UNIVERSITY OF LEEDS

Machine Translation: Challenges and Approaches

Professor Eric Atwell: Hello, this is Eric Atwell, and this lecture is an introduction to machine translation, including some of the linguistic challenges and some of the approaches in practice to machine translation.

For more details of current research methods and algorithms for machine translation, see the chapter on machine translation in the Jurafsky and Martin textbook on speech and language processing.

This particular talk is based on some presentation by Professor Nizar Habash. He's professor of computer science at New York University, Abu Dhabi.

And he has developed an Arabic computing research laboratory there called the CAMeL lab. So, if you Google Nizar Habash CAMeL lab, you'll find out more details of this.

OK, so you probably know about Google Translate. And if you have a piece of text and need it translating... in fact, currently on British television, they have some really good adverts for Google Translate which include speech-to-speech translation.

But even via the web, Google offers translations in, as you can see, many, many languages, including several I've never even heard of before. But it can translate between any of these things. So, Google is not interested in developing language models specifically for English, or just for Chinese, or just for Hindi. They only develop... They're developing language models which work in general with text of, not just the Roman alphabet, but many other alphabets too. And they want to have generic models of language which are statistical language models.

Just as an aside, another interesting language-related research project, Leeds University is the World Mapper project. And here's their URL. And World Mapper draws maps of the world. But they are transformed by...For each country, taking some figure, some number, which represents some feature in that country.

For example, here we have the map of the world where each country is morphed by the number of Arabic speakers in that country. So, in other words, it's a map of the Arabic speaking world.

Focusing on the big lump in the middle, I think, is probably Egypt. And nearby, you can see Saudi Arabia, and Syria, and Jordan, and other Arabic speaking countries. OK, that was an aside.

Now let's summarize. What we're going to look at in terms of machine translation, the road map is, first of all, look at some challenges for machine translation. Why is it difficult? Then look at some of the inherent approaches without going into details of these specific algorithms. And then finally, how do we evaluate if a machine translation is good or not? So, why is it difficult?

Well, one fairly immediate problem, if you're going to deal with all these different languages, is that they have different character sets. And the characters or spellings aren't necessarily

directly as straightforward or phonetic. Or very often, there may be ambiguous spellings. So, there are different ways...

A particular combination of letters can mean several different words. Particularly, for example, in Arabic where you typically leave out the vowels. So, there's many different sounds represented by the same sequence of consonants. In other languages, for example, in Chinese, there's the problem of word boundaries. We saw this in an earlier lecture on word modelling and word semantics.

There is also, even in English, a particular word. Even if it's ambiguous as a word like bank, it can have several different meanings. And each of these meanings or senses has to be translated into something different in another language. So, the word bank, in English, can either be a financial bank or the sides of a river... riverbank. And these are translated to two different words in Arabic.

And even words which you might not think are ambiguous can involve two different translations into another language. So, "eat”, you think, just means to consume or to put things in your mouth and chew and swallow them.

Whereas, in German, there are two different senses of eat. “Essen” is a human eating, and “fressen” is an animal eating. So, in the German language, there is a distinct difference between human eating and animal eating, and if you're doing a machine translation system, you have to get this right.

OK, apart from just the words meaning several things or different things. Different senses. there's also a problem with morphology. That is how a word is divided up into pieces or morphemes.

In English, this isn't very difficult. A verb may have a past tense. So, "kill”, the past tense is "killed", "write”, the past tense is "written", "do”, the past tense is "done". And you see the ending is different for each of these words.

In Arabic, often, an affix...What might be an affix in English becomes an infix or even two bits of things.

And tokenization can be even more complicated. Sometimes what is a single word in one language becomes several words in another language. So, for example, "and the cars" in English, it is essentially two words in Arabic rather than three words.

And we see in French. Well, in English, the "cars", "cars" is plural, and that's noted by "s" at the end of car. In the French, "et les voitures". "Et" means "and". "The" is translated into "les". And "les" has to have an "s" on the end to note it's a plural verb. And "voiture" is the French for car, and it also has to have an "s" to show it's a plural verb. So, sometimes the plural markers have to go on more than just one word.

Then you can have divergences in that the way that the meaning is divided up is different in different languages. So, in English, "I am not here" is just four words. In Arabic, the same phrase, "I am not here" is equivalent to two words. The first word means "I am not", and the second word is "here". Or in French, you'd say "je ne suis pas ici". And the "ne pas" mean "not", and you have "am" in between the two. So, it's "I not am not here". And this is quite common.

Even more complicatedly, what may be a verb in English, like "John swam across the river quickly", if you translate it into Spanish, well, you'd actually say something like "John crossed fast the river swimming". So, the verb becomes "crossing" rather than "swimming". Or in Arabic, it would be something like "Sped John crossing the river swimming", so, the verb becomes "speeding" rather than "crossing" or "swimming". And in other languages, it's different things. In Russian, it's "John quickly cross swam river", so "cross swimming" becomes the verb.

So, what counts is the verb is different in different languages for a sentence translated. So, that's some of the linguistic problems.

Another challenge is you have to have some resources if you're going to train machine learning models for machine translation. You obviously have to have a corpus, a data set. You might use something like WebBootCat to collect a web-based corpus. And you also have to have a dictionary, or list of words in the language. But actually, you don't just want to have. If you do build an English to French machine translation system. You're not good just to having a corpus for English and a dictionary of English and a separate corpus of French and a dictionary for French. You really need a parallel corpus with a source.

That's the English... And the target...That's the French... aligned, so you can work out this English sentence translates into that French sentence, or this English word. You have to learn that a line dictionary says for each English word, what is the equivalent French word?

And this is a problem because you can't just go to the web and find web pages in English with their French translations as easily as you can just find English web pages. But it's particularly difficult for languages which have few resources at all, like non-European languages, like Amharic from Africa or Bengali from Asia.

There's many millions of speakers of these languages. Native speakers... but there's very few webpages or parallel corpora. You may be able to get Bengali web pages, but getting English Bengali parallel pages, or Bengali Amharic parallel web pages is very hard. That's some of the challenges for machine translation.

Now, let's look at a very broad sense of, how is it done? What are the approaches?

Well, there's this idea of a machine translation mountain or pyramid. The idea is you have to start with a source. That's what you'll say, if you're translating from English to French, you have source words.

You'd have a source grammar of English and the meanings of English sentences. These have to be translated into the target. If it's into English to French, it would be French meanings, the French grammar, and the French words. So, you analyze or have analysis of the English source sentence, and then you generate the target French sentence. A simple way to do that is just word for word gisting.

Taking each of the words, and…Here's an example from Spanish. "Sobre la base de dichas experiencias se establecio en 1988 una metodologia". See, I don't speak Spanish very well, so I just said 1988. It isn't the Spanish word.

OK, here we have a word for word translation, and in English, it becomes "Envelope her basis out speak experiences then settle at 1988 one methodology”. And that's not a well-formed English sentence.

But maybe you can work out the gist... what it sort of means. So, if you ask a proper translator, a human translator, they would say, it means something like, on the basis of his experiences, a methodology was arrived at in 1988.

And maybe, from a word for word translation, you can work that out, because it's the same words, but moved around a bit. So, long as you know what the rules are for moving things around, then that will work. We'll have a look at a better example of that in a minute.

So, that's word for word. It doesn't work perfectly. Maybe what you should be doing is taking the grammatical structures of the source and transferring them into the grammatical structures of the target.

So, here's a simple example. In Spanish, you might say x. Whenever x and y are variables. "Puso mantequilla en" And that means "Put butter on". But in English, you can say, "He buttered his toast", for example. So, there we have a phrase, a structured phrase, in Spanish, translating into a single word or verb in English.

So, that's the idea of a transfer lexicon. So, if you have trouble with this, you need lots and lots of handwritten examples to do that. Maybe what you actually want to do is work out the meaning. Some representation of the semantics, or meaning...of the English sentence, and then convert that into the meaning of a French sentence, and then generate a French sentence. There have been many attempts to do this sort of thing.

These are fairly complicated graph structures representing meanings or phrases, like "John broke into the room", and their equivalents in other languages. This gets fairly complicated. So, in fact, that didn't really work out, because trying to work out hand-drawn graph structures for individual meanings of sentences never works very well. What you really want to do is, if you can get hold of a large corpus, a large training set, then you could apply machine learning to it.

In this case, it's a parallel corpus...a corpus with lots of English sentences and for each English sentence, the French equivalent. And from that, you can work out, for each English word, what is the French equivalent word, and what are the rules. The mapping between them.

So, you really need to have dictionaries, or lexicons. You have to have a parallel corpus, or parallel dictionary. You also have to have a dictionary of phrases and a dictionary of meanings for all these things. So, lexicon is another word for dictionary.

So, how can you work out how to translate words and phrases from one language into another? Well, a nice example, if you Google this, is Giza plus plus. Giza is a place in Egypt. It's also a statistical machine translation toolkit used to train word alignments. It uses expectation maximization. It essentially tries to find out, what's the most likely combination?

So, here we have, as an example, an English sentence, "Mary did not slap the green witch”. And the Spanish translation is "Maria no dio una bofetada a la bruja verde". And here we have the actual mapping. "Mary" is "Maria". "Did not" is "no". "Slap" is "dio una bofetada"… "give a hit”, if you like. "The" is "la" "Green" is "verde" and "witch" is "bruja". We see the green witch.

The adjective noun is converted into noun adjective in Spanish. So, you have to work that out. And if you get enough examples, you can do that.

IBM developed a model for basically gisting word for word translation, but they worked out, as we saw in the initial example, that you actually have to change the order of some of the words. They came up with various sort of hand-drawn rules for doing this.

So, for example, "Mary did not slap the green witch”. First of all, there's a fertility rule, which says that some words, like "slap”, have to be converted into longer phrases, like "Mary not slap, slap, slap the green witch”. And then there's null insertion.

After slapping... in English, you slap something, whereas in Spanish, you slap to the something, or you give a slap to something. And then you have a translation, translating each word for word, "Maria”, or "Mary”, in this case, "no daba una fotefada a la verde bruja". That's more almost Spanish. Finally, you have this distortion rule, that adjective noun becomes noun adjective. To swap them around. And you end up with "Mary”, or "Maria", "no daba una botefada a la bruja verde".

In other words, you can do word for word translation, with some extra steps in, and that gives you a reasonable translation for sentences which are fairly straightforward for languages which are quite similar.

So, Spanish and English, or German, or French, or most other European languages, have more or less the same sort of grammar, and therefore that works. It doesn't work so well for English to Arabic, or English to Chinese or Japanese, because they are very different.

So, another way around this is what's called phrase based statistical machine translation.

We've already discovered that sometimes, phrases like "green witch”, you can't just simply do word for word. You have to change things around. So, what you can do is try to look for phrases of several words, rather than just individual words. So, you might notice, for example, here's a German sentence. "Morgen fliege ich nach Kanada zur konferenz". And the mapping for this..."Morgen" is tomorrow, "fliege ich", is "I will fly”, so that has to be swapped around. And "nach Kanada zur

konferenz" is to the conference in Canada. You don't say "I will fly in Canada to the conference”. You have to say, "I will fly to the conference in Canada”.

So, there's always swapping things around, but the things that are moved around are whole phrases, not just individual words. So, if you can work out where the phrases are, then you can build this machine translation system much better.

So, how do we work out what the phrases are? Well, if you have. Going back to our example, "Mary did not slap the green witch”. If you know more or less what the word for word translations are, then you can infer larger phrases that might be translated.

So, for example, "slap" is "bofetada". And "the" is "la". But there's this "a" in there which is missing. So, maybe you can combine these together and say, if you're not sure what "slap the" is, then it should be "dio bofetada a la".

Similarly, you can infer large... So, "green witch" becomes "bruja verde" so you've got to swap them around. But you've just always learned that "green witch" is always "bruja verde". And there's no swapping to do. You just learn the whole phrase in English goes into a whole phrase in Spanish. And he's got bigger and bigger phrases.

"Mary did not slap" becomes "Maria no dio una bofetada". "The green witch" becomes "a la bruja verde". Or perhaps the entire sentence. So, that's one way of doing phrase based machine translation.

Many, many mappings can handle non-compositional phrases. You can get a whole complicated phrase, or even a whole sentence, can be translated in one go, and therefore you don't have to worry about moving things around.

And also, you can capture things like "interest" being ambiguous, between monetary interest or "I like you" type of interest. But if you just translate "interest rate" differently from "interest in”, then you don't have to know that interest has got two different senses. "Interest rate" is different from "interest in”. They're no longer a single spelling with two different senses. They've got two different spellings with two different senses.

And furthermore, just as we saw with earlier work by Banko and Brill, who discovered that if you get more and more data, then the classifiers work better and better.

In the same way, the more data you have, the longer learned phrases you can learn, the more long phrases you can learn. And then eventually, you learn to translate whole phrases and sentences, and you don't have to use word for word mapping anymore.

That's basically how Google Translate works. They started off doing word for word translations. Over time... if you go to Google Translate now, it invites you to say, is this a good translation? If not, can you suggest a better one? And from all of these better ones, it's been able to work out more and more longer phrases that it learns. So, that's machine translation approaches.

As I said, if you want to look in more detail about algorithms, then have a look at the chapter in Jurafsky and Martin. They've got a lot more about current research into algorithms for machine translation, which I am...in this module, it's not enough time or space to go into details. But if you wanted to, for example, for your project, you could well try out this Giza system or some of the other approaches that are in the Jurafsky and Martin textbook.

Finally, I want to say, how do you know if it's any good?

So, a machine translation system comes to you, or, what often happens here at Leeds University, we have a Centre for Translation Studies, and companies will come along and say, we're using this machine translation system. Is it any good? Or, we're trying to decide, should we buy x or y. Which one is better?

And it's pretty much an art, or at least it used to be. Well, one thing, as with most IT systems, you can measure things like, is the interface nice? How does it scale? How does it interact with your existing other systems? Things like that. How slow is it? How fast is it?

But maybe what you're actually interested in, in terms of how good is it, is, is it actually correctly translating? Is the translation correct, regardless of how long it takes or how expensive it is? And that's where we want to... that's where most research in machine translation evaluation goes into, rather than just evaluation of IT systems.

Specifically, about machine translation, it's, how good is the translation? And this often involves using people to do it, rather than…humans can do it very well. They can look at a translation output and say, this is good, or I can see some mistakes in it. But they can be quite slow.

Whereas it would be nice to be able to do this automatically, or semi automatically, because you can speed it up, and you can also evaluate a lot more outputs.

So, the way that it's done by humans is, typically, you apply the machine translation system to a test set, and it produces output. Say, 100 sentences in English are input, and outcome 100 French sentences. And you give these to some people, some human translators, and they rank them in terms of fidelity or accuracy. Standards ranges from 5 to 1. 5 means it's pretty much perfect, it might need some minor corrections, whereas 1 means it's completely missed it. The contents of the original sentence are not conveyed.

And in between the two…it's not just a binary thing. You can get 2, 3, or 4, as well. That's in terms of, is it getting the meaning across? Is it capturing... is the French output sentence, does it actually mean the same as the English input sentence?

However, there's another thing to measure, and that's called fluency, or intelligibility. Is the output, the French sentence actually clear, good grammar, good sentence structure? Does it sound natural? Because if it's not natural, then it can hide the meaning, and at worst, you can have a sentence, a word for word translation, for example, which literally does capture the meaning of the English sentence, but it doesn't make any sense. Its meaning is absolutely unclear to an ordinary French speaker.

So, there's two measures. There's fidelity, or accuracy, and separately, there's fluency, or intelligibility. So, machine translation evaluation measures both of these things. And it can be quite expensive to get.

Typically, you want to have two or three people, not just one person, measuring the output for each sentence, and coming up with two scores. One for accuracy and one for intelligibility, and then somehow combining these to give a score overall. That's quite time consuming.

So, maybe another approach is to semi-automate it. IBM came up with a metric called BLEU. "Bleu" is the French word for "blue”, and blue is what IBM is, so that's why they called it that, I suppose. But it also stands for the bilingual evaluation understudy. Understudy is... In a theatre, it's someone who stands in if the actual actor or actress is ill or is unable to play. So, BLEU helps the humans. It doesn't actually completely take over for humans.

What it does is, given an English sentence, and it's got to be translated by humans into French, there's never one single perfect translation. So, what you do is you get three translators, or several translators, to translate each sentence, and you come up with a bank of translations for each of the inputs. Then, the BLEU metric calculates a score for that.

It's particularly inexpensive and language independent. The score works reasonably well. So, it's a way of generating a score, and the nice thing is the score correlates with human evaluation.

So, if there's two machine translation systems, and BLEU gives one a score of 56 and the other score 49, that means that the 56 one is better.

And typically, people will agree that the one that the BLEU says is better is actually the better one. People may have other...not a straightforward score, but they'll have more subjective opinions about it.

It compares the machine translation against several human translations to give a standardized score.

Let's just give an example of this. Let's say, for a particular French sentence...I can't remember what the French sentence is. The machine translation output was "Colorless green ideas sleep furiously”.

Anybody who's listening to this, who's read the works of Noam Chomsky...Noam Chomsky is a very famous linguistics researcher, and he came up with a book called Syntactic Structures where he used this as an example of a sentence which, in English, is grammatically correct, but doesn't seem to mean anything. Anyway, here's the sentence... "colorless green ideas sleep furiously”.

Now, the humans have taken the French sentence, and the first one translated it "All doll jade ideas sleep irately”. And a second human translated the French sentence into "Drab emerald concepts sleep furiously”. And the third translator took the French sentence and translate into "Colorless immature thoughts nap angrily”.

Now, we see that this is a made up example, just to make it easier to explain the mathematics, but we see that none of the translations are the same. All three are done by humans, therefore, should be OK. But the machine translation output isn't the same as any of them. So, if you want to say machine translation is good if it's the same as a human's... Well, it gets a 0 score.

So, that's no good. And furthermore, the humans don't agree. So, as far as human 2 and human 3 are concerned, human 1 got it wrong. So, requiring an exact match is unreasonable.

What you can do is look at bits...That's unigrams and bigrams, individual words and pairs of words, to see how well they are translated. And that's what the BLEU metric does.

First of all, there's a unigram score, or unigram precision. In the test sentence, how many of the words, the individual words, appear in any of the outputs, or the gold standard references by the human translators? And "colorless" appears in one of them. "Ideas" appears in another one. "Sleep"

appears in two of them. "Sleep" does appear there. "Furiously" is in one. "Sleep" appears in twice, but it doesn't get a double score.

But four out of the five words in the test sentence from the machine translation output are included, so that's a unigram precision of 4 out of 5.

And then you also look at the bigram position. Well, "Colorless green ideas sleep furiously" maps onto four bigrams... "colorless green", "green ideas", "ideas sleep", and "sleep furiously". And "ideas sleep" does appear in one of the translations, and "sleep furiously" does appear in another translation. Therefore, it gets the bigram score of 2 out of 4.

And then the BLEU score is simply, multiply together all the ngram scores and take the n-th root of the ngram scores. In this case, we're just looking at unigram and bigram. You can look at trigrams as well, but the simple model is just to take the unigram score and the bigram score.

That's 0.8 times 0.5, and the square root of that is 0.6325, or 63.25%. And that's the score. You get a score of 63 for this machine translation system, and then you try another machine translation system, and whatever its output is may get another score...Let's say 52...And that tells you that this machine translation system is better than the other one.

It's not that the humans will come up with a score of 63, but rather there's a correlation between the ranking given by the BLEU score and the ranking given by human evaluators. But the human evaluators will take a lot longer to make this decision, whereas BLEU gives you a metric which works, and it seems to be more or less correlated to humans but does it consistently and efficiently.

OK, then, to summarize, machine translation is difficult for various reasons.

Languages have got different writing systems. The words are divided up differently in morphology. That is, dividing up words into segments or morphemes is different. Even for a whole sentence, it can be divided up into different sentences or words differently in different languages, and the ordering of words is different in different languages.

Also, there's the point of needing resources. You don't just need to collect a corpus using Sketch Engine or WebBootCat. You need a parallel corpus. You need to have...For a pair of translations for English to French, so you have to have an English corpus, and each sentence, its French translation. And that could be quite difficult.

There are sources like the United Nations, where every document is translated into English, French, Chinese, Hindi, Arabic. Something else? No, I think that was it. That means there is a huge parallel corpus available for those languages of United Nations legal texts.

But if you want to have other languages or other domains, other subject areas, then it's difficult.

You also need dictionaries, and not just dictionaries as lists of words, but translation dictionaries. You have to have an Arabic to English and an English to Arabic translation dictionary if you want to translate between those two languages.

You also need a bank of other tools, like segmenters and parses and things.

So, the machine translation approaches, there are simple word for word mapping. To do it a bit more sophisticatedly, you have to have word for word plus rules for messing about the order, reordering things in special cases.

There are also transfer models, where you transfer entire phrases rather than just individual words, and you can build this up over time. You have a translation memory of things you've translated before, which includes not just the words, but the phrases and the sentences.

You can try an interlingual where you have some sort of semantic representation of the input sentence, and that's mapped onto the semantic representation in the target language, and then you generate the target sentence. That's maybe sort of a neural network type approach, as an alternative.

And finally, we have this issue of evaluating. How do you decide if a machine translation system is good?

Well, it's not just an accuracy score, but there's also fluency that has to be taken into account. These are two separate things, a bit like precision and recall in measuring how good a classifier is. So, accuracy is how much of the original meaning is carried across in the output translation, and fluency is, given the output translation, can an ordinary native speaker of that language actually understand it?

And you may have very accurate, in terms of word for word translations, which are very difficult to understand and very low fluency. Doing this by humans, getting people to do it, is quite time consuming and therefore costly. IBM BLEU is another method. And there are other ones too, but BLEU is the famous one, the first one to come up with this automated method or semi-automated method.

If you get a test set of sentences translated by humans, and each one has to be translated by several translators, say, three or so, then you can count the unigrams and bigrams overlap between the translation output and the human translations, and then use this to calculate this score.

And this score correlates very well with human evaluations of machine translation systems more generally.

OK.

As I said, if you want more details of the algorithms, then go to your Jurafsky and Martin textbook. In fact, please do read that textbook, because there may be questions on that in the exam coming up, or the test coming up at the end of the module.

Enjoy, and try out some of these tools yourself, if you want to.

And that's it for me for now. Bye.