Statistische Methoden in der Computerlinguistik

1 Probability Theory

1.1 What is “Probability”?

1.1.1 Classical Definition

The probability $P(A)$ of an event $A$ is defined as the ratio of the number $n(A)$ of occurrences of instances of that event to the number $n$ of possible instances:

$$P(A) := \frac{n(A)}{n}$$

Problem(s):

- *a posteriori*: empirically motivated definition
- circular: simple events $A$ must be equiprobable

1.1.2 Statistical Definition

Probability is defined in terms of relative frequency: the ratio $h(A)$ of the number of event instances to the total number of events in a concrete experiment:

$$h(A) := \frac{n(A)}{n}$$

$$P(A) := \lim_{n \to \infty} h(A)$$

Problem(s):

- empirically motivated, yet not empirically groundable
- also works out to be circular

1.1.3 Axiomatic Definition (Kolmogoroff)

Basic Building Blocks:

- $\Omega$: the sample space, a set of basic outcomes.
- $S \subseteq 2^\Omega$: the event space, a $\sigma$-field on $\Omega$.
A probability distribution for an event space \( S \) is a (total) function: \( P : S \rightarrow \mathbb{R} \) which fulfills Axioms (1) through (3):

1. For arbitrary events \( A \in S : A \in S : P(A) \geq 0 \)
2. \( P(\Omega) = 1 \)
3. For arbitrary sequences of pairwise disjoint events \( A_1, A_2, \ldots \in S : \)
\[
P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)
\]

**Theorem 1 (Empty Event)** \( P(\emptyset) = 0 \)

**Theorem 2 (Finite Sums)** For arbitrary finite sequences of pairwise disjoint events \( A_1, A_2, \ldots, A_n : \)
\[
P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i)
\]

**Theorem 3 (Subtraction Rule)** For arbitrary \( A \in S : \)
\[
P(\overline{A}) = 1 - P(A)
\]

**Theorem 4 (Probability Range)** For arbitrary \( A \in S : \)
\[
0 \leq P(A) \leq 1
\]

**Theorem 5 (Subevent Probability)** For \( A, B \in S : \)
\[
A \subseteq B \Rightarrow P(A) \leq P(B)
\]

**Theorem 6 (Addition Rule)** For arbitrary \( A, B \in S : \)
\[
P(A \cup B) = P(A) + P(B) - P(A \cap B)
\]

**Law of Large Numbers**: Given the axioms (and theorems) above, we can approach the statistical definition of probability (see Section 1.1.2) from the other side: since the relative frequency \( \frac{n(A)}{n} \) can vary between experimental runs, we can speak of the probability that it lies in a given interval. If the actual probability of the event is \( P(A) \), then we can say for an arbitrary \( \varepsilon \in \mathbb{R} \) that:
\[
\lim_{n \to \infty} P\left(\left|\frac{n(A)}{n} - P(A)\right| < \varepsilon\right) = 1
\]
This assertion is known as the **Law of Large Numbers**.
1.2 Stochastic (in)Dependence

Definition 1 (Stochastic Independence) Two events $A$ and $B$ are said to be independent iff:

$$P(A \cap B) = P(A) \times P(B)$$

Theorem 7 If two events $A$ and $B$ are independent, then $A$ and $\overline{B}$, $\overline{A}$ and $B$, as well as $\overline{A}$ and $\overline{B}$ are also independent.

1.2.1 Conditional Probability

Definition 2 Given $P(B) > 0$, the conditional probability of $A$ given $B$ $P(A|B)$ is defined as:

$$P(A|B) := \frac{P(A \cap B)}{P(B)}$$

Theorem 8 Assuming $P(B) > 0$, two events $A$ and $B$ are independent iff

$$P(A|B) = P(A)$$

Theorem 9 (Multiplication Rule) For all $A, B \in S$:

$$P(B)P(A|B) = P(A \cap B) = P(A)P(B|A)$$

Theorem 10 (Chain Rule) For $n \in \mathbb{N}$, $A_1, \ldots, A_n \in S$:

$$P(A_1 \cap \cdots \cap A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2) \cdots P(A_n|A_1 \cap \cdots \cap A_{n-1})$$

... somewhat prettier:

$$P\left( \bigcap_{i=1}^{n} A_i \right) = \prod_{i=1}^{n} P\left( A_i \bigg| \bigcap_{j=1}^{i-1} A_j \right)$$

Theorem 11 (Bayes’ Theorem)

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$