

# Distributional Semantic Concept Models for Entity Relation Discovery

<b>Jay Urbain</b> Milwaukee School of Engineering CTSI, MCW 1025 N. Broadway Ave. Milwaukee, WI, USA urbain@msoe.edu	<b>Glenn Bushee</b> Clinical Translational Science Institute Medical College of WI 8701 Watertown Plank Milwaukee, WI, USA gbushee@mcw.edu	<b>Paul Knudson</b> Clinical Translational Science Institute Medical College of WI 8701 Watertown Plank Milwaukee, WI, USA knudson@mcw.edu	<b>George Kowalski</b> Clinical Translational Science Institute Medical College of WI 8701 Watertown Plank Milwaukee, WI, USA gkowalski@mcw.edu	<b>Brad Taylor</b> Clinical Translational Science Institute Medical College of WI 8701 Watertown Plank Milwaukee, WI, USA btaylor@mcw.edu
--	--	--	---	---

## Abstract

We present an ad hoc concept modeling approach using distributional semantic models to identify fine-grained entities and their relations in an online search setting. Concepts are generated from user-defined seed terms, distributional evidence, and a relational model over concept distributions. A dimensional indexing model is used for efficient aggregation of distributional, syntactic, and relational evidence. The proposed semi-supervised model allows concepts to be defined and related at varying levels of granularity and scope. Qualitative evaluations on medical records, intelligence documents, and open domain web data demonstrate the efficacy of our approach.

## 1 Introduction

Knowledge discovery could be facilitated with the ability to define concepts ad hoc, and from these concepts identify semantically related named entities and entity relations. In an online search setting, identification of specific named entities may not be available, or may not have the granularity to support specific information needs. Attempting to provide models for all possible entity and relation types is computationally intractable, so there is a need for a more flexible, fine-grained, user-driven approach.

These needs are in contrast to named entities identified by models defined in advance from labeled training data, knowledge bases, or embedded in a set of rules. Entities identified from these models may be too general, e.g., person *versus* terrorist, or disease *versus* diabetes; or domain specific, e.g., protein type in a dietary *versus* a molecular biology sense. This can be an impediment to search and discovery since

many discoveries are serendipitous in nature and are found by identifying linkages between more specialized concepts within and across domains. Using a flexible dimensional index for efficient aggregation of distributional statistics and a distributional relational model over concept distributions, we propose a new, more flexible approach for creating fine-grained, user-driven concept models for identification of semantically related entity relations.

First, we present an information-seeking scenario to motivate our approach. This is followed by a presentation of our proposed distributional semantic concept model and qualitative results.

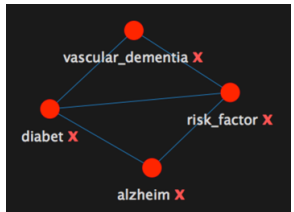
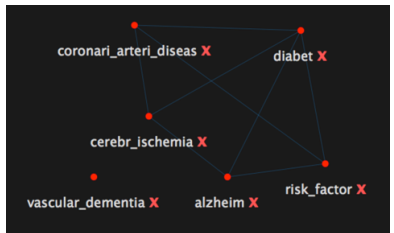
### 1.1 Ad hoc information seeking scenario

Interactive knowledge discovery can be modeled using a dual representation of concepts and relations (Bollegala, et al., 2010). Concepts can be defined by the relations they participate in, and by their lexical and semantic similarity. Relations can be defined by their participating concepts, and by semantically similar relations. In the following scenario, we are interested in identifying relations between *Alzheimer’s disease (AD)* and other diseases. We’ve heard of studies linking *Type 2 Diabetes Mellitus (T2DM)* with *AD*, so we start with the query “*Diabetes related to Alzheimers.*” The system extracts candidate *entity* instances and *relations* from the query (Table 1).

<p>Query: <i>Diabetes related to Alzheimers</i></p> <p>Concepts:</p> <p><b>Diabetes</b> - id: /en/diabetes_mellitus, type: /medicine/disease</p> <p><b>Alzheimers</b> - id: /en/alzheimers_disease, type: /medicine/disease</p> <p>Relations - (concept1, relation 1, 2,..., concept2):</p> <p>diabetes; related to; alzheimers -&gt; disease; related to; disease</p>
--

**Table 1. Parsed query with semantically related entities.**

A structured representation of the query is generated that integrates syntactic and lexical evidence with distributional semantic concept models of each candidate entity. Sentences semantically *relevant* to the query are retrieved and rank ordered. A sample search result for “Diabetes related to Alzheimer’s” with extracted concepts and relations are shown in Table 2(a). Table 2(b) shows a entity-relation-entity graph of the query and a retrieved sentence.

<p><b>a) Retrieved Sentence with concepts &amp; relational dependencies</b></p>	<p><b>Diabetes is a risk factor for vascular dementia.</b></p> <p><i>Dependency relations: (concept1; relation 1, 2,...;concept2)</i></p> <p>diabetes; ; risk_factor</p> <p>risk_factor; for; vascular_dementia</p> <p>diabetes; risk_factor_for; vascular_dementia</p>
<p><b>b) Concept-relation graph: Query + Sentence</b></p>	
<p><b>c) Semantic similarity graph: query: (vascular dementia; risk factor; *).</b></p>	

**Table 2.** (a) Concept-relation search result for query: *Diabetes related to Alzheimer’s*. (b) Graph of query and sentence result. (c) Concept-relation graph search results for query: *(vascular dementia; risk factor; \*)*.

The search result and query provide a relational lattice linking diabetes, vascular dementia, and Alzheimer’s with risk factors. From analyzing the results of the query, the user may be interested in identifying other concepts related to risk factors and vascular dementia. For example, the user may expand the scope of the search space by querying for any concept related as a risk factor to vascular dementia.

A dimensional index is used for efficiently aggregating distributional statistics and relating evidence of concepts and relations within the search index with information from the query. Table 2(c) shows the results using a force-directed graph. The user can now identify new concepts participating in some form of risk factor relation. From these results, other relations for one or more concepts or any combination of concept relation could be explored. Table 4 lists the ranked retrieval process.

<ol style="list-style-type: none"> <li>1. The user presents a natural language query.</li> <li>2. The NLP engine parses the query, extracts candidate entities, dependency relations, syntax, and textual context.</li> <li>3. A structured query is generated from the evidence extracted by the NLP engine.</li> <li>4. A distributional semantic model is generated for each entity within the query from the dimensional index.</li> <li>5. Word and phrase search within the context of individual sentences and documents.</li> <li>6. Query model (4) applied to the top ranking sentences from (step 5).</li> <li>7. User can provide relevance feedback to the system.</li> </ol> <p><i>Iterate over search results.</i></p>
---

**Table 4. Ranked retrieval process (top) and architecture (bottom).**

## 2 Dimensional Indexing

A dimensional indexing model (Kimball, 1996; Gray, et al. 1997) is used for efficient search and aggregation of distributional statistics. The model represents a Vector Space Model (VSM) of distributional statistics for defining concepts, and a data warehousing style (dimensional data model) inverted index of words, phrases, named entities, relations, and sentences. The grain of the index is the individual word with attributes for position, part-of-speech, and phrase. Semantic concepts are defined over word distributions from the index. An *entity-relation-entity* index is also created during indexing to link candidate

entity instances (noun phrases) with their shortest path dependency relation within sentences. The same NLP is used for query processing, and sentence parsing during indexing.

Importantly, the dimensional index facilitates efficient OLAP style SQL queries for aggregating distributional statistics, and for executing relational queries over concepts. The index also supports aggregation over word, phrase, entity, relation, sentence, or document. A variation on this indexing approach has been scaled to several hundred Gigabytes for chemical patent retrieval (Urbain, et al. 2009). Indexes can be created from local collections and integrated with indexes created from online web search results.

### 3 Distributional Semantic Model

Distributional semantics quantifies and categorizes semantic similarities between linguistic terms based on their distributional properties in large samples of text. The central assumption is that the context surrounding a given word provides important information about its meaning (Church et al., 1989, 1991; Firth, 1968; Harris, 1954; Turney and Pantel, 2010). VSMs provide a mechanism for representing term, concept, relation, or sentence meaning by using distributional statistics. The semantic properties of words are captured in a multi-dimensional space by vectors that are constructed from large bodies of text by observing the distributional patterns of co-occurrence with their neighboring words. These vectors can then be used as measures of text similarity between words, phrases, abstract concepts, entities, relations, or snips of arbitrary text.

We base our distributional measures of semantic similarity using pointwise mutual information (PMI). PMI measures the pointwise mutual information between two objects as the log ratio of the joint probability of two objects co-occurring relative to the probability of those objects occurring independently. PMI using information retrieval (PMI-IR) was suggested by Turney (2001) as an unsupervised measure for the evaluation of the semantic similarity of words (Eq. 1). Turney defined words as words co-occurring if they co-occurred within a 10-word window.

$$PMI(w1, w2) = \log_2 \left( \frac{p(w1, w2)}{p(w1)p(w2)} \right) \quad (1)$$

Multiple evaluations have demonstrated the effectiveness of PMI on semantic similarity benchmarks (Mihalecea, 2006; Eneko, 2012). We are also attracted to its simplicity and efficiency for generat-

ing distributional concept models online within our dimensional data model. Tables 6 and 7 show the PMI of words for the concepts Diabetes and CHF (Congestive Heart Failure). The distribution of semantically similar words (shown in lexically stemmed form) for each disease can be used to infer the underlying concepts Diabetes and CHF respectively.

Concept	Stem term	PMI
diabet	mellitu	4.12
diabet	depend	3.52
diabet	type	2.67
diabet	retinopathi	2.14
diabet	insulin	2.13
diabet	nephropathi	2.02
diabet	noninsulin	1.84
diabet	hyperlipidemia	1.76
diabet	esrd	1.54
diabet	adult	1.52
diabet	glaucoma	1.42
diabet	hypercholesterolemia	1.10

Table 6. PMI of words for Diabetes.

Concept	Stem term	PMI
chf	exacerb	2.34
chf	ef	1.5
chf	drainag	1.4
chf	leukocytosi	0.71
chf	lvh	0.47
chf	treat	0.34
chf	secondari	0.33
chf	etiolog	0.31
chf	cad	0.29
chf	diuresi	0.27
chf	evid	0.25
chf	pleural	0.21

Table 7. PMI of words for CHF.

Mihalecea, et al. (2006) extended semantic similarity measurements to two arbitrary text segments. Given a measurement for the semantic similarity of two unordered (bag of words) text segments and a measurement for term specificity, the semantic similarity of two text segments  $C1$  and  $C2$  can be defined using a model that combines the semantic similarities of each text segment in turn with respect to the other text segment. We extended the original bag-of-words text-to-text measurement to include phrases

(candidate entities and their relation dependencies). Using PMI as the underlying measure of semantic similarity, we developed the following 2nd order

PMI-based model for measuring the semantic similarity between concepts  $C_1, C_2$ . (Eq. 2).

$$SemSim(C_1, C_2) = \frac{1}{2} \left( \frac{\sum_{w \in (W_1 \cap W_2)} (PMI(C_1, w) * idf(w) + (PMI(C_2, w) * idf(w)))}{\sum_{w \in (W_1 \cap W_2)} (idf(w))} \right) \quad (2)$$

Concept1	Concept2	Co-term	$PMI(C_1, w) * idf(w)$	$PMI(C_2, w) * idf(w)$	Average
afghanistan	pakistan	india	6.00	6.66	6.33
afghanistan	pakistan	iran	6.10	6.04	6.07
afghanistan	pakistan	china	6.15	5.94	6.05
afghanistan	pakistan	franc	6.03	5.94	5.99
afghanistan	pakistan	russia	5.63	6.04	5.83
afghanistan	pakistan	tajikistan	5.48	6.10	5.79
afghanistan	pakistan	arabia	4.93	5.88	5.41
afghanistan	pakistan	soviet	5.42	5.09	5.25
afghanistan	pakistan	britain	5.63	4.48	5.06

**Table 7. Semantic similarity ( $SemSim$ ) between concepts Afghanistan and Pakistan**

$$RelDepSim(R_1, R_2) = \alpha \sum_{w \in (R_1 \cap R_2)} (NIRDF(w)) + (1 - \alpha) \sum_{i=1}^2 SemSim(e_{r1i}, e_{r2i}) \quad (3)$$

$$LexSim(S_1, S_2) = \alpha_1 \sum_{e \in (E_1 \cap E_2)} (1) + \alpha_2 BM25(S_1, S_2) + \alpha_3 BM25(D_1, D_2) \quad (4)$$

Where  $\alpha_1 > \alpha_2 > \alpha_3$ .

$$AggSim(CR_1, CR_2) = \alpha_1 SemSim(C_1, C_2) + \alpha_2 RelDepSim(R_1, R_2) + \alpha_3 LexSim(S_1, S_2) + \alpha_4 PRSim(S_1, S_2) \quad (5)$$

Where  $\alpha_1 > \alpha_2 > \alpha_3 > \alpha_4$ .

Concept instances used in Eq. 2 may be any text segment. PMI is calculated over the inner product (relational join) of all mutually co-occurring words between  $C_1$  and  $C_2$  is weighted by their respective semantic similarity ( $SemSim$ ) and their normalized inverse document frequency ( $NIDF$ ). This measurement is completely unsupervised and can be used to compare any ordered or non-ordered text segment across any domain. To demonstrate the open domain capability of the semantic similarity measurement, we list the top co-occurring  $PMI * IDF$  measurements for *Afghanistan* and *Pakistan* in a post 9/11 intelligence document collection Table 7.

For reference we provide information retrieval measurements for relational dependency similarity

(Eq.3), lexical similarity (Eq. 4) using Robertson’s *BM25* (2000), and an aggregate similarity measurement integrating semantic, relational dependency, and lexical similarity (Eq. 5).

### 3.1 Learning Semantic Concepts

Figure 1 illustrates the following process for defining semantic concepts.

- 1) Users provide seed terms to bootstrap learning of a semantic concept. In this case, the user defines the semantic concept *CAD*, and seed terms *CAD* and *coronary artery disease*. *Note: Seed terms may be any combination of individual words or phrases.*

**Learn Semantic Concepts**

**Define or query concept**

- Concept name examples: *terrorist*, *CAD*, *diabetes*, etc.
- Terms for representing *terrorist* concept: *Osama bin Ladin*; *KSM*; *Sayid*.
- Terms for representing *CAD\_DM* concept: *diabetes*; *DM*; *CAD*; *CABG*; *coronary artery disease*.
- Alternatively, concepts can be defined using relational algebra defined over existing concepts.

After learning or updating the concept model, the top distributional words are generated using PMI. From that distribution, the top named entities are predicted using 2nd order PMI.

Select concept:

Name:

Definition:

Relation:

Terms:

ATEA Search  
 OIL Search  
 i2b2 Risk Search  
 Web Search (slow)  
 Search Web Session Index Only  
  
 Query Parse Only

Figure 1. Learning semantic concepts

- 2) Concept terms can come from different conceptual areas to meet specific information retrieval needs. For example, terms from *finance* and *terrorism*, or terms identifying medical comorbidities such as *coronary artery disease* and *diabetes*. Additional terms can also be added for increased specificity.

A vector-space model of a concept's distribution is generated from 2<sup>nd</sup> order probabilistic likelihood of co-occurring terms (PMI) (Figure 2):

Termid	Term	Idf	N	Pmi	NPmi
2349	cabg	0.331	2	3.757	1
90	fhx	0.629	1	3.171	0.844
1249	anterosept	0.608	1	2.317	0.617
13612	pnc	0.803	1	2.16	0.575
1675	imi	0.608	1	2.102	0.559
112	known	0.189	2	2.084	0.555
2211	leukemia	0.589	1	1.956	0.521
3410	pvd	0.363	1	1.837	0.489
1635	chf	0.286	1	1.785	0.475
115	coronari	0.159	2	1.76	0.468
93	cad	0.151	1	1.749	0.466
6869	psychosi	0.712	1	1.716	0.457
3646	gastriti	0.53	1	1.646	0.438
2350	septemb	0.43	1	1.525	0.406
276	mi	0.207	2	1.476	0.393
133	arteri	0.142	1	1.453	0.387
139	stent	0.232	2	1.448	0.385
105	hyperlipidemia	0.155	2	1.424	0.379
889	ptca	0.46	1	1.344	0.358
100	hx	0.207	1	1.28	0.341

Figure 2. Distributional concept model for CAD

Qualitative review of concept terms demonstrates the accuracy of this approach. To properly evaluate the semantic model, we should be able to take the model and predict relevant named entities.

- 3) From the concept model CAD, we can predict the likelihood of semantic relatedness of candidate entities (Figure 3). *Note: Candidate entities are noun phrases identified during indexing or query processing.*

Entityid	Term	Idf	SC
86	cad	0.267	1
320	histori	0.163	0.439
204	coronari_arteri_diseas	0.597	0.406
39	htn	0.157	0.363
166	dm	0.182	0.329
321	cabg	0.455	0.328
454	hypertens	0.339	0.261
666	diabet	0.267	0.241
2311	risk	0.666	0.23
40	pt	0.101	0.204
288	chf	0.316	0.2
1614	ag	0.629	0.199
1885	pvd	0.57	0.196
2031	septemb	0.545	0.18

Figure 3. Named entities predicted for concept CAD

- 4) From the semantic concept model, CAD, we can predict the likelihood of generating sentences by using this model for sentence information retrieval (Figure 4).

N	ID	2nd Order PMI	Text
1	215-215-8528	0.122	Ehlers) CABG x 4 (LIMA - LAD/ Sequential graft: SVG1 connects Aorta to D1 then OM1/ SVG2 - LVB2 Incidental anomalous circumflex off the right coronary artery Pre-op EF 50-55% with infero-posterior hypokineses (?post-op) Bicuspid aortic valve (not replaced with CABG according to notes) Mild AS, Mild AI only on pre-CABG echo 2097 Hypertension Dyslipidemia Diabetes Type II Perioperative hyperglycemia post CABG Exsmoker Quit 2097, >50 pack years Peripheral arterial disease Occluded distal aorta (diagnosed after cardiac catheterization attempt in 2097) Small infrarenal abdominal aortic aneurysm 3.6cm Paroxysmal atrial fibrillation (x2 years at least) On coumadin for stroke prevention and amiodarone from maintenance of NSR Currently in NSR BPH Chronic back pain Renal insufficiency (Creatinine 4.1) non-oliguric - ?cause Cr normal November 2097 Gradually increased over November from 1.1-1.6,
2	161-161-6298	0.112	PMH: DM x 20 years, peripheral neuropathy HTN CAD Anteroseptal MI in 03/91, PTCA and stent EF 32% (last echo 2092) Hemorrhoids (normal colo 2095) Iritis, corneal dystrophy Osteoporosis Diverticulosis (?) Meds: Gilipizide 2.5 QD Asa 325 QD Lisinopril 5mg QD Lopressor 50 mg BID Lipitor 10 mg QD Amitriptyline 25 mg QHS Prednisone gtt Caltrate + D (CALCIUM Carbonate 1500 Mg (600 Mg Elem Ca)/ Vit D 200 Iu) 1 TAB PO BID ALL: Acetaminophen--rash SH: Tob: 43 year pack hx, quit 2060 EtOH: denies Illicit: denies FH: Mother: pernicious anemia Father: died in accident at young age Brother: leukemia ROS: As in HPI. - General: no weight loss/gain, no fatigue, no fevers, no chills, no change in appetite - Respiratory: no cough, no SOB, no DOE, no hemoptysis, no wheezing - HEENT: no neck stiffness, no hoarseness, no hearing loss - Cardiac: chest pain/pressure as above, no palpitations, no orthopnea, no PND - Gastrointestinal: no nausea, no vomiting, no diarrhea, no constipation, no bleeding - Neurologic: diminished sensation in LLE bilaterally - Lymph nodes: no enlarged lymph nodes - Musculoskeletal: No back pain, no neck pain, no leg pain, no arm pain - Urologic: No hematuria, no dysuria, no polyuria, no nocturia - Hematologic: No bruising, no bleeding - Exposures: No sick contacts, no recent travel Exam: VS: 97.7 74 18 98% supine 108/54 standing 127/70 HEENT: NG/AT, PERRL, nonicteric.
3	123-123-4822	0.111	PMH: Duchenne's muscular dystrophy, COPD (2.5L home O2), IDDM, recurrent UTIs w/ previous Candida tropicalis fungemia (followed by Dr. Phoebe Abreu), B ureteral obstruction s/p ureteral stent placement, CAD s/p CABG, PVD, SZ disorder, hyperkalemia secondary to hyperaldosteronemia (type III RTA), CRI (baseline 1.5) PSH: CABG; cystoscopy, B RPG, B ureteral stent placement 11/93; cystoscopy, B ureteral stent change 2/94 MED: vancomycin, ceftriaxone, fluconazole, lactulose, senna, kayexalate, tramadol, sarna, nystatin, paxil, lipitor, lantus, humalog ISS, keppra, labetalol, medizine, prilosec ALL: PCN PE: AVSS NAD; alert, responsive, and interactive S/NT/ND Phallus uncircumcised w/ easily retractible foreskin; meatus WNL Testes descended bilaterally and nontender Foley to gravity draining clear urine with sediment LABS: Chem7 (12/24) - 137/6.0/108/16/83/2.7/79

Table 4. Sentence retrieved from the semantic concept model, CAD

### 3.2 Distributional relational model

A distributional relational model can be defined over semantic concept distributions. For example, we may be interested in searching the intersection of concepts *Terrorist* and *Yemen*. So we could define a relational *natural join* operation ( $\bowtie$ ) over *Terrorist* and *Yemen* concept distributions to identify semantically related terms at the intersection of *Terrorist* and *Yemen*. From this result set we could predict the most semantically related entities, relations, or sentences

We may also be interested in major cities in *Afghanistan* and *Pakistan*, i.e., what are the most prominent semantically similar attributes of major cities in Afghanistan Pakistan? In this case, we could formulate a query using relational addition ('+') or UNION. Alternatively, we could use relational subtraction ('-'). For example, what is specific to COPD (Chronic Obstructive Pulmonary Disorder) that is not shared by CAD (Coronary Artery Disease)?

Defining relational operators for addition and subtraction over distributions requires some thought. Given matching terms in separate distributions, how are distributions coalesced? Our approach for defining distributional operators are summarized below:

- *Natural join* ( $\bowtie$ ) – set intersection. Only maintain matching terms in each distribution.

- *Boolean addition*: set UNION. Set semantic similarity coefficient (SSC) to the arithmetic mean of matching terms.
- *Boolean subtraction*: set SUBTRACTION. Remove terms from second operand distribution from first distribution.
- *Distributional addition*: set UNION. Set semantic similarity coefficient (SSC) to sum of matching terms, maximum 1.
- *Distributional subtraction*: set SUBTRACTION. Subtract SSC of matching terms in second operand distribution from first operand distribution, minimum 0.

Relational query operations are defined as a first-order relational algebra and can be of arbitrary complexity. Query expressions are recursively parsed into a postfix expression:

*Expression (Given):*

$((Karachi+Islamabad+Lahore)-Pakistan)+Afghanistan$

*Parse (Output):*

$[ADD, Afghanistan, SUBTRACT, Pakistan, ADD, Lahore, ADD, Islamabad, Karachi]$

The postfix expression is translated to a series of SQL statements, which are executed against concept distribution tables. The result set of the query defines a new concept that can in turn be used as any other distributional concept to predict entities, relations, or sentences.

## 4 Conclusion

We have presented an ad hoc concept modeling approach using distributional semantic models to identify and relate fine-grained entities in an online search setting. We have also presented, a novel distributional relational model for relating semantically similar concepts. The distributional concept and relational models provide a framework for future research. For example, quantitatively determining the most effective concept distribution models and distributional relational operators. What are the best architectures for scaling ad hoc distributional semantics?

## Acknowledgments

This publication and project was supported by the National Center for Advancing Translational Sciences, National Institutes of Health, through Grant Number 8UL1TR000055. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIH.

This material is based on past research sponsored by the Air Force Research Laboratory and Air Force Office of Science and Research Visiting Faculty Research and Summer Faculty Fellowship Programs (2010-2011) agreement number (13.20.02.B4488), and current research being sponsored by the Air Force Research Laboratory under agreement number (FA8750-12-1-0031). The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon.

## References

- Bollegala, D.T., Yutaka M., and Mitsuru I. (2010). Relational duality: Unsupervised extraction of semantic relations between entities on the web. Proceedings of the 19th international conference on World wide web. ACM.
- Church, K.W., Hanks, P. 1989. Word Association Norms, Mutual Information and Lexicography. Proceedings of the 27th Annual Conference of the Association of Computational Linguistics, 76-83.
- Church, K., Gale, W., Hanks, P., Hindle, D. 1991. Using Statistics in Lexical Analysis. In: Uri Zernik (ed.), *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*. New Jersey: Lawrence Erlbaum 115-164.
- Copi, I. 1998. *Introduction to Logic* (1998). Prentice Hall College Div,
- Eneko, A., et al. 2012. Semeval-2012 task 6: A pilot on semantic textual similarity. Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation. Association for Computational Linguistics.
- Finkel, J., Grenager, T., and Manning, C. 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005), pp. 363-370.
- Firth, J.R. A synopsis of linguistic theory 1930-1955. In *Studies in Linguistic Analysis*, 1968. Oxford: Philological Society. (1957). Reprinted in F.R. Palmer (ed.), *Selected Papers of J.R. Firth 1952-1959*, London: Longman.
- Kimball, R. 1996. *Data Warehouse Toolkit: Practical Techniques for Building Dimensional Data Warehouses*. Ralph, John Wiley.
- Gray, J., et al. 1997. Data Cube: A Relational Aggregation Operator Generalizing Group-By, Cross-Tab, and Sub-Totals. *Data Mining and Knowledge Discovery*, Vol. 1, Issue 1.
- Harris, Z. Distributional structure. 1954. *Word* 10 (23): 146-162.
- Mihalcea, R., Corley, C., and Strapparava, C. 2006. Corpus-based and knowledge-based measures of text semantic similarity. AAAI Press. 775-780.
- S. Robertson and S. Walker. Okapi/Keenbow at TREC-8,"NIST Special Publication 500-246, 2000.
- Sahlgren, M. The Distributional Hypothesis. 2008. *Rivista di Linguistica* 20 (1): 33-53.
- Turney, P. 2001. Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL.
- Turney, P., and Pantel, P. 2010. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research* 37.1 141-188.
- Urbain, J., Frieder, O., and Goharian, N. 2009. Passage relevance models for genomics search, *BMC Bioinformatics*, 10 (Suppl 3): S3.
- Urbain, J., and Frieder, O. 2010, Exploring contextual models in chemical patent search. *Advances in Multidisciplinary Retrieval*. Springer Berlin Heidelberg. 60-69.