for Computational Linguistics. Retrieved 20 June 2012.

Literature

- Arnold, D.J. et al. (1993): Machine Translation: an Introductory Guide
- Hutchins, W.J. (1986): Machine Translation: Past, Present, Future

Links

- First International Workshop on Free/Open-Source Rule-Based Machine Translation (http://xixona.dlsi.ua.es/freerbmt09/)
- https://web.archive.org/web/20120306014535/http://www.inf.ed.ac.uk/teaching/courses/mt/lectures/history.pdf
- https://web.archive.org/web/20150914205051/http://www.csse.unimelb.edu.au/research/lt/nlp06/materials/Bond/mt-intro.pdf