A Lexicographer-Friendly Association Score

Pavel Rychlý

Faculty of Informatics, Masaryk University
Botanická 68a, 602 00 Brno, Czech Republic
pary@fi.muni.cz

Abstract. Finding collocation candidates is one of the most important and widely used feature of corpus linguistics tools. There are many statistical association measures used to identify good collocations. Most of these measures define a formula of a association score which indicates amount of statistical association between two words. The score is computed for all possible word pairs and the word pairs with the highest score are presented as collocation candidates. The same scores are used in many other algorithms in corpus linguistics.
The score values are usually meaningless and corpus specific, they cannot be used to compare words (or word pairs) of different corpora. But end-users want an interpretation of such scores and want a score's stability. This paper present a modification of a well known association score which has a reasonable interpretation and other good features.

1 Introduction

Finding collocation candidates is one of the most important and widely used feature of corpus linguistics tools [1]. There are many statistical association measures used to identify good collocations. Most of these measures define a formula of a association score which indicates amount of statistical association between two words. The score is computed for all possible word pairs and the word pairs with the highest score are presented as collocation candidates. The same scores are used in many other algorithms in corpus linguistics, for example to compute collocations in grammatical relations and an importance of grammatical relations in the Sketch Engine [2].

There are two general problems of most association scores:

1. A score is fine-tuned to one particular corpus size and/or key word frequency. If we use a score for a corpus with very different number of tokens the resulting list is not satisfying enough or is completely wrong.
2. The score values are usually meaningless and corpus specific, they cannot be used to compare words (or word pairs) of different corpora. But end-users want an interpretation of such scores and want a score's stability. They want to compare collocation scores of different words and on different corpora or subcorpora.

Petr Sojka, Aleš Horák (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2008, pp. 6–9, 2008.

The article is organized as follows. The following section describe notation and the most widely used association scores. The Section 3 illustrates these two problems on real examples. The next section defines a new score *logDice*, which is a modification of the well known association score *Dice* [3]. The *logDice* score has a reasonable interpretation, scales well on a different corpus size, is stable on subcorpora, and the values are in reasonable range.

2 Association Scores for Collocations

Almost all association score formulas use frequency characteristics from a contingency table, which records the relationship between two words (W_1, W_2). Table 1 shows an example of a contingency table. The numbers in the right-hand column and the bottom row are called marginal frequencies and the number in the bottom right-hand corner is the size of the corpus.

In the rest of this paper we will uses the following symbols (the meaning is also summarized in Table 1):

- f_x = number of occurrences of word X
- f_y = number of occurrences of word Y
- f_{xy} = number of co-occurrences of words X and Y
- $R_x = \frac{f_{xy}}{f_x}$ = relative frequency of word X
- $R_y = \frac{f_{xy}}{f_y}$ = relative frequency of word Y

Table 1. Notation of frequencies of words X and Y

	$W_1 = X$	$W_1 \neq X$	
$W_2 = Y$	f_{xy}	$f_y - f_{xy}$	f_y
$W_2 \neq Y$	$f_x - f_{xy}$	$N - f_{xy}$	$N - f_y$
	f_x	$N - f_x$	N

3 Widely Used Association Scores

This section summarize formulas of some association scores and gives its main characteristics. More scores, motivations, discussion of their mathematical background and full references could be find in [4].

T-score: $\frac{f_{xy} - \frac{f_y f_B}{N}}{\sqrt{f_{xy}}}$

MI-score: $\log_2 \frac{f_{xy} N}{f_x f_y}$

MI^3-score: $\log_2 \frac{f_{xy}^3 N}{f_x f_y}$

Minimum Sensitivity: $\min R_x, R_y$

Dice coefficient: $D = \frac{2f_{xy}}{f_x+f_y}$

MI log Freq: *MI-score* $\times \log f_{xy}$, used as salience in the first version of Word Sketches [2].

Table 2 lists the collocation candidates on lemmas to the verb *break* in the window from 5 tokens to the left to 5 tokens to the right. They were computed on the British National Corpus by the Manatee system [5].

Table 2. Collocation lists for different association scores

	F_{xy}	**T-score**
the	11781	99.223
.	8545	83.897
,	8020	80.169
be	6122	69.439
and	5183	65.918
to	5131	65.798
a	3404	52.214
of	3382	49.851
down	2472	49.412
have	2813	48.891
in	2807	47.157
it	2215	43.314
into	1856	42.469
he	1811	39.434
up	1584	39.038

	F_{xy}	**MI-score**
spell-wall	5	11.698
deadlock	84	10.559
hoodoo	3	10.430
scapulum	3	10.324
Yasa	7	10.266
intervenien	4	10.224
preparedness	21	10.183
stranglehold	18	10.177
logjam	3	10.131
irretrievably	12	10.043
Andernesse	3	10.043
irreparably	4	10.022
Thief	37	9.994
THIEf	4	9.902
non-work	3	9.809

	F_{xy}	**MI^3-score**
the	11781	30.591
down	2472	29.882
.	8545	29.558
,	8020	29.193
be	6122	28.311
to	5131	28.268
and	5183	28.246
into	1856	27.854
up	1584	26.967
a	3404	26.717
have	2813	26.593
of	3382	26.255
in	2807	26.095
it	2215	25.876
out	1141	25.821

	F_{xy}	**Min. Sens.**
down	2472	0.027
silence	327	0.018
leg	304	0.016
law	437	0.014
heart	259	0.014
rule	292	0.013
off	869	0.013
news	236	0.013
into	1856	0.012
barrier	207	0.011
away from	202	0.011
war	294	0.010
ground	182	0.010
record	287	0.010
neck	180	0.010

	F_{xy}	**MI log Freq**
down	2472	57.340
silence	327	48.589
deadlock	84	46.909
barrier	207	46.389
into	1856	46.197
off	869	42.411
up	1584	42.060
leg	304	41.980
neck	180	39.336
law	437	38.805
out	1141	38.783
bone	151	38.263
heart	259	37.327
Thief	37	36.353
news	236	36.296

	F_{xy}	**Dice**
down	2472	0.0449
silence	327	0.0267
into	1856	0.0210
leg	304	0.0203
off	869	0.0201
barrier	207	0.0191
law	437	0.0174
up	1584	0.0158
heart	259	0.0155
neck	180	0.0148
news	236	0.0144
rule	292	0.0142
out	1141	0.0135
away from	202	0.0135
bone	151	0.0130

4 logDice

As one can see from the previous section, *Dice* score gives very good results of collocation candidates. The only problem is that the values of the *Dice* score are usually very small numbers. We have defined *logDice* to fix this problem.

$$logDice = 14 + \log_2 D = 14 + \log_2 \frac{2f_{xy}}{f_x + f_y}$$

Values of the *logDice* have the following features:

- Theoretical maximum is 14, in case when all occurrences of X co-occur with Y and all occurrences of Y co-occur with X. Usually the value is less then 10.
- Value 0 means there is less than 1 co-occurrence of XY per 16,000 X or 16,000 Y. We can say that negative values means there is no statistical significance of XY collocation.
- Comparing two scores, plus 1 point means twice as often collocation, plus 7 points means roughly 100 times frequent collocation.
- The score does not depend on the total size of a corpus. The score combine relative frequencies of XY in relation to X and Y.

All these characteristics are useful orientation points for any field linguist working with collocation candidate lists.

5 Conclusion

In this paper, we have presented the new association score *logDice*. The *logDice* score has a reasonable interpretation, scales well on a different corpus size, is stable on subcorpora, and the values are in reasonable range.

Acknowledgments. This work has been partly supported by the Academy of Sciences of Czech Republic under the projects 1ET200610406, 1ET100300419 and by the Ministry of Education of CR within the Centre of basic research LC536 and National Research Programme 2C06009.

References

1. Smadja, F.: Retrieving Collocations from Text: Xtract. Computational Linguistics **19**(1) (1994) 143–177.
2. Kilgarriff, A., Rychlý, P., Smrž, P., Tugwell, D.: The Sketch Engine. Proceedings of Euralex (2004) 105–116.
3. Dice, L.: Measures of the Amount of Ecologic Association Between Species. Ecology **26**(3) (1945) 297–302.
4. Evert, S.: The Statistics of Word Cooccurrences: Word Pairs and Collocations. Unpublished Ph.D. dissertation, University of Stuttgart (2004).
5. Rychlý, P.: Manatee/Bonito – A Modular Corpus Manager. In: P. Sojka, A. Horák (Eds.): RASLAN 2007 Proceedings (2007), pp. 65–70.