Part-of-Speech Tagging

Dr. Korra Sathya Babu

Assistant Professor
Department of Computer Science & Engineering
National Institute of Technology Rourkela

Table of Contents


NATIONAL INSTITUTE OF TECHNOLOGY
ROURKELA

Part-of-Speech

Part-of-Speech tagging

Part-of-Speech tagging is a well-known task in Natural Language Processing. It refers to the process of classifying words into their parts of speech (also known as words classes or lexical categories). This is a supervised learning approach..

Part-of-Speech

Part-of-Speech tagging

Part-of-Speech tagging is a well-known task in Natural Language Processing. It refers to the process of classifying words into their parts of speech (also known as words classes or lexical categories). This is a supervised learning approach..

Part-of-Speech tagging

Parts-of-speech (POS, word classes, or syntactic categories) are useful because of the large amount of information they give about a word and its neighbors. Knowing whether a word is a noun or a verb tells us a lot about likely neighboring words (nouns are preceded by determiners and adjectives, verbs by nouns) and about the syntactic structure around the word, which makes part-of-speech tagging an important component of syntactic parsing. Parts of speech are useful features for finding named entities like people or organizations in text and other information extraction tasks.

Part-of-Speech

Part-of-Speech

Parts-of-speech can be divided into two broad supercategories: closed class and open class types. Closed classes are those with relatively fixed membership,such as prepositions (new prepositions are rarely coined). By contrast, nouns and verbs are open classes (new nouns and verbs like iPhone or to fax are continually being created or borrowed). Any given speaker or corpus may have different open class words, but all speakers of a language, and sufficiently large corpora, likely share the set of closed class words. Closed class words are generally function words like of, it, and, or you, which tend to be very short, occur frequently, and often have structuring uses in grammar. Four major open classes occur in the languages of the world: nouns, verbs, adjectives, and adverbs.

Part-of-Speech

Noun

The syntactic class noun includes the words for most people, places, or things, but others as well. Open class nouns fall into two classes. Proper nouns, like Regina, Colorado, and IBM, are names of specific persons or entities. In English, they generally aren?t preceded by articles (e.g., the book is upstairs, but Regina is upstairs). The other class, common nouns are divided in many languages, including English, into count nouns and mass nouns. Count nouns allow grammatical enumeration, occurring in both the singular and plural (goat/goats, relationship/relationships) and they can be counted (one goat, two goats). Mass nouns are used when something is conceptualized as a homogeneous group. So words like snow, salt, and communism are not counted

Part-of-Speech

Verb

The verb class includes most of the words referring to actions and processes, including main verbs like draw, provide, and go. English verbs have inflections (non-third-person-sg (eat), third-person-sg (eats), progressive (eating), past participle (eaten)). While many researchers believe that all human languages have the categories of noun and verb, others have argued that some languages, such as Riau Indonesian and Tongan, don?t even make this distinction.

Part-of-Speech

Adjective

The third open class English form is adjectives, a class that includes many terms for properties or qualities. Most languages have adjectives for the concepts of color (white, black), age (old, young), and value (good, bad), but there are languages without adjectives. In Korean, for example, the words corresponding to English adjectives act as a subclass of verbs, so what is in English an adjective ?beautiful? acts in Korean like a verb meaning ?to be beautiful?.

Part-of-Speech

Adverb

What coherence the class has semantically may be solely that each of these words can be viewed as modifying something (often verbs, hence the name ?adverb?, but also other adverbs and entire verb phrases). Directional adverbs or locative adverbs (home, here, downhill) specify the direction or location of some action; degree adverbs (extremely, very, somewhat) specify the extent of some action, process, or property; manner adverbs (slowly, slinkily, delicately) describe the manner of some action or process; and temporal adverbs describe the time that some action or event took place (yesterday, Monday). Because of the heterogeneous nature of this class, some adverbs (e.g., temporal adverbs like Monday) are tagged in some tagging schemes as nouns.

Part-of-Speech

The closed classes differ more from language to language than do the open classes. Some of the important closed classes in English include:

prepositions: on, under, over, near, by, at, from, to, with
determiners: a, an, the
pronouns: she, who, I, others
conjunctions: and, but, or, as, if, when
auxiliary verbs: can, may, should, are
particles: up, down, on, off, in, out, at, by
numerals: one, two, three, first, second, third

Prepositions occur before noun phrases. Semantically they often indicate spatial or temporal relations, whether literal (*on it*, *before then*, *by the house*) or metaphorical (*on time*, *with gusto*, *beside herself*), but often indicate other relations as well, like marking the agent in (*Hamlet was written by Shakespeare*,

Part-of-Speech

A **particle** resembles a preposition or an adverb and is used in combination with a verb. Particles often have extended meanings that aren't quite the same as the prepositions they resemble, as in the particle *over* in *she turned the paper over.*

Part-of-Speech

When a verb and a particle behave as a single syntactic and/or semantic unit, we call the combination a **phrasal verb**. Phrasal verbs cause widespread problems with natural language processing because they often behave as a semantic unit with a **non-compositional** meaning— one that is not predictable from the distinct meanings of the verb and the particle. Thus, *turn down* means something like 'reject', *rule out* means 'eliminate', *find out* is 'discover', and *go on* is 'continue'.

Part-of-Speech

A closed class that occurs with nouns, often marking the beginning of a noun phrase, is the **determiner**. One small subtype of determiners is the **article**: English has three articles: *a*, *an*, and *the*. Other determiners include *this* and *that* (*this chapter*, *that page*). *A* and *an* mark a noun phrase as indefinite, while *the* can mark it as definite; definiteness is a discourse property. Articles are quite frequent in English; indeed, *the* is the most frequently occurring word in most corpora of written English, and *a* and *an* are generally right behind.

Part-of-Speech

Conjunctions join two phrases, clauses, or sentences. Coordinating conjunctions like *and*, *or*, and *but* join two elements of equal status. Subordinating conjunctions are used when one of the elements has some embedded status. For example, *that* in *"I thought that you might like some milk"* is a subordinating conjunction that links the main clause *I thought* with the subordinate clause *you might like some milk*. This clause is called subordinate because this entire clause is the "content" of the main verb *thought*. Subordinating conjunctions like *that* which link a verb to its argument in this way are also called **complementizers**.

Part-of-Speech

Pronouns are forms that often act as a kind of shorthand for referring to some noun phrase or entity or event. **Personal pronouns** refer to persons or entities (*you*, *she*, *I*, *it*, *me*, etc.). **Possessive pronouns** are forms of personal pronouns that indicate either actual possession or more often just an abstract relation between the person and some object (*my, your, his, her, its, one's, our, their*). **Wh-pronouns** (*what, who, whom, whoever*) are used in certain question forms, or may also act as complementizers (*Frida, who married Diego...*).

Part-of-Speech

A closed class subtype of English verbs are the **auxiliary** verbs. Cross-linguistically, auxiliaries mark certain semantic features of a main verb, including whether an action takes place in the present, past, or future (tense), whether it is completed (aspect), whether it is negated (polarity), and whether an action is necessary, possible, suggested, or desired (mood).

English auxiliaries include the **copula** verb *be*, the two verbs *do* and *have*, along with their inflected forms, as well as a class of **modal verbs**. *Be* is called a copula because it connects subjects with certain kinds of predicate nominals and adjectives (*He is a duck*). The verb *have* is used, for example, to mark the perfect tenses (*I have gone*, *I had gone*), and *be* is used as part of the passive (*We were robbed*) or progressive (*We are leaving*) constructions. The modals are used to mark the mood associated with the event or action depicted by the main verb: *can* indicates ability or possibility, *may* indicates permission or possibility, *must* indicates necessity. In addition to the perfect *have* mentioned above, there is a modal verb *have* (e.g., *I have to go*), which is common in spoken English.

Part-of-Speech

English also has many words of more or less unique function, including **interjections** (*oh, hey, alas, uh, um*), **negatives** (*no, not*), **politeness markers** (*please, thank you*), **greetings** (*hello, goodbye*), and the existential **there** (*there are two on the table*) among others. These classes may be distinguished or lumped together as interjections or adverbs depending on the purpose of the labeling.

The Penn Treebank Part-of-Speech Tagset

1. While there are many lists of parts-of-speech, most modern language processing on English uses the 45-tag Penn Treebank tagset. This tagset has been used to label a wide variety of corpora, including the Brown corpus, the Wall Street Journal corpus, and the Switchboard corpus.

The Penn Treebank Part-of-Speech Tagset

Tag	Description	Example	Tag	Description	Example
CC	coordin. conjunction	*and, but, or*	SYM	symbol	+,%, &
CD	cardinal number	*one, two*	TO	"to"	*to*
DT	determiner	*a, the*	UH	interjection	*ah, oops*
EX	existential 'there'	*there*	VB	verb base form	*eat*
FW	foreign word	*mea culpa*	VBD	verb past tense	*ate*
IN	preposition/sub-conj	*of, in, by*	VBG	verb gerund	*eating*
JJ	adjective	*yellow*	VBN	verb past participle	*eaten*
JJR	adj., comparative	*bigger*	VBP	verb non-3sg pres	*eat*
JJS	adj., superlative	*wildest*	VBZ	verb 3sg pres	*eats*
LS	list item marker	*1, 2, One*	WDT	wh-determiner	*which, that*
MD	modal	*can, should*	WP	wh-pronoun	*what, who*
NN	noun, sing. or mass	*llama*	WP$	possessive wh-	*whose*
NNS	noun, plural	*llamas*	WRB	wh-adverb	*how, where*
NNP	proper noun, sing.	*IBM*	$	dollar sign	*$*
NNPS	proper noun, plural	*Carolinas*	#	pound sign	*#*
PDT	predeterminer	*all, both*	"	left quote	' or "
POS	possessive ending	*'s*	"	right quote	' or "
PRP	personal pronoun	*I, you, he*	(	left parenthesis	[, (, {, <
PRP$	possessive pronoun	*your, one's*	)	right parenthesis	],), }, >
RB	adverb	*quickly, never*	,	comma	,
RBR	adverb, comparative	*faster*	.	sentence-final punc	. ! ?
RBS	adverb, superlative	*fastest*	:	mid-sentence punc	: ; ... – -
RP	particle	*up, off*			

Part-of-Speech Tagging

Part-of-speech tagging (**tagging** for short) is the process of assigning a part-of-speech marker to each word in an input text. Because tags are generally also applied to punctuation, **tokenization** is usually performed before, or as part of, the tagging process: separating commas, quotation marks, etc., from words and disambiguating end-of-sentence punctuation (period, question mark, etc.) from part-of-word punctuation (such as in abbreviations like *e.g.* and *etc.*)

The input to a tagging algorithm is a sequence of words and a tagset, and the output is a sequence of tags, a single best tag for each word as shown in the examples on the previous pages.

Tagging is a **disambiguation** task; words are **ambiguous** —have more than one possible part-of-speech— and the goal is to find the correct tag for the situation. For example, the word *book* can be a verb (*book that flight*) or a noun (as in *hand me that book*.

Markov Chains

- Markov Chains and HMM are both extensions of the Finite Automata
- The FA is defined by a set of states and set of transitions between them
- A weighted FSA is a simple augumentation of the FA in which an arc is associated with a probability, indicating how likely that path is taken
- The probability of all the arcs leaving a node must sum to 1
- A markov chain is a special case of a weighted automation in which the input sequence uniquely determines which state the automation go through. As it cant represent inherently ambiguous problems, a Markov chain is only useful for assigning probabilities to unambiguous sequences

Visible Markov Models

- If we want to compute the probability of sequence of states
  $P(X_1, .., X_T) =$
  $P(X_1)P(X_2|X_1)P(X_3|X_1X_2)...P(X_T|X_1...X_{T-1})$
- This is the chain rule
- For bigrams with Markov Assumptions
  $P(X_1, .., X_T) = P(X_1)P(X_2|X_1)P(X_3|X_2)...P(X_T|X_{T-1})$
- So here we ignore all the history but consider only the recent state and the next

Hidden Markov Model

- Observing the sequence of symbols, the sequence of states that led to the generation of the symbols is hidden. HMM Comprise of the following:

  Q = Sequence of states

  O = sequence of observations drawn from vocabulary

  $q_0 q_f$ = start and final states

  A = state transition probabilities

  B = symbol emission probabilities (If you are in certain state, what is the prob. that you are going to emit a certain symbol from that state)

  π = initial state probabilities

  (A,B,π) = Complete probabilistic model

Markov Models

Let's talk about the weather. we have three types of weather: *sunny*, *rainy*, and *foggy*.

Let's assume for the moment that the weather lasts all day, i.e. it doesn't change from rainy to sunny in the middle of the day.

Weather prediction is all about trying to guess what the weather will be like tomorrow based on a history of observations of weather. Let's assume a simplified model of weather prediction: we'll collect statistics on what the weather was like today based on what the weather was like yesterday, the day before, and so forth. We want to collect the following probabilities:

$$P(w_n \mid w_{n-1}, w_{n-2}, \ldots, w_1) \tag{1}$$

Using expression 1, we can give probabilities of types of weather for tomorrow and the next day using n days of history. For example, if we knew that the weather for the past three days was {sunny, sunny, foggy} in chronological order, the probability that tomorrow would be rainy is given by:

$$P(w_4 = Rainy \mid w_3 = Foggy, w_2 = Sunny, w_1 = Sunny) \tag{2}$$

Markov Models

Here's the problem: the larger n is, the more statistics we must collect. Suppose that $n = 5$, then we must collect statistics for $3^5 = 243$ past histories. Therefore, we will make a simplifying assumption, called the *Markov Assumption*:

In a sequence $\{w_1, w_2, \ldots, w_n\}$:

$$P(w_n \mid w_{n-1}, w_{n-2}, \ldots, w_1) \approx P(w_n \mid w_{n-1}) \tag{3}$$

This is called a *first-order* Markov assumption, since we say that the probability of an observation at time n only depends on the observation at time $n-1$. A *second-order* Markov assumption would have the observation at time n depend on $n-1$ and $n-2$. In general, when people talk about Markov assumptions, they usually mean first-order Markov assumptions; I will use the two terms interchangeably.

We can the express the joint probability using the Markov assumption.[1]

$$P(w_1, \ldots, w_n) = \Pi_{i=1}^{n} P(w_i \mid w_{i-1}) \tag{4}$$

Markov Models

So this now has a profound affect on the number of histories that we have to find statistics for— we now only need $3^2 = 9$ numbers to characterize the probabilities of all of the sequences. This assumption may or may not be a valid assumption depending on the situation (in the case of weather, it's probably not valid), but we use these to simplify the situation.

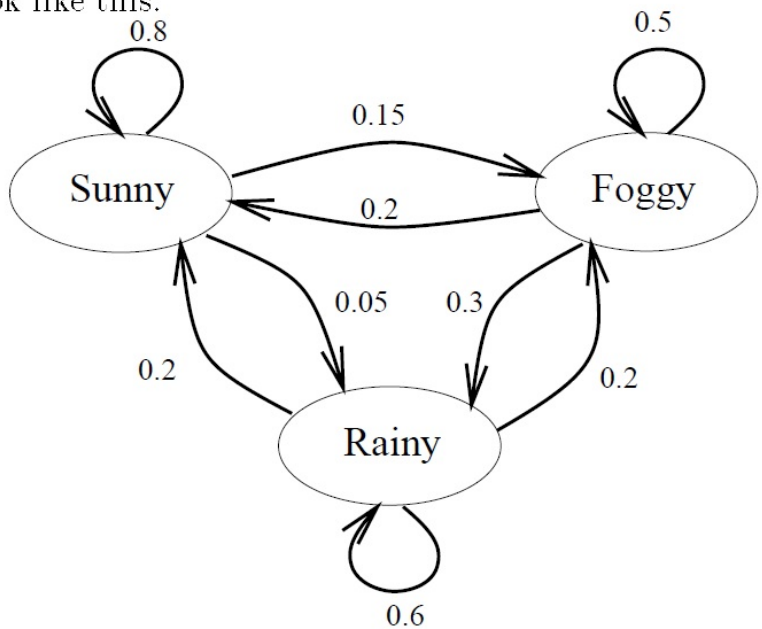
So let's arbitrarily pick some numbers for $P(w_{\text{tomorrow}} \mid w_{\text{today}})$, expressed in Table 1.

		Tomorrow's Weather		
		Sunny	Rainy	Foggy
Today's Weather	Sunny	0.8	0.05	0.15
	Rainy	0.2	0.6	0.2
	Foggy	0.2	0.3	0.5

Table 1: Probabilities of Tomorrow's weather based on Today's Weather

Markov Models

For first-order Markov models, we can use these probabilities to draw a probabilistic finite state automaton. For the weather domain, you would have three states (Sunny, Rainy, and Foggy), and every day you would transition to a (possibly) new state based on the probabilities in Table 1. Such an automaton would look like this:



Markov Models

Given that today is sunny, what's the probability that tomorrow is sunny and the day after is rainy?

Well, this translates into:

$$
\begin{aligned}
P(w_2 = \text{Sunny}, w_3 = \text{Rainy} | w_1 = \text{Sunny}) &= P(w_3 = \text{Rainy} \mid w_2 = \text{Sunny}, w_1 = \text{Sunny}) * \\
&\quad P(w_2 = \text{Sunny} \mid w_1 = \text{Sunny}) \\
&= P(w_3 = \text{Rainy} \mid w_2 = \text{Sunny}) * \\
&\quad P(w_2 = \text{Sunny} \mid w_1 = \text{Sunny}) \\
&= (0.05)(0.8) \\
&= 0.04
\end{aligned}
$$

Markov Models

Given that today is foggy, what's the probability that it will be rainy two days from now?

There are three ways to get from foggy today to rainy two days from now: {foggy, foggy, rainy}, {foggy, rainy, rainy}, and {foggy, sunny, rainy}. Therefore we have to sum over these paths:

$$
\begin{aligned}
P(w_3 = \text{Rainy} \mid w_1 = \text{Foggy}) &= P(w_2 = \text{Foggy}, w_3 = \text{Rainy} \mid w_1 = \text{Foggy}) + \\
&\quad P(w_2 = \text{Rainy}, w_3 = \text{Rainy} \mid w_1 = \text{Foggy}) + \\
&\quad P(w_2 = \text{Sunny}, w_3 = \text{Rainy} \mid w_1 = \text{Foggy}) + \\
&= P(w_3 = \text{Rainy} \mid w_2 = \text{Foggy}) P(w_2 = \text{Foggy} \mid w_1 = \text{Foggy}) + \\
&\quad P(w_3 = \text{Rainy} \mid w_2 = \text{Rainy}) P(w_2 = \text{Rainy} \mid w_1 = \text{Foggy}) + \\
&\quad P(w_3 = \text{Rainy} \mid w_2 = \text{Sunny}) P(w_2 = \text{Sunny} \mid w_1 = \text{Foggy}) \\
&= (0.3)(0.5) + (0.6)(0.3) + (0.05)(0.2) \\
&= 0.34
\end{aligned}
$$

Hidden Markov Models

So what makes a Hidden Markov Model? Well, suppose you were locked in a room for several days, and you were asked about the weather outside. The only piece of evidence you have is whether the person who comes into the room carrying your daily meal is carrying an umbrella or not.

Let's suppose the following probabilities:

	Probability of Umbrella
Sunny	0.1
Rainy	0.8
Foggy	0.3

Table 2: Probabilities of Seeing an Umbrella Based on the Weather

Remember, the equation for the weather Markov process before you were locked in the room was:

$$P(w_1, \ldots, w_n) = \Pi_{i=1}^{n} P(w_i \mid w_{i-1}) \tag{5}$$

Hidden Markov Models

Now we have to factor in the fact that the actual weather is *hidden* from you. We do that by using Bayes' Rule:

$$P(w_1, \ldots, w_n \mid u_1, \ldots, u_n) = \frac{P(u_1, \ldots, u_n | w_1, \ldots, w_n) P(w_1, \ldots, w_n)}{P(u_1, \ldots, u_n)} \tag{6}$$

where u_i is true if your caretaker brought an umbrella on day i, and false if the caretaker didn't. The probability $P(w_1, \ldots, w_n)$ is the same as the Markov model from the last section, and the probability $P(u_1, \ldots, u_n)$ is the prior probability of seeing a particular sequence of umbrella events (e.g. {True, False, True}). The probability $P(u_1, \ldots, u_n | w_1, \ldots, w_n)$ can be estimated as $\Pi_{i=1}^{n} P(u_i | w_i)$, if you assume that, for all i, given w_i, u_i is independent of all u_j and w_j, for all $j \neq i$.

Hidden Markov Models

Suppose the day you were locked in it was sunny. The next day, the caretaker carried an umbrella into the room. Assuming that the prior probability of the caretaker carrying an umbrella on any day is 0.5, what's the probability that the second day was rainy?

$$
\begin{aligned}
P(w_2 = \text{Rainy} | \\ w_1 = \text{Sunny}, u_2 = \text{True}) &= \frac{P(w_2 = \text{Rainy}, w_1 = \text{Sunny} | u_2 = \text{T})}{P(w_1 = \text{Sunny} | u_2 = \text{T})} \\
(u_2 \ and \ w_1 \ independent) &= \frac{P(w_2 = \text{Rainy}, w_1 = \text{Sunny} | u_2 = \text{T})}{P(w_1 = \text{Sunny})} \\
(Bayes' \, Rule) &= \frac{P(u_2 = \text{T} | w_1 = \text{Sunny}, w_2 = \text{Rainy}) P(w_2 = \text{Rainy}, w_1 = \text{Sunny})}{P(w_1 = \text{Sunny}) P(u_2 = \text{T})} \\
(Markov \ assumption) &= \frac{P(u_2 = \text{T} | w_2 = \text{Rainy}) P(w_2 = \text{Rainy}, w_1 = \text{Sunny})}{P(w_1 = \text{Sunny}) P(u_2 = \text{T})} \\
(P(A, B) = P(A|B)P(B)) &= \frac{P(u_2 = \text{T} | w_2 = \text{Rainy}) P(w_2 = \text{Rainy} | w_1 = \text{Sunny}) P(w_1 = \text{Sunny})}{P(w_1 = \text{Sunny}) P(u_2 = \text{T})} \\
(Cancel : P(Sunny)) &= \frac{P(u_2 = \text{T} | w_2 = \text{Rainy}) P(w_2 = \text{Rainy} | w_1 = \text{Sunny})}{P(u_2 = \text{T})} \\
&= \frac{(0.8)(0.05)}{0.5} \\
&= .08
\end{aligned}
$$

Hidden Markov Models

Suppose the day you were locked in the room it was sunny; the caretaker brought in an umbrella on day 2, but not on day 3. Again assuming that the prior probability of the caretaker bringing an umbrella is 0.5, what's the probability that it's foggy on day 3?

$$
\begin{aligned}
P(w_3 = \mathrm{F} \mid w_1 = \mathrm{S}, u_2 = \mathrm{T}, u_3 = \mathrm{F}) = {} & P(w_2 = \text{Foggy}, w_3 = \text{Foggy} \mid w_1 = \text{Sunny}, u_2 = \text{True}, u_3 = \text{False}) + \\
& P(w_2 = \text{Rainy}, w_3 = \text{Foggy} \mid \ldots) + \\
& P(w_2 = \text{Sunny}, w_3 = \text{Foggy} \mid \ldots) \\
= {} & \frac{P(u_3 = F|w_3 = F)P(u_2 = T|w_2 = F)P(w_3 = F|w_2 = F)P(w_2 = F|w_1 = S)P(w_1 = S)}{P(u_3 = F)P(u_2 = T)P(w_1 = S)} + \\
& \frac{P(u_3 = F|w_3 = F)P(u_2 = T|w_2 = R)P(w_3 = F|w_2 = R)P(w_2 = R|w_1 = S)P(w_1 = S)}{P(u_3 = F)P(u_2 = T)P(w_1 = S)} + \\
& \frac{P(u_3 = F|w_3 = F)P(u_2 = T|w_2 = S)P(w_3 = F|w_2 = S)P(w_2 = S|w_1 = S)P(w_1 = S)}{P(u_3 = F)P(u_2 = T)P(w_1 = S)} \\
= {} & \frac{P(u_3 = F|w_3 = F)P(u_2 = T|w_2 = F)P(w_3 = F|w_2 = F)P(w_2 = F|w_1 = S)}{P(u_3 = F)P(u_2 = T)} + \\
& \frac{P(u_3 = F|w_3 = F)P(u_2 = T|w_2 = R)P(w_3 = F|w_2 = R)P(w_2 = R|w_1 = S)}{P(u_3 = F)P(u_2 = T)} + \\
& \frac{P(u_3 = F|w_3 = F)P(u_2 = T|w_2 = S)P(w_3 = F|w_2 = S)P(w_2 = S|w_1 = S)}{P(u_3 = F)P(u_2 = T)} \\
= {} & \frac{(0.7)(0.3)(0.5)(0.15)}{(0.5)(0.5)} + \\
& \frac{(0.7)(0.8)(0.2)(0.05)}{(0.5)(0.5)} + \\
& \frac{(0.7)(0.1)(0.15)(0.8)}{(0.5)(0.5)}
\end{aligned}
$$

argmax

Let's get away from umbrellas and such for a moment and talk about *real* things, like speech. In speech recognition, the basic idea is to find the most likely string of words given some acoustic input, or:

$$\text{argmax}_{\mathbf{w}\in\mathcal{L}} P(\mathbf{w}|\mathbf{y}) \tag{7}$$

Where $\mathbf{w}$ is a string of words, $\mathcal{L}$ is the language you're interested in, and $\mathbf{y}$ is the set of acoustic vectors that you've gotten from your front end processor[2]. To compare this to the weather example, the acoustics are our observations, similar to the umbrella observations, and the words are similar to the weather on successive days.

Remember that the basic equation of speech recognition is Bayes' Rule:

$$\text{argmax}_{\mathbf{w}\in\mathcal{L}} P(\mathbf{w}|\mathbf{y}) = \text{argmax}_{\mathbf{w}\in\mathcal{L}} \frac{P(\mathbf{y}|\mathbf{w})P(\mathbf{w})}{P(\mathbf{y})} \tag{8}$$

For a single speech input (e.g. one sentence), the acoustics ($\mathbf{y}$) will be constant, and therefore, so will $P(\mathbf{y})$. So we only need to find:

$$\text{argmax}_{\mathbf{w}\in\mathcal{L}} P(\mathbf{y}|\mathbf{w})P(\mathbf{w}) \tag{9}$$

The first part of this expression is called the model *likelihood*, and the second part is a prior probability of the word string.

Bigram

In the second part of expression 9, we were concerned with $P(\mathbf{w})$. This is the prior probability of the string of words we are considering. Notice that we can also replace this with a Markov model, using the following equation:

$$P(\mathbf{w}) = \Pi_{i=1}^{n} P(w_i|w_{i-1}) \tag{12}$$

What is this saying? Well, suppose that we had the string "I like cabbages". The probability of this string would be:

$$P(\text{I like cabbages}) = P(\text{I}|\text{START})P(\text{like}|\text{I})P(\text{cabbages}|\text{like}) \tag{13}$$

This type of grammar is called a bigram grammar (since it relates two (bi) words (grams)). Using a higher order Markov assumption allows more context into the grammars; for instance, a second-order Markov language model is a trigram grammar, where the probabilities of word n are given by $P(w_n|w_{n-1}, w_{n-2})$.

Thank You !