Appears in the Proceedings of the Fifteenth National Conference on Artificial Intelligence, July 1998, Madison, WI

Dependent Bigram Identification[*]

Ted Pedersen
Department of Computer Science & Engineering
Southern Methodist University
Dallas, TX 75275–0122
pedersen@seas.smu.edu

Dependent bigrams are two consecutive words that occur together in a text more often than would be expected purely by chance. Identifying such bigrams is an important issue since they provide valuable clues for machine translation, word sense disambiguation, and information retrieval. A variety of significance tests have been proposed (e.g., Church et. al., 1991, Dunning, 1993, Pedersen et. al, 1996) to identify these interesting lexical pairs. In this poster I present a new statistic, *minimum sensitivity*, that is simple to compute and is free from the underlying distributional assumptions commonly made by significance tests.

The challenge in identifying dependent bigrams is that most are relatively rare regardless of the amount of text being considered. This follows from the distributional tendencies of individual bigrams as described by Zipf's Law. If the frequencies of the bigrams in a text are ordered from most to least frequent, $(f_1, f_2, \ldots, f_m)$, these frequencies roughly obey $f_i \propto \frac{1}{i}$.

Consider the following example from a 1,300,000 word sample of the ACL/DCI Wall Street Journal Corpus. A contingency table containing the frequency counts of *oil* and *industry* is shown below. These counts show that *oil industry* occurs 17 times, *oil* occurs without *industry* 240 times, *industry* occurs without *oil* 1001 times, and bigrams other than *oil industry* occur 1,298,742 times. This distribution is sparse and skewed and thus violates a central assumption implicit in significance testing of contingency tables (Read & Cressie 1988).

		W_2		
		industry	¬industry	totals
W_1	oil	$n_{11}=$ 17	$n_{12}=$ 240	$n_{1+}=$ 257
	¬oil	$n_{21}=$ 1001	$n_{22}=$ 1298742	$n_{2+}=$ 1299743
	totals	$n_{+1}=$1018	$n_{+2}=$1298982	$n_{++}=$1300000

Sensitivity is classically defined as the proportion of true results that agree with the true state. For lexical relationships sensitivity is a conditional probability that is the ratio of how often a word occurs in a specific bigram ($w_1 w_2$) to how often it occurs overall.

Sensitivity is computed as follows for each of the two words in a bigram:

$$S_{w_1} = \frac{n_{11}}{n_{+1}} = P(w_1|w_2) \qquad S_{w_2} = \frac{n_{11}}{n_{1+}} = P(w_2|w_1)$$

These values range from 0 to 1 and their minimum serves as the measure of dependence between the two words. The minimum sensitivity is 1 when w_1 and w_2 always, and only, occur together. It is 0 when w_1 and w_2 never occur together. The greater the minimum sensitivity the higher the level of dependence between the two words in a bigram.

From the data in the contingency table, S_{w_1} indicates how sensitive *oil* is to *industry*. Given that *industry* occurs in a text, how often does *oil* precede it? $S_{w_1} = \frac{17}{1018} = .017$. S_{w_2} measures how sensitive *industry* is to *oil*. Given that *oil* occurs, how often does *industry* follow it? $S_{w_2} = \frac{17}{257} = .066$. The former is the minimum sensitivity value and is thus the measure of dependence between *oil* and *industry*.

Experimental results show that minimum sensitivity results in the identification of bigrams that are largely made up of content words. Significance tests frequently identify dependent bigrams where one of the words is a very high frequency non–content word such as *the* or *of*. For example, *the industry* is considered a dependent bigram by the significance tests but not by minimum sensitivity. The tendency of minimum sensitivity to filter out bigrams containing non–content words is an important quality in many practical language processing applications.

Acknowledgments

This research was supported by the Office of Naval Research under grant number N00014-95-1-0776.

References

Read, T., and Cressie, N. 1988. *Goodness of fit Statistics for Discrete Multivariate Data.* New York, NY: Springer-Verlag.