Signifikanzmaße bei der Beurteilung von Kookkurrenzen

In der Statistik wird unter Signifikanz eine Kennzahl verstanden, welche die Wahrscheinlichkeit eines systematischen Zusammenhangs zwischen Variablen, im Falle von Textanalysen also zwischen Teiltexten (z.B. Wörtern), bezeichnet. Die Signifikanz drückt aus, ob ein scheinbarer Zusammenhang rein zufälliger Natur sein könnte oder mit hoher Wahrscheinlichkeit tatsächlich vorliegt.

Zur Berechnung werden abhängig vom Untersuchungsgegenstand unterschiedliche Formeln herangezogen, welche in erster Linie aus der Computerlinguistik stammen. Die Signifikanzmaße sollen dabei helfen, wichtige von unwichtigen Kookkurrenzen zu trennen. Dabei werden statistische Kenngrößen, wie Korpusgröße, Häufigkeit der einzelnen Wörter oder Frequenz des gemeinsamen Auftretens, ins Verhältnis gesetzt.

Eines der einfachsten Signifikanzmaße ist eine frequenzsortierte Kookkurrenzliste, also die Häufigkeit des gemeinsamen Auftretens zweier Worte im Gesamtkorpus. Ein Nachteil frequenzsortierter Listen ist, dass nach dem Zipf'schen Gesetz, dem Beginn der quantitativen Linguistik, sehr viele Wörter sehr selten auftreten. Demzufolge lassen sich mit einem Schwellenwert größer 1, also dem mehrmaligen gemeinsamen Auftreten eines Wortpaares, etwa zwei Drittel der Kookkurrenzen herausfiltern. Berechnet von den eAQUA-Tools sieht dies für ausgewählte Korpora wie folgt aus:

Korpus	Anzahl Kookkurrenzen	Kookkurrenzen freq = 1	in Prozent
BTL [1]	137.486.214	110,876,836	80,65
MPL [2]	580.247.568	398.935.822	68,75
Perseus Shakespeare [3]	6.746.602	5.027.170	74,51
TLG [4]	355.021.014	258.961.566	72,94

Wie aus der kleinen Übersicht zu erkennen ist, sind ein Großteil der gefundenen Kookkurrenzen eher als niedrigfrequent zu bezeichnen. Um daraus die wichtigen zu filtern, sind Berechnungsmethoden erforderlich, von denen hier einige vorgestellt werden.

Dice

Beim Dice-Koeffizienten (auch Sørensen-Dice-Koeffizent, benannt nach den Botanikern Thorvald Sørensen und Lee Raymond Dice) wird die Ähnlichkeit zweier Terme mittels einer Zahl zwischen 0 und 1 angegeben. Berechnungsgrundlage sind sogenannte N-Gramme. Bei N-Grammen wird ein Term bzw. ein Text in gleich große Teile zerlegt. Diese Fragmente können Buchstaben, Phoneme, ganze Wörter oder ähnliches sein.

Ermittelt wird die Anzahl der N-Gramme, die in beiden Termen vorhanden sind, um diese ins Verhältnis zur Gesamtzahl der N-Gramme zu setzen. Berechnet wird nach der Formel $dice_{ab} = \frac{2 \times n_{ab}}{n_a + n_b}$ wobei n_{ab} die Schnittmenge beider Terme und n_a bzw. n_b die Anzahl der gebildeten N-Gramme pro Term angibt.

Beispiel 1: **Ausdruck a = Tür** **Ausdruck b = Tor**	$dice_{ab} = \frac{2 \times n_{ab}}{n_a + n_b}$
Bigramm	Trigramm
a = { §T, Tü, ür, r§ } b = { §T, To, or, r§ }	a = { §§T, §Tü, Tür, ür§, r§§ } b = { §§T, §To, Tor, or§, r§§ }

Beispiel 2 **Ausdruck a = Spiegel** **Ausdruck b = Spargel**	$dice_{ab} = \frac{2 \times n_{ab}}{n_a + n_b}$
Bigramm	Trigramm
a = { §S, Sp, pi, ie, eg, ge, el, l§ } b = { §S, Sp, pa, ar, rg, ge, el, l§ }	a = { §§S, §Sp, Spi, pie, ieg, ege, gel, el§, l§§ } b = { §§S, §Sp, Spa, par, arg, rge, gel, el§, l§§ }

$d_{Spiegel,Spargel} = \frac{2 \times 5}{8+8} = \frac{10}{16} = 0,625$ $d_{Spiegel,Spargel} = \frac{2 \times 5}{9+9} = \frac{10}{18} \approx 0,556$

Bei der Bewertung von Kokkurrenzen kann der Dice-Koeffizient genutzt werden, indem die Häufigkeiten (Frequenzen) der Wörter ins Verhältnis gesetzt werden. n_a und n_b sind dabei die Frequenzen der Terme, n_{ab} die Anzahl des gemeinsamen Auftretens.

Aus der oben angeführten Berechnung ergeben sich relativ einfache Bewertungsmaßstäbe. Je frequenter die beiden Begriffe gemeinsam benutzt werden, um so mehr nähert sich der Wert 1. Treten beide Begriffe nur gemeinsam auf, wird die höchste Signifikanz mit 1 erreicht. Wie oft diese Kookkurrenz im Korpus zu finden ist, spielt dabei keine Rolle. Daraus ergibt sich eine wichtige Eigenschaft des Dice-Koeffizienten: Kookkurrenzen, die selten zusammen auftreten, bei denen ein Wort hoch- und das andere niedrigfrequent sind, werden als unsignifikant bewertet.

Jaccard

Beim Jaccard-Koeffizienten (nach dem Botaniker Paul Jaccard) wird die Ähnlichkeit zweier Terme mittels einer Zahl zwischen 0 und 1 angegeben. Berechnungsgrundlage bei Textmining-Verfahren sind sogenannte N-Gramme. Bei N-Grammen wird ein Term bzw. ein Text in gleich große Teile zerlegt. Diese Fragmente können Buchstaben, Phoneme, ganze Wörter oder ähnliches sein.

Ermittelt wird die Anzahl der N-Gramme, die in beiden Termen vorhanden sind, um diese ins Verhältnis zur Gesamtzahl der N-Gramme zu setzen. Berechnet wird nach der Formel $jaccard_{ab} = \frac{n_{ab}}{n_a + n_b - n_{ab}}$ wobei n_{ab} die Schnittmenge beider Terme und n_a bzw. n_b die Anzahl der gebildeten N-Gramme pro Term angibt.

Beispiel 1: **Ausdruck a = Tür** **Ausdruck b = Tor**	$jaccard_{ab} = \frac{n_{ab}}{n_a + n_b - n_{ab}}$
Bigramm	Trigramm
a = { §T, Tü, ür, r§ } b = { §T, To, or, r§ }	a = { §§T, §Tü, Tür, ür§, r§§ } b = { §§T, §To, Tor, or§, r§§ }

Beispiel 2 **Ausdruck a = Spiegel** **Ausdruck b = Spargel**	$jaccard_{ab} = \frac{n_{ab}}{n_a + n_b - n_{ab}}$
Bigramm	Trigramm
a = { §S, Sp, pi, ie, eg, ge, el, l§ } b = { §S, Sp, pa, ar, rg, ge, el, l§ }	a = { §§S, §Sp, Spi, pie, ieg, ege, gel, el§, l§§ } b = { §§S, §Sp, Spa, par, arg, rge, gel, el§, l§§ }

$d_{\text{Bigramm}} = \frac{5}{11} \approx 0,455$ $d_{\text{Trigramm}} = \frac{5}{13} \approx 0,385$

Für die Bewertung von Kokkurrenzen gilt beim Jaccard-Koeffizienten ähnliches wie beim Dice-Koeffizienten. Beide berechnen den Signifikanzwert ähnlich, die relative Ordnung der Kookkurrenzen bleibt gleich, nur der absolute Signifikanzwert unterscheidet sich marginal. Eine Modell-Berechnung mit mittlerer Frequenz von 100 sieht wie folgt aus:

n_a	n_b	n_{ab}	**Dice**	**Jaccard**
100	100	1	0,01	0,005
100	100	10	0,1	0,05
100	100	50	0,5	0,33
100	100	90	0,9	0,82
100	100	100	1	1

Poisson-Maß

Ein Ansatz zur Berechnung von signifikanten Kookkurrenzen basiert auf der Poisson-Verteilung (benannt nach dem Mathematiker Siméon Denis Poisson), einer diskreten Wahrscheinlichkeitsverteilung $p\left(n,k\right) = \frac{1}{k!}\gamma^k e^{-\gamma}$

Auf der Basis der Poisson-Verteilung geben *Quasthoff / Wolff* [5] das Poisson-Maß mit der Formel

$p\left(n_a,n_b,k,n\right)=\frac{k\times\left(\log k-\log\gamma-1\right)}{\log n_{ab}}$ an, welche beispielsweise für die Berechnung von Korpora im Wortschatz-Portal genutzt wurde, und in der die zwei Faktoren **n** (Anzahl der Sätze im Korpus) und **k** (Häufigkeit des gemeinsamen Auftretens auch n_{ab} bezeichnet) maßgeblich sind.

Nach einer Umstellung und der Grundannahme $\gamma=\frac{n_a\times n_b}{n}$ ergibt sich folgende Berechnung

$$p=\frac{n_{ab}\times\log\frac{n_{ab}\times n}{n_a\times n_b}-n_{ab}}{\log n}$$

Somit ließe sich das Poisson-Maß auf die Differenz zwischen Local Mutual Information und Frequenz reduzieren.

Log-Likelihood-Maß

Eine der populärsten Signifikanzmaße bei der Analyse großer Textcorpora ist nach *Dunning* [6] das Log-Likelihood-Maß, welches basiert auf der Binomialverteilung, eine der wichtigsten diskreten Wahrscheinlichkeitsverteilungen.

$$p\left(K=k\right)=p^k\left(1-p\right)^{n-k}\binom{n}{k}$$

Dunning kommt schließlich zur Berechnung von **log likelihood** zu der Formel:

unter der Voraussetzung $-2\log\lambda=2\left[\log L\left(p_1,k_1,n_1\right)+\log L\left(p_2,k_2,n_2\right)-\log L\left(p_1,k_1,n_1\right)-\log L\left(p_2,k_2,n_2\right)\right]$

$\log L\left(p,n,k\right)=k\log p+\left(n-k\right)\log\left(1-p\right)$

Das Log-Likelihood-Maß kann demzufolge abgeleitet werden

$lgl=2\left[n\log n-n_a\log n_a-n_b\log n_b+n_{ab}\log n_{ab}+\left(n-n_a-n_b+n_{ab}\right)\log\left(n-n_a-n_b+n_{ab}\right)+\left(n_a-n_{ab}\right)\log\left(n_a-n_{ab}\right)+\left(n_b-n_{ab}\right)\log\left(n_b-n_{ab}\right)-\left(n-n_a\right)\log\left(n-n_a\right)-\left(n-n_b\right)\log\left(n-n_b\right)\right]$

Charakteristisch für das Log-Likelihood-Maß ist, im Gegensatz beispielsweise zum Poisson-Maß, die Gleichbehandlung von signifikant häufigen und signifikant seltenen Ereignissen. So finden sich in den Digitalisaten vom TLG in der Version TLG-E bei rund 73,8 Millionen Wörtern etwa 1,3 Millionen Kookkurrenzen, die nur einmal auftreten und trotzdem mit einem lgl-Wert von 30 und ein wenig mehr belegt sind. Einen ähnlich großen Wert von 34,553 haben zum Beispiel **καὶ** und **Τὸ**, die zusammen 14311 Mal gezählt wurden.

---

1) Bibliotheca Teubneriana Latina, Online-Version, Stand vom Februar 2014

2) Patrologia Latina Database, CD-ROM Version, November 1995c

3) William Shakespeare in Perseus Digital Library, Renaissance Materials, Stand vom Mai 2013

4) TLG-E, CD-ROM Version aus dem Jahre 1999

5) [Quasthoff 02]. Uwe QUASTHOFF, Christian WOLFF. The Poisson Collocation Measure and its Applications. In Second International Workshop on Computational Approaches to Collocations, 2002.

6) [Dunning 93]. Dunning, T. „Accurate Methods for the Statistics of Surprise and Coincidenc“. In: Computational Linguistics 19, 1 (1993), 61-74.

From:
http://www.eaqua.net/doku/ - **Wissensdatenbank**

Permanent link:
http://www.eaqua.net/doku/doku.php/signifikanz?rev=1400571064

Last update: **2021/03/31 12:49**