In this video, we are going to talk about machine translation approaches. We are going to focus on those approaches that are more popular nowadays, rule- and corpus-based.

Approaches to machine translation can be classified into different types depending on the criteria that we are using. We have, if we use the criteria of sources of information, we have the rule-based, human-written, the rule-based approaches that use human-written specific rules, and the data-driven approaches that use data to learn.

On the other hand, if we use the criteria of level of representation, we have the direct systems that translate from source to target without any type of analysis, we have the interlingua approaches that use an interlingual language to transform the source language into this language, and then from this language into a target language, and finally, in between the interlingua and the direct systems, we have the transfer-based machine translation approaches, which do some analysis of the source language, and then do a transfer stage, and finally generate the target language.

One question: Singapore has four official languages: English, Chinese, Malay, and Tamil. The government wants to build MT systems in all directions. Choose the correct answers.

An interlingua approach would require building eight MT systems.

A transfer-based approach would require building eight MT systems.

A direct approach would require building 12 MT systems.

Well, the point of this question is seeing that an interlingua approach requires less MT systems than a direct approach, or in having a multilingual scenario. So while an interlingua approach requires two times the number of languages that we have, a direct approach, or a transfer-based approach, requires the number of languages that we have multiplied by this number minus one. So in our case, a transfer-based or direct approach requires 12 machine translation systems, and an interlingua approach requires eight machine translation systems.

The rule-based machine translation paradigm is a paradigm that requires a lot of human dedication. It requires that a human person knows the source and the target language involved in the translation to design linguistic rules that make the transfer between the source language and the target language, and also will require building dictionaries that analyze the source language and that generate the final target language, and a third dictionary that goes from the source language into the target language.

There is an open-source platform called Apertium that makes it easier to build a rule-based machine translation system. It requires only the dictionaries and the transfer rules in appropriate XML format.

The corpus-based approach uses data to learn the translations. Which type of data? Basically, we need collections of parallel texts at the level of sentences. What does it mean? So we have the English sentence, and we have the corresponding translation in the Russian sentence, and like this we have a lot of parallel sentences to train our systems.

One of the first parallel texts was found in the Rosetta Stone, and it contained the same text in three different languages. The first one was Egyptian, the hieroglyphics, then it was the same text in the Demotic script, and finally in ancient Greek. So this Rosetta Stone allowed to decipher the hieroglyphics that we didn't know the meaning.

OK, in the web, we don't have, fortunately, we don't have to use stones now, we have the web. And we find parallel texts, and also comparable texts. Comparable texts are texts that talk about the same topic in different languages. There are a lot of techniques that try to extract comparable -- try to transform comparable texts into parallel texts, that parallel text is what a corpus-based approach needs to train the system.

Corpus-based approach can be classified into example-based or the statistical-based. Within the statistical-based, we have phrase-based, syntax-based, and hierarchical-based. The main difference between example and statistical is that example-based systems translate by analogy, whereas statistical-based translate on the basis of statistical models.

Example-based machine translation, when translating by analogy, it tries to solve a problem of pattern recognition. The simplest case is when I have a sentence that I have to translate that matches exactly 100% my translation memory.

Another question: Choose all properties that apply to statistical machine translation.

It's language independent.

No data needed.

Difficult to deploy.

Good knowledge of the language involved.

The answer is that in general, statistical machine translation is language independent, because we train our systems on parallel data.

Here, we see a scheme of a phrase-based statistical machine translation system. If it's still available, you can go to the YouTube address at the very top of the slide, and you will see a video by Google that gives you an overview of a phrase-based machine translation system. But here we go with the slide.

On the right part of the slide, we have the training, how we train a phrase-based statistical machine translation system. We use parallel corpus, and from each parallel sentence, we align the words. And once we have this word alignment, we do a phrase extraction.

In this case, we would have a phrase that is, "Voy a comprar," "I will buy," and many others. Once we have all our phrases, we have the translation model that focuses on finding the right target words given the source words.

Differently from the translation model, we have the language model that is trained on a monolingual corpus. And this language model assures that the translated words come in the right order.

These two models are combined in the decoding, that is, the left part of the slide, that is, the testing part. From the source-language text, we pass into a target-language text using the decoder. The decoder finds the most probable target sentence given the source sentence, combining the translation model and the language model information.

Differently from the phrase-based system, we have the syntax-based systems that also use parallel texts, and they require syntactic information. They require the parse trees of the sentences.

Also, we have the hierarchical systems that are a combination of the phrase-based systems and the syntax-based systems. Basically, hierarchic systems use rules that allow phrases to contain subphrases.

This slide is just informative to show which type of parallel corpus we can have freely available. We have the European Parliament Plenary Speeches, the Acquis Corpus that talks about European laws, the United Nations data, and the Canadian Hansards. Other sources of information are entities like TAUS or the Linguistic Data Consortium.

And popular statistical machine translation software available on the web are the word alignment, like GIZA, language modeling, like the SRILM, phrase extractors, like THOT, and the popular decoder, Moses.

Advantages of statistical machine translation, well, here we have a little list, that this type of translation is data driven, it's language independent, there's no need for language experts, it's easy to prototype a new system, and there is high coverage and flexibility of matching heuristics.

In the next video, we are going to see the main challenges in statistical machine translation.