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

If we analyze the translation output of a statistical machine translation system, we can classify the errors into the different linguistic levels, which would be orthography, morphology, lexis, syntax, and semantics.

On the orthography level, we have the following challenges:

Spelling mistakes and typographical errors, which may convert an existing word in the training corpus into an out-of-vocabulary word. In this case, a letter in a word is a mistake.

Truecasing and capitalization. It is common in a statistical machine translation to lowercase all training and test data in order to avoid orthographic mismatching, but then we have to make a postprocess that puts the capital letters when necessary. This postprocess sometimes has errors.

Normalization. Some words can be usually written in different ways, leading to orthographic differences with respect to the training corpus. It is better in a text to have the same word written in the same way. Even though it has different correct ways to be written, it's better to maintain coherence.

Then we have the detokenization, splitting a stream of text up into appropriate tokens to facilitate the input for further processing a text. This is what tokenization is about. The thing is that if we have a word and a punctuation mark next to it, we split the word and the punctuation mark to reduce the vocabulary.

And finally we have the transliteration, which consists in the conversion of text strings from one orthography to another, while preserving the phonetics in both languages.

Here we see examples of each one of the errors, spelling mistakes, truecasing, normalization, detokenization, and transliteration.

When we are at the morphological level, we can classify languages into different types, depending on the morphology that they have. For example, fully synthetic languages are those that allow for a word to have several morphemes, and they reach a point that a single word can be a single sentence.

On the other side, we have the isolating languages, that is, a one-to-one correspondence word and morpheme. So words do not have any type of inflection.

In the middle between polysynthetic and isolating languages, we have the agglutinative and fusional languages. The agglutinative are those that are allowed to have morphemes in a word and these morphemes are easily segmentable, so it's easy to identify these morphemes. And then finally, we have the fusional languages that contain morphemes and do not have clear boundaries.

Here, the challenge in MT is moving from one isolating language to an agglutinative language, because we are moving from a language that has no flections to a language that is highly flexive.

At the level of the lexis, we have lexical challenges such as unknown words. Unknown words are words that do not appear in the training set.

Here we play with the in-domain versus out-domain concept. The idea is that if we have trained a statistical machine translation system on a politics domain, if we translate a medical domain, we are going to have more unknown words than if we translate a text about politics, because it would be an in-domain problem, whereas if we translate a medical document, it’s an out-domain problem.

At the level of syntax, the main problem at the level of syntax is word ordering. Statistical machine translation paradigm assumes that the translation is monotonic, but in reality, it is not. We have languages that follow the order of subject-verb-object, and we have languages that follow the order subject-object-verb. This generates long reorderings.

And finally, at the level of semantics, we have the problem, for example, of word sense disambiguation. This would be at the level of lexical semantics, at the meaning of words, deciding which is the translation of the word in the particular context that we are using.

And then, also at the level of semantics, we have what is called the principle of compositionality, that is, that the meaning of a complex expression is a function of its parts. So we have to take into account the structure of the sentence to understand its meaning.

Question: You can mark all possible answers. Syntactic challenges include long reordering:

One, when source and target languages have different structures: subject-verb-object and subject-object-verb.

Two, when adjectives and nouns follow different orders.

Three, when languages involved in the translation are derived from the same family.

And four, when word ordering in one or both languages is flexible.

The answers are the first one and the last one.

Finally, I just wanted to show you a reference for you to have, the works that have been done in trying to resolve these challenges in the statistical machine translation, at the level of orthography, morphology, lexis, syntax, and semantics. Also, you have the reference of a survey on this topic.

In the next video, we are going to talk about rule-based machine translation.