Comparative Evaluation of Collocation Extraction Metrics

Aristomenis Thanopoulos, Nikos Fakotakis, George Kokkinakis

Wire Communications Laboratory
Electrical & Computer Engineering Dept., University of Patras
265 00 Rion, Patras, Greece
{aristom,fakotaki,gkokkin}@wcl.ee.upatras.gr

Abstract

Corpus-based automatic extraction of collocations is typically carried out employing some statistic indicating concurrency in order to identify words that co-occur more often than expected by chance. In this paper we are concerned with some typical measures such as the t-score, Pearson's χ-square test, log-likelihood ratio, pointwise mutual information and a novel information theoretic measure, namely *mutual dependency*. Apart from some theoretical discussion about their correlation, we perform comparative evaluation experiments judging performance by their ability to identify lexically associated bigrams. We use two different gold standards: WordNet and lists of named-entities. Besides discovering that a frequency-biased version of mutual dependency performs the best, followed close by likelihood ratio, we point out some implications that usage of available electronic dictionaries such as the WordNet for evaluation of collocation extraction encompasses.

1. Introduction

Collocational information is important not only for second language learning but also for many natural language processing tasks. Specifically, in natural language generation and machine translation it is necessary to ensure generation of lexically correct expressions; for example, "strong", unlike "powerful", modifies "coffee" but not "computers". From the other hand, in automatic construction of thesauri and ontologies, identification of multiwords representing characteristic entities and concepts of the domain, such as "General Motors" or "general relativity", is an obvious necessity. These two types of collocations are quite different in terms of both compositionality and offset (signed distance between constituents in text). In this paper we are concerned with the latter type of collocations; that is multiwords the meaning of which is not compositionally derivable.

Collocations are abundant in language and vary significantly in terms of length, syntactic patterns and offset. They are also domain-dependent and language-dependent; therefore their automatic extraction from domain specific corpora is of high importance, as far as portability of NLP systems is concerned. Therefore, in this paper we will focus on purely corpus-based automatic extraction of collocations, assuming that other available knowledge sources, such as in (Justeson and Katz, 1995) or in (Pearce, 2001), can be employed additionally.

2. Corpus-based Collocation Extraction

Collocations are recurrent in texts and express lexical selectivity. Therefore two or more words that co-occur in text corpora (much) more often than expected by chance (most) possibly constitute a collocation. In order to test this hypothesis of dependence, several metrics have been adopted by the corpus linguistics community. Typical statistics are the t-score (TSC), Pearson's χ-square test (χ^2), log-likelihood ratio (LLR) and pointwise mutual information (PMI). However, usually no systematic comparative evaluation is accomplished. For example, Kita et al. (1993) accomplish partial and intuitive comparisons between metrics, while Smadja (1993) resorts to lexicographic evaluation of his approach. However, it is obvious that an objective and automated evaluation process offers repeatability, allowing for comparison among different collocation extraction measures and techniques.

The aforementioned metrics estimate lexical selectivity on bigrams. Smadja (1993) proposes an algorithm for joining significant bigrams together to construct significant n-grams. However, since his target task is NLG and therefore he is interested in collocations not necessarily non-compositional, he keeps the larger produced n-grams, while conceptually-oriented tasks such as automatic construction of ontologies and thesauri, require identification of the minimal semantic constituents. For example, "the president of United States" is a collocation but it can be decomposed in two minimal lexico-semantic units of which the straightforward combination produces its meaning: "president" and "United States". However, extraction of significant bigrams is an important task for both applications. In order to eliminate the factor of offset and exploit available recourses for evaluation we focus on sequential collocations (multi-words).

An objective set of multi-words is necessary, as "gold standard", for the comparative experiments. Unfortunately, complete machine-readable databases of collocations are not widely available, even for English. On-line lexical resources, such as WordNet (Miller, 1990), contain such information, although incomplete (Roark and Charniak, 1998) and not pure, as we will note. Obvious, easily obtained and explicitly non-compositional multi-words are also entity names, such as location and organisation names. Terminology lists can also be used in the case of restricted domain corpora.

3. Pairwise significance measures

Since multiwords can be quite long (e.g. "Default Proof Credit Card System Inc"), applying straightforwardly statistics on n-gram counts are computationally prohibitive. The only feasible strategy for extraction of multiwords from large corpora is to initially extract a set of significant bigrams (according to a measure of pairwise dependence) and then, connecting significant bigrams together, to calculate the statistical significance of the larger strings (Smadja, 1993). In order to simplify the task of comparison among measures of significance, the present study is confined to the first stage, i.e. the extraction of significant bigrams.

The most simplistic approach to collocation extraction is to exploit the property of recurrency and therefore to extract the most frequent word co-occurrences. However this has the obvious drawback that the *a priori* word frequencies are not taken into account; therefore such collocations are often fully compositional and thus of no lexical interest (e.g. "analyst said", "three years", etc.). Therefore, measures of dependence taking into account word probabilities (using maximum likelihood estimators) have been widely employed for collocation extraction. In the following sub-sections we define some of the most prominent measures (see (Manning and Schütze, 1999) for more details), we discuss some of their prominent characteristics and we propose a few more measures.

3.1. Information theoretic measures

In one of the premier studies in automatic corpus-based collocation extraction, Church and Hanks (1990) proposed the ***association ratio***, a metric based on the information theoretic concept of mutual information, and specifically to the pointwise mutual information (PMI), which is defined as:

$$I(\mathrm{w}_1, \mathrm{w}_2) = \log_2 \frac{\mathrm{P}(\mathrm{w}_1 \mathrm{w}_2)}{\mathrm{P}(\mathrm{w}_1) \cdot \mathrm{P}(\mathrm{w}_2)}$$

However, PMI is actually a measure of independence rather than of dependence (Manning and Schütze, 1999). Considering perfectly dependent bigrams, i.e. $P(w_1) = P(w_2) = P(w_1w_2)$, we obtain: $I(w_1,w_2) = -\log2(P(w_1))$, which shows that PMI exhibits preference to rare events, because they a priori contain higher amount of information, in comparison to frequent events. This suggests that dependence can actually be identified subtracting from PMI the information that the whole event bears, which is its self-information (Gallager, 1968): $I(x) = -\log(P(x))$. We call the proposed measure **Mutual Dependency** (MD):

$$\mathrm{D}(\mathrm{w}_1, \mathrm{w}_2) = \mathrm{I}(\mathrm{w}_1, \mathrm{w}_2) - \mathrm{I}(\mathrm{w}_1 \mathrm{w}_2) = \\ = \log_2 \frac{\mathrm{P}^2(\mathrm{w}_1 \mathrm{w}_2)}{\mathrm{P}(\mathrm{w}_1) \cdot \mathrm{P}(\mathrm{w}_2)}$$

which is obviously maximized for perfectly dependent bigrams, without dependence to their frequency. However, intuition suggests that a slight bias towards frequency can be beneficial reflecting statistical confidence; that is among similarly dependent bigrams the most frequent ones should be favored. Therefore we also tested a few such measures (for example combining MD with the t-score), of which Log-Frequency biased MD (LFMD) proved experimentally the most satisfactory:

$$D_{LF}(w_1w_2) = D(w_1,w_2) + \log_2 P(w_1w_2)$$

3.2. Hypothesis testing

From a statistical point of view, our problem can be expressed as to determine if the word co-occurrence indicates lexical correlation or it is due to chance. The latter case, which constitutes the *null hypothesis*, is that the considered words w_1 and w_2 are generated independently in the corpus and thus their co-occurrence probability can be estimated as: $P(w_1w_2) = P(w_1)\cdot P(w_2)$. Word probabilities are calculated using maximum likelihood estimators (MLE).

The most widely used statistical tests employed to estimate the divergence of the observed co-occurrence frequency from the one according to the null hypothesis are the *t-score* (TSC), the Pearson's *χ-square test* (XSQ) and the *log-likelihood ratio* (LLR). The null hypothesis can be rejected with a certain confidence when the used statistic surpasses a certain threshold.

3.2.1. T-score

The t statistic is defined as:

$$t = \frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{N}}}$$

where $\bar{x}$ is the sample mean, s^2 is the sample variance, N is the sample size and μ is the mean of the distribution which generation of words follows. If the t statistic is large enough the null hypothesis can be rejected. The t-test is typically calculated supposing a normal distribution. That is, for a bigram *uv* we have:
$\bar{x} = P(uv)$, $\mu = P(u)\cdot P(v)$ and
$s^2 = P(uv)\cdot(1-P(uv)) \approx P(uv)$.
In terms of frequency counts, we have:
$t(u,v)=1-f_u\cdot f_v/f_{uv}$

Then the relation between the t-score and mutual information can be easily derived:

$$t = \sqrt{f_{uv}}\,(1 - 2^{-I(u,v)})$$

This formula reveals that even loosely related bigrams (e.g. PMI ≈ 3) are ranked roughly according to their frequency; which indicates that the t-score is rather frequency biased.

3.2.2. Pearson's χ-square test

Pearson's *χ-square test*, bypasses the arbitrary assumption of normality employed for the calculation of the t-test. It sums the squared differences between the expected and observed frequencies scaled by the expected frequencies in all combinations of co-occurrence or not of the words under consideration:

$$\chi^2 = \frac{(f_{uv} - f_u f_v)^2}{f_u f_v} + \frac{(f_{u\bar{v}} - f_u f_{\bar{v}})^2}{f_u f_{\bar{v}}} + \frac{(f_{\bar{u}v} - f_{\bar{u}} f_v)^2}{f_{\bar{u}} f_v} + \frac{(f_{\overline{uv}} - f_{\bar{u}} f_{\bar{v}})^2}{f_{\bar{u}} f_{\bar{v}}}$$

where *uv* represents the sequence of events *u* and *v* and *ū* the event (word or bigram) *not(u)*.

In the case of statistical data from large corpora, frequency of occurrence is always many orders of magnitude higher than frequency of non-occurrence, and therefore the 3 latter additives are insignificant, compared to the former, which is actually a different formulation of the MD measure. Indeed, the ranking lists of the two measures are quite similar.

3.2.3. Likelihood ratio

A widely accepted measure of statistical significance is the logarithm of the ratio between the likelihoods of the hypotheses of dependence and independence (LLR):

$$-2\log\lambda = 2 \cdot \log\frac{\mathrm{L(H_1)}}{\mathrm{L(H_0)}}$$

where L(H) is the likelihood of hypothesis H based on observed data assuming that occurrence of words follows binomial distribution (Dunning, 1993), which is a more plausible assumption than the normality assumption.

4. Experimental evaluation

We conducted comparative evaluation experiments of the discussed measures using English language for the obvious reason of availability of both training text corpora and, mainly, lexical resources for automatic evaluation. Since the involved linguistic analysis is minimal we deem that our results and conclusions are valid for any language.

Statistical data were acquired from the Wall Street Journal corpus (WSJC) of 1988, comprised of about 11 million content words. Since Sentence Boundary Detection is, although not trivial, a rather solved NLP task and in WSJC sentence boundaries are annotated, we considered only intra-sentential co-occurrences. Moreover, we eliminate bigrams containing functional words or numbers and bigrams appearing less than 3 times in the corpus. Therefore our statistical data consist of roughly 4.4 million bigrams, that is 800000 distinct bigrams were considered.

The 30-best scoring bigrams for each measure are shown in Table 1. We can see that the t-score produces exactly the same hits (ranked slightly different) as plain frequency, some of which are compositional and uninteresting (e.g. "company said", "says mr"), while LFMD and LLR produce some interesting and important collocations (e.g. merrill lynch, los angeles, dow jones). MI and MD (or χ^2) 30-best lists contain exclusively named-entity bigrams (for all, it is $f_u = f_v = f_{uv}$); the former, unlike the latter, ranks higher the low frequency bigrams.

For the automatic evaluation we employed two completely unrelated gold standards. The one is WordNet (Miller, 1990); one of the most copious lexical resources in electronic form for English. WordNet contains semantic relations between lexical entities representing entities and concepts and therefore this set of lexical entities was used as a "gold standard". Since we study bigram dependency measures, we kept only lexical entities comprised of two words. We didn't extract bigrams from WordNet n-gram entities with n>2, because in many cases they are analytical descriptions of synsets or semantic categories (e.g. "capital_of_the_United_Kingdom", "ABO_blood_group_system") rather than lexical collocations.

The second source is comprised of named entities which appear in abundance in American newswire texts and were gathered mainly from the Internet for this specific purpose. Namely they are 5700 US cities (2000 of them periphrastic), 11000 US companies and 11000 person names which recurrently occur in newswire texts in the domains of politics, business, sports, arts, etc. In this case we maintained all extracted bigrams, excluding only words which are designators of the respective category (e.g. *city, corp, company, ltd, sir*, etc.) and therefore their contribution to the meaning of the phrase is compositional. In total there were 23000 distinct bigrams the 5400 of which occur in our corpus.

R	FREQUENCY	T-score	LLR	LFMD	MD, χ^2	PMI
1	vice president	vice president	vice president	vice president	bala cynwyd	leonie rysanek
2	stock exchange	stock exchange	stock exchange	wall street	zoete wedd	lineas aereas
3	chief executive	chief executive	chief executive	chief executive	ralston purina	yand renjun
4	~~year earlier~~	~~year earlier~~	~~cents share~~	~~cents share~~	bateman eichler	yue-kong pao
5	~~cents share~~	~~cents share~~	~~year earlier~~	stock exchange	corpus christi	fayez sarofim
6	york stock	york stock	executive officer	executive officer	gallium arsenide	steer-mom pop's
7	~~million shares~~	executive officer	wall street	~~year earlier~~	kaposi's sarcoma	tech-ops landauer
8	~~company said~~	composite trading	composite trading	dow jones	rotan mosle	sunder rajan
9	executive officer	million shares	york stock	los angeles	cahora bassa	tiang siew
10	composite trading	~~company said~~	~~net income~~	composite trading	mager dietz	tercel ez
11	~~spokesman said~~	~~spokesman said~~	exchange composite	real estate	vazquez rana	toa nenryo
12	wall street	wall street	interest rates	shearson lehman	ku klux	rabi blancos
13	interest rates	net income	real estate	hong kong	kwik kopy	rodrigo borja
14	net income	interest rates	dow jones	merrill lynch	nissho iwai	rubik's cube
15	exchange composite	exchange composite	~~shares outstanding~~	net income	deja vu	roussel uclaf
16	~~trading yesterday~~	~~trading yesterday~~	tender offer	exchange composite	fii fyffes	schiapparelli farmaceutici
17	common shares	common shares	~~years ago~~	interest rates	epeda bertrand	tu bishvat
18	~~shares outstanding~~	~~shares outstanding~~	los angeles	tender offer	revolucionario institucional	usinor sacilor
19	~~years ago~~	~~years ago~~	shearson lehman	burnham lambert	minas gerais	~~ils sont~~
20	~~inc said~~	~~inc said~~	~~million shares~~	~~shares outstanding~~	ds bancor	immanuel kant
21	~~corp said~~	tender offer	~~trading yesterday~~	~~years ago~~	yom kippur	jebel kusha
22	~~says mr~~	real estate	common shares	york stock	munoz ledo	jovito salonga
23	stock market	~~says mr~~	merrill lynch	leveraged buy-out	kumagai gumi	koninklijke nedlloyd
24	tender offer	stock market	hong kong	lehman hutton	aerolineas argentinas	josip broz
25	real estate	~~corp said~~	~~years old~~	morgan stanley	mats wilander	lanthanum gallate
26	holding company	holding company	~~spokesman said~~	san francisco	aer lingus	katsuya takanashi
27	~~shares closed~~	~~shares closed~~	fourth quarter	~~years old~~	khalifa al-sabah	~~sont partis~~
28	~~million year~~	dow jones	white house	white house	modus operandi	sotto voce
29	dow jones	~~years old~~	holding company	drexel burnham	dextran sulfate	hau pei-tsun
30	~~said mr~~	exchange commission	west german	seasonally adjusted	twyla tharp	harve benard

Table 1: The 30-best bigrams for all tested measures of bigram association. Bigrams lacking interesting (i.e. lexical) association are marked with a diagonal line, while lexically associated bigrams appearing in any of the gold standards are indicated with grey fill. Note that some named entity fragments (e.g. "ku klux" from "Ku Klux Klan" and "york stock" from "New York Stock Exchange") have not matched WordNet entities due to the exclusion of n-grams with n>2.

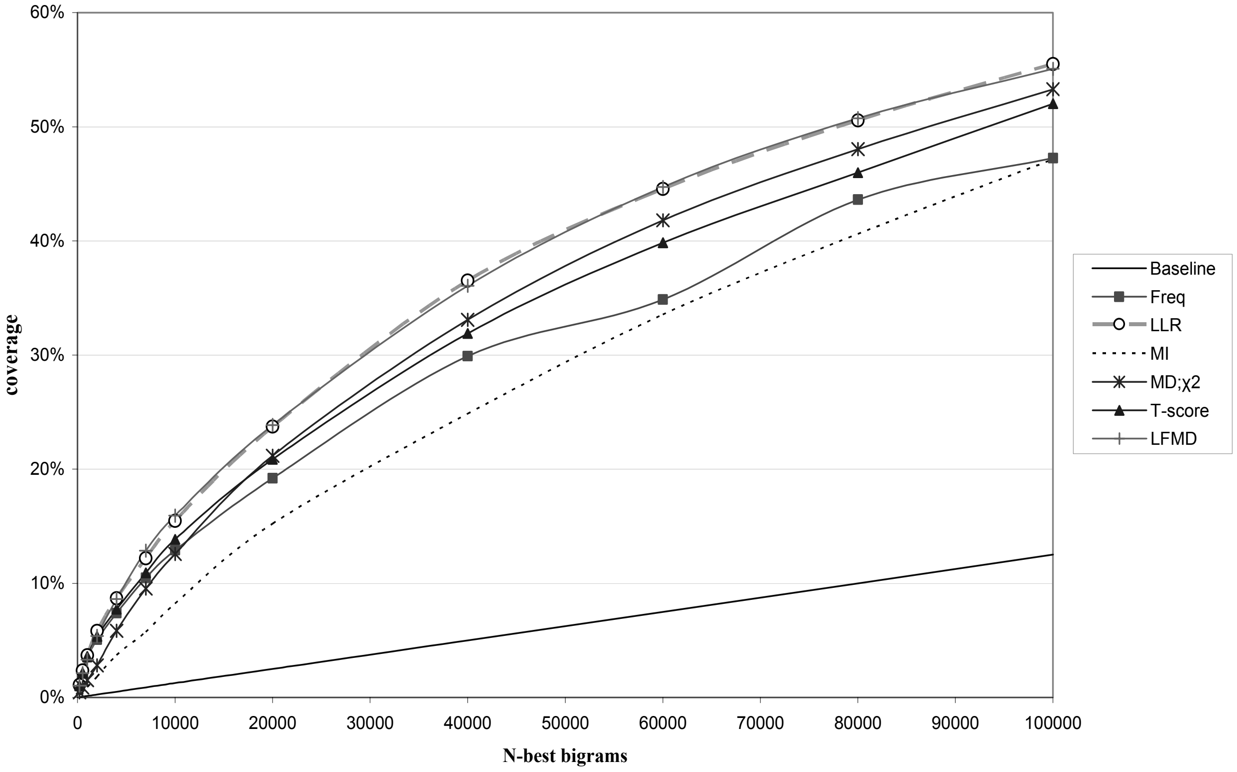

Figure 1: Comparative evaluation against WordNet entities.

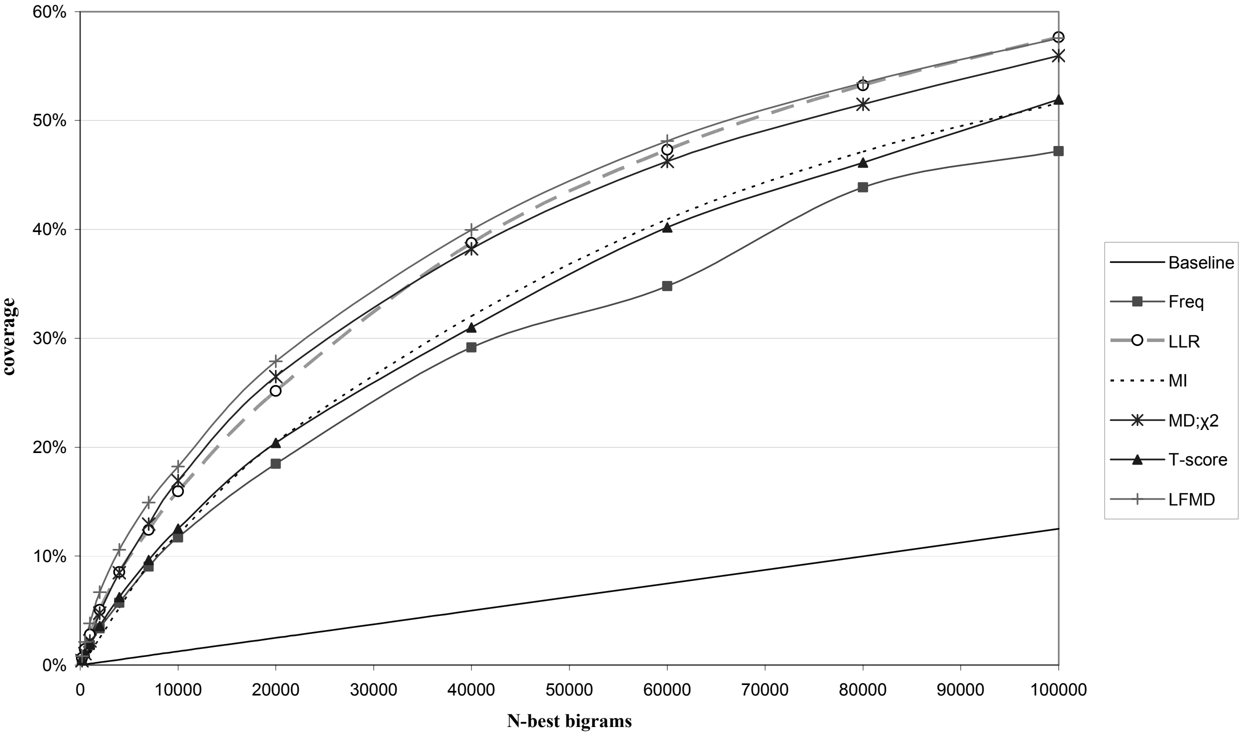

Figure 2: Comparative evaluation against Named Entities

Since for a certain level of confidence different statistics may provide different numbers of suggestions, resulting in different levels of coverage and precision, we follow a simple but objective evaluation scheme: We rank bigrams according to every measure and we depict the percentage of pairs retrieved correctly among the N-best candidates, with N ranging from 200 to 100000. As we noted in Section 3, the performance of Pearson's χ-square test is almost identical to that of Mutual Dependency. Baseline is drawn supposing random selection of bigrams.

The evaluation results show that LFMD and LLR outperform the other measures in both gold standard evaluations. It is interesting however that evaluation against WordNet favours not only the t-score but plain frequency as well, at the expense of information theoretic metrics. We argue that this demonstrates that employing straightforwardly WordNet as a "gold standard" multiword repository is not quite infallible for two reasons:

1. Many WordNet entities are rather analytical descriptions of lexical entities and therefore by no means non-compositional multiwords of interest. For example both "Japanese_capital" and "capital_of_Japan" are included in WordNet as well as many phrases of the same pattern. A preprocesing filter should probably be applied to exclude compositional expressions using the WordNet hypernym relations. Indeed, compositionality of "capital_of_Japan" can be easily identified as since there are numerous "capital_of_X" entities where X has the same hypernym as "Japan". Moreover, the specific level of the hierarchy the entity occurs can be of assistance towards the same end. That is, the lower the level, the more possible the WordNet word sequence to be a non-compositional lexical entity rather than a synset description.
2. No systematic attempt to include (or omit) all named entities in WordNet (as probably in most dictionaries) have been done. Therefore only the most frequent named entities are included; introducing a preference to frequency-biased measures of association.

5. Conclusion

We have studied certain pairwise word association measures, typically applied for collocation extraction. We have discussed some associations among them, both analytically and experimentally. Although length and offset of collocations varies significantly, our study was restricted on bigrams in order to eliminate other intervening factors and to exploit available lexical resources for evaluation.

In specific, the strong bias of the t-score towards frequency, makes it incapable to identify rare collocations, which comprise a large portion of terminology, as Zipf's law suggests. From the other hand, the inverse-frequency bias of pointwise mutual information can be corrected taking into consideration the self-information of the co-occurrence. Introducing a slight bias towards frequency results in a top-performing measure, surpassing even likelihood ratio.

The evaluation procedure needs also special attention. Available lexical resources such as WordNet are both impure and incomplete regarding non-compositional collocations. Therefore, enrichment with terminological information and elaborated preprocessing for the exclusion of descriptive expressions seem necessary prior to evaluation experiments.

6. References

Church K. and Hanks P., 1990. Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics*, 16/1: 22-29.

Dunning T., 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics,* 19/1: 61-74.

Gallager R., 1968. *Information theory and reliable communication.* John Wiley & Sons, Inc.

Justeson J., Katz S., 1995. Principled Disambiguation: Discriminating Adjective Senses with Modified Nouns. *Computational Linguistics,* 21/1: 1-27.

Kita K., Kato Y., Omoto T., Yano Y., 1993. "A Comparative Study of Automatic Extraction of Collocations from Corpora: Mutual Information vs. Cost Criteria. *Journal of Natural Language Processing* 1/1: 21-32.

Manning C. and Schütze H., 1999. *Foundations of Statistical Natural Language Processing*. Cambridge: MIT Press.

Miller G., 1990. WordNet: An on-line lexical database. *International Journal of Lexicography*, 34.

Pearce, D., 2001. Synonymy in collocation extraction. *Proceedings of the NAACL'01 Workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations.* Pittsburgh, PA.

Smadja F., 1993. Retrieving Collocations from text: Xtract, *Computational Linguistics*, 19/1: 143-177.

Roark B., Charniak E., 1998. Noun-phrase co-occurrence statistics for semi-automatic semantic lexicon construction. *COLING-ACL 1998*, 1:1110-1116.