Probability for (Computational) Linguists[1]

Linguistics 409 · Computational Linguistics

Rice University

February 1, 2013

[1] some parts adapted from materials by Rens Bod

Probability has become central to Computational Linguistics

- Speech recognition: find the most probable string of words given an acoustic signal
- Probabilistic Context Free Grammars: not all ambiguous parses are equally likely
- Word Sense Disambiguation: the likeliest word sense given the preceding word
- Part of Speech tagging
- Named entity recognition
- Machine translation (probability of an English expression given a French expression)
- etc., etc., etc.

But it is also important for Linguistics proper!

- Linguistic behavior is inherently gradient, e.g.:
    - language acquisition
    - code switching
    - sound change
    - phonological acceptability
    - morphological productivity
    - syntactic wellformedness
    - semantic interpretation judgments
    - etc.
- Probability theory provides a tool to model and account for the gradiency of linguistic behavior
- And reasserts the importance of the mundane!

So what are probabilities?

As Abney says, we don’t actually know for certain. There are two main viewpoints:

Frequentist (aka objectivist) Probabilities are real aspects of the world that can be measured by relative frequencies of outcomes of experiments.

Bayesian (aka subjectivist) Probabilities are descriptions of an observer’s degree of belief or uncertainty rather than having any external significance.

Both views are relevant for (computational)? linguistics. Fortunately, the laws of probability theory remain the same under both interpretations.

Frequencies matter in linguistics

There is an extremely large psycholinguistic literature, showing that:

1. People register frequencies and differences in frequencies (e.g. Mehler & Carey 1968; Hasher & Chromiak 1977; Tanenhaus 1995)
2. People's judgments of words and sentences are very well predicted by the combined frequencies of their subparts (e.g. Baayen et al. 1997; Coleman & Pierrehumbert 1997; Hay 2001)
3. Probability theory offers a framework which can compute the probability of a *whole* from the probabilities of its *parts*.

A working definition...

- The probability of an event e is computed as the relative frequency with which *e* occurs in a sequence of *n* identical experiments:
- That is, the probability *P* of an event *e* is computed as:

$$P(e) = \frac{n_e}{n} \tag{1}$$

- n_e is the number of occurrences of the event *e* in *n* identical experiments.
- In *practice*, we compute the probability of an event as the relative frequency with which it occurs in a large corpus of linguistic data.

“But the frequency is not the same as the probability, rather it is a consequence of the probability.”

Example...

Example: Suppose we have a simple corpus of 50 words consisting of:

- 25 nouns
- 20 verbs
- 5 adjectives

Then the probability of sampling a word is:

- P(noun) = 25/50 = 0.5
- P(verb) = 20/50 = 0.4
- P(adjective) = 5/50 = 0.1

Example: note that...

Then the probability of sampling a word is:

- P(noun) = 25/50 = 0.5
- P(verb) = 20/50 = 0.4
- P(adjective) = 5/50 = 0.1

- A probability is a number between 0 and 1
- The probabilities of an exhaustive set of outcomes must sum to 1
- $P(\{noun\} \cup \{verb\} \cup \{adjective\}) = 1$
- What is the probability that one of the words in this corpus is the definite article?

Joint Probabilities

Joint probabilities describe the probability of two (or more) events occurring.

> *E.g., in describing the probability that a sentence has a certain structure we want to know the joint probability of the rules generating the structure.*

Coming back to our simple corpus (25 nouns, 20 verbs and 5 adjectives): In an experiment where we sample two words, what is the probability of sampling a noun and a verb?

We write this as: $P(\{noun\}, \{verb\})$ or as: $P(\{noun\} \cap \{verb\})$

Joint Probabilities

Joint probabilities describe the probability of two (or more) events occurring.

Remember that we have already computed the single probabilities:

- $P(\{noun\}) = 0.5$
- $P(\{verb\}) = 0.4$

Joint probabilities

Intuitively: We sample a noun in 50% of the cases and a verb in 40% of the cases. Thus, we sample them together in 50% of 40% of experiments, i.e. in 20% of the cases.

Formally: $P(\{noun\}, \{verb\}) = P(\{noun\}) x P(\{verb\}) = 0.5x0.4 = 0.2$
We can do this simple multiplication because we tacitly assumed that sampling a *verb* is "independent" of sampling a *noun*.

In general, for two *independent* events:
$P(e_1, e_2) = P(e_1) x P(e_2)$ if e_1 and e_2 are independent.

Conditional probabilities

It is often the case that events are **dependent**

Suppose that in our example corpus a *noun* is always followed by a *verb*. Then, in an experiment where we sample two *consecutive* words, the probability that the word at position $n + 1$ is a *verb* given that we have first sampled a *noun* at position n is 1.

Definition: The probability of an event e_2 given that we have seen event e_1 is called the conditional probability of e2 given e1, and is written as $P(e_2|e_1)$.

Example: Using the example corpus and experiment described above, what is the probability of sampling a *noun* and a *verb*?

- $P(\{noun\}) = 0.5$
- $P(\{verb\}|\{noun\}) = 1$

Conditional probabilities

Their joint probability is then the product:

$P(\{noun\}, \{verb\})$ = $P(\{noun\})$ x $P(\{verb\}|\{noun\})$ = 0.5 x 1 = 0.5

Because in our example corpus a *noun* is always followed by a *verb*. In an experiment where we sample two *consecutive* words, the probability of sampling a *verb* given that we have first sampled a *noun* is 1: $P(e_2 = verb|e_1 = noun) = 1$

In general: $P(e_1, e_2) = P(e_1)xP(e_2|e_1)$

Multiplication Rule

Multiplication Rule

In general: $P(e_1, e_2) = P(e_1)xP(e_2|e_1)$

We can rewrite the multiplication rule as a general definition for conditional probability of two events *e* and *f*:

Bayes' rule

$$P(e|f) = \frac{P(e,f)}{P(f)} = \frac{P(e)xP(f|e)}{P(f)}$$

The Chain Rule

The multiplication rule is generalized to multiple events by the **chain rule**:

$$P(e_1, e_2, e_3, e_4, e_n) = P(e_1)xP(e_2|e_1)xP(e_3|e_2, e_1)x...xP(e_n|e_{n-1}, e_{n-2}, ..., e_1)$$

This long product is usually expressed:

$$P(e_1, e_2, ..., e_n) = \prod_{i=1}^{n} P(e_i|e_{i-1}, e_{i-2}, ..., e_1)$$

Which is much simpler if the events are independent:

$$P(e_1, e_2, ..., e_n) = \prod_{i=1}^{n} P(e_i)$$

Or in case each event depends only on the preceding event (a 1st order Markov model):

$$P(e_1, e_2, ..., e_n) = \prod_{i=1}^{n} P(e_i|e_{i-1})$$

A brief digression: logs

As an aside, in computatonal linguistics we're often going to use **log probabilities** rather than real numbers between 0 and 1 in our calculations and tools. This is for a number of practical reasons:

- We can calculate the product of probabilities with addition of logs rather than the more expensive/slower multiplication.
- We can avoid the accumulation of floating point rounding errors inherent in computing with very small or very large real numbers.
- Many probabilities in natural language are logarithmically distributed anyway so the comparison of logs is often more intuitable.

Example: the probability of a sentence

Colorless	green	ideas	sleep	furiously
3	94	143	65	12

For next time:

For next time:

1. Friday: **N-Gram models**