A Pattern Dictionary for Natural Language Processing

- Patrick Hanks and James Pustejovsky, 2005

Purpose

- Examine the current WSD resources available
  - WordNet, FrameNet, Levin classes
- Propose an alternate (radical!!?) approach to conventional WSD resources

The Problem

- Current resources focus too much on getting every possible sense
  - In words with multiple senses, generally one sense accounts for over 80% of use (Hanks 2002)
- Organization and implementation is left to the intuition of the compiler

The Solution

- Focus on patterns of verbs and valencies rather than assigning a word a meaning in isolation.
  - CPA
  - Primary implicature
  - Benchmark the likely meaning
- Skip the “exploitations of norms” – only cover normal usage
CORPUS PATTERN ANALYSIS (CPA) PROJECT AT BRANDEIS

- Aims to “link word use to word meaning in a machine-tractable way.”
- Links a pattern to a prototypical meaning
- Based on British National Corpus data
- Focus is on verbs

QUANTUM WORD SENSES

- “Words in isolation...do not have specific meaning; rather they have multifaceted potential” (64)
- Contextual patterns of word use are very regular (ignoring those usages that are for rhetorical effect – “exploitations of norms”)

CPA PROJECT PROCESS

- Take large samples of verb usage data from BNC
- Analyze valencies (subject, object, etc.)
- Assign semantic values (types and roles) to each valency
  - Semantic Type: Susan is a [[Person]]
  - Semantic Role (linked to Semantic Type): [[Person=Doctor]] [[Person=Patient]]
- Result: A dictionary linking word use to word meaning based on empirical data

CPA PROJECT “FIRE” PATTERNS

1. [[Person]] fire [[LENSET Projectile]] (off) (from [[LENSET Firearm]]) (1st [[PhysObj]]) [[ADV (Direction)]] (20%)
   DIRECTION: [[Person]] causes [[Firearm]] to discharge [[Projectile]] toward [[PhysObj=Target]]
   COMMENT: Often passive
   LENSET [[Projects-Artifact]]: bullet, round, shell, shot, volley, fire, rocket, blast, burst, valve, broadband, barrage, torpedo, grenade, missile, Exocet, blank, (Verey *’s)
   LENSET [[Firearms-Artifact]]: See Pattern 1.
   EX: But at Millington a shot was fired from a 12-gauge shotgun.
   He fired off a volley of shots from his semi-automatic rifle.
   Each time a single shot was fired.
   Leopard tortoises fired a missile at the top security Cumolin Broad jail in Belfast last night.
   One man ... fired two shots from a handguns into the officer’s chest.
2. [[Person]] fire [[Firearm]] (1st [[PhysObj]]) (in [[HumanGroup]]) [[ADV (Direction)]] (20%)
   DIRECTION: [[Person]] causes a gun or other firearm to discharge a projectile (in a given direction)
   COMMENT: This is an “unexpressed object” alternation of 1.
   EX: He ordered his men to fire.
   He more or less admits that he fired first.
A SURVEY OF OTHER RESOURCES (AND WHAT IS WRONG WITH THEM)

- Discussed:
  - WordNet, FrameNet, Levin classes

- Not discussed:
  - Electronic versions of print dictionaries
  - PropBank, NomBank, VerbNet

WORDNET (FELLBAUM, 1998)

- What it’s good for:
  - Provides a full inventory of English words

WORDNET: WHAT IT’S NOT GOOD FOR

- Problem #1:
  - Many of the synsets (synset == sense) do not actually distinguish a different sense of a word (65)

1. write, compose, pen, indite – (produce a literary work; She composed a poem; He wrote four novels)
2. write – (communicate or express by writing; Please write to me every week)
3. publish, write – (have one’s written work issued for publication; How many books did Georges Simenon write?; She published 23 books during her long career)
4. write, drop a line – (communicate (with) in writing; Write her soon, please!
5. write – (communicate by letter; He wrote that he would be coming soon)
6. compose, write – (write music; Beethoven composed nine symphonies)
7. write – (mark or trace on a surface; The artist wrote Chinese characters on a big piece of white paper)
8. write – (record data on a computer; Boot-up instructions are written on the hard disk)
9. spell, write – (write or name the letters that comprise the conventionally accepted form of (a word or part of a word); He spelled the word wrong in this letter)
10. write (create code, write a computer program; She writes code faster than anybody else.
WordNet Problem #2

- WordNet’s synsets are built into a giant hierarchical ontology
  - Unfortunately they’re not very useful.
  - The nodes don’t seem to represent semantic classes or indicate whether they fill particular slots in verb argument structure.

The superordinates of “Write”

1. create verbally
2. communicate, intercommunicate
3. create verbally
4. correspond
5. create verbally
6. make, create (which is itself a superordinate of “create verbally”)
7. trace, draw, line, describe, delineate
8. record, tape
9. [No superordinate]
10. create code, write a computer program

The superordinates of “Write”

1. create verbally
2. communicate, intercommunicate
3. create verbally
4. correspond
5. create verbally
6. make, create (which is itself a superordinate of “create verbally”)
7. trace, draw, line, describe, delineate
8. record, tape
9. [No superordinate]
10. create code, write a computer program

The superordinates of “Write”

1. create verbally
2. communicate, intercommunicate
3. create verbally
4. correspond
5. create verbally
6. make, create (which is itself a superordinate of “create verbally”)
7. trace, draw, line, describe, delineate
8. record, tape
9. [No superordinate]
10. create code, write a computer program
**FRAMENET**
- FrameNet uses corpus data for its frames, but “relies on the intuitions of its researchers to populate each frame with words” (67).
- Some frames overlap redundantly
- Some entries are marked as complete when only rare senses have been covered
  - ex.: Spoil
    - Covers rotting and desiring, but not “spoil a child,” one of the most common usages

**LEVIN CLASSES**
- “Many of Levin’s assertions about the behaviour (and sometimes also the meaning) of particular verbs in her verb classes are idiosyncratic or simply wrong” (68).
- Levin’s comments on diathesis alternations apply to some but not all members of the classes.
- Deliberately omits verbs that take sentential complements.
  - “Tempt” only listed as “amuse”.
    - Common usage “We were tempted to laugh” omitted.
- Covers 3,000 verbs, and leaves out many major ones
- Not all senses of verbs that are included are covered
- Yet Levin classes are still widely cited in the NLP community

**IMPROVEMENTS BY THE CPA PROJECT**
- Levin discusses **diathesis alternations** of verbs
  - CPA covers **semantic alternation** as well.
    - Ex.: For the medical sense of “treat”
      - [[Person=Doctor]] alternates with [[Medicament]]
      - [[Person=Patient]] alternates with [[Injury]] and [[Ailment]]
- CPA also covers **lexical alternation**.
  - “Grasping/clutching at straws”
A DIFFERENT WAY OF VIEWING MEANING

- Levin claims that the behavior of a verb is largely determined by its meaning.
  - Is this useful?

- Word behavior is observable whereas word meaning is “imponderable, a matter of introspection, conjecture, and unsubstantiated assertion” (68).

- Flip that statement around and you have a sound empirical starting point

CONTEXT IN CPA PROJECT

- The semantic value of a verb’s valencies can disambiguate word-sense.
  - “Fire a gun” vs “Fire a person”

- What about “shoot a person”?  
  - Camera or gun?

- Thus the CPA Project also specifies relevant, recurrent clues  
  - “Shoot a person dead”  
  - “Shoot and injure a person”

- A central group of clues is recorded for each verb.

OTHER CPA PROJECT METHODS

- Also records relative frequency of each pattern to provide a default basis for likelihood of meaning

- Goal:  
  - Build up an inventory of normal syntagmatic behavior for use in WSD, message understanding, natural text generation, etc.

RELEVANCE TO PROJECT

- We’re using the CPA resource described here to cluster verbs with tools built by Octavian Popescu

  Part I
  - Get things installed on other things (Daniel’s bit)
  - Map OntoNotes Named Entities onto SUMO types

  Part II
  - Cluster verbs with a hierarchical Dirichlet process (ask Daniel about that bit)
  - Go through final clusters and note errors and types of errors