Unsupervised Word Sense Discrimination By Clustering Similar Contexts

Amruta Purandare
Advisor: Dr. Ted Pedersen
07/08/2004

**Research Supported by National Science Foundation
Faculty Early Career Development Award (#0092784)**

Overview

shells exploded in a US diplomatic complex in Liberia
shell scripts are user interactive
artillery guns were used to fire highly explosive **shells**
the biggest shop on the shore for serious **shell** collectors
shell script is a series of commands written into a file that Unix executes
she sells sea **shells** by the sea shore
sherry enjoys walking along the beach and collecting **shells**
firework **shells** exploded onto usually dark screens in a variety of colors
shells automate system administrative tasks
we specialize in low priced corals, starfish and **shells**
we help people in identifying wonderful sea **shells** along the coastlines
shop at the biggest **shell** store by the shore
shell script is much like the ms dos batch file

---

shells exploded in a US diplomatic complex in Liberia
firework **shells** exploded onto usually dark screens in a variety of colors
artillery guns were used to fire highly explosive **shells**

sherry enjoys walking along the beach and collecting **shells**
we specialize in low priced corals, starfish and **shells**
we help people in identifying wonderful sea **shells** along the coastlines
shop at the biggest **shell** store by the shore
she sells sea **shells** by the sea shore
the biggest shop on the shore for serious **shell** collectors

shell script is much like the ms dos batch file
shell script is a series of commands written into a file that Unix executes
shell scripts are user interactive
shells automate system administrative tasks

Our Approach

- Strong Contextual Hypothesis
  - Sea Shells => (sea, beach, ocean, water, corals)
  - Bomb Shells => (kill, attack, fire, guns, explode)
  - Unix Shells => (machine, OS, computer, system)
- Corpus—Based Machine Learning
- Knowledge—Lean
  - Portable – Other languages, domains
  - Scalable – Large Raw Text
  - Adaptable – Fluid Word Meanings

Methodology

- Feature Selection
- Context Representation
- Measuring Similarities
- Clustering
- Evaluation

Feature Selection

- What Data ?
- What Features ?
- How to Select ?

What Data ?

- Training Vs Test
  - Training => Features
  - Test => Cluster
- Training = Test
  - Amount of Training crucial !
- Separate Training
  - Test ⊆ Training

Local Training

Pectens or Scallops are one of the few bivalve **shells** that actually swim. This is accomplished by rapidly opening & closing their valves, sending the shell backward.

Fire marshals hauled out something that looked like a rifle with tubes attached to it, along with several bags of bullets and **shells**.

If you hear a snapping sound when you're in the water, chances are it is the sound of the valves hitting together as it opens and shuts its **shell**.

Teenagers tried to make a bomb or some kind of homemade fireworks by taking the bullets and shotgun **shells** apart and collecting the black powder.

Bivalve **shells** are mollusks with two valves joined by a hinge. Most of the 20,000 species are marine including clams, mussels, oysters and scallops.

There was an explosion in one of the **shells**, it flamed over the top of the other shells and sealed in the fireworks, so when they ignited, it made it react like a pipe bomb."

These edible oysters are the most commonly known throughout the world as a popular source of seafood. The **shell** is porcelaneous and the pearls produced from these edible oysters have little value.

Global Training

U.S. researchers said sea shells may be the product of a geological accident that flooded ancient oceans with calcium, thereby diversifying marine life. Researchers at the U.S. Geological Survey have found the amount of calcium in sea water shot up between the end of the Proterozoic era (about 544 million years ago) and the early Cambrian period (515 million years ago). This increase, they suggested, allowed soft-bodied marine organisms to create hard shells or body parts from the calcium minerals. The researchers studied the chemical composition of liquids trapped in the cavities of salty rocks called halites, which provide samples of prehistoric oceans.

John Kerry is a man who knows how to keep a secret. The Democratic White House hopeful was so obsessed with making sure the name of John Edwards, his vice presidential running mate, remained under wraps until the announcement that he had vendors who printed up placards and T-shirts sign a non-disclosure agreement. Kerry himself telephoned his plane charter company at 6 p.m. on Monday night to let them in on his decision in time to have the red, white and blue aircraft's decal changed to read "Kerry-Edwards A Stronger America." Edwards did not travel to Pittsburgh to attend the rally at which his name was announced, which also might have alerted the media. After months of speculation, first reports began emerging less than 90 minutes before Kerry made his public announcement at 9 a.m.

Surface Lexical Features

- Unigrams
- Bigrams
- Co-occurrences

Unigrams

in today's world the scallop is a popular design in architecture and is well known as the shell gasoline logo if you hear a snapping sound when you're in the water chances are it is the sound of the valves hitting together as it opens and shuts its shell

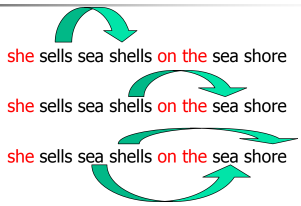
---

Bigrams

she sells sea shells on the sea shore

Selected	Rejected
sells<>sea	she<>sells
sea<>shells	shells<>on
sea<>shore	on<>the
	the<>sea

Bigrams in Window



she sells sea shells on the sea shore

she sells sea shells on the sea shore

she sells sea shells on the sea shore

Window3	Window4	window5
sells<>shells	shells<>sea	sea<>sea
		shells<>shore

Co-occurrences

Scallops are bivalve shells that actually swim

Teenagers tried to make a bomb or some kind of homemade fireworks by taking the bullets and shotgun shells apart

bivalve shells are mollusks with two valves joined by a hinge

shells can decorate an aquarium

Feature Matching

- Exact, No Stemming
- Unigram Matching
  sells doesn't match sell or sold
- Bigram Matching
  - No Window
    sea shells doesn't match sea shore sells or shells sea
  - Window
    sea shells matches sea creatures live in shells
- Co-occurrence Matching

1st Order Context Vectors

C1: if she sells shells by the sea shore, then the shells she sells must be sea shore shells and not firework shells

C2: store the system commands in a unix shell and invoke csh to execute these commands

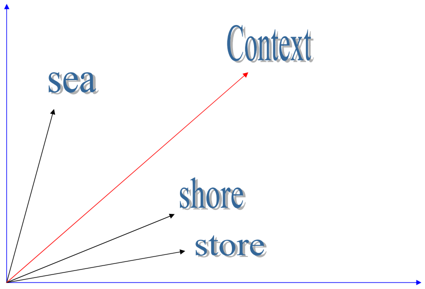
	sea	shore	system	execute	firework	unix	commands
C1	2	2	0	0	1	0	0
C2	0	0	1	1	0	1	2

2nd Order Context Vectors

The largest shell store by the sea shore

	Sells	Water	North-West	Sandy	Bombs	Sales	Artillery
Sea	18.5533	3324.98	30.520	51.7812	8.7399	0	0
Shore	0	0	29.576	136.0441	0	0	0
Store	134.5102	205.5469	0	0	0	18818.55	0
O2 context	51.021	1176.84	20.032	62.6084	2.9133	6272.85	0

2nd Order Context Vectors



Measuring Similarities

c1: {file, unix, commands, system, store}
c2: {machine, os, unix, system, computer, dos, store}

- Matching = |X Π Y|
  {unix, system, store} = 3

- Cosine = |X Π Y|/(|X|*|Y|)
  3/(√5*√7) = 3/(2.2361*2.646) = 0.5070

Cosine in Int/Real Space

	file	Unix	commands	system	store	machine	os	comp	admin	dos
C1	**2**	**1**	**3**	**1**	**2**	**0**	**0**	**0**	**0**	**0**
C2	**0**	**2**	**0**	**1**	**2**	**1**	**2**	**1**	**0**	**1**

COS(c1,c2)
= (2+1+4)/ (√19*√16)
= 7/(4.3589*4)
= 7/ 17.4356 = 0.4015

Limitations

Kill	Murder	Destroy	Fire	Shoot	Missile	Weapon
2.53	0	1.28	0	3.24	0	28.72
0	4.21	0	0.92	0	52.27	0

Burn	CD	Fire	Pipe	Bomb	Command	Execute
2.56	1.28	0	72.7	0	2.36	19.23
34.2	0	22.1	46.2	14.6	0	17.77

Latent Semantic Analysis

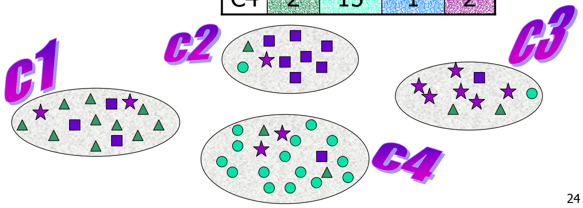
- Singular Value Decomposition
- Resolves Polysemy and Synonymy
- Conceptual Fuzzy Feature Matching
- Word Space to Semantic Space

Clustering

- UPGMA
  - Hierarchical : Agglomerative
- Repeated Bisections
  - Hybrid : Divisive + Partitional

Evaluation (before mapping)

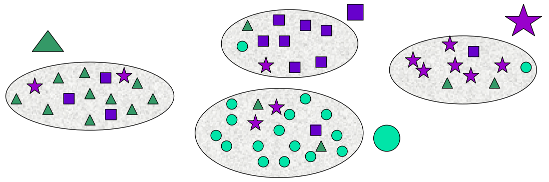
	▲	○	■	★
C1	10	0	3	2
C2	1	1	7	1
C3	2	1	1	6
C4	2	15	1	2



Evaluation (after mapping)

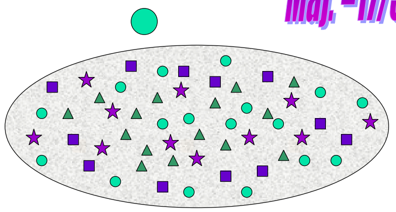
	▲	■	★	●	
C1	**10**	3	2	0	15
C2	1	**7**	1	1	10
C3	2	1	**6**	1	10
C4	2	1	2	**15**	20
	15	12	11	**17**	55

Accuracy=38/55=0.69



Majority Sense Classifier

Maj. =17/55=0.31



Data

- Line, Hard, Serve
  - 4000+ Instances / Word
  - 60:40 Split
  - 3-5 Senses / Word
- SENSEVAL-2
  - 73 words = 28 V + 29 N + 15 A
  - Approx. 50-100 Test, 100-200 Train
  - 8-12 Senses/Word

Experiment 1: Features and Measures

- Features
  - Unigrams
  - Bigrams
  - Second-Order Co-occurrences
- 1st Order Contexts
- Similarity Measures
  - Match
  - Cosine
- Agglomerative Clustering with UPGMA
- Senseval-2 Data

Experiment 1: Results
POS wise

29 NOUNS

	COS	MAT
SOC	6	7
BI	5	3
UNI	7	8

28verbs

	COS	MAT
SOC	11	6
BI	5	5
UNI	13	9

15 adjs

	COS	MAT
SOC	1	1
BI	0	0
UNI	1	0

No of words of a POS for which experiment obtained accuracy more than Majority

---

Experiment 1: Results
Feature wise

SOC

	COS	MAT
N	6	7
V	11	6
ADJ	1	1

32

BI

	COS	MAT
N	5	3
V	5	5
ADJ	0	0

18

UNI

	COS	MAT
N	7	8
V	13	9
ADJ	1	0

38

Experiment 1: Results
Measure wise

COS

	SOC	BI	UNI
N	6	5	7
V	11	5	13
ADJ	1	0	1

49

MAT

	SOC	BI	UNI
N	7	3	8
V	6	5	9
ADJ	1	0	0

39

Experiment 1: Conclusions

- Single Token Matching better
- Scaling done by Cosine helps
- 1st order contexts very sparse
- Similarity space even more sparse

Published in HLT-NAACL 2003,
Student Research Workshop

Experiment 2:
2nd Order Contexts and RBR

Pedersen & Bruce **(1st Order Contexts)**	Schütze **(2nd Order Contexts)**
PB1 **Co-occurrences, UPGMA, Similarity Space**	**SC1** **Co-occurrence Matrix, SVD RB, Vector Space**
PB2 **PB1 except RB, Vector Space**	**SC2** **SC1 except UPGMA, Similarity Space**
PB3 **PB1 with Bi-gram Features**	**SC3** **SC1 with Bi-gram Matrix**

Experiment 2: Sval2 Results
Bi-grams Vs Co-occurrences

	N	A	V	
PB1	7	1	2	Bi-gram > COC
Vs	6	4	2	Bi-gram < COC
PB3	1	1	0	Bi-gram = COC
SC1	9	3	3	Bi-gram > COC
Vs	4	1	1	Bi-gram < COC
SC3	1	2	0	Bi-gram = COC

Experiment 2: Sval2 Results
RB Vs UPGMA

	N	A	V	
PB1	9	4	1	RB > UPGMA
Vs	4	0	2	RB < UPGMA
PB2	1	2	1	RB = UPGMA
SC1	8	1	3	RB > UPGMA
Vs	2	5	0	RB < UPGMA
SC2	4	0	1	RB = UPGMA

Experiment 2: Sval2 Results
Comparing with MAJ

	N	A	V	Total
SC3 > MAJ	8	3	1	12
SC1 > MAJ	6	2	2	10
PB2 > MAJ	7	2	0	9
SC2 > MAJ	6	1	2	9
PB1 > MAJ	4	1	1	6
PB3 > MAJ	3	0	2	5

Experiment 2: Results
Line, Hard, Serve (TOP 3)

	1st	2nd	3rd
Line.n	PB1	PB3	PB2
Hard.a	PB3	PB1	SC2
Serve.v	PB3	PB1	PB2

Experiment 2: Conclusions

Nature of Data	Recommendation
Smaller Data (like SENSEVAL-2)	2nd order, RB
Large, Homogeneous (like Line, Hard, Serve)	1st order, UPGMA

Published in CONLL 2004

Experiment 4:
Local Vs Global Training

- Same as Experiment 2
- Global Training
  - Associated Press Worldstream English Service (APW)
  - Nov1994 - June2002 by LDC, UPenn
  - 539,665,000 words

Experiment 4: Results

	G	L	X
PB1	**12**	**10**	**5**
PB2	**8**	**19**	**0**
PB3	**17**	**9**	**1**
SC1	**2**	**19**	**6**
SC2	**10**	**10**	**7**
SC3	**5**	**12**	**1**

- Global helps UPGMA
- Global improves PB3 (1st order + Bigrams + UPGMA)
- Overall Local Better

Experiment 3:
Incorporating Dictionary Meanings

COCs (bomb) = *{atomic, nuclear, blast, attack, damage, kill}*
Gloss (bomb) = *{attack, denote, explosive, vessel}*
COCs+Gloss= *{atomic, nuclear, blast, attack, damage, kill, denote, explosive, vessel}*

- WordNet Glosses into Feature Vectors
- 2nd Order Contexts
- SVD (retain 2%)
- Agglomerative Clustering with UPGMA

Experiment 3: Results

SVAL2	GL>NOGL	GL=NOGL	GL<NOGL
N	17	0	12
A	9	4	2
V	17	4	7
	43	8	21

LINE, HARD, SERVE NO IMPROVEMENT

Overall Conclusions

- Smaller Data
    - 2nd Order + RBR
- Larger Local Data
    - 1st Order + UPGMA
- Global Data
    - 1st Order Bigrams, UPGMA
- Incorporating Dictionary Content

Contributions

- Systematic Comparison
    - Pedersen & Bruce (1997)
    - Schütze (1998)
- Discrimination Parameters
    - Features
    - Context Representations
    - Clustering Approaches

Contributions contd…

- Training Variations
  - Local
  - Global
- Relative Comparison
  - Raw Corpus
  - Corpus + Dictionary
- Software
  - http://senseclusters.sourceforge.net

Future Work: Refinements

- Training
  - Local + Global
  - Large Local from Newswire, BNC, Web
- Features
  - Syntactic
  - Stemming, Fuzzy Matching
- Context Representations
  - 1st order + 2nd Order
- Right #Clusters

Future Work: New Additions

- Sense Labeling
  - Unsupervised Word Sense Disambiguation
- Applications
  - Synonymy Identification
  - Name Discrimination
  - Email Foldering
  - Ontology Acquisition

Why discriminate ?

Search Google for Ted Pedersen

Software

- SenseClusters - http://senseclusters.sourceforge.net/
- Cluto - http://www-users.cs.umn.edu/~karypis/cluto/
- SVDPack - http://netlib.org/svdpack/
- N-gram Statistic Package - http://www.d.umn.edu/~tpederse/nsp.html