In this video, we are going to see the rule-based machine technology more in detail.

The first rule-based machine translation systems were developed in the early 70s. In fact, one of the first companies to use this technologies was Systran, founded by Peter Toma in 1968. Also, there were some Japanese MT systems, like Toshiba, NEC, Fujitsu, that used rule-based technology.

One of the most important projects during that time was the EUROTRA Project, founded by the European Commission from 1978 to 1992. This project used rule-based technology, and it consisted in developed systems over 7 to 9 official languages.

Nowadays, there is one popular toolkit that makes it easier to develop rule-based machine translation systems. It's the Apertium platform. This platform provides a language-independent machine translation engine, a tool to manage the linguistic data necessary to build a machine translation system for a given language pair, and linguistic data for a growing number of language pairs.

Initially, Apertium was developed for language pairs that were similar, so for example, Spanish-Catalan, and so on. Nowadays, it has been extended to translate pairs of languages more distant, like English to Spanish, or even Chinese to Spanish.

About licenses: Apertium is developed under a GPL license, which means that users can see the code, modify it, and it remains open source. Systran is a commercial system, so there is no access to the sources of this one.

The components of a rule-based machine translation system: Basically, a rule-based machine translation system is a transfer-based machine translation approach that has three steps, the analysis, the transfer, and the generation.

In the analysis part, there is a morphological and a syntactic analysis. So for example, the morphological analysis consists in, if we have, "Houses are expensive," this is the source sentence to be translated, what we first do is an analysis, a morphological analysis, that is, "house" "be" "expensive". This gives us the lemma of each word, and also the morphological information. Here instructs that "houses" is plural, "are" is present, and so on. Also, the syntactic analysis, what it gives us is that "house" is the subject, "be" is the verb, and "expensive" is the object.

Once we have this information analyzed, we go to the transfer stage that transfers the source information, that analyzes source information, into unanalyzed target information, and finally there is the generation step, that from unanalyzed target language generates the final forms by doing morphological generation and syntactic generation.

Here we have the block diagram of Apertium. We are going to see each block in detail. So first of all, the de-formatter separates the text from format information, is based on finite-state techniques, and it's available for plain text, html, etc.

So if we have this sentence, with the html information, we process it, we just get rid of the tags of html, and we get only the text, "this is a test".

After the de-formatter, we have the morphological analyzer that segments the source text in surface forms, as we have seen in the example. From full forms, surface forms, we go to the lemma and assign to each surface form a lexical form, lemma, lexical category, morphological inflection, information. Also, it processes contractions, like “can’t” is transformed into “can not”, for example.

After the morphological analyzer, we have the categorial disambiguator. So one word that can be polysemic, for example, “play”. Play might be a noun, like the role you play, the role you do in a play, or a verb, for example. So the idea here is to assign a category to the word.

This block uses hidden Markov models and is trained using representative corpora for the source language.

After the categorial disambiguator, we have the structural transfer. The structural transfer basically is where the rules that transfer information from the source to a target language are placed.

Rules have a pattern-action form, it detects the lexical form patterns to be processed using a left-to-right longest-match strategy, and it executes the actions associated to each pattern in the rule file to generate the corresponding lexical form pattern for the target language.

When we have languages that differ in reordering and have long reorderings to be done, here we require a more complex structural transfer. And we work with patterns of chunks.

Also, we have the lexical transfer that basically reads each source language lexical form and generates the corresponding target language lexical form.

Finally we have the morphological generation that generates from each lexical form a target surface form after adequately inflecting it, and also performs some target orthographical transformations, such as contractions.

The re-formatter integrates format information into the translated text.

Here we have some examples of dictionaries. We have a monolingual dictionary for the source language, here it’s Spanish, this is the source language, a monolingual dictionary for the target language, which would be a generation dictionary, and we have the bilingual dictionary that goes from the source to a target language.

The monolingual dictionary, basically here we have the word, the entry, “*cósmico*”, and we have another adjective that you wonder, what is that? So this means that “*cósmico*” follows the paradigm of inflection, of morphological inflection, of “*absoluto*”. The same in Catalan, we have the entry, “còsmic”, and we have the paradigm, which is “*académico*” in this case -- “*acadèmic*” in Catalan, in this case.

And the bilingual dictionary simply contains the entry of the monolingual dictionary, in the left part is the source language, and the right part is the target language. This is all for dictionaries, and we go to an example of transfer rule.

In this case, we are doing a rule that goes from Chinese to Spanish. And this rule, what it does is it reorders adjective + noun into noun + adjective. Also, we see here that we have the gender and number, it forces that the noun and adjective in the target language agree. If one is plural, the other one's plural, if one is feminine, the other one is feminine.

Question: Given the following rule, choose one interpretation.

This rule does:

Agreement between determinant, adjective, and noun in terms of number.

Agreement between determinant, adjective, and noun in terms of number and gender.

Reordering between noun and adjective and agreement between determinant, adjective, and noun in terms of number and gender.

The answer is the third one. Why? Because we have as input determinant, adjective, and noun, and then we get determinant, noun, and adjective, so there is a reordering. Also, there is an agreement, because we force gender and number to agree between determinant, noun, and adjective.

In the next video, we are going to see the theory behind statistical machine translation.