# Scaled log likelihood ratios for the detection of abbreviations in text corpora

Tibor Kiss
Sprachwissenschaftliches Institut
Ruhr-Universität Bochum
D-44780 Bochum
tibor@linguistics.ruhr-uni-bochum.de

Jan Strunk
Sprachwissenschaftliches Institut
Ruhr-Universität Bochum
D-44780 Bochum
strunk@linguistics.ruhr-uni-bochum.de

**Abstract**

We describe a language-independent, flexible, and accurate method for the detection of abbreviations in text corpora. It is based on the idea that an abbreviation can be viewed as a collocation, and can be identified by using methods for collocation detection such as the *log likelihood ratio.* Although the log likelihood ratio is known to show a good recall, its precision is poor. We employ scaling factors which lead to a strong improvement of precision. Experiments with English and German corpora show that abbreviations can be detected with high accuracy.

**Introduction**

The detection of abbreviations in a text corpus forms one of the initial steps in tokenization (cf. Liberman/Church 1992). This is not a trivial task, since a tokenizer is confronted with ambiguous tokens. For English, e.g., Palmer/Hearst (1997:241) report that *periods* (•) can be used as decimal points, abbreviation marks, end-of-sentence marks, and as abbreviation marks at the end of a sentence. In this paper, we will concentrate on the classification of the period as either an abbreviation mark or a punctuation mark. We assume that an abbreviation can be viewed as a *collocation* consisting of the abbreviated word itself and the following •. In case of an abbreviation, we expect the occurrence of • following the previous 'word' to be more likely than in a case of an end-of-sentence punctation. The starting point is the *log likelihood ratio* (log λ, Dunning 1993).

If the null hypothesis *($H_0$)* – as given in (1) – expresses that the occurrence of a period is independent of the preceeding word, the alternative hypothesis *($H_A$)* in (2) assumes that the occurrence of a period is not independent of the occurrence of the word preceeding it.

(1) $H_0$: $P(\bullet|w) = p = P(\bullet|\neg w)$
(2) $H_A$: $P(\bullet|w) = p_1 \neq p_2 = P(\bullet|\neg w)$

The log λ of the two hypotheses is given in (3). Its distribution is asymptotic to a $\chi^2$ distribution and can hence be used as a test statistic (Dunning 1993).

(3) $$\log \lambda = -2 \log \left( L(H_0) \Big/ L(H_A) \right)$$

**1 Problems for an unscaled log λ approach**

Although log λ identifies collocations much better than competing approaches (Dunning 1993) in terms of its *recall,* it suffers from its relatively poor *precision rates.* As is reported in Evert et al. (2000), log λ is very likely to detect all collocations contained in a corpus, but as more collocations are detected with decreasing log λ, the number of wrongly classified items increases. The table in (4) is a sample from the *Wall Street Journal (1987).*[1] According to the asymptotic $\chi^2$ distribution all the pairs given in (4) count as candidates for abbreviations. Some of the 'true' abbreviations are either ranked lower than non-abbreviations or receive the same log λ values as non-abbreviations. Candidates which should not be analyzed as abbreviations are indicated in boldface.

(4) *Candidates for abbreviations from WSJ*

[1] As distributed by ACL/DCI. We have removed all annotations from the corpora before processing them.

*(1987)*

| *Candidate* | *C(w, •)* | *C(w, ¬•)* | *log λ* |
|---|---|---|---|
| L.F | 5 | 0 | 29.29 |
| N.H | 5 | 0 | 29.29 |
| **holiday** | 7 | 4 | 27.02 |
| **direction** | 8 | 8 | 25.56 |
| **ounces** | 4 | 0 | 23.43 |
| Vt | 4 | 0 | 23.43 |
| **debts** | 7 | 7 | 22.36 |
| **Frankfurt** | 5 | 2 | 21.13 |
| U.N | 3 | 0 | 17.57 |
| **depositor** | 3 | 0 | 17.57 |

In the present sample, the likelihood of a period being dependent on the word preceeding it should be 99.99 % if its log λ is higher than 7.88.[2] But, as has been illustrated in (4), even this figure leads to a problematic classification of the candidates, since many non-abbreviations are wrongly classified as being abbreviations. This means that an unmodified log λ approach to the detection of abbreviations will produce many errors and thus cannot be employed.

## 2 Scaling log likelihood ratios

Since a pure log λ approach falsely classifies many non-abbreviations as being abbreviations, we use log λ as a basic ranking which is scaled by several factors. These factors have been experimentally developed by measuring their effect in terms of precision and recall on a training corpus from WSJ.[3] The result of the scaling operation is a much more compact ranking of the true positives in the corpus. The effect of the scaling methods on the data presented in (4) are illustrated in (5).

By applying the scaling factors, the asymptotic relation to the $\chi^2$ distribution cannot be retained. The threshold value of the classification is hence no longer determined by the $\chi^2$ distribution, but determined on the basis of the classification results derived from the training corpus. The scaling factors, once they have been determined on the basis of the training corpus, have not been modified any further. In this sense, the method described here can be characterized as a corpus-filter method, where a given corpus is used to filter the initial results (cf. Grefenstette 1999:128f.).

(5) *Result of applying scaling factors*

| *Candidate* | *log λ* | *S(log λ)* |
|---|---|---|
| L.F | 29.29 | 216.43 |
| N.H | 29.29 | 216.43 |
| **holiday** | 27.02 | 0.03 |
| **direction** | 25.56 | 0.00 |
| ***ounces*** | 23.43 | 3.17 |
| Vt | 23.43 | 173.14 |
| **debts** | 22.36 | 0.00 |
| **Frankfurt** | 21.13 | 0.01 |
| U.N | 17.57 | 17.57 |
| **depositor** | 17.57 | 0.04 |

In the present setting, applying the scaling factors to the training corpus has led to to a threshold value of 1.0. Hence, a value above 1.0 allows a classification of a given pair as an abbreviation, while a value below that leads to an exclusion of the candidate. An ordering of the candidates from table (5) is given in (6), where the threshold is indicated through the dashed line.

(6) *Ranking according to S(log λ)*

| *Candidate* | *log λ* | *S(log λ)* |
|---|---|---|
| L.F | 29.29 | 216.43 |
| N.H | 29.29 | 216.43 |
| Vt | 23.43 | 173.14 |
| Thurs | 29.29 | 29.29 |
| U.N | 17.57 | 17.57 |
| **ounces** | 23.43 | 3.17 |
| depositor | 17.57 | 0.04 |
| holiday | 27.02 | 0.03 |
| Frankfurt | 21.13 | 0.01 |
| direction | 25.56 | 0.00 |
| debts | 22.36 | 0.00 |

As can be witnessed in (6), the scaling methods are not perfect. In particular, *ounces* is still wrongly considered as an initial element of an abbreviation, poiting to a weakness of the approach which will be discussed in section 5.

## 3 The scaling factors

We have employed three different scaling factors, as given in (7), (8), and (9).[4] Each scaling

[2] This is the corresponding $\chi^2$ value for a confidence degree of 99.99 %.

[3] The training corpus had a size of 6 MB.

[4] The use of *e* as a base for scaling factors $S_1$ and $S_2$ reflects that log λ can also be expressed as $H_A$ being $e^{log\ \lambda/2}$ more likely than $H_0$ (cf. Manning/Schütze

factor is applied to the log λ of a candidate pair. The weighting factors are formulated in such a way that allows a tension between them (cf. section 3.4). The effect of this tension is that an increase following from one factor may be cancelled out or reduced by a decrease following from the application of another factor, and vice versa.

(7) $S_1(\log \lambda)$: $\log \lambda \cdot e^{C(word, \bullet)/C(word, \neg\bullet)}$.

(8) $S_2(\log \lambda)$: $\log \lambda \cdot \frac{C(word,\bullet) - C(word,\neg\bullet)}{C(word,\bullet) + C(word,\neg\bullet)}$.

(9) $S_3(\log \lambda)$: $\log \lambda \cdot 1/e^{\text{length of word}}$.

### 3.1 Ratio of occurrence: $S_1$

By employing scaling factor (7), the log λ is additionally weighted by the ranking which is determined by the occurrence of pairs of the form *(word, •)* in relation to pairs of the form *(word, ¬•)*. If events of the second type are either rare or at least lower than events of the first type, the scaling factor leads to an increase of the initial log λ value.[5]

### 3.2 Relative difference: $S_2$

The second scaling factor is a variation of the *relative difference*. Depending on the figures of *C(word, •)* and *C(word, ¬•)*, its value can be either positive, negative, or 0.

(10) If *C(word, •)* > *C(word, ¬•)*, $0 < S_2 \leq 1$.
(11) If *C(word, •)* = *C(word, ¬•)*, $S_2 = 0$.
(12) If C(word, •) < C(word, ¬•), $-1 \leq S_2 < 0$.

If *C(word, ¬•) = 0*, $S_2$ reaches a maximum of 1. Hence, $S_2$ in general leads to a reduction of the initial log λ value. $S_2$ also has a significant effect on log λ if the occurrence of *word* with • equals the occurrence of *word* without •. In this case, $S_2$ will be 0. Since the log λ values are multiplied with each scaling factor, a value of 0 for $S_2$ will lead to a value of 0 throughout. Hence the pair *(word, •)* will be excluded from being an abbreviation. This move seems extremely plausible: if *word* occurs approximately the same time with and without a following •, it is quite unlikely that the pair *(word, •)* forms an abbreviation.[6] Similarly, the value of $S_2$ will be negative if the number of occurrences of *word* without • is higher than the number of occurrences of *word* with •. Again, the resulting decrease reflects that the pair *(word, •)* is even more unlikely to be an abbreviation.

Both the relative difference ($S_2$) and the ratio of occurrence ($S_1$) allow a scaling that abstracts away from the absolute figure of occurrence, which strongly influences log λ.[7]

### 3.3 Length of abbreviations: $S_3$

Scaling factor (9), finally, leads to a reduction of log λ depending on the length of the word which preceeds a period. This scaling factor follows the idea that an abbreviation is more likely to be short.

### 3.4 Interaction of scaling factors

As was already mentioned, the scaling factors can interact with each other. Consequently, an increase by a factor may be reduced by another one. This can be illustrated with the pair (U.N, •) in (6). The application of the scaling factors does not change the value as the initial log λ calculation.

(13) $S_1(\text{U.N}, \bullet) = e^3$, $S_2(\text{U.N}, \bullet) = 1$,
$S_3(\text{U.N}, \bullet) = \frac{1}{e^3}$

Since the length of *word* actually equals its occurrence together with a •, and since U.N never occurs without a trailing •, $S_1$ leads to an increase by a factor of $e^3$, which however is fully compensated by the application of $S_3$.

---

1999:172f.).

[5] If *C(word, ¬•) = 0*, $S_1(\log \lambda) = \log \lambda \cdot e^{C(word,\bullet)}$, reflecting an even higher likelihood that the pair should actually count as an abbreviation.

[6] Obviously, this assumption is only valid if the absolute number of occurrence is not too small.

[7] As an illustration, consider the pairs *(outstanding, •)* and *(Adm, •)*. The first pair occurs 260 times in our training corpus, the second one 51 times. While *(outstanding, ¬•)* occurs 246 times, *(Adm, ¬•)* never occurs. Still, the log λ value for *(outstanding, •)* is 804.34, while the log λ value for *(Adm, •)* is just 289.38, reflecting a bias for absolute numbers of occurrence.

## 4 Experiments

The scaling methods described in section 3 have been applied to test corpora from English (Wall Street Journal, WSJ) and German (Neue Zürcher Zeitung, NZZ). The scaled log λ was calculated for all pairs of the form (word, •). The test corpora were annotated in the following fashion: If the value was higher than 1, the tag <A> was assigned to the pair. All other candidates were tagged as <S>.[8] The automatically classified corpora were compared with their hand-tagged references.

(14) Annotation for test corpora

| *Tag* | *Interpretation* |
|---|---|
| <S> | End-of-Sentence |
| <A> | Abbreviation |
| <A><S> | Abbreviation at end of sentence |

We have chosen two different types of test corpora: First, we have used two test corpora of an approximate size of 2 and 6 MB, respectively. The WSJ corpus contained 19,776 candidates of the form (word, •); the NZZ corpus contained 37,986 such pairs. Second, we have tried to determine the sensitivity of the present approach to data sparseness. Hence, the approach was applied to ten individual articles from each WSJ and NZZ. For English, these articles contained between 7 and 26 candidate pairs, for German the articles comprised between 16 and 52 pairs.
The reference annotation allowed the determination of a baseline which determines the percentage of correctly classified end-of-sentence marks if each pair *(word, •)* is classified as an end-of-sentence mark.[9] The baseline varies from corpus to corpus, depending on a variety of factors (cf. Palmer/Hearst 1997). In the following tables, we have reported two measures: first, the *error rate*, which is defined in (15), and second, the *F measure* (cf. van Rijsbergen 1979:174), which is a weighted measure of *precision and recall*, as defined in (16).[10]

(15) *Error rate*[11]

$$\frac{C(<A> \rightarrow <S>) + C(<S> \rightarrow <A>)}{C(all\ candidates)}$$

(16) *F measure:* $\frac{2PR}{(R+P)}$

### 4.1 Results of first experiment

The results of the classification process for the larger files are reported in table (17). *F(B)* and *F(S)* are the F measure of the baseline, and the present approach, respectively. *E(B)* is the error rate of the baseline, and *E(S)* is the error rate of the scaled log λ approach.

(17) *Results of classification for large files*

| | *F(B)* | *F(S)* | *E(B)* | *E(S)* |
|---|---|---|---|---|
| WSJ | 81.11 | 99.57 | 31.78 | 0.59 |
| NZZ | 95.05 | 99.71 | 9.44 | 0.29 |

As (17) shows, the application of the scaled log λ leads to significant improvements for both files. In particular, the error rate has dropped from over 30 % to 0.6 % in the WSJ corpus. For both files, the accuracy is beyond 99 %.

### 4.2 Results of second experiment

The results of the second experiment are reported in table (18) for the articles from the *Wall Street Journal*, and in table (19) for the articles from *the Neue Zürcher Zeitung.* The scaled log λ approach generally outperforms the baseline approach. This is reflected in the F measure as well as in the error rate, which is reduced to a third. For one article (WSJ_1) the present approach actually performs below the baseline (cf. section 5).

---

[8] A tokenizer should treat pairs which have been annotated with <A> as single tokens, while tokens which have been annotated with <S> should be treated as two separate tokens. Three-dot-ellipses are currently not considered. Also <A><S> tags are not considered in the experiments (cf. section 5).

[9] Following this baseline, we assume that correctly classified end-of-sentence marks count as *true positives* in the evaluations.

[10] Manning/Schütze (1999:269) criticize the use of a*ccuracy* and *error* if the number of *true negatives* – *C(<A> → <A>)* in the present case – is large. Since the number of true negatives is small here, *accuracy* and *error* escape this criticism.

[11] *C(<X> → <Y>)* is the number of X which have been wrongly classified as Y. In (16), P stands for the *precision,* and R for the *recall.*

(18) *Results of classification for single articles from WSJ*

| | *F(B)* | *F(S)* | *E(B)* | *E(S)* |
|---|---|---|---|---|
| WSJ_1 | 88.00 | 77.78 | 21.43 | 28.57 |
| WSJ_2 | 83.87 | 100.00 | 27.78 | 0.00 |
| WSJ_3 | 100.00 | 100.00 | 0.00 | 0.00 |
| WSJ_4 | 81.82 | 97.30 | 30.77 | 3.85 |
| WSJ_5 | 66.67 | 85.71 | 50.00 | 16.67 |
| WSJ_6 | 89.66 | 96.30 | 18.75 | 6.25 |
| WSJ_7 | 100.00 | 100.00 | 0.00 | 0.00 |
| WSJ_8 | 88.00 | 90.00 | 21.43 | 14.29 |
| WSJ_9 | 47.06 | 72.73 | 69.23 | 23.08 |
| WSJ_10 | 83.33 | 100.00 | 28.57 | 0.00 |
| $\mu$ | 82.84 | 91.98 | 26.80 | 9.27 |

(19) *Results of classification for single articles from NZZ*

| | *F(B)* | *F(S)* | *E(B)* | *E(S)* |
|---|---|---|---|---|
| NZZ_1 | 95.08 | 100.00 | 9.38 | 0.00 |
| NZZ_2 | 93.02 | 97.56 | 13.04 | 4.35 |
| NZZ_3 | 96.00 | 98.97 | 7.69 | 1.92 |
| NZZ_4 | 96.15 | 100.00 | 7.41 | 0.00 |
| NZZ_5 | 93.18 | 98.80 | 12.77 | 2.13 |
| NZZ_6 | 96.84 | 98.92 | 6.12 | 2.04 |
| NZZ_7 | 97.50 | 97.37 | 4.88 | 4.88 |
| NZZ_8 | 89.66 | 100.00 | 18.75 | 0.00 |
| NZZ_9 | 96.97 | 97.14 | 5.88 | 2.86 |
| NZZ_10 | 93.94 | 99.71 | 11.43 | 0.29 |
| $\mu$ | 94.83 | 98.18 | 9.73 | 1.82 |

In general, the articles from NZZ contained fewer abbreviations, which is reflected in the comparatively high baseline scores. Still, the present approach is able to outperform the baseline approach. Particularly noteworthy are the articles NZZ_1, NZZ_4, and NZZ_8, where the error rate is reduced to 0. In general, the error rate has been reduced to a fifth.

## 5 Weaknesses and future steps

We have noted in section 2 that the scaling factors do not lead to a perfect classification. This is particularly reflected in the application of S(log $\lambda$) to WSJ_1 and NZZ_7, which actually show the same problem: In the training corpus, *ounces* was always followed by •. In WSJ_1, the word *said* was always followed by •, and this also happened in NZZ_7 for *kann*. Without the inclusion of additional metrics, non-abbreviations which exclusively occur at the end of sentences are wrongly classified. The table in (20) illustrates, however, that the error rate for false negatives drops significantly if plausible corpus sizes are considered.

(20) *False negatives (f.n.) and corpus size*

| | <S> | f.n. = <S> → <A> | Error % |
|---|---|---|---|
| NZZ | 34,400 | 81 | 0.23 |
| WSJ | 13,492 | 56 | 0.41 |
| NZZ_7 | 39 | 2 | 5.12 |
| WSJ_1 | 11 | 4 | 36.36 |

We have also ignored abbreviation occuring at the end of the sentence. The next step will be to integrate methods for the detection of abbreviations at the end of the sentence, e.g. by integrating additional phonotactic information, and also to cover the problematic cases reported above.

## Conclusion

We have presented an accurate and comparatively simple method for the detection of abbreviations which makes use of scaled log likelihood ratios. Experiments have shown that the method works well with large files and also with small samples with sparse data. We expect further improvements once additional classification schemata have been integrated.

## References


Dunning T. (1993) *Accurate methods for the statistics of surprise and coincidence.* Computational Linguistics, 19/1, pp. 61—74.

Evert S., U. Heid and W. Lezius (2000) *Methoden zum qualitativen Vergleich von Signifikanzmaßen zur Kollokationsidentifikation.* ITG Fachbericht 161, pp. 215—220.

Grefenstette G. (1999) *Tokenization.* "Syntactic Wordclass Tagging", H. van Halteren, ed., Kluwer Academic Publishers, pp. 117—133.

Liberman M.Y. and K.W. Church (1992) *Text analysis and word pronunciation in text-to-speech synthesis.* In "Advances in Speech Signal Processing", S. Furui & M.M. Sondhi, ed., M. Dekker Inc., pp. 791—831.

Manning, C.D. and H. Schütze (1999) *Foundations of statistical natural language processing.* The MIT Press, Cambridge/London.

Palmer D.D. and M.A. Hearst (1997) *Adaptive multilingual sentence boundary disambiguation.* Computational Linguistics, 23/3, pp. 241—267.

van Rijsbergen C.J. (1979) *Information Retrieval*. Butterworths, London.