Significance tests for the evaluation of ranking methods

Stefan Evert
Institut für maschinelle Sprachverarbeitung
Universität Stuttgart
Azenbergstr. 12, 70174 Stuttgart, Germany
evert@ims.uni-stuttgart.de

Abstract

This paper presents a statistical model that interprets the evaluation of ranking methods as a random experiment. This model predicts the variability of evaluation results, so that appropriate significance tests for the results can be derived. The paper concludes with an empirical validation of the model on a collocation extraction task.

1 Introduction

Many tools in the area of natural-language processing involve the application of ranking methods to sets of candidates, in order to select the most useful items from an all too often overwhelming list. Examples of such tools range from syntactic parsers (where alternative analyses are ranked by their plausibility) to the extraction of collocations from text corpora (where a ranking according to the scores assigned by a lexical association measure is the essential component of an extraction "pipeline").

To this end, a scoring function g is applied to the candidate set, which assigns a real number $g(x) \in \mathbb{R}$ to every candidate x.[1] Conventionally, higher scores are assigned to candidates that the scoring function considers more "useful". Candidates can then be selected in one of two ways: (i) by comparison with a pre-defined threshold $\gamma \in \mathbb{R}$ (i.e. x is accepted iff $g(x) \geq \gamma$), resulting in a γ-*acceptance set*; (ii) by ranking the entire candidate set according to the scores $g(x)$ and selecting the n highest-scoring candidates, resulting in an n-*best list* (where n is either determined by practical constraints or interactively by manual inspection). Note that an n-best list can also be interpreted as a γ-acceptance set with a suitably chosen cutoff threshold $\gamma_g(n)$ (determined from the scores of all candidates).

Ranking methods usually involve various heuristics and statistical guesses, so that an empirical evaluation of their performance is necessary. Even when there is a solid theoretical foundation, its predictions may not be borne out in practice. Often, the main goal of an evaluation experiment is the comparison of different ranking methods (i.e. scoring functions) in order to determine the most useful one.

A widely-used evaluation strategy classifies the candidates accepted by a ranking method into "good" ones (*true positives*, TP) and "bad" ones (*false positives*, FP). This is sometimes achieved by comparison of the relevant γ-acceptance sets or n-best lists with a gold standard, but for certain applications (such as collocation extraction), manual inspection of the candidates leads to more clear-cut and meaningful results. When TPs and FPs have been identified, the precision Π of a γ-acceptance set or an n-best list can be computed as the proportion of TPs among the accepted candidates. The most useful ranking method is the one that achieves the highest precision, usually comparing n-best lists of a given size n. If the full candidate set has been annotated, it is also possible to determine the recall R as the number of accepted TPs divided by the total number of TPs in the candidate set. While the evaluation of extraction tools (e.g. in information retrieval) usually requires that both precision and recall are high, ranking methods often put greater weight on high precision, possibly at the price of missing a considerable number of TPs. Moreover, when n-best lists of the same size are compared, precision and recall are fully equivalent.[2] For these reasons, I will concentrate on the precision Π here.

As an example, consider the identification of collocations from text corpora. Following the methodology described by Evert and Krenn (2001), German PP-verb combinations were extracted from a chunk-parsed version of the Frankfurter Rundschau Corpus.[3] A cooccurrence frequency threshold of

[1] Some systems may directly produce a sorted candidate list without assigning explicit scores. However, unless this operation is (implicitly) based on an underlying scoring function, the result will in most cases be a partial ordering (where some pairs of candidates are incomparable) or lead to inconsistencies.

[2] Namely, $\Pi = n_{\text{TP}} \cdot R/n$, where n_{TP} stands for the total number of TPs in the candidate set.

[3] The Frankfurter Rundschau Corpus is a German newspaper corpus, comprising ca. 40 million words of text. It is part of the ECI Multilingual Corpus 1 distributed by ELSNET. For this

$f \geq 30$ was applied, resulting in a candidate set of $5\,102$ PP-verb pairs. The candidates were then ranked according to the scores assigned by four association measures: the *log-likelihood* ratio G^2 (Dunning, 1993), Pearson's *chi-squared* statistic X^2 (Manning and Schütze, 1999, 169–172), the *t-score* statistic t (Church et al., 1991), and mere cooccurrence *frequency* f.[4] TPs were identified according to the definition of Krenn (2000). The graphs in Figure 1 show the precision achieved by these measures, for n ranging from 100 to $2\,000$ (lists with $n < 100$ were omitted because the graphs become highly unstable for small n). The baseline precision of 11.09% corresponds to a random selection of n candidates.

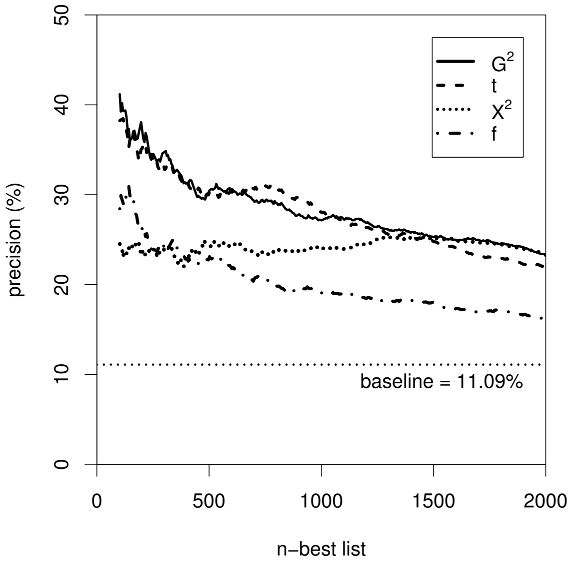

Figure 1: Evaluation example: candidates for German PP-verb collocations are ranked by four different association measures.

From Figure 1, we can see that G^2 and t are the most useful ranking methods, t being marginally better for $n \approx 800$ and G^2 for $n \geq 1\,500$. Both measures are by far superior to frequency-based ranking. The evaluation results also confirm the argument of Dunning (1993), who suggested G^2 as a more robust alternative to X^2. Such results cannot be taken at face value, though, as they may simply be due to chance. When two equally useful ranking methods are compared, method A might just happen to perform better in a particular experiment, with B taking the lead in a repetition of the experiment under similar conditions. The causes of such random variation include the source material from which the candidates are extracted (what if a slightly different source had been used?), noise introduced by automatic pre-processing and extraction tools, and the uncertainty of human annotators manifested in varying degrees of inter-annotator agreement. Most researchers understand the necessity of testing whether their results are *statistically significant*, but it is fairly unclear which tests are appropriate. For instance, Krenn (2000) applies the standard χ^2-test to her comparative evaluation of collocation extraction methods. She is aware, though, that this test assumes independent samples and is hardly suitable for different ranking methods applied to the same candidate set: Krenn and Evert (2001) suggest several alternative tests for related samples. A wide range of exact and asymptotic tests as well as computationally intensive randomisation tests (Yeh, 2000) are available and add to the confusion about an appropriate choice.

The aim of this paper is to formulate a statistical model that interprets the evaluation of ranking methods as a *random experiment*. This model defines the degree to which evaluation results are affected by random variation, allowing us to derive appropriate significance tests. After formalising the evaluation procedure in Section 2, I recast the procedure as a random experiment and make the underlying assumptions explicit (Section 3.1). On the basis of this model, I develop significance tests for the precision of a single ranking method (Section 3.2) and for the comparison of two ranking methods (Section 3.3). The paper concludes with an empirical validation of the statistical model in Section 4.

2 A formal account of ranking methods and their evaluation

In this section I present a formalisation of rankings and their evaluation, giving γ-acceptance sets a geometrical interpretation that is essential for the formulation of a statistical model in Section 3.

The scores computed by a ranking method are based on certain *features* of the candidates. Each candidate can therefore be represented by its *feature vector* $x \in \Omega$, where Ω is an abstract *feature space*. For all practical purposes, Ω can be equated with a subset of the (possibly high-dimensional) real Euclidean space $\mathbb{R}^m$. The complete *set of candidates* corresponds to a discrete subset $C \subseteq \Omega$ of the feature space.[5] A ranking method is represented by

experiment, the corpus was annotated with the partial parser YAC (Kermes, 2003).

[4] See Evert (2004) for detailed information about these association measures, as well as many further alternatives.

[5] More precisely, C is a multi-set because there may be multiple candidates with identical feature vectors. In order to simplify notation I assume that C is a proper subset of Ω, which

a real-valued function $g : \Omega \rightarrow \mathbb{R}$ on the feature space, called a *scoring function* (SF). In the following, I assume that there are no candidates with equal scores, and hence no ties in the rankings.[6]

The γ-acceptance set for a SF g contains all candidates $x \in C$ with $g(x) \geq \gamma$. In a geometrical interpretation, this condition is equivalent to $x \in A_g(\gamma) \subseteq \Omega$, where

$$A_g(\gamma) := \{x \in \Omega \mid g(x) \geq \gamma\}$$

is called the *γ-acceptance region* of g. The γ-acceptance set of g is then given by the intersection $A_g(\gamma) \cap C =: C_g(\gamma)$. The selection of an *$n$-best list* is based on the γ-acceptance region $A_g(\gamma_g(n))$ for a suitably chosen *n-best threshold* $\gamma_g(n)$.[7]

As an example, consider the collocation extraction task introduced in Section 1. The feature vector x associated with a collocation candidate represents the cooccurrence frequency information for this candidate: $x = (O_{11}, O_{12}, O_{21}, O_{22})$, where O_{ij} are the cell counts of a 2×2 contingency table (Evert, 2004). Therefore, we have a four-dimensional feature space $\Omega \subseteq \mathbb{R}^4$, and each association measure defines a SF $g : \Omega \rightarrow \mathbb{R}$. The selection of collocation candidates is usually made in the form of an n-best list, but may also be based on a pre-defined threshold γ.[8]

For an evaluation in terms of precision and recall, the candidates in the set C are classified into *true positives* C_+ and *false positives* C_-. The precision corresponding to an acceptance region A is then given by

$$\Pi_A := |C_+ \cap A| \,/\, |C \cap A| \,, \tag{1}$$

i.e. the proportion of TPs among the accepted candidates. The precision achieved by a SF g with threshold γ is $\Pi_{C_g(\gamma)}$. Note that the numerator in Eq. (1) reduces to n for an n-best list (i.e. $\gamma = \gamma_g(n)$), yielding the n-best precision $\Pi_{g,n}$. Figure 1 shows graphs of $\Pi_{g,n}$ for $100 \leq n \leq 2\,000$, for the SFs $g_1 = G^2$, $g_2 = t$, $g_3 = X^2$, and $g_4 = f$.

3 Significance tests for evaluation results

3.1 Evaluation as a random experiment

When an evaluation experiment is repeated, the results will not be exactly the same. There are many causes for such variation, including different source material used by the second experiment, changes in the tool settings, changes in the evaluation criteria, or the different intuitions of human annotators. Statistical significance tests are designed to account for a small fraction of this variation that is due to random effects, assuming that all parameters that may have a systematic influence on the evaluation results are kept constant. Thus, they provide a lower limit for the variation that has to be expected in an actual repetition of the experiment. Only when results are significant can we expect them to be reproducible, but even then a second experiment may draw a different picture.

In particular, the influence of qualitatively different source material or different evaluation criteria can never be predicted by statistical means alone. In the example of the collocation extraction task, randomness is mainly introduced by the selection of a source corpus, e.g. the choice of one particular newspaper rather than another. Disagreement between human annotators and uncertainty about the interpretation of annotation guidelines may also lead to an element of randomness in the evaluation. However, even significant results cannot be generalised to a different type of collocation (such as adjective-noun instead of PP-verb), different evaluation criteria, a different domain or text type, or even a source corpus of different size, as the results of Krenn and Evert (2001) show.

A first step in the search for an appropriate significance test is to formulate a (plausible) model for random variation in the evaluation results. Because of the inherent randomness, every repetition of an evaluation experiment under similar conditions will lead to different candidate sets C_+ and C_-. Some elements will be entirely new candidates, sometimes the same candidate appears with a different feature vector (and thus represented by a different point $x \in \Omega$), and sometimes a candidate that was annotated as a TP in one experiment may be annotated as a FP in the next. In order to encapsulate all three kinds of variation, let us assume that C_+ and C_- are randomly selected from a large set of hypothetical possibilities (where each candidate corresponds to many different possibilities with different feature vectors, some of which may be TPs and some FPs).

For any acceptance region A, both the number of TPs in A, $T_A := |C_+ \cap A|$, and the number of FPs

can be enforced by adding a small amount of random *jitter* to the feature vectors of candidates.

[6] Under very general conditions, random jittering (cf. Footnote 5) ensures that no two candidates have equal scores. This procedure is (almost) equivalent to breaking ties in the rankings randomly.

[7] Since I assume that there are no ties in the rankings, $\gamma_g(n)$ can always be determined in such a way that the acceptance set contains *exactly* n candidates.

[8] For instance, Church et al. (1991) use a threshold of $\gamma = 1.65$ for the *t-score* measure corresponding to a nominal significance level of $\alpha = .05$. This threshold is obtained from the limiting distribution of the t statistic.

in A, $F_A := |C_- \cap A|$, are thus *random variables*. We do not know their precise distributions, but it is reasonable to assume that (i) T_A and F_A are always independent and (ii) T_A and T_B (as well as F_A and F_B) are independent for any two disjoint regions A and B. Note that T_A and T_B cannot be independent for $A \cap B \neq \emptyset$ because they include the same number of TPs from the region $A \cap B$. The total number of candidates in the region A is also a random variable $N_A := T_A + F_A$, and the same follows for the precision Π_A, which can now be written as $\Pi_A = T_A/N_A$.[9]

Following the standard approach, we may now assume that Π_A approximately follows a normal distribution with mean π_A and variance σ_A^2, i.e. $\Pi_A \sim N(\pi_A, \sigma_A^2)$. The mean π_A can be interpreted as the *average precision* of the acceptance region A (obtained by averaging over many repetitions of the evaluation experiment). However, there are two problems with this assumption. First, while Π_A is an unbiased estimator for π_a, the variance σ_A^2 cannot be estimated from a single experiment.[10] Second, Π_A is a discrete variable because both T_A and N_A are non-negative integers. When the number of candidates N_A is small (as in Section 3.3), approximating the distribution of Π_A by a continuous normal distribution will not be valid.

It is reasonable to assume that the distribution of N_A does not depend on the average precision π_A. In this case, N_A is called an *ancillary statistic* and can be eliminated without loss of information by conditioning on its observed value (see Lehmann (1991, 542ff) for a formal definition of ancillary statistics and the merits of conditional inference). Instead of probabilities $P(\Pi_A)$ we will now consider the conditional probabilities $P(\Pi_A \,|\, N_A)$. Because N_A is fixed to the observed value, Π_A is proportional to T_A and the conditional probabilities are equivalent to $P(T_A \,|\, N_A)$. When we choose one of the N_A candidates at random, the probability that it is a TP (averaged over many repetitions of the experiment) should be equal to the average precision π_A. Consequently, $P(T_A \,|\, N_A)$ should follow a binomial distribution with success probability π_A, i.e.

$$P(T_A = k \,|\, N_A) = \binom{N_A}{k} \cdot (\pi_A)^k \cdot (1 - \pi_A)^{N_A - k} \qquad (2)$$

for $k = 0, \ldots, N_A$. We can now make inferences about the average precision π_A based on this binomial distribution.[11]

As a second step in our search for an appropriate significance test, it is essential to understand exactly what question this test should address: What does it mean for an evaluation result (or result difference) to be significant? In fact, two different questions can be asked:

A: *If we repeat an evaluation experiment under the same conditions, to what extent will the observed precision values vary?* This question is addressed in Section 3.2.

B: *If we repeat an evaluation experiment under the same conditions, will method A again perform better than method B?* This question is addressed in Section 3.3.

3.2 The stability of evaluation results

Question A can be rephrased in the following way: *How much does the observed precision value for an acceptance region A differ from the true average precision π_A?* In other words, our goal here is to make inferences about π_A, for a given SF g and threshold γ. From Eq. (2), we obtain a binomial confidence interval for the true value π_A, given the observed values of T_A and N_A (Lehmann, 1991, 89ff). Using the customary 95% confidence level, π_A should be contained in the estimated interval in all but one out of twenty repetitions of the experiment. Binomial confidence intervals can easily be computed with standard software packages such as R (R Development Core Team, 2003). As an example, assume that an observed precision of $\Pi_A = 40\%$ is based on $T_A = 200$ TPs out of $N_A = 500$ accepted candidates. Precision graphs as those in Figure 1 display Π_A as a maximum-likelihood estimate for π_A, but its true value may range from 35.7% to 44.4% (with 95% confidence).[12]

[9] In the definition of the n-best precision $\Pi_{g,n}$, i.e. for $A = C_g(\gamma_g(n))$, the number of candidates in A is constant: $N_A = n$. At first sight, this may seem to be inconsistent with the interpretation of N_A as a random variable. However, one has to keep in mind that $\gamma_g(n)$, which is determined from the candidate set C, is itself a random variable. Consequently, A is *not* a fixed acceptance region and its variation counter-balances that of N_A.

[10] Sometimes, cross-validation is used to estimate the variability of evaluation results. While this method is appropriate e.g. for machine learning and classification tasks, it is not useful for the evaluation of ranking methods. Since the cross-validation would have to be based on random samples from a single candidate set, it would not be able to tell us anything about random variation between different candidate sets.

[11] Note that some of the assumptions leading to Eq. (2) are far from self-evident. As an example, (2) tacitly assumes that the success probability is equal to π_A regardless of the particular value of N_A on which the distribution is conditioned, which need not be the case. Therefore, an empirical validation is necessary (see Section 4).

[12] This confidence interval was computed with the R command `binom.test(200,500)`.

Figure 2 shows binomial confidence intervals for the association measures G^2 and X^2 as shaded regions around the precision graphs. It is obvious that a repetition of the evaluation experiment may lead to quite different precision values, especially for $n < 1\,000$. In other words, there is a considerable amount of uncertainty in the evaluation results for each individual measure. However, we can be confident that both ranking methods offer a substantial improvement over the baseline.

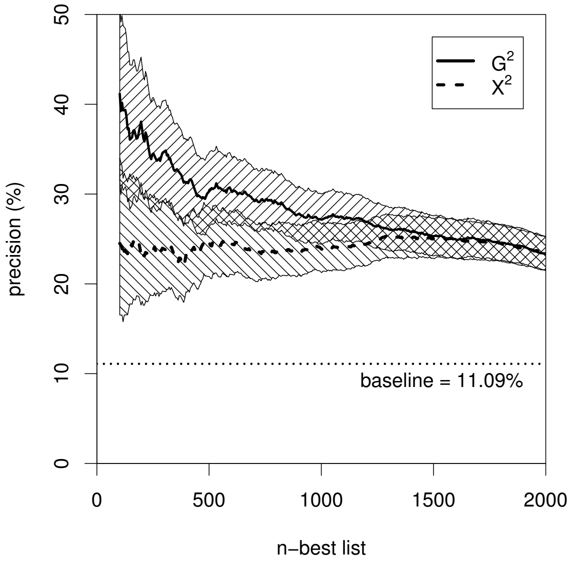

Figure 2: Precision graphs for the G^2 and X^2 measures with 95% confidence intervals.

For an evaluation based on n-best lists (as in the collocation extraction example), it has to be noted that the confidence intervals are estimates for the average precision π_A of a fixed γ-acceptance region (with $\gamma = \gamma_g(n)$ computed from the observed candidate set). While this region contains exactly $N_A = n$ candidates in the current evaluation, N_A may be different from n when the experiment is repeated. Consequently, π_A is not necessarily identical to the average precision of n-best lists.

3.3 The comparison of ranking methods

Question B can be rephrased in the following way: *Does the SF g_1 on average achieve higher precision than the SF g_2?* (This question is normally asked when g_1 performed better than g_2 in the evaluation.) In other words, our goal is to test whether $\pi_A > \pi_B$ for given acceptance regions A of g_1 and B of g_2.

The confidence intervals obtained for two SF g_1 and g_2 will often overlap (cf. Figure 2, where the confidence intervals of G^2 and X^2 overlap for all list sizes n), suggesting that there is no significant difference between the two ranking methods. Both observed precision values are consistent with an average precision $\pi_A = \pi_B$ in the region of overlap, so that the observed differences may be due to random variation in opposite directions. However, this conclusion is premature because the two rankings are *not independent*. Therefore, the observed precision values of g_1 and g_2 will tend to vary in the same direction, the degree of correlation being determined by the amount of overlap between the two rankings. Given acceptance regions $A := A_{g_1}(\gamma_1)$ and $B := A_{g_2}(\gamma_2)$, both SF make the same decision for any candidates in the intersection $A \cap B$ (both SF accept) and in the "complement" $\Omega \setminus (A \cup B)$ (both SF reject). Therefore, the performance of g_1 and g_2 can only differ in the regions $D_1 := A \setminus B$ (g_1 accepts, but g_2 rejects) and $B \setminus A$ (vice versa). Correspondingly, the counts T_A and T_B are correlated because they include the same number of TPs from the region $A \cap B$ (namely, the set $C_+ \cap A \cap B$),

Indisputably, g_1 is a better ranking method than g_2 iff $\pi_{D_1} > \pi_{D_2}$ and vice versa.[13] Our goal is thus to test the null hypothesis $H_0 : \pi_{D_1} = \pi_{D_2}$ on the basis of the binomial distributions $P(T_{D_1} \mid N_{D_1})$ and $P(T_{D_2} \mid N_{D_2})$. I assume that these distributions are independent because $D_1 \cap D_2 = \emptyset$ (cf. Section 3.1). The number of candidates in the difference regions, N_{D_1} and N_{D_2}, may be small, especially for acceptance regions with large overlap (this was one of the reasons for using conditional inference rather than a normal approximation in Section 3.1). Therefore, it is advisable to use Fisher's exact test (Agresti, 1990, 60–66) instead of an asymptotic test that relies on large-sample approximations. The data for Fisher's test consist of a 2 × 2 contingency table with columns (T_{D_1}, F_{D_1}) and (T_{D_2}, F_{D_2}). Note that a two-sided test is called for because there is no *a priori* reason to assume that g_1 is better than g_2 (or vice versa). Although the implementation of a two-sided Fisher's test is not trivial, it is available in software packages such as R.

Figure 3 shows the same precision graphs as Figure 2. Significant differences between the G^2 and X^2 measures according to Fisher's test (at a 95% confidence level) are marked by grey triangles.

[13] Note that $\pi_{D_1} > \pi_{D_2}$ does not necessarily entail $\pi_A > \pi_B$ if N_A and N_B are vastly different and $\pi_{A \cap B} \gg \pi_{D_i}$. In this case, the winner will always be the SF that accepts the smaller number of candidates (because the additional candidates only serve to lower the precision achieved in $A \cap B$). This example shows that it is "unfair" to compare acceptance sets of (substantially) different sizes just in terms of their overall precision. Evaluation should therefore either be based on n-best lists or needs to take recall into account.

Contrary to what the confidence intervals in Figure 2 suggested, the observed differences turn out to be significant for all n-best lists up to $n = 1\,250$ (marked by a thin vertical line).

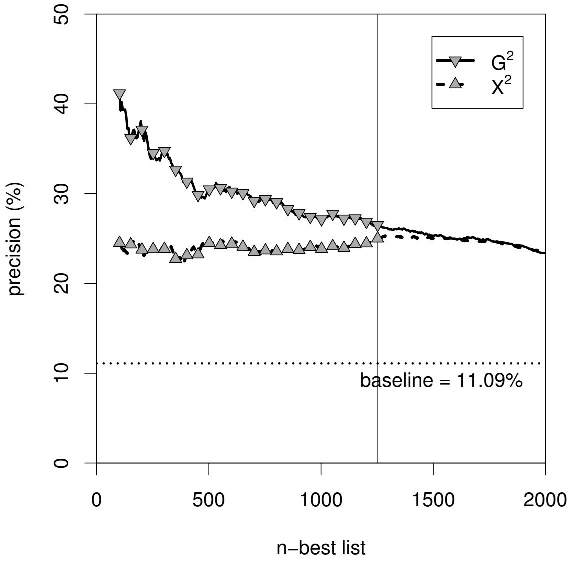

Figure 3: Significant differences between the G^2 and X^2 measures at 95% confidence level.

4 Empirical validation

In order to validate the statistical model and the significance tests proposed in Section 3, it is necessary to simulate the repetition of an evaluation experiment. Following the arguments of Section 3.1, the conditions should be the same for all repetitions so that the amount of purely random variation can be measured. To achieve this, I divided the Frankfurter Rundschau Corpus into 80 contiguous, non-overlapping parts, each one containing approx. 500k words. Candidates for PP-verb collocations were extracted as described in Section 1, with a frequency threshold of $f \geq 4$. The 80 samples of candidate sets were ranked using the association measures G^2, X^2 and t as scoring functions, and true positives were manually identified according to the criteria of (Krenn, 2000).[14] The true average precision π_A of an acceptance set A was estimated by averaging over all 80 samples.

Both the confidence intervals of Section 3.2 and the significance tests of Section 3.3 are based on the assumption that $P(T_A \,|\, N_A)$ follows a binomial distribution as given by Eq. (2). Unfortunately, it is impossible to test the conditional distribution directly, which would require that N_A is the same for all samples. Therefore, I use the following approach based on the unconditional distribution $P(\Pi_A)$. If N_A is sufficiently large, $P(\Pi_A \,|\, N_A)$ can be approximated by a normal distribution with mean $\mu = \pi_A$ and variance $\sigma^2 = \pi_A(1 - \pi_A)/N_A$ (from Eq. (2)). Since μ does not depend on N_A and the standard deviation σ is proportional to $(N_A)^{-1/2}$, it is valid to make the approximation

$$P(\Pi_A \,|\, N_A) \approx P(\Pi_A) \qquad (3)$$

as long as N_A is relatively stable. Eq. (3) allows us to pool the data from all samples, predicting that

$$P(\Pi_A) \sim N(\mu, \sigma^2) \qquad (4)$$

with $\mu = \pi_A$ and $\sigma^2 = \pi_A(1 - \pi_A)/N$. Here, N stands for the *average* number of TPs in A.

These predictions were tested for the measures $g_1 = G^2$ and $g_2 = t$, with cutoff thresholds $\gamma_1 = 32.5$ and $\gamma_2 = 2.09$ (chosen so that $N = 100$ candidates are accepted on average). Figure 4 compares the empirical distribution of Π_A with the expected distribution according to Eq. (4). These histograms show that the theoretical model agrees quite well with the empirical results, although there is a little more variation than expected.[15] The empirical standard deviation is between 20% and 40% larger than expected, with $s = 0.057$ vs. $\sigma = 0.044$ for G^2 and $s = 0.066$ vs. $\sigma = 0.047$ for t. These findings suggest that the model proposed in Section 3.1 may indeed represent a lower bound on the true amount of random variation.

Further evidence for this conclusion comes from a validation of the confidence intervals defined in Section 3.2. For a 95% confidence interval, the true proportion π_A should fall within the confidence interval in all but 4 of the 80 samples. For G^2 (with $\gamma = 32.5$) and X^2 (with $\gamma = 239.0$), π_A was outside the confidence interval in 9 cases each (three of them very close to the boundary), while the confidence interval for t (with $\gamma = 2.09$) failed in 12 cases, which is significantly more than can be explained by chance ($p < .001$, binomial test).

5 Conclusion

In the past, various statistical tests have been used to assess the significance of results obtained in the evaluation of ranking methods. There is much confusion about their validity, though, mainly due to

[14] I would like to thank Brigitte Krenn for making her annotation database of PP-verb collocations (Krenn, 2000) available, and for the manual annotation of 1 913 candidates that were not covered by the existing database.

[15] The agreement is confirmed by the Kolmogorov test of goodness-of-fit, which does not reject the theoretical model (4) in either case.

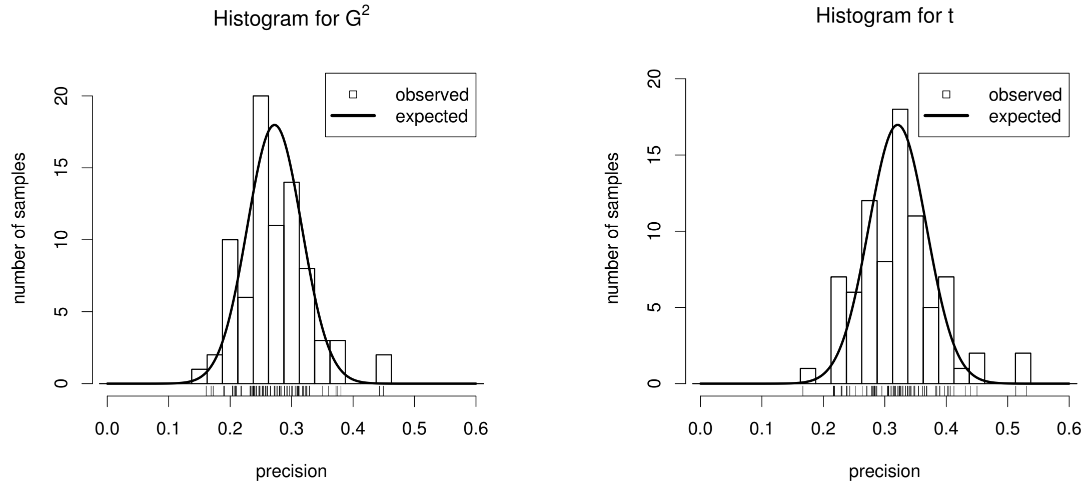

Figure 4: Distribution of the observed precision Π_A for γ-acceptance regions of the association measures G^2 (left panel) and t (right panel). The solid lines indicate the expected distribution according to Eq. (2).

the fact that assumptions behind the application of a test are seldom made explicit. This paper is an attempt to remedy the situation by interpreting the evaluation procedure as a random experiment. The model assumptions, motivated by intuitive arguments, are stated explicitly and are open for discussion. Empirical validation on a collocation extraction task has confirmed the usefulness of the model, indicating that it represents a lower bound on the variability of evaluation results. On the basis of this model, I have developed appropriate significance tests for the evaluation of ranking methods. These tests are implemented in the UCS toolkit, which was used to produce the graphs in this paper and can be downloaded from `http://www.collocations.de/`.

References

Alan Agresti. 1990. *Categorical Data Analysis*. John Wiley & Sons, New York.

Kenneth Church, William Gale, Patrick Hanks, and Donald Hindle. 1991. Using statistics in lexical analysis. In *Lexical Acquisition: Using On-line Resources to Build a Lexicon*, pages 115–164. Lawrence Erlbaum.

Ted Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74.

Stefan Evert and Brigitte Krenn. 2001. Methods for the qualitative evaluation of lexical association measures. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pages 188–195, Toulouse, France.

Stefan Evert. 2004. An on-line repository of association measures. `http://www.collocations.de/AM/`.

Hannah Kermes. 2003. *Off-line (and On-line) Text Analysis for Computational Lexicography*. Ph.D. thesis, IMS, University of Stuttgart. Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung (AIMS), volume 9, number 3.

Brigitte Krenn and Stefan Evert. 2001. Can we do better than frequency? a case study on extracting pp-verb collocations. In *Proceedings of the ACL Workshop on Collocations*, pages 39–46, Toulouse, France, July.

Brigitte Krenn. 2000. *The Usual Suspects: Data-Oriented Models for the Identification and Representation of Lexical Collocations.*, volume 7 of *Saarbrücken Dissertations in Computational Linguistics and Language Technology*. DFKI & Universität des Saarlandes, Saarbrücken, Germany.

E. L. Lehmann. 1991. *Testing Statistical Hypotheses*. Wadsworth, 2nd edition.

Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA.

R Development Core Team, 2003. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-00-3. See also `http://www.r-project.org/`.

Alexander Yeh. 2000. More accurate tests for the statistical significance of result differences. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000)*, Saarbrücken, Germany.