Concept Bag: A New Method for Computing Concept Similarity in Biomedical Data

Richard L. Bradshaw, Ramkiran Gouripeddi, and Julio C. Facelli(✉)

Department of Biomedical Informatics,
University of Utah, Salt Lake City, UT, USA
{rick.bradshaw, ram.gouripeddi, julio.facelli}@utah.edu

Abstract. Biomedical data are a rich source of information and knowledge, not only for direct patient care, but also for secondary use in population health, clinical research, and translational research. Biomedical data are typically scattered across multiple systems and syntactic and semantic data integration is necessary to fully utilize the data's potential. This paper introduces new algorithms that were devised to support automatic and semi-automatic integration of semantically heterogeneous biomedical data. The new algorithms incorporate both data mining and biomedical informatics methods to create "concept bags" in the same way that "word bags" are used in data mining and text retrieval. The methods are highly configurable and were tested in five different ways on different types of biomedical data. The new methods performed well in computing similarity between medical terms and data elements - both critical for semi/automatic data integration operations.

Keywords: Concept bag · Semantic integration · Biomedical data · Data federation

1 Introduction

Biomedical data are a potentially rich source for information and knowledge discovery. Biomedical data collected for patient care or specific research protocols are valuable for reuse to address broader clinical, translational, comparative effectiveness (CER), population health, or public health research questions. However, reusing biomedical data for these purposes is not trivial. Data from multiple biomedical data systems are typically required to answer research questions and integrating these data is complex. Biomedical data are modeled and stored using various formats, syntaxes, and value sets to represent clinical or biomedical observations or facts about patients, research subjects or other artifacts. For instance, to answer a translational research study question, one would likely need demographic data from one system, diagnostic data from another system, and data from bioinformatics pipelines [1, 2].

Biomedical data integration generally requires homogenization of semantically and syntactically heterogeneous data. Semantic heterogeneity occurs when the same domain is modeled differently and data elements and values have different meanings [3], whereas syntactic heterogeneity occurs when the structural formats between data sets are different. OpenFurther [4], and i2b2/SHRINE (Informatics for Integrating Biology and

I. Rojas et al. (Eds.): IWBBIO 2019, LNBI 11466, pp. 15–23, 2019.
https://doi.org/10.1007/978-3-030-17935-9_2

the Bed-side/Shared Health Research Information Network) [5] are example state-of-the-art biomedical data integration tools. Each employ different integration strategies, but both require experts to perform the semantic and syntactic integration. Between the two forms of heterogeneity, semantic integration is the more challenging and costly aspect of biomedical data integration [6]. Expensive terminologists and/or highly trained knowledge engineers are needed to perform the work [7, 8]. At the rate and scale that biomedical data are created, there are simply not enough humans with these skills to massively scale integration efforts. Moreover much of biomedical data are available as free text and in semi-structured formats that are often not leveraged in semantic integration efforts that may affect results from translation research studies [9]. Semiautomatic, and ultimately automatic semantic integration tools, are required to achieve massive biomedical data integration.

Automatic data integration techniques are based on computing semantic "alignments" between data sets, but none of the existing methods achieve the level of performance needed for biomedical investigations [10]. Here we present new algorithms that incorporate both data mining and biomedical informatics methods to improve on and add to existing biomedical data integration methods. The new methods are based on "concept bags" that are similar to "word bags" used for data mining and text retrieval.

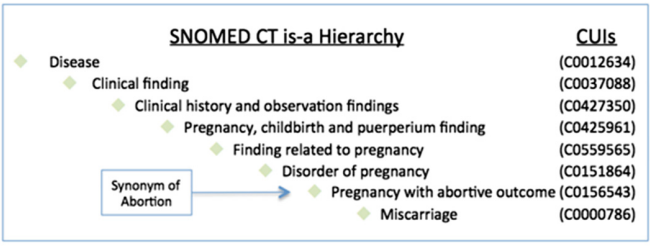
2 Methods

The concept bag (CB) method is based on the n-gram method [11]; however, instead of comparing textual elements, the CB method operates by comparing concept codes. The CB method is defined by two essential steps: (1) Construction of the CB - convert textual or named entities to a representative set of concept codes, and (2) Similarity Measurement – systematically compare CBs using a set or vector-based analysis method. This is a very general definition that leaves a great deal of flexibility in its implementation. Any reasonable selection of a text-to-concept code method or set/vector n-gram-like analysis method could be considered for implementation. This stepwise process describes how the CBs were constructed for the studies that follow:

1. Assign each textual element a unique identifier.
2. Process textual elements and extract representative concept codes. We used the Name-Entity Recognition (NER) software [12] named MetaMap [13] that identifies Unified Medical Language System (UMLS, https://www.nlm.nih.gov/research/umls/) Concept Unique Identifiers (CUI) in the text. MetaMap is a fully integrated NER software which includes, all necessary natural language processes including spell checking for misprints.
3. Associate the distinct set of concept codes from each textual element from step 2 with its unique identifier from step 1.
4. A CB for each textual element is constructed using associations created in step 3. All concept codes associated with each textual identifier comprise the CB for that textual element.

The CB method was originally designed to resolve semantic similarity of strings such as “SBP” and “systolic BP” to the same concept recognizing synonymy between the two strings. However, the method does not consider similarity between words such as “abortion” and “miscarriage” that are not exactly synonyms, but are semantically related [3]. Observing the flexibility of the CB and the potential for ontological relatedness of textual elements, we enhanced CBs by adding UMLS CUIs from the underlying ontological relationships. CUIs from SNOMED’s “is-a” hierarchy were extracted and added for each original CB CUI. We named this strategy the Hierarchical Concept Bag (HCB) method. Further details of the method are given in Bradshaw’s Dissertation [14].

In order to measure similarities between CBs, we selected the Jaccard similarity algorithm [15]. The Jaccard formula computes a decimal value between 0 and 1, where 0 represents no conceptual similarity, and 1 represents a perfect match, by calculating the ratio between the cardinal numbers of the intersection and union of the CBs of the two entities under comparison. Using the CB to compute the similarity between “abortion” and “miscarriage” literally returned 0.0 similarity. With HCBs, the ontological CUI sets representing “abortion” and “miscarriage” share several CUIs in common. After adding the HCB codes, the similarity between “abortion” and “miscarriage” was upgraded from the CB’s 0.0 to 0.89, see Fig. 1 for details.



CB_1 = *ConceptBag(“abortion”) = C0156543*
CB_2 = *ConceptBag(“miscarriage”) = C0000786*
HCB_1 = *HierarchicalConceptBag(“abortion”) = C0037088, C0012634, C0427350, C0425961, C0559565, C0151864, C0156543*

HCB_2 = *HierarchicalConceptBag(“miscarriage”) = C0037088, C0012634, C0427350, C0425961, C0559565, C0151864, C0156543, C0000786*

Jaccard(CB_1, CB_2) = 0
Jaccard(HCB_1, HCB_2) = 8/9 = 0.89

Fig. 1. Examples of the CB and HCB with hierarchical concept codes (CUIs) from the SNOMED CT is-a hierarchy comparing “miscarriage” and “abortion.” The “Parents” figure shows the hierarchies (SNOMED CT is-a hierarchy is poly-hierarchical) and all of the UMLS clinical concepts for both “miscarriage” and “abortion.” The Jaccard similarity scores illustrate the differences between the two methods.

To test the performance of the new methods, we considered three test cases that measured the alignment between, (1) heterogeneous data elements (DE) from a controlled vocabulary, (2) DE from an uncontrolled vocabulary, and (3) medical terms. In the following, we present a brief description of the data sets and how they have been used in these test-cases.

- DEs from a controlled vocabulary: 17 DEs from three domains of the UMLS were selected for this study, seven from demographics, five from vital signs, and five echocardiogram measures.
- DEs from an uncontrolled vocabulary: 899,649 DEs extracted from REDCap [16] from 5 sites engaged in clinical and translational research.
- Medical terms: A benchmark containing a set of 30 medical term pairs that have been curated and judged by physicians and terminologists [17].

The textual features of the data sets were evaluated [14] and each had different textual characteristics are depicted in Table 1. The text size averages varied from 13.1 to 74.9 characters per element, while the mean number of concepts per word varied from 1.5 to 3.2. A more detailed description can be found in [14].

Table 1. Descriptive statistics for the studied data sets, UMLS-selected DEs, REDCap DEs, and the medical terms reference.

	Datasets		
	UMLS DEs	REDCap DEs	Medical terms
Element counts	315	899649	60
Mean characters/Element	24 ± 13.1	43.1 ± 74.9	13.4 ± 6.0
Mean words/Element	3.1 ± 1.5	7.1 ± 11.8	1.6 ± 0.7
Concepts/Data set	380	4187	133
Hierarchical concepts/Data set	753	6929	740
Mean Concepts/Element	9.8 ± 6.7	10.4 ± 10.0	2.6 ± 1.9
Mean concepts/Word	3.2 ± 1.8	2 ± 1.4	1.5 ± 0.6
Mean hierarchical concepts/Word	15.6 ± 11.0	12.8 ± 14.5	20.0 ± 14.3

2.1 DE Alignment Measurement

We measured the alignments between the DEs using similarity measurement algorithms. These algorithms returned real values between 0.0 and 1.0 where 1.0 is perfect similarity and 0.0 is no similarity. Alignment decisions were determined by alignment measurement cutoff values. Similarity values do not represent the percentage of semantic alignment or statistical p-values where one chooses a standard cutoff point. Cutoff values that ultimately define alignment decisions are algorithm-specific and needed to be determined for each of the tested algorithms; therefore, cutoff values for each algorithm were calculated using the Youden method [18], which select cutoff values that maximized the specificity and sensitivity of the decision for each algorithm [14].

All DEs were compared to each other to examine all possible alignments. This process required n(n – 1)/2 comparisons for each set, or approximately 50,000 comparisons for the UMLS controlled vocabulary concepts for test case 1, and half a trillion comparisons of REDCap DE alignments for test case 2. For the REDCap data set preliminary exploration of the computed CB similarity scores indicated that match candidates were very infrequent, ~3 per 1,000 pairs, indicating that an unreasonably large random sample would have been required for human review while maintaining both an accurate sample distribution and conclusive confidence interval. Therefore, a stratified random sample was assembled with 12 buckets based on the computed CB similarity scores, from which a set of 1,200 DE pairs were then randomly sorted and manually reviewed for semantic matches by two professional clinical data architects. Disagreements were reviewed by the two architects to reach consensus. Details of the construction of the data sets used in these comparisons are given in Ref. [14].

Four configurations of the CB and HCB were tested for measuring the alignment between Medical Terms benchmark.

1. CBs created using concepts produced by MetaMap (with default settings) and Jaccard similarity algorithm. This configuration was completely unsupervised.
2. Same as step 1 except the CBs were augmented to HCBs using the SNOMED CT "is-a" hierarchy. This configuration was also completely unsupervised.
3. Same as step 2 but restricting MetaMap to SNOMED CT, and retaining only the single highest-ranking concept for each term. The authors resolved rank ties, which should be considered a minor supervision.
4. Same as step 3 except HCBs were converted to vectors and compared using cosine similarity.

Each of the four CB configurations produced similarity values for each of the 30 pairs from the benchmark data, which were scaled to the same range used by the experts judging the original data set. Correlation between the four CB algorithm configurations given above, seven other published algorithms, and benchmark experts were calculated using Pearson's correlation.

3 Results and Discussion

Receiver operator characteristic (ROC) curves were used to examine the alignment compliance of the DEs from the controlled (UMLS) vocabulary and uncontrolled (REDCap) environment for each data set and each algorithm, see Figs. 2 and 3. The CB and HCB performed very well at the task of aligning UMLS DE as indicated by the AUC = 0.92/0.89, F-measure = 0.79/67 and ROC curves. The performance numbers are slightly lower for the REDCap DE alignment performance AUC = 0.92/0.91, F-measure = 0.55/0.53. This was not surprising due to the nature of REDCap uncontrolled environment where arbitrary abbreviations and local jargon are allowed and supported. None-the-less, even with the added complexities of REDCap DE, the CB and HCB still had much lower combined false positive and false negative rates than the other algorithms considered here.

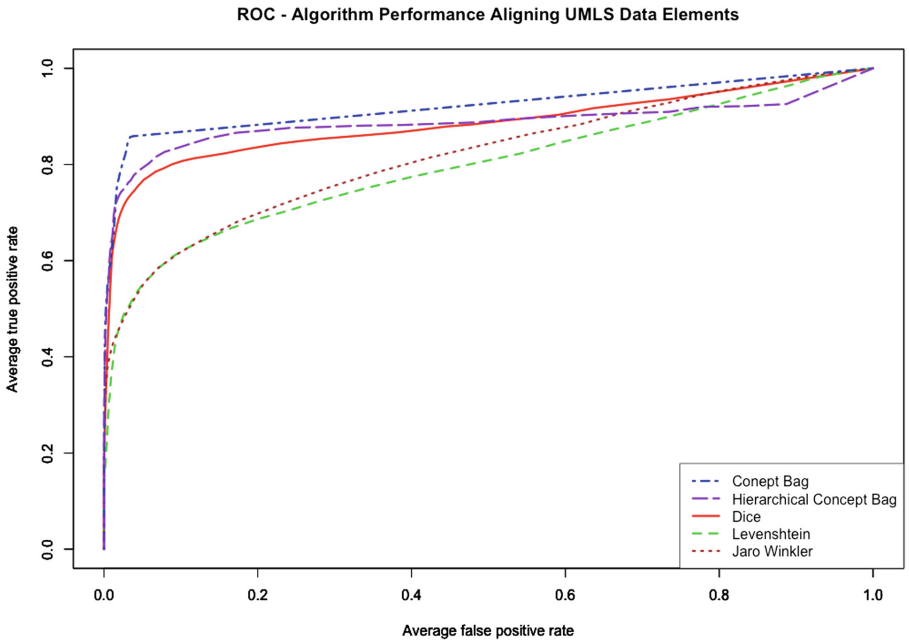


Fig. 2. ROC curve of the UMLS DE alignment algorithm performance. Curves closest to the top left corner are the best performers.

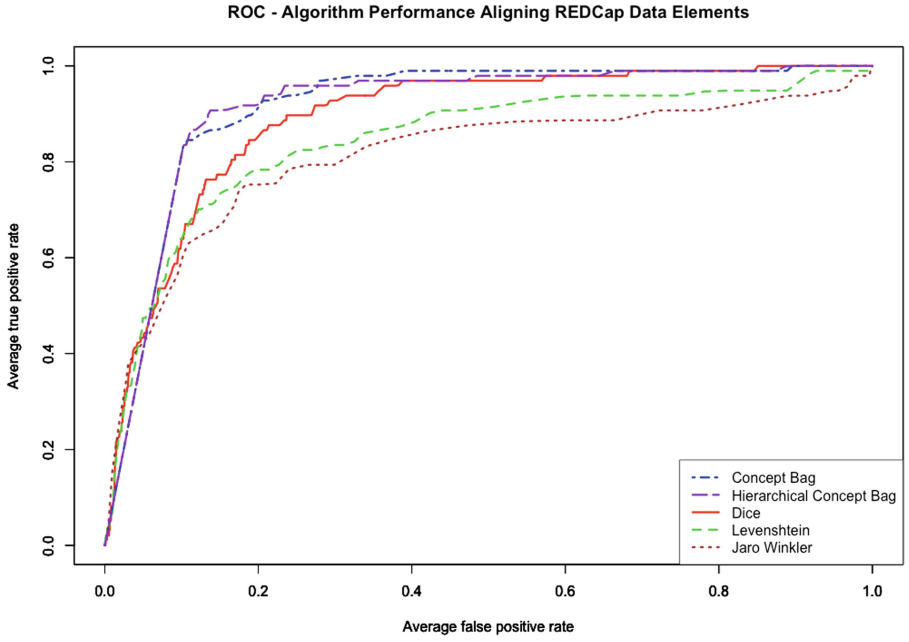


Fig. 3. ROC curve of the REDCap DE alignment algorithm performance. Curves closest to the top left corner are the best performers.

When the goal is automatic DE alignment for semantic integration, alignment decisions are binary, the algorithm either aligns or does not align. In this binary case, degrees of similarity are not tested; therefore, to test the algorithms' ability to assess degree of similarity, we used a medical term similarity benchmark. The CB and HCB algorithms were tested using the medical term similarity benchmark. Table 2 contains the results of the medical term similarity evaluation.

Overall, the HCB correlation scores matched or exceeded 31 other published algorithms [17], for which the top seven are included in Table 2. Of the four CB algorithms tested, the HCB using SNOMED CT with the "is-a" hierarchy measured using Jaccard similarity, tied with the highest correlation with medical terminologists, 0.76, and was the second highest correlation with physicians, 0.72. The tie for the highest correlation with terminologists was with Personalized PageRank algorithm [19]. The highest correlation with physicians, 0.84, was attributed to Pedersen's Context Vector [3] that was successfully augmented with physician-based information content (IC) from a large physician-created corpus. The methods that used terminologist-created SNOMED is-a hierarchy correlated highest with terminologists' similarity judgements, and the method that used physician-created IC correlated highest with physicians.

Table 2. Correlation of the similarity scores obtained with Concept Bag (CB), Hierarchical Concept Bag (HCB), Leacock and Chadorow (LC) [20], Wu and Palmer (WP) [21], Personalized PageRank (PPR) [19], and Context Vector [3].

Method	Physicians	Terminologists
SNOMED HCB	0.72	0.76
SNOMED HCB-Vector	0.65	0.67
UMLS CB	0.46	0.59
UMLS HCB	0.46	0.57
SNOMED LC	0.50	0.66
UMLS LC	0.60	0.65
SNOMED WP	0.54	0.66
UMLS WP	0.66	0.74
SNOMED PPR	0.49	0.61
UMLS PPR	0.67	0.76
Context vector	0.84	0.75

The CB and HCB methods using UMLS (USABase library) and Jaccard similarity both had correlation values of 0.46 with physicians, while the correlation with terminologists was higher for both, 0.59 and 0.57, respectively. The lower correlations are likely due to the broader concept coverage contained in UMLS for which SNOMED CT is a subset, i.e. UMLS produces larger concept bags than it does when there is a reduced set of source vocabularies resulting in more sensitive and less specific similarity measurements. Adding additional hierarchical concept codes magnifies this effect.

4 Conclusions

Automatic alignment of heterogeneous biomedical data is a challenging problem due to the sophisticated semantics of biomedical data. In this paper, we introduced a new class of methods that introduces the idea of using "concept bags" to represent the semantics of textual data elements, described how they can be used to evaluate semantic similarity, and then demonstrated how the similarity measures were tested to automatically align biomedical data elements and compute medical term similarity. Several configurations of the new similarity algorithms were tested for each application and the new methods performed as well or better than well-established methods. Unlike bag of words and n-gram methods, CB and HCB are capable of measuring semantic similarities between synonymous and non-similarly spelled words. We believe we have established the face validity of the CB and HCB methods and recommend them as viable options for computing semantic similarity as demonstrated.

Acknowledgements. This work was partially supported by NCATS award 1ULTR002538 (JCF) and NIH bioCADDIE award 1U24AI117966-01. The authors acknowledge Bernie LaSalle for securing the REDCap data from the CTSA funded institutions and their informatics leaders for providing the data. Computational resources were provided by the Utah Center for High Performance Computing, which has been partially funded by the NIH Shared Instrumentation Grant 1S10OD021644-01A1.

References

1. Gouripeddi, R., Warner, P., Mo, P.: Federating clinical data from six pediatric hospitals: process and initial results for microbiology from the PHIS+ Consortium. In: AMIA Annual Symposium Proceedings (2012)
2. Narus, S.P., Srivastava, R., Gouripeddi, R., Livne, O.E., Mo, P., Bickel, J.P., et al.: Federating clinical data from six pediatric hospitals: process and initial results from the PHIS + Consortium. In: AMIA Annual Symposium Proceedings, pp. 994–1003 (2011). PubMed PMID: 22195159; PubMed Central PMCID: PMCPMC3243196
3. Pedersen, T., Pakhomov, S.V.S., Patwardhan, S., Chute, C.G.: Measures of semantic similarity and relatedness in the biomedical domain. J. Biomed. Inform. **40**(3), 288–299 (2007). https://doi.org/10.1016/j.jbi.2006.06.004
4. Lasalle, B., Varner, M., Botkin, J., Jackson, M., Stark, L., Cessna, M., et al.: Biobanking informatics infrastructure to support clinical and translational research. AMIA Jt. Summits Transl. Sci. Proc. 132–5 (2013). PubMed PMID: 24303252; PubMed Central PMCID: PMC3845745
5. Murphy, S.N., Weber, G., Mendis, M., Gainer, V., Chueh, H.C., Churchill, S., et al.: Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J. Am. Med. Inform. Assoc. (JAMIA) **17**(2), 124–130 (2010). https://doi.org/10.1136/jamia.2009.000893. PubMed PMID: 20190053; PubMed Central PMCID: PMC3000779
6. Chute, C.G., Beck, S.A., Fisk, T.B., Mohr, D.N.: The enterprise data trust at Mayo clinic: a semantically integrated warehouse of biomedical data. J. Am. Med. Inform. Assoc. (JAMIA) **17**(2), 131–135 (2010). https://doi.org/10.1136/jamia.2009.002691. PubMed PMID: 20190054; PubMed Central PMCID: PMC3000789

7. Fan, J.W., Friedman, C.: Deriving a probabilistic syntacto-semantic grammar for biomedicine based on domain-specific terminologies. J Biomed. Inform. **44**(5), 805–814 (2011). https://doi.org/10.1016/j.jbi.2011.04.006. PubMed PMID: 21549857; PubMed Central PMCID: PMC3172402
8. Jezek, P., Moucek, R.: Semantic framework for mapping object-oriented model to semantic web languages. Front Neuroinform. **9**, 3 (2015). https://doi.org/10.3389/fninf.2015.00003. PubMed PMID: 25762923; PubMed Central PMCID: PMC4340193
9. Price, S.J., Stapley, S.A., Shephard, E., Barraclough, K., Hamilton, W.T.: Is omission of free text records a possible source of data loss and bias in clinical practice research datalink studies? A case–control study. BMJ open **6**(5) (2016). https://bmjopen.bmj.com/content/6/5/e011664
10. Dhombres, F., Charlet, J.: As ontologies reach maturity, artificial intelligence starts being fully efficient: findings from the section on knowledge representation and management for the yearbook 2018. Yearb. Med. Inform. **27**(1), 140–145 (2018). https://doi.org/10.1055/s-0038-1667078. PubMed PMID: 30157517
11. Baayen, R.H., Hendrix, P., Ramscar, M.: Sidestepping the combinatorial explosion: an explanation of n-gram frequency effects based on naive discriminative learning. Lang. Speech **56**(Pt 3), 329–347 (2013). PubMed PMID: 24416960
12. Patrick, J., Li, M.: High accuracy information extraction of medication information from clinical notes: 2009 i2b2 medication extraction challenge. J. Am. Med. Inform. Assoc. (JAMIA) **17**(5), 524–527 (2010). https://doi.org/10.1136/jamia.2010.003939. PubMed PMID: 20819856; PubMed Central PMCID: PMC2995676
13. Aronson, A.R., Lang, F.M.: An overview of MetaMap: historical perspective and recent advances. J. Am. Med. Inform. Assoc. (JAMIA) **17**(3), 229–236 (2010). https://doi.org/10.1136/jamia.2009.002733. PubMed PMID: 20442139; PubMed Central PMCID: PMCPMC2995713
14. Bradshaw, R.: Concept bag: a new method for computing similarity. University of Utah (2015)
15. Jaccard, P.J.: Similarity and shingling. Data mining (2015). http://www.cs.utah.edu/∼jeffp/teaching/cs5955/L4-Jaccard+Shingle.pdf
16. Harris, P.A., Taylor, R., Thielke, R., Payne, J., Gonzalez, N., Conde, J.G.: Research electronic data capture (REDCap)–a metadata-driven methodology and workflow process for providing translational research informatics support. J. Biomed. Inform. **42**(2), 377–381 (2009). https://doi.org/10.1016/j.jbi.2008.08.010. PubMed PMID: 18929686; PubMed Central PMCID: PMC2700030
17. Garla, V.N., Brandt, C.: Semantic similarity in the biomedical domain: an evaluation across knowledge sources. BMC Bioinform. **13**, 261 (2012). https://doi.org/10.1186/1471-2105-13-261. PubMed PMID: 23046094; PubMed Central PMCID: PMCPMC3533586
18. Lopez-Raton, M., Rodriguez-Alvarez, M., Cadarso-Suarez, C., Gude-Sampedro, F.: Optimal cutpoints: an R package for selecting optimal cutpoints in diagnostic tests. J. Stat. Softw. **61**(8), 1–36 (2015)
19. Aquire, E., cuadros, M., Rigua, G., Soroa, A. (eds.): Exploring knowledge bases for similarity. In: Proceedings of the Seventh International Conference on Language Resources and Evaluation. European Language Resources Association, Valleta (2010)
20. Leacock, C., Chodorow, M.: Using corpus statistics and wordnet relations for sense identification. In: Fellbaum, C. (ed.) Wordnet: An Electronic Lexical Database, pp. 265–283. MIT Press, Cambridge (1998)
21. Wu, Z., Palmer, M.: Verbs semantics and lexical selection. In: Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, Las Cruces (1994)