Rule-based machine translation

Basic concepts of RBMT. Overview of resources needed for RBMT

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.

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

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.

Collection of rules and a bilingual or multilingual lexicon are the resources used in RBMT. The transfer model involves three stages: analysis, transfer and generation. Figure 1 shows the complete work flow of translation in the form of a pipeline.

During analysis phase linguistic analysis is performed on the input source sentence in order to extract information in terms of morphology, parts of speech, phrases, named entity and word sense disambiguation. During the lexical transfer phase, there are two steps namely word translation and grammar translation. In word translation, source language root word is replaced by the target language root word with the help of a bilingual dictionary and in grammar translation, suffixes are getting translated. In generation phase genders of the translated words are corrected and it will be followed by short distance and long-distance agreements performed by intrachunk and the inter-chunk module. These ensure that the gender, number and person of local groups of phrases agree as also the gender of the subject's verbs or objects reflect those of the subject. [1]

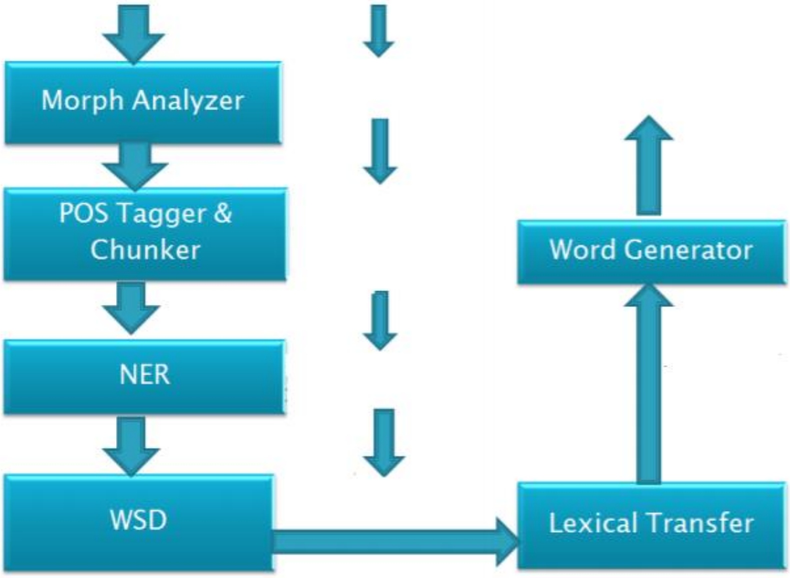

Figure 1. RBMT work flow [1]

Components

The RBMT system contains:

- 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.

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.

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. [2]

Types of RBMT

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

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

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.

Dictionary-based machine translation

Machine translation can use a method based on dictionary entries, which means that words will be translated the same way as in a dictionary - word by word, as a rule, without a special semantic relationship between them. Dictionary search can be performed with morphological analysis, with or without word normalization. While this approach to machine translation is probably the least complex, *dictionary-based machine translation* is ideal for translating long lists of phrases of incomplete sentences, such as inventories or simply catalogs of products and services.

It can also be used to speed up translation of manuals if the person doing it is fluent in both languages and therefore able to correct the syntax and grammar. [3]

Transfer-based machine translation

Transform-based machine translation is a type of machine translation. It is based on the idea of an intermediate language (interlingua) and is currently one of the most widely used machine translation methods.

Machine translation is based on processing and machine translation is based on the intermediate language are one and the same idea: to make a transfer, you must have an intermediate representation, which includes the "value" of the original proposal in order to obtain the correct ne revoda. The IP-based intermediate language representation of the intermediate fraction wives be independent of the source language, whereas in the IP-based pre formations, it has some dependence on the language pair.

The way in which work machine translation systems based on pre formations varies considerably, but in general they follow the same pattern: they apply a set of linguistic rules, which are defined as the correspondence between the structure of the source language and the target language. The first stage involves analyzing the morphology and syntax (and sometimes semantics) of the input text to create an internal representation. The translation is generated from this representation using bilingual dictionaries and grammar rules.

With this translation strategy, it is possible to obtain translations of a sufficiently high quality, with an accuracy in the region of 90% (although this strongly depends on the pair of languages).

In a rule-based machine translation system, the source text is first analyzed morphologically and syntactically to obtain a syntactic representation. This view can be refined at a more abstract level, with emphasis on translation-related parts and ignoring other types of information. The transfer process then converts this final representation (still in the original language) to a representation of the same level of abstraction in the target language. These two representations are referred to as "intermediate" representations. Then, for presentation in the target language, the same steps are applied in reverse order.

Various methods of analysis and transformation can be used to obtain the final result. Along with this, statistical approaches can be used to create hybrid systems. The methods and emphasis chosen are highly dependent on the system design, however, most systems include at least the following steps:

- ***Morphological analysis***. The words of the input text are classified by parts of speech (noun, verb, etc.) and subcategories (number, gender, tense, etc.). At this stage, all possible "analyzes" for each word are usually given, along with the initial form of the word.

- ***Lexical categorization***. In the text, some words may have more than one meaning, which leads to ambiguity in the analysis. Lexical categorization looks at the context of a word to try to determine the correct meaning in the context of the input. This may include the analysis of parts of speech, and the meaning of words.

- ***Lexical transformation***. This is mainly a dictionary translation, the initial form of words of the source language (possibly with semantic information) is searched for in a bilingual dictionary, and its translation is determined.

- ***Structural transformation***. If the previous stages work with words, then this stage - with larger components, for example, with phrases and phrases. Characteristic features of this stage include reconciliation of gender and number, as well as changing the order of words or phrases.

- ***Morphological synthesis***. With the help of the output stage structure -temperature conversion, formed the word in the target language.

Transfer types. One of the main features of transfer-based machine translation systems is a phase that "transfers" an intermediate representation of the text in the original language to an intermediate representation of text in the target language. This can work at one of two levels of linguistic analysis, or somewhere in between. The levels are:

- ***Superficial transfer (or syntactic)***. This level is characterized by transferring "syntactic structures" between the source and target languages. It is suitable for languages in the same family or of the same type, for example in the Romance languages between Spanish, Catalan, French, Italian, etc.

- ***Deep transfer (or semantic)***. This level constructs a semantic representation that is dependent on the source language. This representation can consist of a series of structures which represent the meaning. In these transfer systems predicates are typically produced. The translation also typically requires structural transfer. This level is used to translate between more distantly related languages (e.g. Spanish-English or Spanish-Basque, etc.). [4]

Interlingual machine translation

Interlingual machine translation is one of the classic approaches to machine translation. With this approach, the text in the input language is converted to an intermediate language, i.e. abstract, language-independent representation. Then the target language is generated from the intermediate language. Within the rule-based machine translation paradigm, the cross-language approach is an alternative to the direct approach and the transformative approach.

In the direct approach, words are translated directly, without complements or tion presentation. In the approach to the transformation of the original language of the pre formed into an abstract, less dependent on the specific language of submission. Linguistic rules that are specific to the language pair, then transformed representation of the original language into an abstract before representation of the target language, and it is formed by using the target language sentence (Figure 2).

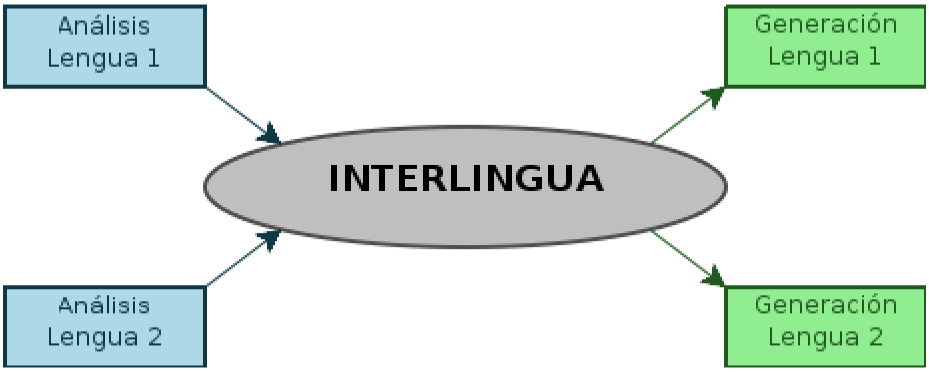

Figure 2. Demonstration of the languages which are used in the process of translating using a bridge language [5]

Machine translation with intermediate language has its advantages and disadvantages. The advantage of being able to translate multilingual machine translation is that you do not have to deal with component transformations for each language pair. The obvious disadvantage is that the definition of an interlingua is difficult and maybe even impossible for a wider domain. The ideal context for interlingual machine translation is thus multilingual machine translation in a very specific domain.

With this method of translation, the intermediate language can be considered as a way of describing the analysis of a text written in the original language, such that it is possible to transform its morphological, syntactic, semantic characteristics, and, consequently, its meaning into the target language. Such pro Mezhuyev precise language is able to describe all the characteristics of all the languages that must be translated, not just translate from one language to another (Figure 3).

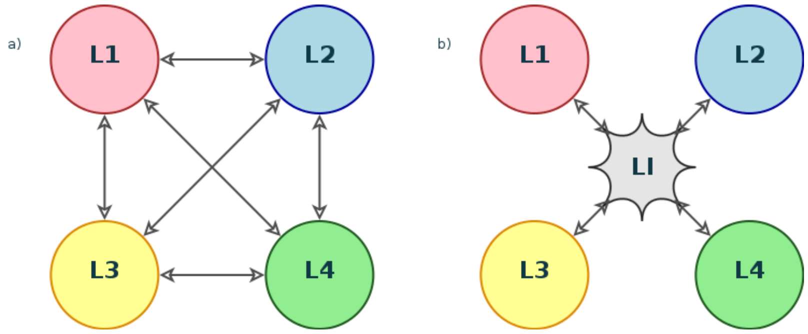

Figure 3. a) Translation graph required for direct or transfer-based machine translation (12 dictionaries are required); b) Translation graph required when using a bridge language (only 8 translation modules are required) [5]

Sometimes two intermediate languages are used in translation (Figure 4). It is possible that one of the two covers more characteristics of the source language, while the other has more characteristics of the target language. Translation then occurs by converting phrases from the input language into sentences closer to the target language in two steps. The system can also be configured so that the second intermediate language uses more specific vocabulary, which is closer or more similar to the target language, and this can improve the quality of the translation.

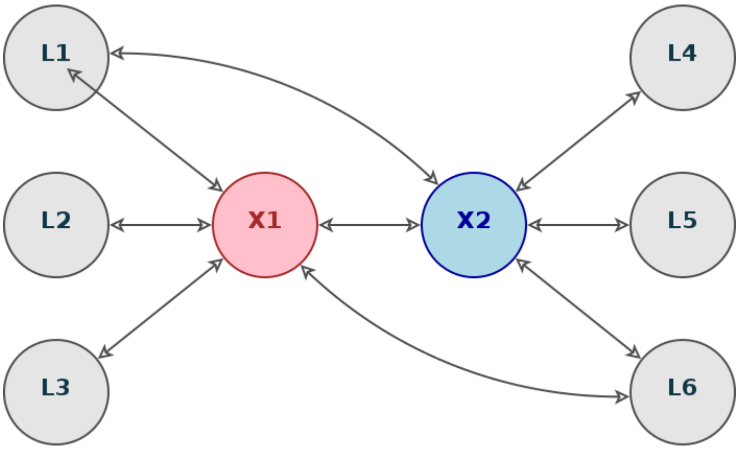

Figure 4. Graph of translation using two intermediate languages [5]

The above-mentioned system is based on the idea of using linguistic proximity to improve the translation quality from a text in one original language to many other structurally similar languages from only one original analysis. This principle is also used in pivot point machine translation, in which natural language is used as a "bridge" between two more distant languages. For example, in the case of translation into English from Ukrainian with Russian as an intermediary language.

In machine translation systems with an intermediate language, there are two one speaking component: analyzing the source language and the intermediate language and the synthesis of the intermediate language and the target language. At the same time, it is necessary to distinguish between systems with an intermediate language that use only syntactic methods and systems based on artificial intelligence.

The following resources are required for an intermediate language machine translation system:

- Dictionaries for analysis and synthesis (specific to the field and the languages used).
- Conceptual vocabulary (specific to the area), which is the knowledge base of events and formations known in the domain.

- A set of projection rules (specific to the area and the languages used).
- Grammar for the analysis and synthesis of the languages used.

One of the problems with knowledge-based machine translation systems (Figure 5) is the inability to create databases for domains larger than very specific domains. Another problem is that the processing of these databases is very computationally intensive.

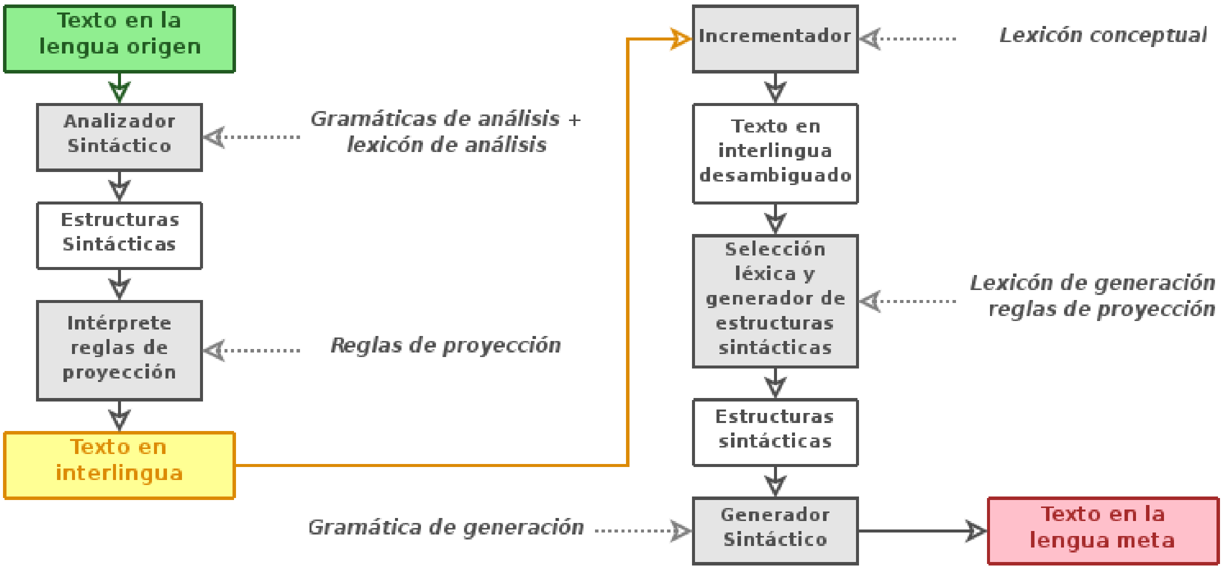
Texto en la lengua origen
Analizador Sintáctico
Gramáticas de análisis + lexicón de análisis
Estructuras Sintácticas
Intérprete reglas de proyección
Reglas de proyección
Texto en interlingua
Incrementador
Lexicón conceptual
Texto en interlingua desambiguado
Selección léxica y generador de estructuras sintácticas
Lexicón de generación y reglas de proyección
Estructuras sintácticas
Gramática de generación
Generador Sintáctico
Texto en la lengua meta

Figure 5. Machine translation in a knowledge-based system [5]

One of the main advantages of this approach is that it provides a cost-effective way to create multilingual translation systems. With an intermediate language, it becomes unnecessary to do a couple of translations between every pair of languages in the system. Thus, instead of creating n (n - 1) language pairs, where n is the number of languages in the system, it is only necessary to create 2n pairs between n languages and an intermediate language.

The main disadvantage of this approach is the difficulty of creating an adequate intermediate language. It should be both abstract and independent of the source and target languages. The more languages are added to the translation system, and the more they differ, the more powerful the intermediate language must be in order to express all possible directions of translation. Another problem is that it is difficult to extract meaning from texts in the input languages to create an intermediate representation. [5]

References

1. Sreelekha, S., (2016). Statistical Vs Rule Based Machine Translation: A Case Study on Indian Language Perspective. World Journal of Computer Application and Technology. 4(4): 46-57.

2. Rule-based machine translation. URL: https://en.wikipedia.org/wiki/Rule-based_machine_translation.

3. Dictionary-based machine translation. URL: https://en.wikipedia.org/wiki/Dictionary-based_machine_translation.

4. Transfer-based machine translation. URL: https://en.wikipedia.org/wiki/Transfer-based_machine_translation.

5. Interlingual machine translation. URL: https://en.wikipedia.org/wiki/Interlingual_machine_translation.