Multi-word expressions

Tom Pickard: Hi. I'm Tom Pickard. I'm a research fellow at the University of Leeds, in the School of Computing, and the main thrust of my research at the moment is a project called the EDUBOTS project which is an EU-funded project looking at chatbots and conversational AI in higher education. So we're specifically interested in how chatbots, conversational agents, can be used to help teachers and students and support staff in universities deliver the best quality teaching and experiences that we can for students.

I'm going to talk about Multi Word Expressions, which is essentially a linguistic phenomenon, really, but I'm going to talk about them in the context of the consequences that they have for natural language processing with computers.

So, first of all, what are Multi Word Expressions? Essentially a Multi Word Expression, or an MWE, is a group of more than one word which functions linguistically as a single unit of meaning. So I've got some examples on the screen here of things like the phrase red herring, or spill the beans, a pain in the neck-- which are phrases which we understand and we are familiar with as a unit of multiple words, rather than just as single individual words. I've got at least one example here, l'espirit d'escalier, which is French, so this is not specifically a phenomenon in English. It's common to most if not all natural languages that are studied in the world, and that includes constructed languages like Klingon and Dothraki and elvish and things-- they also tend to create and include Multi Word Expressions. And not only do these things appear in lots of languages, but they're also relatively common as well. You may not be consciously aware of it, especially if you're speaking your native language, but they're very, very frequent. So up to about half, depending on who you ask and when they go and look, of the entries in word net, which is essentially a dictionary with some extra information in it, consist of more than one word. And this paper that I've cited there from Schneider, they looked at a particular corpus of, I think it was, reviews or comments from the internet, and they found well over half of sentences that were used contained at least one of these Multi Word Expressions. So they're frequent and they're a big deal. And they occur in lots of languages.

One of the properties that Multi Word Expressions have is something called non-compositionality, which is, effectively, that, even if you are familiar with each of the component words of an expression, like one-armed bandit-- you know the number one, you know what arms are, you know what bandits are-- unless you've been given the meaning of that phrase as a whole unit specifically, you don't know that that means a slot machine or a gambling machine that you might get in a pub or a Casino. You just have to-- that's external knowledge, that's independent of the meaning of the component words.

That's not just a yes or no thing, so some Multi Word Expressions are easier to understand than others. Or some you can sort of get some of the meaning. It kind of lies on a bit of a scale like I've got here. So something like the phrase Iron Curtain or Pandora's box, are very opaque-- they're almost impossible to understand without the extra context, even if you know what the word iron and the word curtain mean. Whereas a swimming pool or a fairy tale, you can get the gist of what those expressions mean. Or swimming pool is very easy to understand, if you know what swimming is and you know what a pool is. So we can imagine a foreign visitor, or maybe an alien, or a robot which has

perfect knowledge of every individual word in a dictionary, but doesn't know about these Multi Word Expressions, and they will very quickly run into trouble if they try to interact with users of the language.

So, as you might be able to understand and suspect, quite a lot of people here probably speak English or at least one of the languages as a second language, you've probably encountered some of these when learning languages. So Multi Word Expressions do tend to cause problems for language learners. And also they're difficult for natural language processing. The instinct, when you're getting a computer to process language, is just to break up the text into individual words, languages which use words in the same way, and look at them separately and then construct the meaning. But Multi Word Expressions give you a way in which that doesn't matter. And in particular, of course, if you think of machine translation as the combination of learning a language and getting a computer to process the text, you're effectively trying to teach a computer to work from one language to another. These become very, very difficult because the same Multi Word Expression-- there may not be an equivalent one in the target language as there is in the source, and so on. So they get really difficult. So I'll start talking a bit more about how we actually go about handling Multi Word Expressions. The first thing to note is you can't just create a list of all of the ones that exist in a given language, and just refer back to it and then solve the problem that way. They are-- it's an ongoing problem of needing to identify or discover new Multi Word Expressions.

People coin Multi Word Expressions frequently, especially if you're looking at social media corpuses, and Twitter, that kind of place. These examples I've got on the screen here for fake news, social distancing, chef's kiss-- a very recent coinage it is, but very, very familiar hopefully to anyone who's on Twitter. So there's an ongoing need to find Multi Word Expressions, and then work out what to do with them. So the starting point really, for trying to actually identify these things, is-- we look for what we call collocation, so that is words which occur together, close together, in sentences, or right next to each other in a sentence, more often than we might expect them to, if we assume that words are just sort of randomly distributed through text. So if we're looking for something that is occurring more often than we might expect it to by chance, that's the kind of thing that would imply that a statistical measure of some sort is what we want to use here. And we want that statistical measure to account for the fact that some words are more common just in general than other words are.

Conjunctions and words are very, very common. And there are loads of words, especially in a language like English, which you just don't encounter very frequently in text. And so you need to adjust for that. There are a number of different measures that people have come up with for doing this. I'm just going to very briefly introduce the idea. Don't worry about the maths, but the idea of Pointwise Mutual Information, or PMI, which is a sort of occasionally-used statistical measure for this. But essentially, the idea is-- so we look at some text, a big bulk of text, all of Wikipedia or a big chunk of the internet or something like that, ideally. We look for how often do we see a pair of words, x and y, fake and news, occurring together, compared with how often we see those words occurring on their own, individually. So news is a reasonably common word, probably shows up relatively frequently in English text, especially on the internet, I suppose. But the pair fake news is probably more frequent than we would expect it to be if we just looked at the frequency of the word fake and the word news. And we can do some maths using that and construct something like this Pointwise Mutual Information measure and then, when that number is big, that implies that these words are co-occurring more often than we'd expect them to, which tells us there's probably something interesting about the combination of the words. However, that isn't enough for them to constitute a Multi Word Expression.

There are pairs of words which occur together frequently, which don't actually make up a single unit of meaning. So I've got a couple of examples here like strong and coffee, or black coffee. The verb to kick will occur often with the object, a ball. That's just part of the way in which language is used. So you need to do something else beyond just finding these interesting co-occurrences to actually look for Multi Word Expressions. And one of the ways of attacking this, and something that I looked at for my dissertation research, to try to use semantic embeddings or word vector embeddings. Hopefully this isn't entirely new to you, but you can use tools like word2vec or GloVe, which are the ones that I use, there's various models, to try to capture, with maths as vectors, something about the meaning of a word. So you replace words with a vector in this large multi-dimensional space. And by doing so, through the way in which the model has been trained, we capture something about the meaning of the word, or certainly in the relationships between them anyway.

OK, so if we're looking for non-compositional Multi Word Expressions, that is, we know that the meaning of the expression as a whole differs from the meanings of the individual component words. So if we're talking about the phrase green thumbs, the meaning of that phrase is very different to the meanings of the word green and of the word thumbs. If you add the word green and the word thumbs together, you shouldn't expect to get the meaning of the phrase green thumbs. So maybe we can use vector embeddings to help us find where this is true-- where the phrase as a whole has a different meaning to the components. So if we take the two separate words where they occur in some text and join them together into a single word or a single token, and then build some word vector embedding models, and then we can look at the vector that represents green thumbs, in this case. And the vector for the word green, the vector for the word thumbs, and look at the differences between them and see how similar they are and if they're less similar if these vectors are further apart in this multi-dimensional space. That tells us that our phrase is less compositional, and therefore, that we should treat it as a Multi Word Expression.

So the intuition is that we should be able to look at pairs of words which are statistically interesting in our text, and see if the meaning of the pair differs from the meaning of the component words that make it up. So in order to do that, I took a relatively large corpus, I took the text from the English Wikipedia. Wikipedia is lovely, and allow you to just download all of the content from there on their extracts that they produce once a month or so. And because I was doing this as an MSC project and working on just my own personal computer, I ended up taking a 10% sample of sentences in Wikipedia, just to reduce the size of the space and give me something that was a bit more manageable, and then looked for these pairs of words, or actually pairs and triples of words, that occurred together. Calculated the co-location measure, like Pointwise Mutual Information, took the most interesting to the highest-ranked half a million items from on that list, and then divided them up into multiple batches and replaced the pairs or the triples of words with a single token, and trained vector embedding models on that text to try to capture the meaning of the potential Multi Word Expressions.

I did this two different ways. I did it with a word2vec model and with GloVe. Hopefully you've encountered both of those at least briefly. And then for each of the models, compared the vector for the potential Multi Word Expression with the components, calculating the cosine similarity. And all of this was based on a methodology by Will Roberts and Markus Egg from a couple of years ago. And the idea behind all of this was to allow me to compare word2vec and GloVe, which are different approaches to making vector embeddings, and see which one was more useful or more effective for this task. And my kind of hypothesis going in was, so Roberts and Egg did this with word2vec, maybe

GloVe will produce better results for some reason. It's a different way of constructing the word embeddings, maybe it will be better for this particular application.

And it turns out, it's not.

I found the opposite, in fact. So GloVe produced either lower quality results, so-- results where the ranking that I produced for how similar things were, and therefore how likely they were to be a Multi Word Expression, didn't correlate well at all with what humans thought of those same expressions. So there are reference data sets that you can use, where humans have been asked to evaluate how non-compositional an MWE is. Or you could get GloVe to give you similar results to word2vec, but the computational costs were higher, so I needed to run for longer, or took up more memory on my computer, that kind of thing. So from a practical point of view word2vec was the better option. I did also conclude that-- I mentioned earlier-- that I took a 10% sample of the sentences in English Wikipedia to work from, and that yielded very reasonable and quite effective results. 10% of Wikipedia was enough to give you a good idea of what would be-- of what was effective. You don't need to bother processing all of Wikipedia or something that large, and you can still get an idea of the results here. And there is a link here to code and my final data outputs, if anyone wanted to look at those and to share those with the community.

And then there are some open questions, then, or some areas in which this could be sort of expanded upon, or looked into, in order to understand a bit more about this, or try to maybe push this idea a bit further. So I was working with Multi Word Expressions which were sets of groups of two or three words which occurred next to each other in the text. But there are other kinds of Multi Word Expression that we find in language, where there might be other words in the sentence or the structure might be more complicated. Could one expand this idea to try to look at those? Obviously I was working in English specifically, but you could take this idea and apply it to other languages and so on. It's obviously worth talking briefly about this idea of words and Multi Word Expressions with different languages.

So English tends, for the most part, to separate words with spaces. It's a pretty good approximation, if you just look at English as a sequence of words with spaces in between them. But that doesn't hold true nearly as much for some other languages, so languages like German, and other Germanic languages, tend to combine nouns together into single compound nouns with lots of-- several words effectively sort of smushed together. So you need to worry about how to separate those first. And there are other languages like Chinese and Arabic and things, where just word boundaries sort of work very differently in the language. So there's lots of extra sort of challenges to just taking this method and applying it to a different language. It's potentially quite inefficient. It's quite expensive.

I was training a word2vec model, or something like that, or 10 of them in fact, for every list of candidates that I was trying to explore. So there's potentially some ways that you could be done to improve the efficiency of such an approach. As I mentioned, it's an ongoing problem discovering MWEs, and you probably want to be able to do it reasonably efficiently. Something else I'm particularly interested in is looking at the way in which people use Multi Word Expressions in particular domains as well. So it's relatively easy to get hold of lots of general text for English. So you can grab things from the internet or Wikipedia or social media, and you'll get a general sense of how people use English just in conversation.

But if you're interested in the way in which Multi Word Expressions function, or even just the Multi Word Expressions which exist in quite specific domains to specific topics of text, like academia, or like data science modules, or medicine is a really good example-- it's harder to get those specialised text corpuses and work with those. And there's a big question for a lot of word embedding models and things, about how well do they transfer? They're often built on very large data sets of general language, how well do they transfer and apply to those more specific applications?

And finally, how do Multi Word Expressions interact with more complex or more contemporary word embedding approaches like BERT. So I think Eric mentioned BERT at some point earlier in the course, that's a different kind of neural network for learning the word embeddings, and I'm interested in trying to maybe explore, A, does this same kind of approach affect the way BERT works? Does telling it that a pair of words is actually a single unit of meaning, then make it-- allow it to process that word better? Or does something about the architecture, and the way that BERT already works already give it some of that capability? And if the latter is true, if it is already capable of handling Multi Word Expressions to some extent, then can we flip the problem on its head, and can we use BERT-- or by investigating its architecture and its weights, can it tell us about Multi Word Expressions which it encountered in the text. Is there something in the structure of BERT that says, Oh, hang on-- green plus thumbs together means something different from green and thumbs. So that's a potential area for research, which I'm not familiar with any sort of detailed work that's already been done.

And then the other thing I've got is a bit of suggested reading if anyone is interested in finding out a bit more about this. And the first thing I was going to recommend here is not so much reading as viewing. There's an episode of Star Trek-- The Next Generation which, I don't know if somebody might have already seen, called Darmok, which is not only one of the best episodes of Star Trek that's ever been made, but it's also a really, really good expression of-- an example of Multi Word Expressions and the challenges that they provide for both for humans trying to kind of make themselves understood and work in multiple languages, but also for machine translation, trying to handle the same problem.

There's a very readable paper by Sag and his colleagues called Multi Word Expressions: A Pain In The Neck for NLP, which really sets out, from 2002, really sets out the sorts of problems that people working on Multi Word Expressions have been trying to solve for the last 10 years or so.

And then if you really want to dive a bit more deeply into them, or especially if you were interested in looking at something in MWEs for a project, I would recommend Carlos Ramish's book. Carlos Ramish organises lots of conferences and things on Multi Word Expressions, and his book covers a lot of detail about how going about discovering and working with them.

And then finally I've put in a little plug for myself, so there's a link there to my paper that I wrote last year describing these results and sort of sharing them with the community.

[END]