SemSim: Resources for Normalized Semantic Similarity Computation Using Lexical Networks

Elias Iosif, Alexandros Potamianos

Department of Electronics & Computer Engineering, Technical University of Crete, Chania 73100, Greece
{iosife,potam}@telecom.tuc.gr

Abstract

We investigate the creation of corpora from web-harvested data following a scalable approach that has linear query complexity. Individual web queries are posed for a lexicon that includes thousands of nouns and the retrieved data are aggregated. A lexical network is constructed, in which the lexicon nouns are linked according to their context-based similarity. We introduce the notion of semantic neighborhoods, which are exploited for the computation of semantic similarity. Two types of normalization are proposed and evaluated on the semantic tasks of: (i) similarity judgement, and (ii) noun categorization and taxonomy creation. The created corpus along with a set of tools and noun similarities are made publicly available.

Keywords: lexical networks, distributional semantic models, semantic similarity

1. Introduction

Semantic similarity is the building block for numerous applications of natural language processing, such as grammar induction (Meng and Siu, 2002) and affective text categorization (Malandrakis et al., 2011). Distributional semantic models (DSMs) (Baroni and Lenci, 2010) are based on the distributional hypothesis of meaning (Harris, 1954) assuming that semantic similarity between words is a function of the overlap of their linguistic contexts. DSMs can be categorized into unstructured (unsupervised) that employ a bag-of-words model (Iosif and Potamianos, 2010) and structured that employ syntactic relationships between words (Grefenstette, 1994; Baroni and Lenci, 2010). DSMs are typically constructed from co-occurrence statistics of word tuples that are extracted on existing corpora or on corpora specifically harvested from the web. The main utility of DSMs is the computation of semantic similarity between word pairs. A popular method for web corpus creation that has been shown to perform quite well for this task (Iosif and Potamianos, 2010) is to search for conjunctive AND web queries in search of documents where word pairs co-occur. However, this methodology suffers from scalability issues, since it requires a quadratic number of queries with respect to the size of the lexicon. In this work, we investigate the estimation of semantic similarity using lexical networks, following a corpus-based approach. In particular, a web corpus is created using individual queries, i.e., "w_i". Individual queries have linear complexity with respect to the lexicon size, and thus they are scalable to large lexicons (unlike conjunctive AND queries). Creating a corpus with large lexical coverage is critical for semantic models that estimate the similarity of polysemous words as the similarity of their closest senses. In order to improve the lexical (and sense) coverage of our corpus, we propose the aggregation of data harvested by a large number of individual queries. In addition, we encode the corpus information by constructing a network of words, in particular nouns, linked according to their semantic similarity. Using a network two main advantages are provided: (i) associations are revealed through network edges that can not be directly identified, and (ii) it is a parsimonious representation of the corpus. Semantic neighborhoods are exploited for the computation of semantic similarity. Two types of normalization are proposed that are shown to significantly improve performance. Two semantic tasks were adopted for the evaluation of the computation of semantic similarity: (i) judgement of noun similarity, and (ii) noun categorization and taxonomy creation. In addition, our data, including the created corpus, the noun similarities and a set of tools are available for downloading.

2. Semantic Similarity Computation

The basic idea here is the computation of semantic similarity between words, for the construction of a lexical network. The similarities were estimated according to the unsupervised paradigm of DSMs, where no linguistic knowledge is required. The fundamental assumption here is that *similarity of context implies similarity of meaning*: we expect words that share similar lexical contexts will be semantically related (Harris, 1954). A common representation of contextual features is the "bag-of-words" model that assumes independence between features (Sebastiani and Ricerche, 2002).

For context-based metrics, a contextual window of size $2H + 1$ words is centered on the word of interest w_i and

lexical features are extracted. For every instance of w_i in the corpus, the H words left and right of w_i are taken into consideration, i.e.,

$$[f_{H,l} \ ... \ f_{2,l} \ f_{1,l}] \ w_i \ [f_{1,r} \ f_{2,r} \ ... \ f_{H,r}],$$

where $f_{k,l}$ and $f_{k,r}$ represent the feature k positions to the left of right of w_i. For a given value of H, the feature vector for w_i is built as $T_{w_i,H} = (t_{w_i,1}, t_{w_i,2}, ..., t_{w_i,Z})$, where $t_{w_i,k}$ is a non-negative integer. The feature vector has length equal to the vocabulary size Z. Non-zero feature value $t_{w_i,k}$ indicates the occurrence of vocabulary word t_k within the left or right context of w_i. Note that the value of $t_{w_i,k}$ is set by considering all occurrences of w_i in the corpus. The value of $t_{w_i,k}$ can be defined according to a binary scheme (Iosif and Potamianos, 2010). This scheme assigns 1 to $t_{w_i,k}$ if vocabulary word t_k occurs within H positions left or right of word w_i, otherwise, $t_{w_i,k} = 0$. The context-based semantic similarity metric s^H between words w_i and w_j is computed as the cosine distance between their corresponding feature vectors:

$$s^H(w_i, w_j) = \frac{\sum_{k=1}^{Z} t_{w_i,k} \ t_{w_j,k}}{\sqrt{\sum_{k=1}^{Z} t_{w_i,k}^2} \sqrt{\sum_{k=1}^{Z} t_{w_j,k}^2}}, \quad (1)$$

for context size H and vocabulary size Z. For words w_i, w_j that share no common context (completely dissimilar words) the corresponding semantic similarity score is 0. Also $s^H(w, w) = 1$. In this work, the pairwise similarities of 8,752 nouns were computed according to (1) for several values of H.

3. Corpus Creation using Web Queries

There are two main types of web queries that can be used for corpus creation: (i) conjunctive AND queries, and (ii) individual (IND) queries. Assuming N words in our lexicon, in the first case all pairwise AND conjunctions are formed and the corresponding queries are posed to a web engine, e.g., "w_i AND w_j". Corpus creation via AND queries leads to quadratic query complexity $\mathcal{O}(N^2)$ in the number of words in the lexicon. Alternatively, one can download documents or snippets with linear query complexity $\mathcal{O}(N)$ using IND queries, i.e., "w_i".

The main advantage of AND queries is that they construct a corpus that is conditioned on word-pairs, explicitly requesting the co-occurrence of word-pairs in the same document. Co-occurrence is a strong indicator of similarity and corpora created via AND queries have been shown to provide very good semantic similarity estimates (Iosif and Potamianos, 2010). To better understand the role of co-occurrence as a feature in semantic similarity computation, we need to revisit the very definition of semantic similarity, as it pertains to words and their senses. According to the information-theoretic approach proposed in (Resnik, 1995), the similarity of two concepts can be estimated as the similarity of their two closest senses. This is also in agreement with our "common sense" (cognitive) model of semantic similarity, when two words are mentioned, their closest senses are activated [1]. We believe that an important contribution of the co-occurrence feature to semantic similarity computation is that *co-occurrence acts as a semantic filter that only retains the two closest senses*.

Unfortunately the attempt to build corpora and DSMs using conjunctive AND queries does not scale to thousands of words due to the quadratic query complexity. We are thus forced to investigate the alternative of using IND queries and face the sense disambiguation issues associated with such corpora. Corpora created via IND queries are similar to a typical text corpus with one important difference: the frequency of occurrence of the words in our lexicon is somewhat normalized, assuming that the same number of snippets is downloaded for each word in the lexicon. Given the requested number of snippets, we expect that rare words will be well-represented within the corpus. In addition, the information content of the corpus pertaining to the words in the lexicon is expected to increase, i.e., the entropy rate of a unigram (zeroth order Markov process) model.

4. Lexical Network

Using a web corpus created via IND queries on a lexicon L we construct next a semantic network encoding the relevant corpus statistics. The links between words in this network are determined and weighted according to the pairwise semantic similarity. The network is defined as an undirected (under a symmetric similarity metric) graph $G = (N, E)$ whose the set of vertices N includes the members of the lexicon L, and the set of edges E contains the links between the vertices.

The network is a parsimonious representation of corpus statistics as they pertain to the estimation of semantic similarities between word-pairs in the lexicon. In addition, the semantic network can be used to discover relations that are not directly observable in the data; such relations emerge via the *systematic covariation of features and similarity metrics*. Semantic neighborhoods play an important role in this process. The members of the semantic neighborhoods of two words are expected to contain features of these words capturing diverse information at the syntactic, semantic and pragmatic level.

The identification of semantic features is also a way for performing sense discovery. Word senses play a central role in semantic similarity estimation. However, sense

[1] The maximum sense similarity assertion is widely employed by many similarity metrics, such as the WordNet-based metrics (Budanitsky and Hirst, 2006), achieving good results.

discovery through semantic neighborhoods is not feasible if the corpus has limited lexical coverage. In this work, we alleviate this issue by aggregating data that are harvested for a lexicon L containing thousands of words. Also, given a large L, instances of a word w_i can be found implicitly, i.e., within data retrieved for w_j, where $w_i \neq w_j$. This enables the discovery of less frequent senses for polysemous words, as well as, relations in which rare words participate.

4.1. Semantic Neighborhoods

For each word (reference word) that is included in the lexicon, $w_i \in L$, we consider a subgraph of G, $G_i = (N_i, E_i)$, where the set of vertices N_i includes in total n members of L, which are linked with w_i via edges E_i. The G_i subgraph is referred to as the semantic neighborhood of w_i. The members of N_i (neighbors of w_i) are selected from L according to the semantic similarity metric, defined by (1), with respect to w_i, i.e., the n most similar words to w_i are selected.

Reference noun	Neighbors selected by $s^{H=1}$ metric
automobile	**auto**, truck, **vehicle**, **car**, engine, bus, boat, [aviation], tractor, [lighting]
car	truck, **vehicle**, travel, service, price, business, home, city, game, quality
food	water, health, family, service, industry, product, market, life, quality, home
slave	nigger, slavery, servant, manumission, beggar, [nationalism], society, [democracy], [aristocracy]

Table 1: Excerpt of semantic neighborhoods.

Some of the neighbors for four nouns computed according to (1) using $H = 1$ are presented in Table 1. The neighbors that are emphasized using bold fonts denote (lexicalized) senses of the respective reference nouns. In general, the neighborhoods are semantically diverse, capturing word senses, as well as, other types of semantic relations. We observe that the discovery of a number of senses via its neighborhoods is feasible for some nouns, e.g., "automobile" and "car". However, this is not true for other nouns ('food" and "slave"), for which their respective senses can not be easily described by single words. In addition to synonymy, taxonomic relations are encoded within the neighborhoods, e.g., IsA(vehicle, car), PartOf(automobile, engine). Relations of associative nature, e.g., ProducedBy(industry, food), are also denoted by some neighbors. Given that the neighborhoods are computed according to contextual similarity, there is no need for the neighbors to co-occur with the reference nouns. In practice, the majority of them co-occur at the sentence level. The exceptions are enclosed in square brackets in Table 1. In such cases, the respective relations seem to have a broader semantic/pragmatic scope, e.g., the concept of slave is somehow related with democracy.

5. Normalization of Neighborhoods

The semantic network is not a metric space under semantic similarity (1) because the triangle inequality is not satisfied. Moreover, we expect that different words will have different neighborhood statistics. Based on our assumption that the neigborhoods capture (to some extent) the semantics of words, we suggest that the neighborhood differences should be taken into account during the computation of semantic similarity. We investigated two normalization schemes in order to address this issue.

Local Normalization. Motivated by similar approaches from the area of multimedia (Lagrange and Tzanetakis, 2011) we applied the N-normalization (or local scaling) (Zelnik-Manor and Perona, 2004), defined as

$$s_N(n_1, n_2; H) = \frac{s(n_1, n_2; H)}{\sqrt{s(n_1, n_{1,N}; H) s(n_2, n_{2,N}; H)}}, \quad (2)$$

where $s(n_i, n_j; H)$ is the similarity score between n_i and n_j for a contextual window of size H (computed by (1)), N is the number of neighbors included in the neighborhood, and $n_{i,N}$ is the N^{th} neighbor of n_i.

Global Normalization. Z-normalization (Cohen, 1995) is employed as a type of global normalization, by considering all the nouns of the network as members of the semantic neighborhood. The Z-normalized similarity between two nouns is defined as

$$s_Z(n_1, n_2; H) = \frac{s(n_1, n_2; H) - \mu_1}{\sigma_1}, \quad (3)$$

where μ_1 and σ_1 are the arithmetic mean and the standard deviation, respectively, of the similarity scores between n_1 and the rest nouns of the network. Also, $s(n_i, n_j; H)$ is the similarity score between n_i and n_j for a contextual window of size H (computed by (1)). The similarity computed by (3) is not symmetric, i.e., $s_Z(n_1, n_2; H) \neq s_Z(n_2, n_1; H)$, since

$$s_Z(n_2, n_1; H) = \frac{s(n_1, n_2; H) - \mu_2}{\sigma_2}, \quad (4)$$

where μ_2 and σ_2 are the arithmetic mean and the standard deviation, respectively, of the similarity scores between n_2 and the rest nouns of the network. The similarity score $s(n_1, n_2; H)$ is identical to the score used in (3).

In this work, a symmetric similarity score was defined as

$$s_Z^M(n_1, n_2; H) = \max\{s_Z(n_1, n_2; H), s_Z(n_2, n_1; H)\}. \quad (5)$$

6. Experimental Procedure and Parameters

The experimental procedure consists of the following steps. 1) Query formulation and corpus creation. As a lexicon we used $8,752$ English nouns taken from the SemCor3 [2] corpus. For each noun an individual query was formulated and the $1,000$ top ranked results (document snippets) were retrieved using the Yahoo! Search API (2 Feb.'11). The corpus was created by aggregating the snippets for all nouns. 2) Computation of semantic similarity. The pairwise noun similarities were computed according to (1) for $H = 1, 2, 3, 5$. 3) Network creation. The semantic neighborhoods of nouns were computed, following the procedure described in Section 4.1. 4) Similarity computation using normalization. Local and global normalization schemes were applied. We experimented with $H = 1, 2, 3, 5$, and with various values of N ranging from 10 up to 200.

7. Evaluation

In this section, we evaluate the performance of the normalization schemes with respect to two tasks: (i) similarity judgement between nouns, and (ii) noun categorization. The normalization-based approaches are also compared to the baseline method for semantic similarity computation.

7.1. Similarity Judgement

The baseline and the normalized similarity scores were evaluated against human ratings using two standard datasets of noun pairs, MC (Miller and Charles, 1998), and RG (Rubenstein and Goodenough, 1965). The first dataset consists of 28 noun pairs, while for the second dataset we used 57 nouns pairs, also in SemCor3. The Pearson's correlation coefficient was used as evaluation metric. The performance of the normalized similarities (solid line) in comparison with the baseline performance (dashed line) is shown in Fig.1. The correlation results for the case of s_N (local norm.) for $H = 1$ are depicted in Fig.1(a) and (b), for MC and RG datasets, respectively. The correlation is plotted as a function of the number of neighbors (N). The performance of similarity scores normalized by s_Z^M (global) with respect to MC and RG datasets, is presented in Fig.1(c) and (d), respectively. The correlation scores are plotted against different values of H. Overall, the performance of similarities normalized by the global scheme is significantly higher compared to baseline, and similarities normalized by the local scheme [3] . The reported correlation scores are lower compared to the state-of-the-art results: (i) 0.88 for CM (Iosif and Potamianos, 2010), where conjunctive AND queries are used, and (ii) 0.85 for RG (Baroni and Lenci, 2010), where linguistic knowledge is exploited. To our knowledge, these are the best reported results using individual queries, i.e., with linear query complexity.

[2] `http://www.cse.unt.edu/~rada/downloads.html`

7.2. Noun Categorization and Taxonomy Creation

The performance of the similarities computed by the baseline metric and the global normalization schemes were evaluated on noun categorization and taxonomy creation tasks. The similarity scores were used for the construction of a similarity matrix upon which the k-means clustering algorithm was applied. The experimental datasets are presented in Table 2. Regarding noun categorization we used the Battig (Baroni et al., 2010) and the AP (Almuhareb and Poesio, 2005) datasets. We experimented with those nouns included in the set of $8,752$ nouns: 49 nouns classified into 10 classes for the Battig dataset, and 21 classes including 240 nouns for the AP dataset. For the task of taxonomy creation we used the ESSLLI dataset (Baroni et al., 2008), which is a three-level hierarchy ($2-3-6$ classes). The lowest level of the hierarchy (6 classes) is presented in Table 2. The middle level includes the classes *animals*, *vegetables*, and *artifacts*, while the upper level is distinguished in *living beings*, and *objects*. We considered 31 nouns included in the set of $8,752$ nouns.

The purity of clusters, P, was used as evaluation metric, defined as (Baroni and Lenci, 2010):

$$P = \frac{1}{c}\sum_{i=1}^{k}\max_j(c_i^j), \quad (6)$$

where c_i^j is the number of nouns assigned to the i^{th} cluster that belong to the j^{th} groundtruth class. The number of clusters is denoted by k, while c is the total number of nouns included in the dataset. Purity expresses the fraction of nouns that belong to the true class, which is most represented in the cluster (Baroni and Lenci, 2010), taking values in the range $[0, 1]$, where 1 stands for perfect clustering. The results are presented in Table 3 for the baseline similarities and the normalized similarities according to s_Z^M (global norm.) for several values of H. The performance of the normalized similarities is consistently better than the performance of the baseline similarities. These results are close enough to the state-of-the-

[3] Also, we experimented with various linear combinations of the similarity scores computed by the two normalization schemes without any significant improvement in performance.

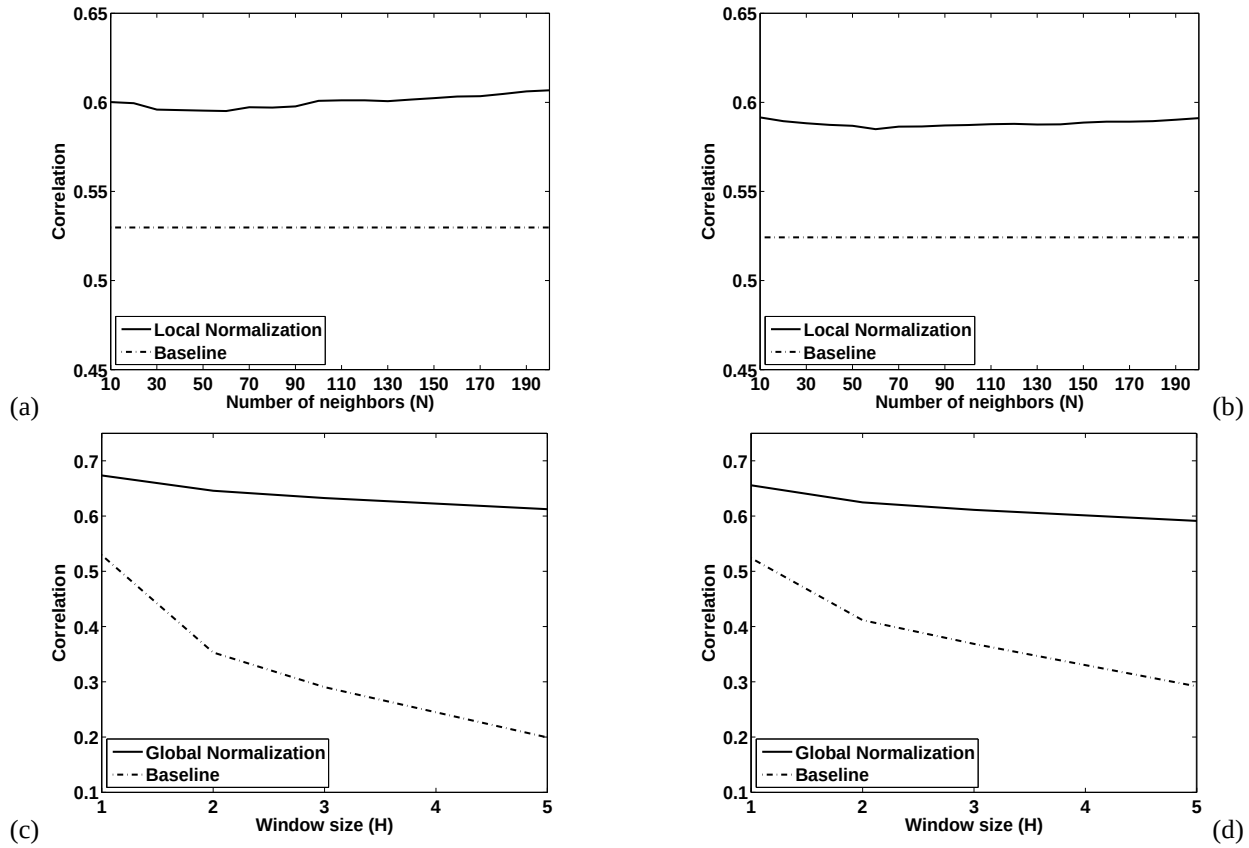

Figure 1: Correlation for the task of similarity judgement. Baseline and normalized similarities using s_N (local) for: (a) MC, (b) RG datasets. Baseline and normalized similarities using s_Z^M (global) for: (c) MC, (d) RG datasets.

Dataset	# of nouns	# of classes	Description of classes
Battig	49	10	mammals, birds, fish, vegetables, fruit tress, vehicles, clothes, tools, kitchenware
AP	240	21	animal, assets, atmospheric phenomenon, chemical element, creator, district, edible fruit, feeling, game, illness1, illness2, legal document, monetary unit, pain, physical property, social occasion, social unit1, social unit2, solid, tree, vehicle
ESSLLI	31	6 (lowest level)	birds, land animals, fruit, greens, vehicles, tools

Table 2: Datasets for noun categorization and taxonomy creation.

art results (Battig: 0.96, AP: 0.79, ESSLLI: $1-1-0.91$, (Baroni and Lenci, 2010)), which are obtained by methods exploiting linguistic knowledge.

8. Data, Tools and Resources

In this section, we briefly describe the data that are made publicly available [4].

Data (SemSim Corpus). This is the corpus of snippets that were aggregated for the $8,752$ English nouns. Overall, the SemSim corpus consists of approximately $8,752,000$ snippets that correspond to $12,435,600$ sentence fragments. In general, a snippet may include more than one sentence fragment. The vocabulary size is $1,413,775$, while in total the corpus contains $199,510,174$ tokens.

Tools (CParse & CosSim). CParse parses the SemSim corpus and creates the context feature vectors. CosSim is fed with the feature vectors and computes similarities in a computational efficient manner (18K/s on a 2.66GHz Pentium). Both tools are re-usable, e.g., for enriching the existing pool of similarities, or for other corpora.

[4] `http://www.telecom.tuc.gr/~iosife/downloads.html`

Contextual Window Size (H)	Dataset				
	Battig	AP	ESSLLI Taxonomy		
			Level 0	Level 1	Level 2
	(10 classes)	(21 classes)	(2 classes)	(3 classes)	(6 classes)
1	0.86/0.96	0.53/0.53	0.65/1	0.87/0.90	0.77/0.84
2	0.80/0.94	0.45/0.48	0.65/0.87	0.84/0.84	0.74/0.81
3	0.67/0.92	0.41/0.45	0.65/0.87	0.81/0.81	0.74/0.77
5	0.71/0.86	0.38/0.41	0.58/0.81	0.74/0.81	0.61/0.68

Table 3: Purity of classes: baseline similarities/similarities normalized by s_Z^M (global).

Resources (SemSim Repository). This is a repository that includes the pairwise semantic similarities of the $8,752$ nouns. The baseline similarities were computed according to (1) for $H = 1, 2, 3, 5$. Also, the repository includes the normalized (local and global) similarities, for a total of $919,170,048$ scores.

9. Conclusions

In this work, we followed an unsupervised approach for the computation of semantic similarity using individual queries (linear query complexity). More importantly, we showed how to construct a large lexical network that can reveal useful information regarding the linked words. Also we investigated two normalization schemes showing significant performance improvement. Last but not least, we make available large resources and tools, fostering their re-usability.

10. Acknowledgements

Elias Iosif was partially funded by the Basic Research Programme, Technical University of Crete, Project Number 99637: "Unsupervised Semantic Relationship Acquisition by Humans and Machines: Application to Automatic Ontology Creation".

11. References

A. Almuhareb and M. Poesio. 2005. Finding attributes in the web using a parser. In *Proc. of Corpus Linguistics Conference*.

M. Baroni and A. Lenci. 2010. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721.

M. Baroni, S. Evert, and A. Lenci. 2008. Bridging the gap between semantic theory and computational simulations. In *Proc. of ESSLLI Distributional Semantic Workshop*.

M. Baroni, B. Murphy, E. Barbu, and M. Poesio. 2010. Strudel: A corpus-based semantic model based on properties and types. *Cognitive Science*, 34(2):222–254.

A. Budanitsky and G. Hirst. 2006. Evaluating WordNet-based measures of semantic distance. *Computational Linguistics*, 32:13–47.

P. R. Cohen. 1995. *Empirical Methods for Artificial Intelligence*. MIT Press.

G. Grefenstette. 1994. *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Publishers.

Z. Harris. 1954. Distributional structure. *Word*, 10(23):146–162.

E. Iosif and A. Potamianos. 2010. Unsupervised semantic similarity computation between terms using web documents. *IEEE Transactions on Knowledge and Data Engineering*, 22(11):1637–1647.

M. Lagrange and G. Tzanetakis. 2011. Adaptive n-normalization for enhancing music similarity. In *Proc. ICASSP*.

N. Malandrakis, A. Potamianos, E. Iosif, and S. Narayanan. 2011. Kernel models for affective lexicon creation. In *Proc. Interspeech*.

H. Meng and K.-C. Siu. 2002. Semi-automatic acquisition of semantic structures for understanding domain-specific natural language queries. *IEEE Transactions on Knowledge and Data Engineering*, 14(1):172–181.

G. Miller and W. Charles. 1998. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28.

P. Resnik. 1995. Using information content to evaluate semantic similarity in a taxanomy. In *Proc. of International Joint Conference for Artificial Intelligence*, pages 448–453.

H. Rubenstein and J.B. Goodenough. 1965. Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633.

F. Sebastiani and C. N. D. Ricerche. 2002. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47.

L. Zelnik-Manor and P. Perona. 2004. Self-tuning spectral clustering. In *Proc. Conference on Neural Information Processing Systems*.