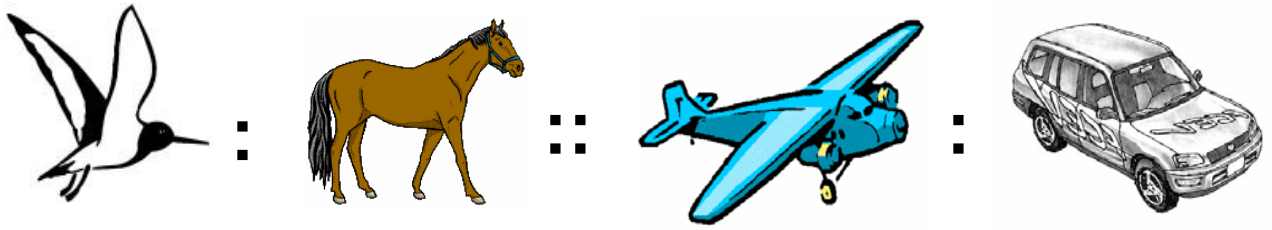
NRC · CNRC

Corpus-Based Learning of Analogies and Semantic Relations



Peter Turney
National Research Council of Canada
(work done with Michael Littman, Rutgers)

February 2004

Outline

- introduction
- motivation and applications
  - ubiquity of metaphor
  - classifying semantic relations
- related work
- solving analogy problems
  - Vector Space Model
  - experiments
- noun-modifier semantic relations
  - 30 classes of semantic relations
  - experiments
- future work
- conclusion

Introduction

Introduction

- verbal analogy has form A:B::C:D
  - A is to B as C is to D
- example
  - mason:stone::carpenter:wood
  - mason is to stone as carpenter is to wood
- analogies have been studied at least since Aristotle
  - Nicomachean Ethics, Book V, Section 3
  - but still not well understood
- idea
  - use SAT verbal analogy questions to guide research on computational approaches to analogies

SAT Analogy Question

Stem:		mason:stone
Choices:	(a)	teacher:chalk
	(b)	carpenter:wood
	(c)	soldier:gun
	(d)	photograph:camera
	(e)	book:word
Solution:	(b)	carpenter:wood

SAT Analogy Question

- answering SAT analogy questions requires knowledge of semantic relations between words
  - semantic relations are often implicit
- noun-modifier semantic relations always implicit
  - "sleeping dog"
    - noun: dog, modifier: sleeping
    - the dog is *in a state of* sleeping
  - "sleeping pill"
    - the pill *causes* sleep
  - "sleeping area"
    - an area *for* sleeping in
- want to automatically identify semantic relations

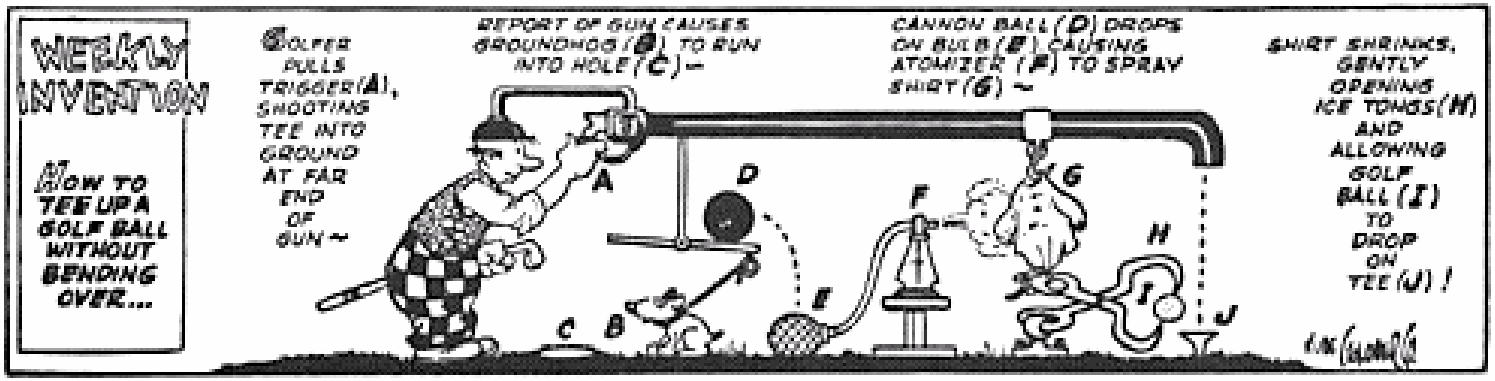
SAT Analogy Question

		word pair	semantic relation
Stem:		mason:stone	builds with
Choices:	(a)	teacher:chalk	writes with
	(b)	carpenter:wood	builds with
	(c)	soldier:gun	fights with a
	(d)	photograph:camera	is produced by a
	(e)	book:word	contains a
Solution:	(b)	carpenter:wood	builds with

Our Approach to Verbal Analogies

- corpus-based learning
  - use Web as very large corpus of text
- vector of phrase frequencies to characterize semantic relation
  - statistical “signature” of relationship
- measure similarity of semantic relations by cosine of angle between vectors
  - Vector Space Model (VSM) of Information Retrieval
  - used by all major search engines to rank hits

Motivation and Applications



RUBE GOLDBERG (TM) RGI 134

Ubiquity of Metaphor

- metaphorical language is very common
- metaphors can be understood as verbal analogies
  - He *shot down* all of my *arguments*.
    - aircraft:shoot_down::argument:criticize
  - You need to *budget* your *time*.
    - money:budget::time:schedule
  - I *gave* you that *idea*.
    - object:give::idea:communicate
  - *Life* has *cheated* me.
    - charlatan:cheat::life:disappoint
  - The Michelson-Morely *experiment gave birth* to a new physical theory.
    - mother:give_birth::experiment:initiate

Evolution of Language

- language evolution is often metaphorical
- etymology can often be understood as verbal analogy
  - *Bias:* a partiality that prevents objective consideration of an issue or situation. From the French *biais*, a slant, slope; hence, inclination to one side.
    - bias:person::slant:line
  - *Disseminate:* cause to become widely known. From the Latin *disseminare*, to scatter seed.
    - disseminate:information::scatter:seed
  - *Insult:* treat, mention, or speak to rudely. From the Latin *insultare*, to leap upon.
    - insult:character::leap_upon:body

Noun-Modifier Semantic Relations

- algorithm for SAT verbal analogies could classify noun-modifier semantic relations
- nearest neighbour supervised learning
  - given set of noun-modifier pairs, hand-labeled with semantic relations
    - training data
  - given new noun-modifier pair, unknown semantic relation
    - testing data
  - classify by looking for *most analogous* noun-modifier pair in training set
    - most analogous = nearest neighbour

Noun-Modifier Semantic Relations

- applications for noun-modifier classification
  - machine translation
    - translate “electron microscope” to another language
    - is semantic relation *purpose* or *instrument*?
  - information extraction
    - extract parties involved from news about wars
    - in “cigarette war”, relation is *topic*, not *agent*
  - word sense disambiguation
    - “plant” might be industrial plant or living plant
    - helpful to know that relation in “plant food” is *beneficiary*, not *source*

Related Work






Metaphor and Analogy

- French (2002)
  - general survey of literature on analogy and metaphor
  - all work in survey involved hand-built knowledge-bases
  - no prior work in machine learning cited
- Dolan (1995)
  - extracting knowledge automatically from a dictionary
  - discovered “conventional” metaphors
  - no systematic evaluation
- Marx et al. (2002)
  - clustering algorithm; could discover analogies between clusters of words, but not between individual words
- Veale (2003)
  - extracts analogies of form adjective:noun::adjective:noun from WordNet

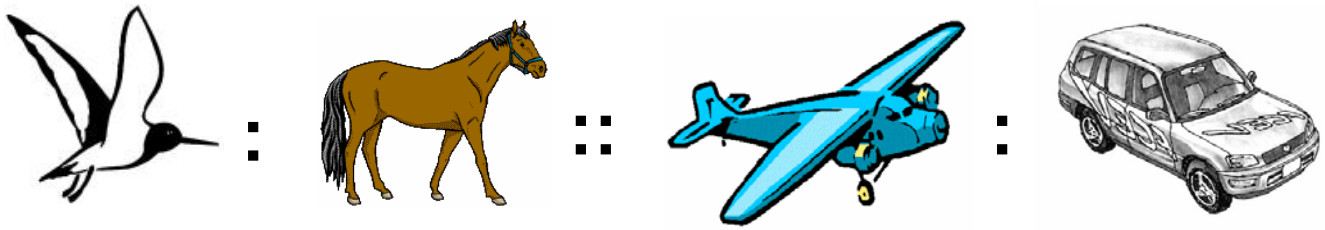
Vector Space Model

- VSM first developed in Information Retrieval
  - similarity of query to document measure by cosine of angle between query vector and document vector
  - Salton and McGill (1983), Salton (1989)
- cosine also used to measure word similarity
  - Lesk (1969), Ruge (1992), Pantel and Lin (2002)
- our use of cosine for analogies is new
  - we measure similarity of word pairs (similarity semantic relations), not individual words (similarity of concepts)

Noun-Modifier Semantic Relations

- Nastase and Szpakowicz (2003)
  - use supervised learning to classify 600 noun-modifier pairs
  - same data as we use here
  - algorithm uses features from WordNet, rather than corpus-based features
  - still in "exploratory" phase of research
- Rosario and Hearst (2001) and Rosario et al. (2002)
  - semantic relations in medical text
    - domain-specific
  - use features from MeSH (Medical Subject Headings) and UMLS (Unified Medical Language System)
- no prior corpus-based approach

Solving Analogy Problems



Solving Analogy Problems

- assign scores to candidate analogies A:B::C:D
  - for multiple-choice questions, guess highest scoring choice
- quality of analogy depends on degree of similarity between semantic relation R_1 of A:B and semantic relation R_2 of C:D
  - difficulty is that R_1 and R_2 are implicit
- attempt to learn R_1 and R_2 using unsupervised learning from a very large corpus

Vector Space Model

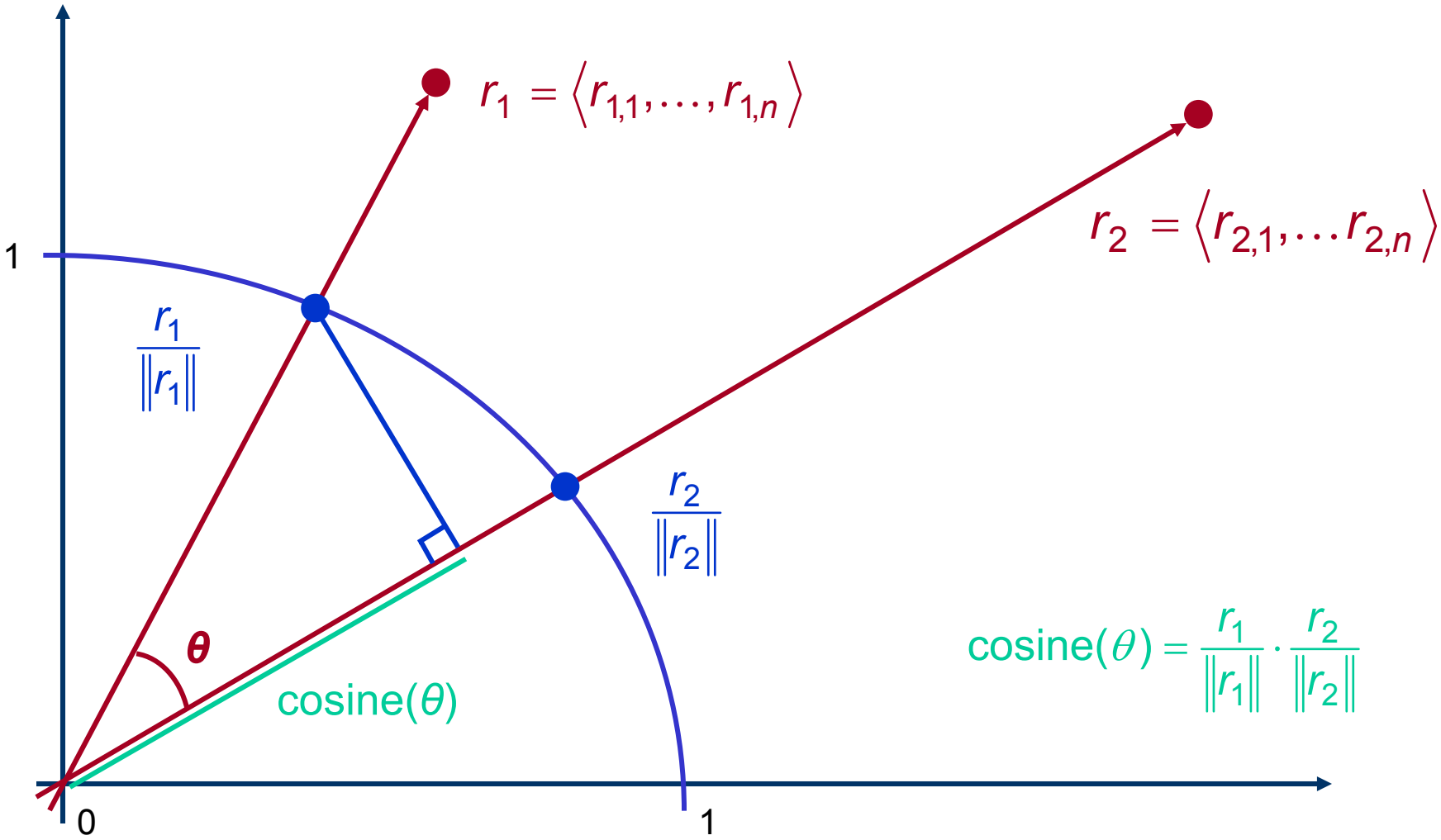
- create vectors, r_1 and r_2, that represent features of R_1 and R_2

$$r_1 = \langle r_{1,1}, \ldots, r_{1,n} \rangle \qquad r_2 = \langle r_{2,1}, \ldots r_{2,n} \rangle$$

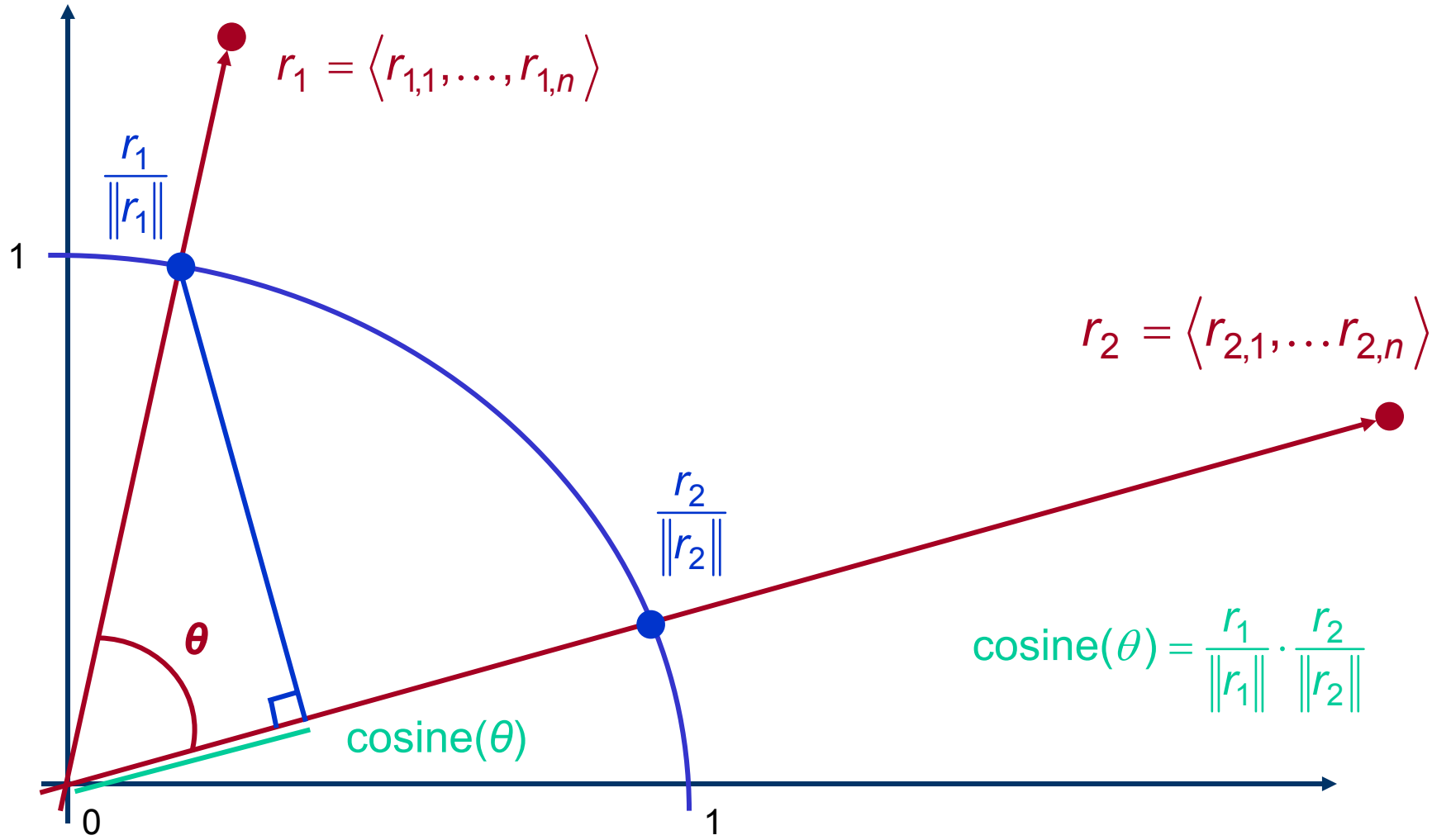
- measure the similarity of R_1 and R_2 by the cosine of the angle θ between r_1 and r_2

$$\text{cosine}(\theta) = \frac{\sum_{i=1}^{n} r_{1,i}\ r_{2,i}}{\sqrt{\sum_{i=1}^{n} (r_{1,i})^2 \ \cdot \ \sum_{i=1}^{n} (r_{2,i})^2}} = \frac{r_1 \cdot r_2}{\sqrt{r_1 \cdot r_1} \cdot \sqrt{r_2 \cdot r_2}} = \frac{r_1 \cdot r_2}{\|r_1\| \cdot \|r_2\|}$$

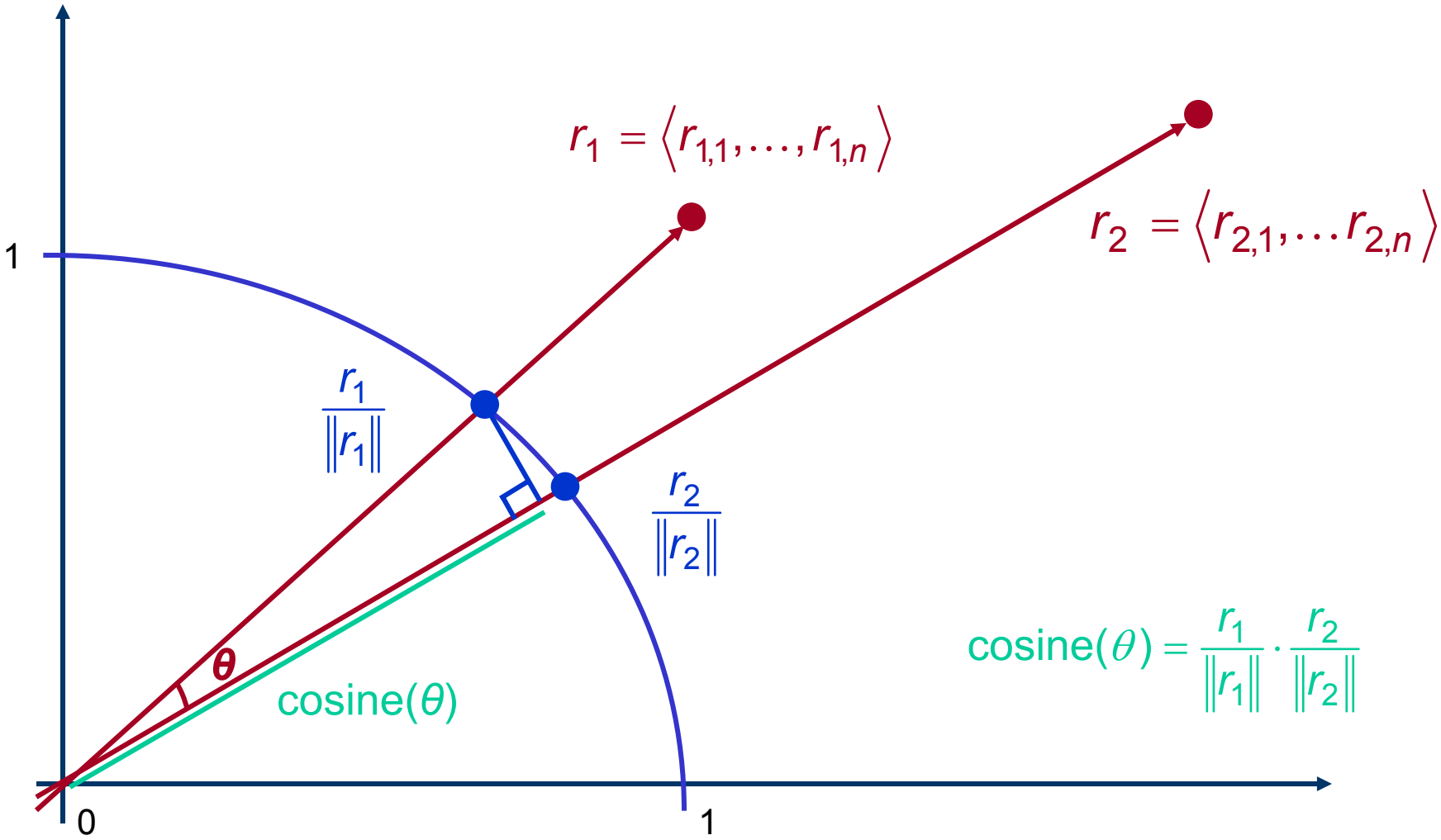
Vector Space Model



Vector Space Model



Vector Space Model



Generating Vectors

- given word pair X:Y
  - use joining term J to make phrases “X J Y” and “Y J X”
  - search web for frequencies of phrases “X J Y” and “Y J X”
  - N joining terms results in vector of 2N numbers
  - take logarithm of frequencies
- example
  - word pair: “mason:stone”
  - joining terms: “with”, “to”, “for”, “of”, …
  - search AltaVista: “mason with stone”, “stone with mason”, …
  - note number of hits (matching web pages)
  - vector of logs of hits
  - 64 joining terms; 128 elements in vector

Algorithm

- given candidate analogy A:B::C:D
  - traffic:street::water:riverbed
- generate vector for A:B and vector for C:D
  - r_1 for A:B and r_2 for C:D
- calculate cosine of angle between vectors
  - cosine(r_1, r_2)
- cosine is score for candidate analogy
  - score(traffic:street::water:riverbed) = cosine(r_1, r_2)
- similar pattern of frequencies implies small angle between vectors, implies large cosine
  - note importance of vector length normalization
  - frequent words result in longer vectors
  - we care about *direction*, not *length*

Algorithm

- example
  - traffic:street::water:riverbed

query	traffic:street	water:riverbed
“X in the Y”	615 hits	91 hits
“Y on X”	6 hits	0 hits
“Y with X”	478 hits	11 hits
“X from the Y”	136 hits	14 hits
“X when Y”	2 hits	0 hits
total	1237 hits	116 hits

Algorithm

- example
  - traffic:street::water:riverbed
  - one of the SAT questions

Stem pair:		traffic:street	Cosine
Choices:	(a)	ship:gangplank	0.31874
	(b)	crop:harvest	0.57234
	(c)	car:garage	0.68757
	(d)	pedestrians:feet	0.49725
	(e)	water:riverbed	0.69265

Evaluation Metrics

- **374 SAT analogy questions, 5 choices each**

$$\text{precision} = \frac{\text{number of correct guesses}}{\text{total number of guesses made}}$$

$$\text{recall} = \frac{\text{number of correct guesses}}{\text{maximum possible number correct}}$$

$$F = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

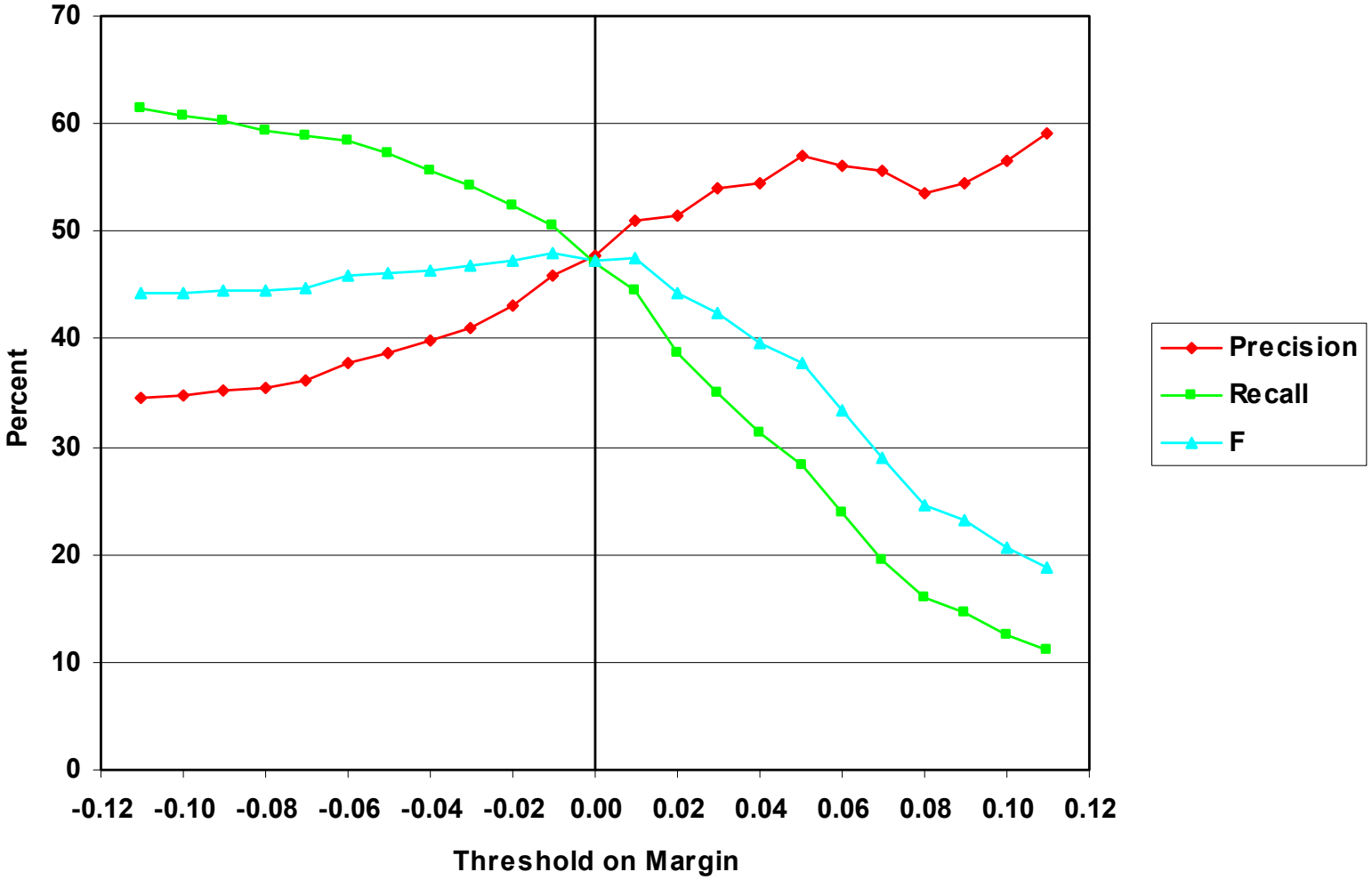
Results on 374 SAT Questions

	Number	Percent
Correct	176	47.1%
Incorrect	193	51.6%
Skipped	5	1.3%
Total	374	100.0%
Precision	176 / 369	47.7%
Recall	176 / 374	47.1%
F		47.4%

Human Performance on SAT

Note	Percent correct (no skipping)	SAT I raw score verbal	SAT I scaled score verbal	Percentile rank
	100%	78	800	100.0
	92%	70	740	98.0
	82%	60	645	88.5
	71%	50	580	74.0
College-bound mean	57%	36	504	48.0
VSM algorithm	47%	26	445	29.0
	41%	20	410	18.5
	30%	10	335	5.5
Random guessing	20%	0	225	0.5

Precision versus Recall



Generating Analogies

- SAT test is about *recognizing* analogies
- what about *generating* analogies?
- maybe reduce *generation* to *recognition*?
  - randomly create candidate word pairs
  - see which pair is most similar to given stem pair
- step towards generation
  - 374 questions, 5 skipped = 369 not skipped
  - merge all 369 correct answer pairs
  - for each of 369 stem pairs, select correct answer pair from set of 369 choices
  - how often is correct choice among top 10 choices?
  - random guessing: 10/369 = 2.7%

Generating Analogies

Rank #	Matches #	Matches %	Cumulative #	Cumulative %
1	31	8.4%	31	8.4%
2	19	5.1%	50	13.6%
3	13	3.5%	63	17.1%
4	11	3.0%	74	20.1%
5	6	1.6%	80	21.7%
6	7	1.9%	87	23.6%
7	9	2.4%	96	26.0%
8	5	1.4%	101	27.4%
9	5	1.4%	106	28.7%
10	3	0.8%	109	29.5%

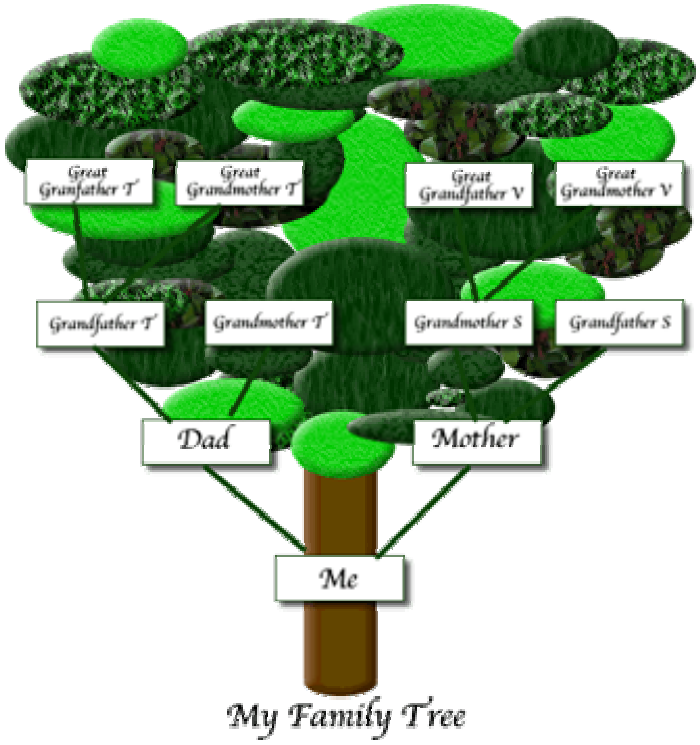
Generating Analogies

Rank	Word pair	Cosine	Question #
Stem	tourniquet:bleeding		46
1	antidote:poisoning	0.7540	308
2	belligerent:fight	0.7482	84
3	chair:furniture	0.7481	107
4	mural:wall	0.7430	302
5	reciprocate:favor	0.7429	151
6	menu:diner	0.7421	284
7	assurance:uncertainty	0.7287	8
8	beagle:dog	0.7210	19
9	canvas:painting	0.7205	5
10	ewe:sheep	0.7148	261

Execution Time

- experiments presented here required 287,232 queries to AltaVista
  - 374 analogy questions
  - × 6 word pairs per question
  - × 128 queries per word pair
  - = 287,232 queries
- as courtesy to AltaVista, inserted a five second delay between each query
  - processing 287,232 queries took about seventeen days

Noun-Modifier Semantic Relations



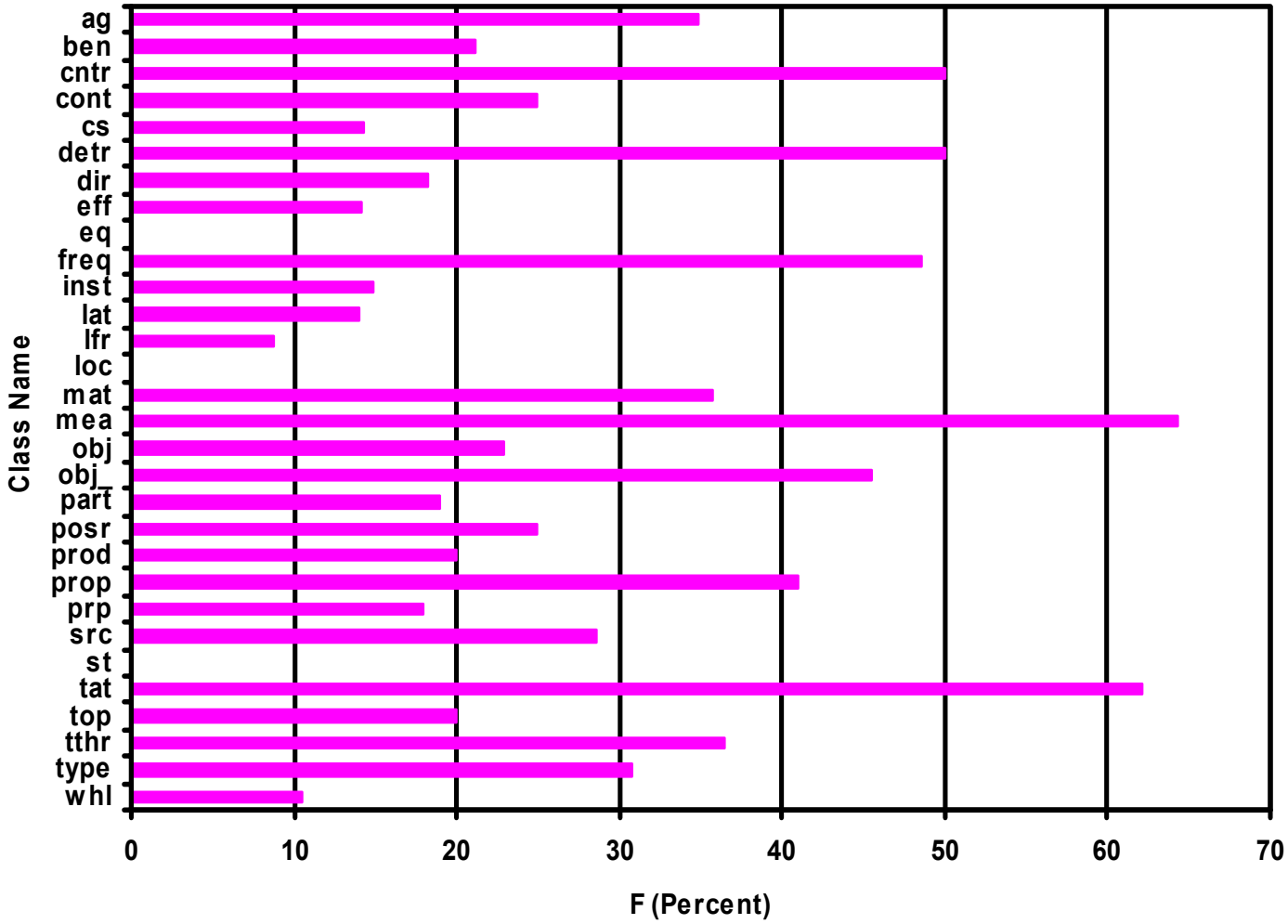
Noun-Modifier Semantic Relations

- nearest neighbour supervised learning
  - given set of noun-modifier pairs, hand-labeled with semantic relations (Nastase and Szpakowicz, 2003)
    - training data
  - given new noun-modifier pair, unknown semantic relation
    - testing data
  - classify by looking for *most analogous* noun-modifier pair in training set
    - most analogous = nearest neighbour
- nearest neighbour = cosine
  - cosine(training pair, testing pair)
  - vector of 128 elements, same joining terms as before

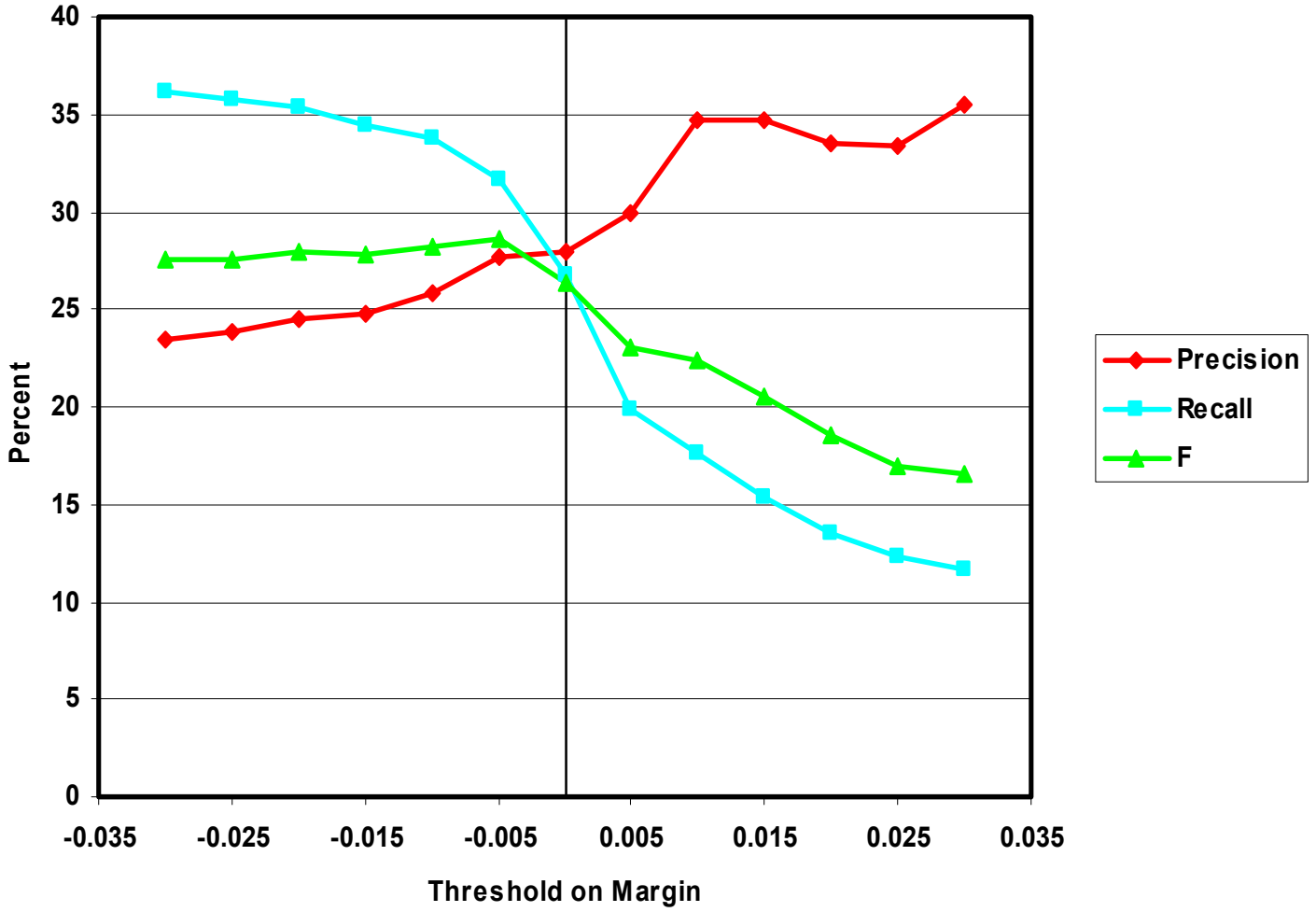
30 Semantic Relations

	Relation	Example		Relation	Example
1	cause	flu virus	16	object property	sunken ship
2	effect	exam anxiety	17	part	printer tray
3	purpose	concert hall	18	possessor	national debt
4	detraction	headache pill	19	property	blue book
5	frequency	daily exercise	20	product	plum tree
6	time at	morning exercise	21	source	olive oil
7	time through	six-hour meeting	22	stative	sleeping dog
8	direction	outgoing mail	23	whole	daisy chain
9	location	home town	24	container	film music
10	location at	desert storm	25	content	apple cake
11	location from	foreign capital	26	equative	player coach
12	agent	student protest	27	material	brick house
13	beneficiary	student discount	28	measure	expensive book
14	instrument	laser printer	29	topic	weather report
15	object	metal separator	30	type	oak tree

F for the 30 Classes



Precision versus Recall



30 Semantic Relations

- F when precision and recall are balanced
  - 26.5%
- F for random guessing
  - 3.3%
- much better than random guessing
  - but still much room for improvement
- 30 classes is hard
  - too many possibilities for confusing classes
- try 5 classes instead
  - group classes together

30 Semantic Relations

	Relation	Example		Relation	Example
1	cause	flu virus	16	object property	sunken ship
2	effect		17	part	printer tray
3	purpose		18	possessor	national debt
4	detraction	headache pill	19	property	blue book
5	frequency		20	product	
6	time at		21	source	olive oil
7	time through	six-hour meeting	22	stative	sleeping dog
8	direction	outgoing mail	23	whole	daisy chain
9	location		24	container	film music
10	location at	m	25	content	apple cake
11	location from	foreign capital	26	equative	ach
12	agent	student protest	27	material	se
13	beneficiary		28	measure	expensive book
14	instrument		29	topic	weather report
15	object	metal separator	30	type	oak tree

Causality

Temporality

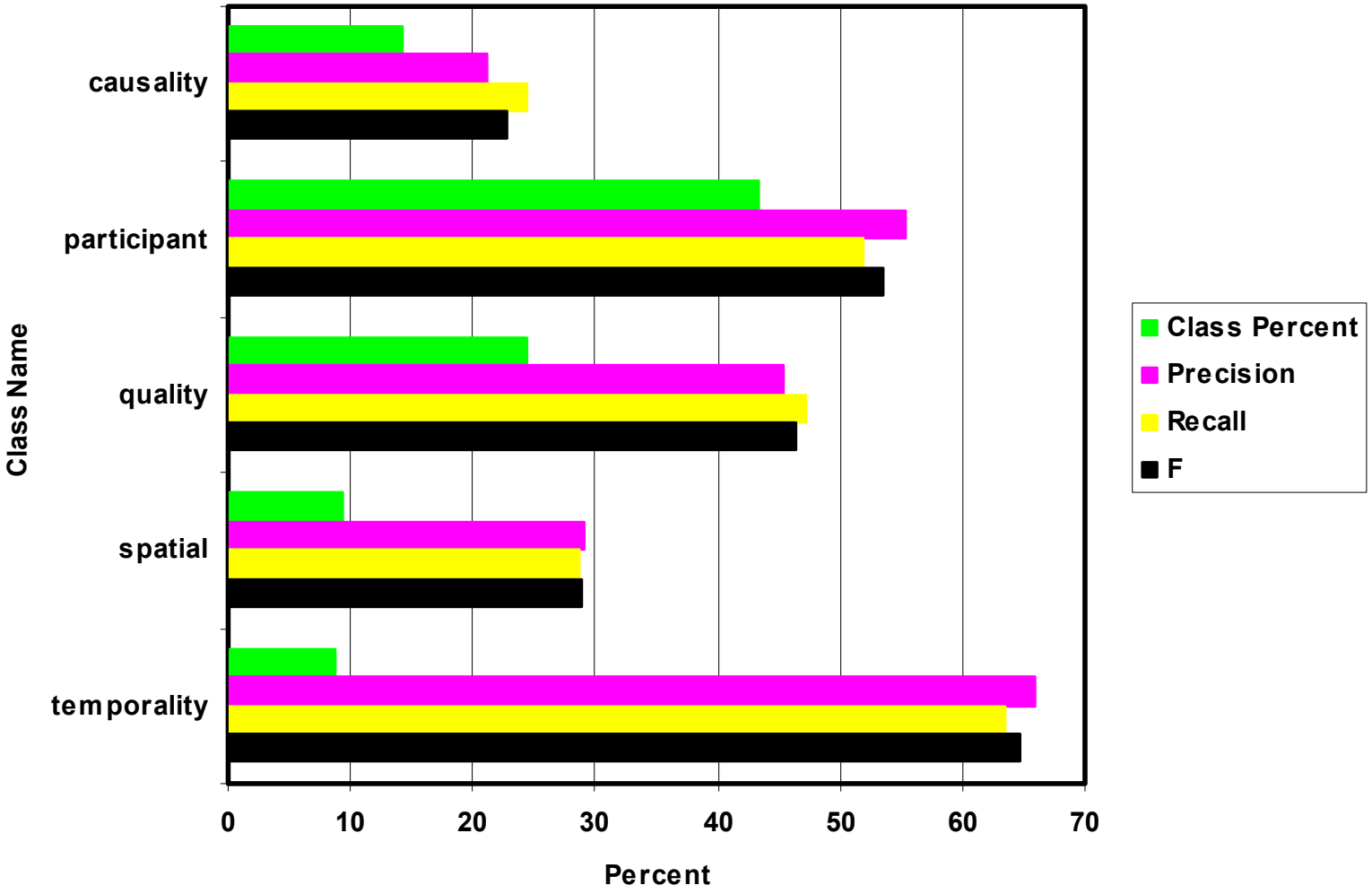
Spatial

Participant

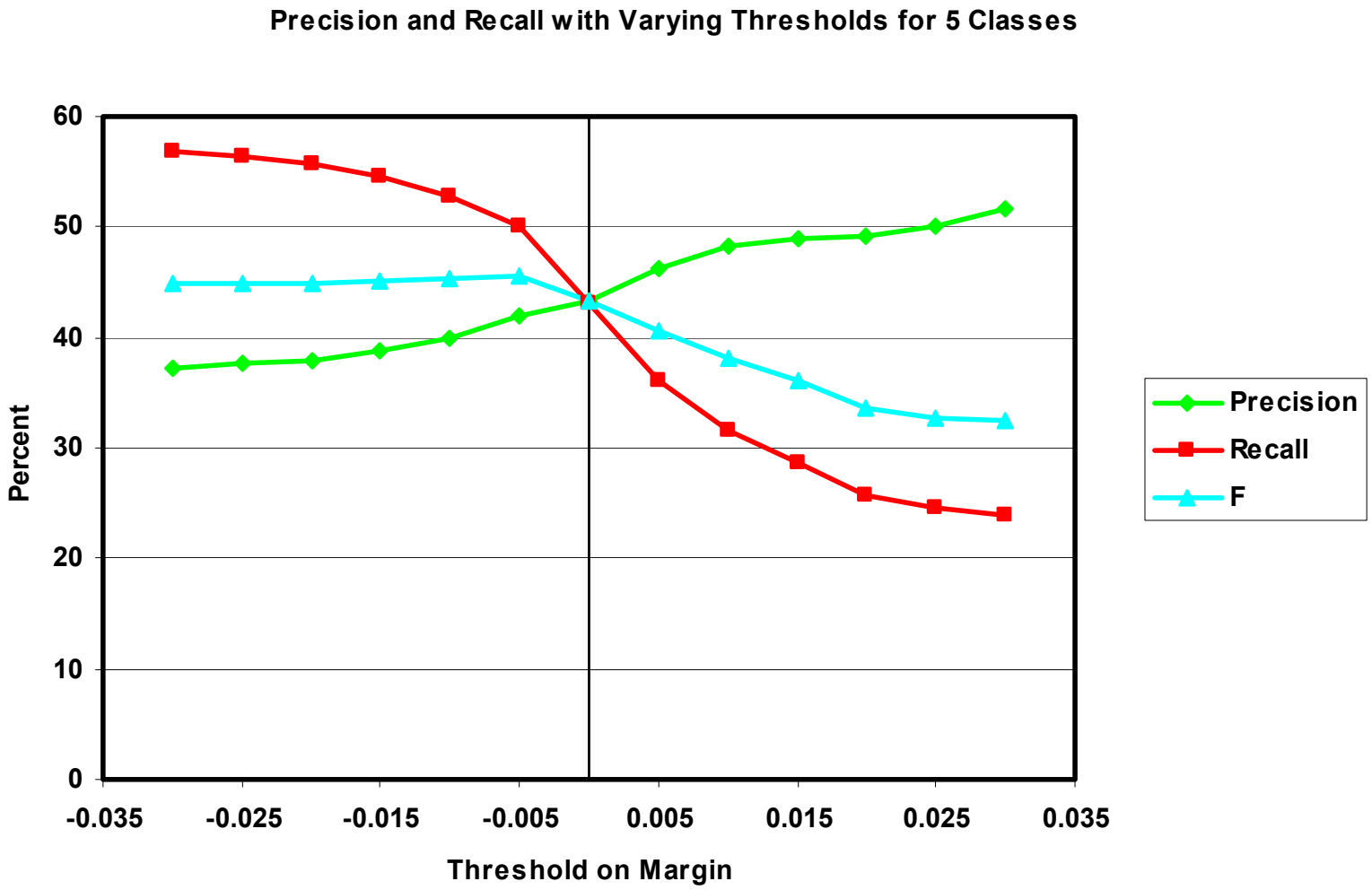
Participant

Quality

F for the 5 Classes



Precision versus Recall



5 Semantic Relations

- F when precision and recall are balanced
  - 43.2%
- F for random guessing
  - 20.0%
- better than random guessing
- better than 30 classes
  - 26.5%
  - but still room for improvement

Execution Time

- experiments presented here required 76,800 queries to AltaVista
  - 600 word pairs
  - × 128 queries per word pair
  - = 76,800 queries
- as courtesy to AltaVista, inserted a five second delay between each query
  - processing 76,800 queries took about five days

Future Work

Future Work

- much room for experimentation in choice of joining terms
  - but experiments take long time to run
- progress in hardware will allow searching local database of AltaVista size
  - recently acquired 16 CPU Beowulf Cluster
  - terabyte corpus from University of Waterloo
- variations on VSM
  - LSA, GVSM, term weighting schemes, …

Conclusion

Conclusion

- analogy and metaphor play a central role in human cognition and language
  - Lakoff and Johnson (1980), Hofstadter *et al.* (1995), French (2002)
- SAT-style analogy questions are a simple but powerful and objective tool for investigating these phenomena
  - can express many metaphors as verbal analogies
- promising first attempt at corpus-based learning of analogies
  - first objective evaluation on human-level tests
  - not yet ready for real-world applications, but soon
    - classifying semantic relations in noun-modifier pairs