Tagging POS and NER

Professor Eric Atwell: Hello. This is Eric Atwell, and in this lecture, I'm going to tell you about sequence labelling for part of speech tagging and named entity recognition.

And for more details, please see the chapter in Speech and Language Processing by Jurafsky and Martin.

A lot more details on this, but I'm going to try and give you an overview of the challenges. First of all, we'll look at part of speech tagging. That is, adding to every word in a text its part of speech.

And then we're looking at named entity recognition. Labelling entities or names or things or people or places. And both of these are examples of libelling words in a sequence of words in a text.

So, parts of speech. This is one area where, if you've learned English as a second or foreign language, then you should well be familiar with this. If you've just spoken English from birth and never really learned English formally, then you may not be so aware of what these are.

But it's actually part of a very old English tradition, and before English existed, other languages also used this idea that words, in text or in speech, are divided up into categories, and these categories are called parts of speech or word classes or pause or pause tags.

At least, in terms of literature, the earliest documents describing parts of speech that were known about are from around the time of Christ, was Dionysus Thrax of Alexandria. A scholar then wrote about describing Greek and Latin, and words were categorised into nouns, verbs, pronouns, prepositions, adverbs, conjunctions, participles, and articles.

And these categories, because they were used in describing Greek and Roman, Latin, and then they went on to describe other languages which developed from there – including English - they're still widely used today.

Notice that this list doesn't cover everything, for example, everything you might be familiar with. There's no adjectives in this list, and that's partly because adjectives were not seen as being as important somehow.

OK. So, there's eight classes, but there's actually two broad classes of words, the closed class versus the open class. So, closed class, often known as function words, are the short, very frequent words which have some sort of grammatical function, and they don't really have much meaning attached to them.

So, for English, you've got determined as - the pronouns like she, he, and I, and prepositions like on, under, over, and near. And others as well. Notice these are fairly fixed.

You don't usually come up with new pronouns, for example. And you can do - so for example, quite recently, to deal with gender differentiation, rather than - I've been advised that in reviewing research proposals to avoid gender stereotypes.

I should not use the word she or he, but I should use the pronoun they whenever I'm describing - referring to a person. So, there is actually a plural pronoun, but it is coming to be used as a singular pronoun for a person to replace he or she.

Well, that's unusual. On the other hand, open class or content words, they keep on coming up with new words. So, in a corpus - in a test set, if you have out-of-vocabulary words, words which are not found in the training set, almost always these are content words.

These are nouns, verbs, adjectives, or adverbs. And also, possibly interjections like oh, ouch, uh-huh. Little words which are used for phatic communication, that don't really have entities or actions attached to them. And you can have, all the time, new nouns and verbs being generated and used.

iPhone isn't particularly new anymore. Google used to be a noun, and now you can google. I want to google something. It's being used as a verb.

So, just an overview, and probably you know some of this already. This is not really something that you might think this isn't computing, as opposed to linguistics, but at least you have to be aware of these categories if you're going to build AI systems for analysing them.

So, there's nouns, and nouns are subdivided into proper nouns and names of people in places like China and Italy.

And common nouns, which don't denote a particular person or thing, like cat or mango.

And then there's verbs. But before we go onto verbs, I want to briefly say, adjectives, adverbs, and interjections are fairly straightforward open, and you can have more and more of those things.

Verbs are mainly open class because you can keep on - you can generate new verbs, like to google something. But there are a few verbs, like I can do things, which is a sort of verb.

It's often called an auxiliary verb. And words like can and be and have are - there's only a small number of them, so they're more or less closed class.

But most of the closed class are fairly clearly defined like determiners, conjunctions, pronouns, prepositions, and particles.

And possibly others too. Just notice that that's pretty much everything. But in ordinary text, there are also a few other things like numbers. And numbers sort of halfway between the two, because there's only a set number of digits, or a set number of 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12 in English.

And then many of the other numbers are made up by some combination of those things. So, they're sort of open-ish.

But pretty much, you can make up a number like 122,312. But there's sort of grammar for generating these, and the parts, the individual component, is like 100, 20, 2, and so on. This is a closed class of those.

OK. So, that's that.

So, part of speech tagging involves assigning a speech, or tag, to each word in the text. It's made a little bit more difficult because you might just say, take each word, look it up in a dictionary.

What is the part of speech? Whatever it is, add that to the text.

But words can have more than one part of speech.

We saw Google before - book is another example. A book, you might think, is hand me that book. That's a noun. But you can also book a flight, which is a verb. So, it becomes a little bit more difficult than simply taking each word and looking it up in a dictionary because, in the dictionary, you may have two or more parts of speech for whatever word.

And then you have to take the context into account to work out which one it is in this particular context. And that means you're going to map a sequence of words- x1, x2, and so on - onto a sequence of part of speech tags. And the location within the sequence of the word will determine what is the appropriate part of speech in that location as well.

There are many different tag sets which have been developed. We saw the early Greek and Latin one. There is an attempt by researchers to come up with a universal dependencies tag set.

Universal, in the sense of it, applies to all languages. And these are sort of cheap, because you've got the obvious ones, the open classes, adjective, adverb, noun, verb, and a special one, proper noun, to allow for the difference between common noun and proper noun.

So, noun just means common noun. A prop noun means | noun. And then you've got the closed classes, which again, we've seen before. And to allow for the possibility of others, there is another class as well.

For example, punctuation doesn't fit in there, but it is - punctuation is important in text, so, it gets its own symbol, or own tag.

In some texts, you could have other symbols like dollar signs or emojis. So, therefore, there's an extra tag symbol. And there's also another tag, X, for anything else.

Because in whatever language it is, there may be additional things. For example, URLs, which don't really figure amongst all these other ones.

So, it is a universal tag set, but it allows for variations in different languages by allowing other tags to be added for particular languages or particular types of text. So, here's an example, applied to English. And typically, for each word, you have a tag.

And one way of writing this is just to write the word, and then a slash, and then the tag. Or, in Sketch Engine, it will be a word, and then a space and a tag. So, you can do tagging in Sketch Engine, as we saw previously.

And this is just the universal tag set. There are lots of other tag set, and these can be dependent on a particular language. So, it leads - I've been involved in the corpus of contemporary Arabic development, and there are tag set for Arabic, which aren't quite the same as the tag set for English.

And even for English, there are many different tag sets.

You may remember I mentioned the Brown Corpus of American English and the Lancaster-Oslo/Bergen corpus of British English that I worked on. And Professor Leech of linguistics at Lancaster decided that the Brown tag set wasn't quite right, so you had a slightly different tag set for LOB Corpus.

And then, Professor Greenbaum at London University, developed the International Corpus of English, and he decided that the Brown and LOB tag sets weren't quite right, so he'd make it a bit better. So, we have an ICE tag set.

And then at University of Pennsylvania, they developed a different tag set again. So, they carried on.

These are more or less similar, but they maybe have different labels and some variation in the boundaries of the categories.

We even had a project called Amalgam to try to work out a way of mapping between these different tag sets, so you could merge the different data sets that were tagged in these different ways.

So, why would you want to do part-of-speech tagging?

It's interesting for linguists. It's interesting for language teachers, if you have to teach that nouns can be used in one way and verbs can be used in another way.

And it's useful to know which are the nouns and which are the verbs. If you want to work out the grammatical structure, either as part of language learning, or with a syntactic parsing program, then if you want to work out the structure, the first thing you need to do is work out which are the nouns, and which are the verbs, and so on.

In machine translation, this can be quite important because, for example, in translating from Spanish to English, in Spanish, you might have a noun followed by an adjective. And in English, you typically have an adjective before the noun.

So, you have to swap them around, and that's one extra thing. So, a simple word-to-word mapping for similar languages, like Spanish and English, works reasonably well.

But you have to do some reordering for adjectives and nouns.

Other specific tasks may want to focus on particular word categories. For example, if you're doing sentiment analysis, then often the adjectives are particularly important because the adjectives typically tell things like good and happy and nice or bad and rubbish as an adjective.

And it's also important, to some extent, in doing text-to-speech or speech-to-text because English spelling is not purely phonetic. You can't tell the pronunciation exactly from how it's spelled.

So, they're spelling L-E-A-D. If I want to buy a lead for my dog, or I want to lead the dog into the garden, that's one way.

And the roof has lead on it. That's L-E-A-D as well.

Or, this is an object. That's a noun.

Whereas, I object to your behavior. That's a verb.

So, that's a sort of general use for it.

And you might - in that case, let's say we want some text, and we don't want to do part-of-speech technically for text.

Is it really difficult? I mean, it used to be difficult simply because we had a million words of text, and the computers of the 1980s couldn't deal with that. But apart from that, there are sort of theoretical issues. You might say, about - if you go to a dictionary and look up words, about 15% of them are ambiguous, have more than one category.

But most of them, 85%, are unambiguous. For example, Janet is always a proper noun. Hesitantly is always an adverb.

However, the 15% or so which are ambiguous tend to be the very common words.

So, actually, in a corpus of word tokens, over half of them are going to be ambiguous. For example, back is a very common word, and it's more typically ambiguous.

Earning gross took a backseat. There it's an adjective.

A small building in the back, that's a noun.

A clear majority of senators back the bill. There, it's a verb.

Enable the country to buy back debt. Buy back - back there is a participle adding to the meaning of the word buy.

I was 21 back then, a long time ago. Well, back there is an adverb, as it happens.

So, back is an example. Another example is like.

You can think this for yourself, but like also has lots of different parts of speech. So, how do you measure?

So, that's the problem, and you want to have an accurate tag. Or how do you tell how good it is? Well, you can count up in a piece of text, after the taggers run on it, how many effort words are correctly tagged.

And it turns out that fairly sophisticated models get around 97% accuracy. There's always a very few cases, about 3% of words, which are just very difficult to get right.

And this hasn't changed much over the past - it's just over 10 years. I was working on part of speech tagging in the 1980s, and then we were getting 95%, 97% on the LOB corpus tag, of the Lancaster-Oslo/Bergen Corpus tagger.

So, even sophisticated models like Hidden Markov models and BERT, we'll see later on, they don't really change that.

However, human accuracy is about the same. If you ask two different people to go and tag a piece of text, they will have problems with about 3% of the words, and they may disagree about those 3% of the words.

You saw in that back example, if I asked you to do it, you'd probably find some cases you're not sure about.

Or like, some cases you're not too sure about. So, the best you can do is the human performance.

But actually, those are sophisticated models.

The baseline model, which is a stupid model, which is computationally very fast but simple, gets even 92%. The baseline model is basically looking a word up in the dictionary. If it's got a tag, then - if it's got one tag, then give it that tag. If it's got more than one tag, then choose the most frequent tag.

So, for example, Google is probably a noun most of the time. So, just always tag it noun. And then you'll get a very high accuracy.

And if it's unknown words, if it's not in a dictionary, just say they're nouns. Because most new words are made up are nouns, and that gives you a very high accuracy.

So, therefore, it's sort of easy because many words are unambiguous.

And if you want to go into part-of-speech tagging research, it becomes research into how to deal with 3% of the words. How do you get higher than 97%? Which is a bit abstract.

OK, so what does the part-of-speech tagger do? Well, essentially, it looks up each word in a dictionary. If a word is ambiguous, has got more than one part of speech, then you have to list the possible parts of speech and work out from the context which one it is.

So, the context, is first of all, the prior probabilities of a word. That is, for a given word in the dictionary, what is the probability of it being auxiliary versus noun versus verb?

Will, for example, is almost always an auxiliary. It can be a noun or a verb, but hardly ever is. So, if you just choose always auxiliary, you'll get a very high accuracy.

There is also the neighbouring words. So, for example, bill could be a noun or a verb.

But if the word the is before it, it's almost certainly a noun, and it's not likely to be a verb. You don't get the verb.

So, those are the sorts of information sources that are used. And you essentially can count up these or use these in a corpus. There's also another hint. If you have a word which isn't in your dictionary, then it's worth looking at the suffixes or prefixes.

So, unfollowed by something is probably an adjective.

Or some word ending in L-Y is probably an adjective. Importantly. It could be an adverb. That's a slight problem as well.

Also, in English, you could look at the first letter. If the first letter is a capital, then it's probably a proper noun unless it's the first word of a sentence.

In other languages, this doesn't always work. In Arabic, for example, there are no capital letters. In Chinese letters, there's no capital letters. So, that doesn't help. But in English, it is.

So, the algorithms behind this - what I just showed you, for example, if the word the appears before, then you can rule out that it is a verb. And whatever's left is likely to be correct.

And that's the basis of the constraint grammar. And you can come up with hand-crafted, rule-based systems.

So, Helsinki University, they developed this-- and essentially, go and be a PhD student there, and spend three or four years developing a constraint grammar.

The first one was for English, and then they did it for Finnish and for Arabic and for most other languages.

So, if you speak some obscure language, and you want a PhD, then go to Helsinki University and get a PhD developing a rule-based system for your language. The trouble with that is, obviously, it requires someone to spend a lot of time doing linguistic research, to work on what the rules are and gradually adding these.

So, now, much more common in AI is to have supervised machine-learning algorithms, and the number of Hidden Markov models, or Engram models, is the obvious one.

And there are various more sophisticated models, like maximum entropy Markov models and so on. Have a look in the book for more details on this.

Also, over the past few years, neural network models have come up, like transformers and large deep-learning models like BERT. And these have also been applied to part-of-speech tagging. All of them need a hand-labelled training set.

So, they have to learn from a corpus where every word has a part of speech. And where do you get these part of speech from?

Well, you can't just run a tagger because the tagger is not perfect. Well, you could run a tagger, but it wouldn't be perfect training data. So, there's still a need to develop a hand-labelled training set.

So, for the LOB corpus, for example, we part-of-speeched – we developed a rule-based system to tagged part of it, we applied a Hidden Markov model on top of that to come up with a better tagger, and then we used that to tag the whole of the corpus.

And then Professor Leech and I had to go through and proofread it and correct all the mistakes. And we did this at least a couple of times because - then we ended up with a training set which was 100% perfect.

But all of this requires a very large effort to get a tagger which is 97% accurate anyway.

So, maybe, it's not that worthwhile anymore.

Nowadays, there's no longer such a demand for a large, hand-labelled training set. They all make use of this sort of information, for Hidden Markov models and so on.

There is a move now to use deep learning or neural network models to learn the models themselves and self-improve. So, a current research areas, how can you automatically spot the mistakes and correct them to produce a fully 100% training data set.

So, that's part-of-speech tagging.

I now want to look at named entity recognition as another sort of thing you might want to do to label the text. So, named entity is a tag like person, location, organisation, or geopolitical entity, place.

And all of these are proper nouns. They tend to be labelled as proper nouns, but the problem is, they can be more than one word and multi-word phrases like New York or New York City.

And the term can also be applied to dates and times and prices, even though these aren't really proper names, because they're also useful entities in a text.

So, it's a bit like part-of-speech tagging, but a little bit more complicated because you've got to find the span of text that constitutes the proper name and then tag the type of this entity.

So, it's two steps.

First of all, identify which words are the entity, and also work out what is the type of this entity.

And here, we have an example citing high fuel prices. United Airlines said Friday, it had increased fares by $6 per roundtrip.

Blah, blah, blah.

And all of these are labelled, and we can see that, for example, United Airlines is two words, or American Airlines is two words. And United Airlines happens to start with capital letters. So, that helps.

But United Airlines is an organisation, whereas Tim Wagner, which also starts with capital letters, is a person.

So, you've got segmentation, but you've also got a libelling task. And both of these are difficult.

So, if you come to Leeds, and you go to the bus station, you will find there's the New York Street, and there's even a New York diner, I think. And you'll find that that's actually not New York, but there is a York Street, and then - which goes from Leeds to York - and then, later on, they built a new street next to it, and they call that the new York Street.

So, New York, in that case, is not an entity. It's York, which is the entity.

So, you can see, it's not straightforward.

Why would you want to do that?

Well, there's lots of applications. A very common application is in sentiment analysis.

And you want to analyse the sentiment, and you want to identify the company or the product or the name of the person that they're talking about. Or, in question answering - often, when you ask a question, it's about an entity.

So, if you identify the entity there, that helps you to find the answer.

Or in information extraction. And we'll see later on tools and methods for information extraction.

Typically, you want to find the entities in a text and the relationships between those entities.

So, it's hard, in part-of-speech tagging, there's no segmentation problem because we just basically take each word - and there are some special cases like New York.

If there's a space in it, you do want to label it as one proper name, as opposed to two.

But that's-- a way around that is to say that New York is two nouns, New and York in this case.

You have to find and segment - and there's a problem of ambiguity, just as in part-of-speech tagging.

For example, Washington could be a person or an organisation or a location or a geopolitical entity, a state, so on.

So, what we'd like to be able to do is to use all these algorithms we already know work for part-of-speech tagging. But the trouble is, part-of-speech tagging only works on one tag per word.

And now, we have several words together.

So, one way around this is it's called BIO tagging. BIO tagging means having tags like person, organisation, location, and so on.

But in addition, you say, do they begin or, are they inside - so Jane is begin person. Villanova is inside person. Of is other. United is begin organisation. Airlines is inside organisation. Holding is inside organisation. Discussed his is other.

There is other. And so on.

So, now, we only have one tag per token, so you can look up a word, and the idea is Jane, for example, is ambiguous between begin person and inside person.

But typically, begin person will only occur after another. It will never occur after an inside tag, for example.

So, we have something that looks like part-of-speech tagging. B token begins a span, I token is inside a span, and O token is outside of the span. So, we have tags.

Person, organisation, location, that's three tags.

But you also have to have another tag as well. So, that's an extra tag.

And then you also have begin and inside for each of the different possible tags.

So, we have one other tag, n begin tanks and n other tags, where n would be three in this case. So, there's a total of 2n plus 1.

In other words, whatever the tag set is, you basically doubled it, plus one. So, there's more categories to label. But it still works. You can still do that.

And there are- as I said, there are different tag sets. There's also variants on the BIO tagging.

So, you might include an extra tag for begin, inside, and other. Or, you might just say, forget about begin. We'll just have inside and other, so I tags and O tags. Or you might have begin, inside, and end as well.

So, B-I-O tags. So, for example, United Airlines Holding becomes begin United, Airlines is inside, and Holding is end.

And there are pros and cons for each of these things.

BIO seems to be the standard, but these other ones have got applications to it.

So, that means the standard algorithms that worked for part-of-speech tagging will also work for named entity recognition.

Later on, we'll look at some more examples of information extraction. The standard algorithms are things like Hidden Markov models for supervised machine learning.

You can also have rule-based systems for named entity extraction, and I'll go into some more details of an example later on in the course.

That's enough for now.

Those are two different sorts of tagging. There are other sorts of tagging too, but just to give you a flavour of the issues. If you have a text, then before you do anything else to it, in terms of information extraction or machine translation or whatever, it may be helpful to label each of the words with its part of speech and to label each of the entities with its entity category.

Because that, then, helps in-- you've essentially added additional features, and these additional features can be used later on in machine learning applications like information extraction.

OK. Thank you very much, and enjoy the rest of your day.

[END]