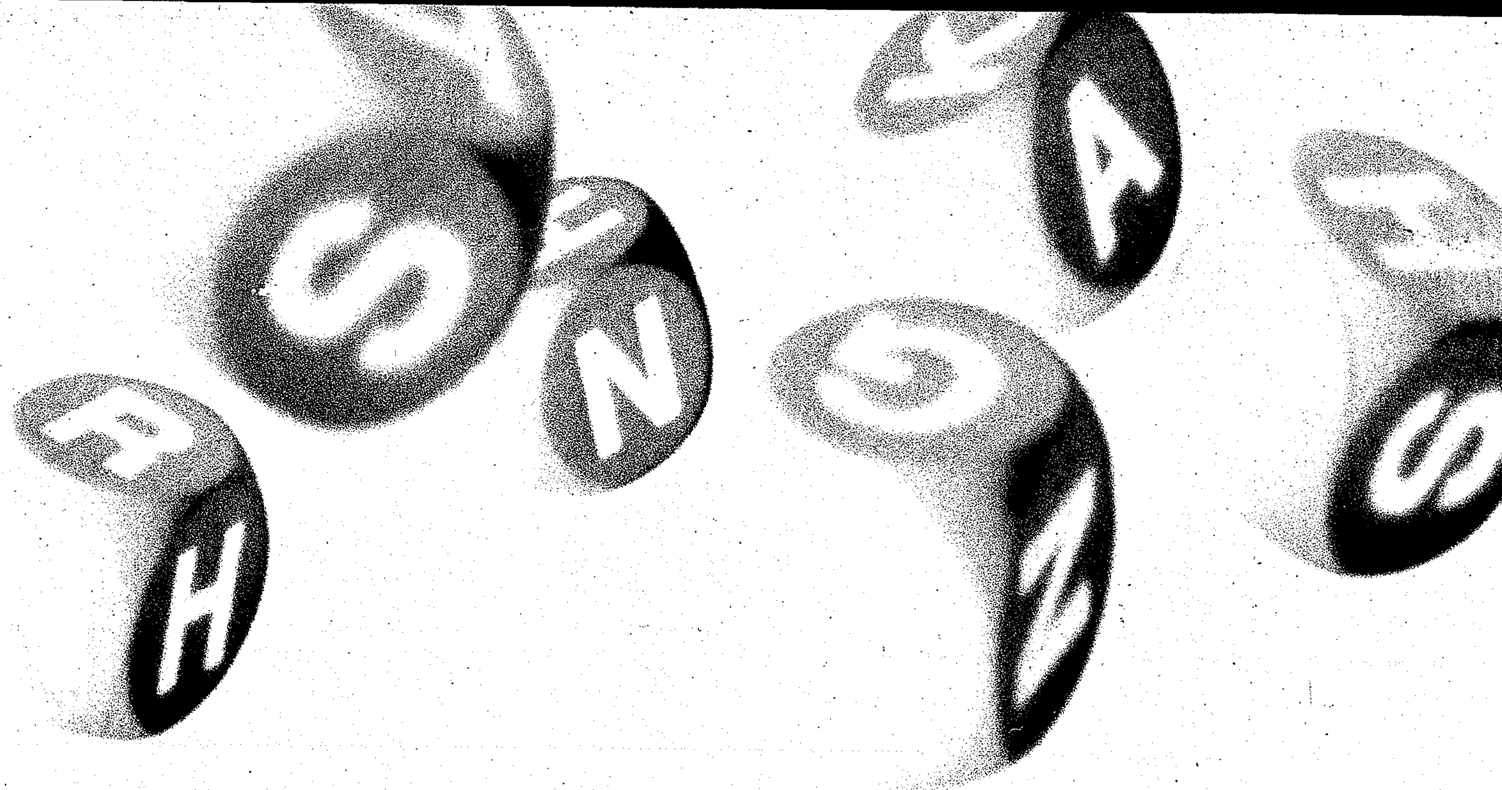
EXHIBIT E



FOUNDATIONS OF STATISTICAL NATURAL LANGUAGE PROCESSING

CHRISTOPHER D. MANNING AND
HINRICH SCHÜTZE

Typeset in 10/13 Lucida Bright by the authors using $\LaTeX\,2_\varepsilon$.
Printed and bound in the United States of America.

Library of Congress Cataloging-in-Publication Information

Manning, Christopher D.
Foundations of statistical natural language processing / Christopher D. Manning, Hinrich Schütze.
p. cm.
Includes bibliographical references (p.) and index.
ISBN 978-0-262-13360-9 (hc.:alk.paper)
1. Computational linguistics—Statistical methods. I. Schütze, Hinrich. II. Title.
P98.5.S83M36 1999
410'.285—dc21 99-21137
CIP

10

- $P(\Omega) = 1$

DISJOINT

- Countable additivity: For *disjoint* sets $A_j \in \mathcal{F}$ (i.e., $A_j \cap A_k = \emptyset$ for $j \neq k$)

$$P\Big(\bigcup_{j=1}^{\infty} A_j\Big) = \sum_{j=1}^{\infty} P(A_j) \tag{2.1}$$

We call $P(A)$ the probability of the event A. These axioms say that an event that encompasses, say, three distinct possibilities must have a probability that is the sum of the probabilities of each possibility, and that since an experiment must have some basic outcome as its result, the probability of that is 1. Using basic set theory, we can derive from these axioms a set of further properties of probability functions; see exercise 2.1.

PROBABILITY SPACE

A well-founded *probability space* consists of a sample space Ω, a σ-field of events $\mathcal{F}$, and a probability function P. In Statistical NLP applications, we always seek to properly define such a probability space for our models. Otherwise, the numbers we use are merely ad hoc scaling factors, and there is no mathematical theory to help us. In practice, though, corners often have been, and continue to be, cut.

Example 1: A fair coin is tossed 3 times. What is the chance of 2 heads?

Solution: The experimental protocol is clear. The sample space is:

$$\Omega = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$$

UNIFORM DISTRIBUTION

Each of the basic outcomes in Ω is equally likely, and thus has probability $1/8$. A situation where each basic outcome is equally likely is called a *uniform distribution*. In a finite sample space with equiprobable basic outcomes, $P(A) = \frac{|A|}{|\Omega|}$ (where $|A|$ is the number of elements in a set A). The event of interest is:

$$A = \{HHT, HTH, THH\}$$

So:

$$P(A) = \frac{|A|}{|\Omega|} = \frac{3}{8}$$

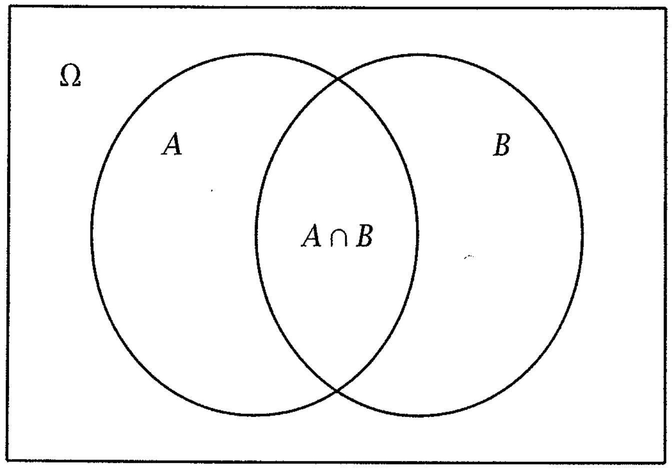


Figure 2.1 A diagram illustrating the calculation of conditional probability $P(A|B)$. Once we know that the outcome is in B, the probability of A becomes $P(A \cap B)/P(B)$.

2.1.2 Conditional probability and independence

CONDITIONAL PROBABILITY

PRIOR PROBABILITY

POSTERIOR PROBABILITY

Sometimes we have partial knowledge about the outcome of an experiment and that naturally influences what experimental outcomes are possible. We capture this knowledge through the notion of *conditional probability*. This is the updated probability of an event given some knowledge. The probability of an event before we consider our additional knowledge is called the *prior probability* of the event, while the new probability that results from using our additional knowledge is referred to as the *posterior probability* of the event. Returning to example 1 (the chance of getting 2 heads when tossing 3 coins), if the first coin has been tossed and is a head, then of the 4 remaining possible basic outcomes, 2 result in 2 heads, and so the probability of getting 2 heads now becomes $\frac{1}{2}$. The conditional probability of an event A given that an event B has occurred ($P(B) > 0$) is:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{2.2}$$

Even if $P(B) = 0$ we have that:

$$P(A \cap B) = P(B)P(A|B) = P(A)P(B|A) \qquad \text{[The multiplication rule]} \tag{2.3}$$

We can do the conditionalization either way because set intersection is symmetric ($A \cap B = B \cap A$). One can easily visualize this result by looking at the diagram in figure 2.1.

The generalization of this rule to multiple events is a central result that will be used throughout this book, the *chain rule*:

CHAIN RULE

(2.4) $$P(A_1 \cap \ldots \cap A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2) \cdots P(A_n| \cap_{i=1}^{n-1} A_i)$$

▼ The chain rule is used in many places in Statistical NLP, such as working out the properties of Markov models in chapter 9.

INDEPENDENCE

DEPENDENCE

CONDITIONAL INDEPENDENCE

Two events A, B are *independent* of each other if $P(A \cap B) = P(A)P(B)$. Unless $P(B) = 0$ this is equivalent to saying that $P(A) = P(A|B)$ (i.e., knowing that B is the case does not affect the probability of A). This equivalence follows trivially from the chain rule. Otherwise events are *dependent*. We can also say that A and B are *conditionally independent* given C when $P(A \cap B|C) = P(A|C)P(B|C)$.

2.1.3 Bayes' theorem

BAYES' THEOREM

Bayes' theorem lets us swap the order of dependence between events. That is, it lets us calculate $P(B|A)$ in terms of $P(A|B)$. This is useful when the former quantity is difficult to determine. It is a central tool that we will use again and again, but it is a trivial consequence of the definition of conditional probability and the chain rule introduced in equations (2.2) and (2.3):

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

NORMALIZING CONSTANT

The righthand side denominator $P(A)$ can be viewed as a *normalizing constant*, something that ensures that we have a probability function. If we are simply interested in which event out of some set is most likely given A, we can ignore it. Since the denominator is the same in all cases, we have that:

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

However, we can also evaluate the denominator by recalling that:

$$\begin{aligned} P(A \cap B) &= P(A|B)P(B) \\ P(A \cap \overline{B}) &= P(A|\overline{B})P(\overline{B}) \end{aligned}$$

So we have:

$$\begin{aligned} P(A) &= P(A \cap B) + P(A \cap \overline{B}) \qquad \text{[additivity]} \\ &= P(A|B)P(B) + P(A|\overline{B})P(\overline{B}) \end{aligned}$$