Word Sense Annotation of Polysemous Words by Multiple Annotators

Rebecca J. Passonneau, Ansaf Salleb-Aoussi, Vikas Bhardwaj
Columbia University, New York, NY
Nancy Ide
Vassar College, Poughkeepsie, NY
Outline

• Word senses
• MASC word sense annotation
• Interannotator agreement: word/pos dependent
• Exploring the data
  – InterSense Similarity Measures (ISSM)
  – Association rules among annotators
• Future work
Word Senses: Theoretical Issues

• Synchronic variation
  – Selected for by the sentence/utterance context
  – Generative (Pustejovsky)
  – Many contexts are essentially the same (Kilgariff)

• Diachronic variation
  – Changes in senses over time
  – Changes in sense frequency over time

• Situational/sociolinguistic variation
  – Different usage likelihoods in distinct corpora
  – Differences across language users

5/21/2010  LREC 2010 Malta
Annotation Issues

• How much context is enough?
• How much training for annotators?
• How much agreement is possible among annotators? (Fellbaum; Ng; Pedersen; Palmer)

• Sense inventories
  Con: Arbitrary
  Con: No degrees of specificity, e.g., “a long chapter”
  – Other methods (Erk & McCarthy, ratings of all senses)
  Pro: Understandable
  Pro: Convenient annotation labels
  - Explore label usage among many annotators
MASC Word Sense Annotation

• MASC Corpus (May release): Ide et al. 2010 ACL

• Word Sense annotation goals:
  – Harmonize WordNet/FrameNet senses
  – Provide manually annotated data for supervised WSD

• Five rounds to date, a sixth underway
  – MASC subcorpus from OANC: open, heterogeneous
  – WordNet sense labels on 1000 sentences/word
  – Sentences in context (annotator can adjust)
  – Trained annotators at Vassar, Columbia
  – Annotation tool: SATANiC
Round 2.2

- 10 polysemous words (9.5 senses per word on avg.)
- Balanced for POS
  - 3 Adj
  - 3 Nouns
  - 4 Verbs
- Sample of 100 sentences
  - Three Columbia undergraduates
  - Three Vassar undergraduates
  - Same training, same annotation tool
- Interannotator agreement: Krippendorff’s Alpha
  - Wide range of agreement results
  - Word dependent
## Interannotator Agreement

| Word-POS | Senses in WN | Senses Assigned | Annotators | Alpha  |
|----------|--------------|-----------------|------------|--------|
| LONG-J   | 9            | 4               | 6          | 0.67   |
| FAIR-J   | 10           | 6               | 5          | 0.54   |
| QUIET-J  | 6            | 5               | 6          | 0.49   |
| TIME-N   | 10           | 8               | 5          | 0.68   |
| WORK-N   | 7            | 7               | 5          | 0.62   |
| LAND-N   | 11           | 9               | 6          | 0.49   |
| SHOW-V   | 12           | 10              | 5          | 0.46   |
| TELL-V   | 8            | 8               | 6          | 0.46   |
| KNOW-V   | 11           | 10              | 6          | 0.37   |
| SAY-V    | 11           | 10              | 6          | 0.37   |
Observations on IA

• Agreement is less good on V than N and J
• Most senses are used; sense frequency does not correlate exactly with WN predictions
• Agreement does not degrade as number of senses increases
• Within each part-of-speech, IA varies with no discernible cause other than the word itself
• Words differ with respect to concreteness (e.g., “long” versus “fair” – SEW 2009)
Intersense Similarity

• Hypothesis: words more confusable senses have lower IA

• Measure sense relatedness: Lesk Similarity (Banerjee & Pedersen 2002)

• \( ISM_w(S_1,S_2) = \text{Lesk similarity}(S_1,S_2) \)

• Confusion threshold CT for \( w \):
  \[
  CT_w = \mu ISM_w + \sigma ISM_w
  \]

• Only partial correlation (for adjectives \( \varrho = 0.73 \), but very few datapoints; overall correlation: \( \varrho = 0.59 \))
### ISMs Round 2 Words

| Word-POS | Pairs of Senses | Alpha | % > CT |
|----------|-----------------|-------|--------|
| LONG-J   | 36              | 0.67  | 0.17   |
| FAIR-J   | 45              | 0.54  | 0.18   |
| QUIET-J  | 15              | 0.49  | 0.20   |
| TIME-N   | 45              | 0.68  | 0.11   |
| WORK-N   | 21              | 0.62  | 0.14   |
| LAND-N   | 54              | 0.49  | 0.07   |
| SHOW-V   | 28              | 0.46  | 0.07   |
| TELL-V   | 66              | 0.46  | 0.12   |
| KNOW-V   | 55              | 0.37  | 0.18   |
| SAY-V    | 55              | 0.37  | 0.09   |
Association Rules

• Association rules express relations among instances in a dataset, based on their attributes (Agrawal et al. 1993; Borgelt’s Apriori)

• An association rule is an expression C1 → C2, where C1 and C2 express conditions on features describing the instances

Measuring strength of association rules:

• Supp(C) is the fraction of instances satisfying C
• Supp(C1 → C2) = Supp(C1)
• Conf(C1 → C2) = Supp(C1 ∧ C2)/Supp(C1)
Association Rules: Annotators & Senses

- The word sense data is a 3D matrix of instances, annotators, senses
- Flatten the data to a 2D form with Annotator_SenseLabel as an attribute
- Mine association rules among annotators’ choices of senses
- Mining agreement on ‘time’ (IA=0.68): strongest rules for sense 3
  - 101.S3 → 105.S3 with 36% supp. and 77.8% conf.
  - 105.S3 → 101.S3 with 34% supp. and 82.4% conf.
Long (IA=0.67)

| \(\text{Ann}_i.S_j\) | \(\text{Ann}_m.S_n\) | Supp | Conf |
|----------------------|----------------------|------|------|
| Long                 | Long                 |      |      |
| 102.Coll             | 108.S1               | 60.0 | 55.0 |
| 108.S2               | 102.Coll             | 37.0 | 89.2 |

• If 102 assigns a collocation, 108 assigns sense 1 primarily temporal sense; being or indicating a relatively great or greater than average duration or passage of time or a duration as specified: "a long life"; "a long boring speech"; . . .

• If 108 assigns sense 2, 102 assigns a collocation primarily spatial sense; of relatively great or greater than average spatial extension or extension as specified: "a long road"; "a long distance"
**Fair (IA=0.54)**

| $Ann_i.S_j \rightarrow$ | $Ann_m.S_n$ | Supp | Conf |
|--------------------------|-------------|------|------|
|                          | Fair        |      |      |
| 107.S2                   | 102.S1      | 56.0 | 28.6 |
| 102.S1                   | 107.S2      | 31.0 | 51.6 |

- If 107 assigns sense 2, 102 assigns sense 1
- If 102 assigns sense 1, 107 assigns sense 2

**Sense 1:** Free from favoritism or self-interest or bias or deception; conforming with established standards or rules: "a fair referee"; "fair deal"; "on a fair footing"; "a fair fight"; "by fair means or foul"

**Sense 2:** Not excessive or extreme: "a fairish income"; "reasonable prices"
### Quiet (IA=0.49)

| Ann\textsubscript{i}.S\textsubscript{j} | Ann\textsubscript{m}.S\textsubscript{n} | Supp | Conf |
|-------------------------------------|-------------------------------------|------|------|
| 107.S3                              | 103.S1                              | 58.0 | 34.5 |
| 103.S1                              | 107.S3                              | 36.0 | 55.6 |

- If 107 assigns sense 3, 103 assigns sense 1
- If 103 assigns sense 1, 107 assigns sense 3

**Sense 1:** characterized by an absence or near absence of agitation or activity: "a quiet life"; "a quiet throng of onlookers"; "quiet peace-loving people"; "the factions remained quiet for almost 10 years"

**Sense 3:** not showy or obtrusive: "clothes in quiet good taste"
Conclusions and Future Work

• Good agreement among annotators on word senses can be achieved for polysemous words

• Two annotators may be insufficient

• Disagreements can include systematic patterns of difference due to, e.g., subjectivity in meaning

• Future work:
  – Measurement (LAW IV)
    • Drop outliers (e.g., 102 for “long”)
    • Identify confusable senses
    • Identify systematic differences among subsets of annotators
    • Compare trained and a larger number of untrained annotators
  – Allow annotators to assign multiple senses