ECE8813
Statistical Natural Language Processing

Lecture 11: Word Sense Disambiguation

Chin-Hui Lee

School of ECE, Georgia Tech

Atlanta, GA 30332, USA

chl@ece.gatech.edu

Problem Definitions

- Word sense disambiguation
  - Goal: to determine which of the multiple senses of an ambiguous word is invoked in a particular use of the word
  - Word sense is a long-standing problem in traditional language community as a vehicle for deep message "understanding" as opposed to "shallow understanding" done in statistical language processing
- Lexical acquisition
  - Goal: to develop algorithms and statistical techniques for filling in holes in machine-readable dictionaries by looking at co-occurrence patterns of words in large text corpora
  - Lexical definitions depend on the particular grammars and their corresponding languages being used (multilinguality)
- Only consider statistical techniques

An Example: Bank (18 Senses)

- (Noun-883) depository financial institution, bank, banking concern, banking company -- (a financial institution that accepts deposits and channels the money into lending activities; "he cashed a check at the bank"; "that bank holds the mortgage on my home")
- (Noun-99) bank -- (sloping land (especially the slope beside a body of water); "they pulled the canoe up on the bank"; "he sat on the bank of the river and watched the currents")
- (Verb-2) bank -- (tip laterally; of boats and aircraft)
- (Verb-1) bank -- (enclose with a bank; "bank roads")

WordNet: An Electronic Lexical Database

- WordNet® is a large lexical database of English, developed under the direction of George A. Miller
- Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations. The resulting network of meaningfully related words and concepts can be navigated with the browser
- WordNet is freely and publicly available for download (http://wordnet.princeton.edu/obtain)
- WordNet's structure makes it a useful tool for natural language processing and computational linguistics
- *WordNet: An Electronic Lexical Database* (MIT Press)

Overview of Our Discussion

- **Methodology**
  - **Supervised Disambiguation:** based on a labeled training set
  - **Dictionary-Based Disambiguation:** based on lexical resources such as dictionaries and thesauri
  - **Unsupervised Disambiguation:** based on unlabeled corpora

Methodological Preliminaries

- **Supervised vs Unsupervised Learning**: In supervised learning (classification), the sense label of each word occurrence is provided in the training set; whereas, in unsupervised learning (clustering), it is not provided
- **Pseudowords**: used to generate artificial evaluation data for comparison and improvements of text-processing algorithms, e.g., replace each of two words (e.g., *bell* and *book*) with a psuedoword (e.g., *bell-book*)
- **Upper and Lower Bounds on Performance**: to find out how well an algorithm performs wrt the difficulty of the task
  - **Upper:** human performance
  - **Lower:** baseline using highest frequency alternative (best of 2 vs 10)

Supervised Disambiguation

- **Training set**: exemplars where each occurrence of the ambiguous word *w* is annotated with a semantic label. This becomes a statistical classification problem; assign *w* some sense s_k in context c_l
- **Approaches**:
  - Bayesian Classification: the context of occurrence is treated as a bag of words without structure, but it integrates information from many words in a context window
  - Information Theory: only looks at the most informative feature in the context, which may be sensitive to text structure
  - There are many more approaches (see Chapter 16 or a text on Machine Learning (ML)) that could be applied

Supervised Bayesian Classification

- **(Gale et al, 1992):** look at the words around an ambiguous word in a large context window. Each content word contributes potentially useful information about which sense of the ambiguous word is likely to be used with it. The classifier does no feature selection; it simply combines the evidence from all features, assuming they are independent.
- **Bayes decision rule**: Decide *s'* if P(*s'*|*c*) > P(s_k|*c*) for $s_k \neq s'$
  - Optimal because it minimizes the probability of error; for each individual case it selects the class with the highest conditional probability (and hence lowest error rate)
  - Error rate for a sequence will also be minimized

Bayesian Classification Framework

- We do not usually know the posterior probability, P(s_k|c), but we can use Bayes' Rule to compute it:
  - P(s_k|c) = (P(c|s_k)/P(c)) × P(s_k)
  - P(s_k) is the prior probability of s_k, i.e., the probability of instance s_k without any contextual information
  - When updating the prior with evidence from context (i.e., P(c|s_k)/P(c)), we obtain the posterior probability P(s_k|c)
  - If all we want to do is select the correct class, we can ignore P(c). Also use logs to simplify computation
- Assign word w sense s' = argmax_{s_k} P(s_k|c)

Computing Posterior Probability

- We want to assign the ambiguous word *w* to the sense *s'*, given context *c*, where:

$$s' = \underset{s_k}{\arg\max}\; P(s_k/c)$$

$$= \underset{s_k}{\arg\max}\; \frac{P(c/s_k)}{P(c)} P(s_k)$$

$$= \underset{s_k}{\arg\max}\; [\log P(c/s_k) + \log P(s_k)]$$

$$= \underset{s_k}{\arg\max}\; P(c/s_k)\, P(s_k)$$

Naïve Bayes Approach

- Naïve Bayes:
  - is widely used in ML due to its ability to efficiently combine evidence from a wide variety of features
  - can be applied if the state of the world we base our classification on can be described as a series of attributes
  - in this case, we describe the context of *w* in terms of the words v_j that occur in the context
- Naïve Bayes assumption:
  - The attributes used for classification are conditionally independent: $P(c|s_k) = P(\{v_j|\ v_j \text{ in } c\}|s_k) = \Pi_{v_j \text{ in } c} P(v_j \mid s_k)$
  - Two consequences:
    - **The structure and linear ordering of words is ignored: bag of words model**
    - **The presence of one word is independent of another, which is clearly untrue in text**

Naïve Bayes Disambiguation

- Assumption: attributes (clues) used for word description are all conditionally independent (?)
  - Although the Naïve Bayes assumption is incorrect in the context of text processing, it often does quite well

$$P(c \mid s_k) = P(\{v_j \mid v_j \text{ in } c\} \mid s_k) = \prod_{v_j \text{ in } c} P(v_j \mid s_k)$$

- Bayes Decision Rule

$$\text{Decide } \hat{s} \text{ if } \hat{s} = \operatorname{argmax}_{s_k} [\sum_{v_j \text{ in } c} \log P(v_j \mid s_k) + \log P(s_k)]$$

  - Applying ML estimation for the probabilities

$$P(v_j \mid s_k) = \frac{\mathrm{C}(v_j, s_k)}{\sum_l \mathrm{C}(v_l, s_k)}, P(s_k) = \frac{\mathrm{C}(s_k \mid w)}{\sum_k C(s_k \mid w)}$$

Training and Disambiguation Algorithms

Training:

```
for all senses $s_k$ of $w$ do
    for all $v_j$ in vocabulary do
        $P(v_j|s_k) = C(v_j,s_k)/C(s_k)$
    end
end
for all senses $s_k$ of $w$ do
    $P(s_k) = C(s_k)/C(w)$
end
```

Disambiguation:

```
for all senses $s_k$ of $w$ do
    $\text{score}(s_k) = \log P(s_k)$
    for all $v_j$ in context window $c$ do
        $\text{score}(s_k) = \text{score}(s_k) + \log P(v_j|s_k)$
    end
end
choose $\text{argmax}_{s_k}$ score $(s_k)$
```

(Gale, Church, and Yarowsky) obtain 90% correct disambiguation on 6 ambiguous nouns in Hansard corpus using this approach (e.g., *drug* as a medication vs. illicit substance

An Information-Theoretic Approach

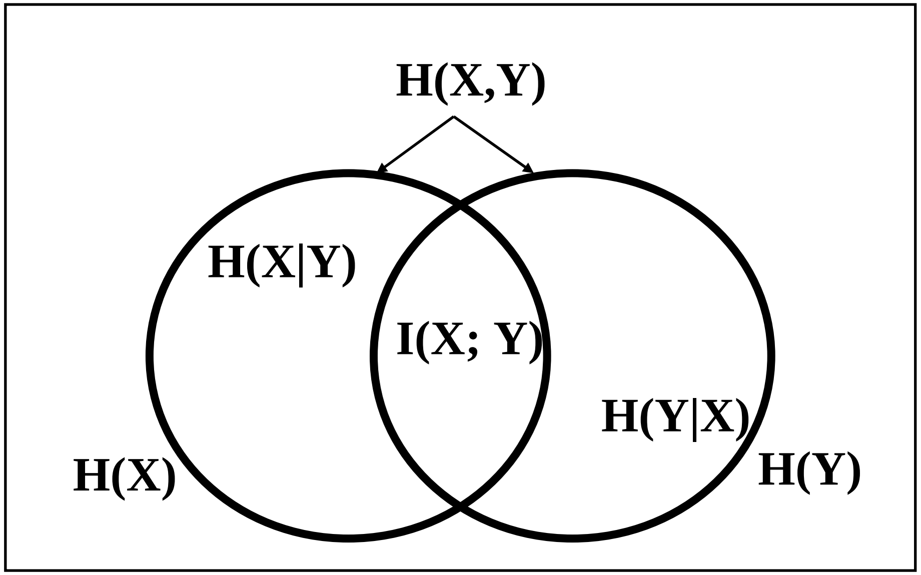
- (Brown et al., 1991) attempt to find a single contextual feature that reliably indicates which sense of an ambiguous word is being used
- For example, the French verb *prendre* has two different readings that are affected by the word appearing in object position (*mesure* → *to take*, *décision* → *to make*), but the verb *vouloir*'s reading is affected by tense (present → *to want*, conditional → *to like*)
- To make good use of an informant, its values need to be categorized as to which sense they indicate (e.g., *mesure* → *to take*, *décision* → *to make*); Brown et al. use the Flip-Flop algorithm to do this

An Information-Theoretic Framework

- Let $t_1, \ldots, t_m$ be translations for an ambiguous word and $x_1, \ldots, x_n$ be possible values of the indicator
- The Flip-Flop algorithm is used to disambiguate between the different senses of a word using mutual information:
  - $I(X;Y)=\Sigma_{x \in X} \Sigma_{y \in Y}\, p(x,y) \log p(x,y)/(p(x)p(y))$
  - See Brown et al. for an extension to more than two senses
- The algorithm works by searching for a partition of senses that maximizes the mutual information. The algorithm stops when the increase becomes insignificant

Mutual Information

- I(X; Y)=***H(X)-H(X|Y)=H(Y)-H(Y|X)***, the ***mutual information*** between *X* and *Y*, is the reduction in uncertainty of one random variable due to knowing about another, or, in other words, the amount of information one random variable contains about another



Mutual Information (Cont.)

$$I(X;Y) = H(X) - H(X \mid Y) = H(Y) - H(Y \mid X)$$

- *I(X; Y)* is symmetric, non-negative measure of the common information of two variables
- Some see it as a measure of dependence between two variables, but better to think of it as a measure of independence
  - *I(X; Y)* is 0 only when *X* and *Y* are independent: *H(X|Y)=H(X)*
  - For two dependent variables, I grows not only according to the degree of dependence but also according to the entropy of the two variables
- *H(X)=H(X)-H(X|X)=I(X; X)* ⇒ Why entropy is called self-information

Flip-Flop Algorithm

```
Find random partition $P=\{P_1, P_2\}$ of translations $\{t_1, \ldots, t_m\}$
while (there is a significant improvement) do
    find partition $Q=\{Q_1, Q_2\}$ of indicators $\{x_1, \ldots, x_n\}$ that maximizes $I(P;Q)$
    find partition $P=\{P_1, P_2\}$ of translations $\{t_1, \ldots, t_m\}$ that maximizes $I(P;Q)$
end
```

$$I(X; Y) = \mathrm{Sum}_{x\in X}\ \mathrm{Sum}_{y\in Y}\ p(x,y) \log (p(x,y)/(p(x)p(y)))$$

- Mutual information increases monotonically in the Flip-Flop algorithm, so it is reasonable to stop when there is only an insignificant improvement

An Example

- Assume we want to translate *prendre* based on its object and have {t_1, …, t_m}={*take, make, rise, speak*} and {x_1, …, x_n} ={*mesure, note, exemple, décision, parole*}, and that *prendre* is used as *take* when occurring with the objects *mesure, note,* and *exemple*; otherwise used as *make, rise,* or *speak*
  - Suppose the initial partition is P_1={*take, rise*} and P_2={*make, speak*}
  - Then choose partition of Q of indicator values that maximizes I(P;Q), say Q_1={*mesure, note, exemple*} and Q_2={*décision, parole*} (selected if the division gives us the most information for distinguishing translations in P_1 from translations in P_2)
  - *prendre la parole* is not translated as *rise to speak* when it should be; repartition as P_1={*take*} and P_2={*rise, make, speak*}, and Q as previously. This is always correct for *take* sense
  - To distinguish among the others, we would have to consider more than two senses

Flip-Flop Algorithm (Cont.)

- A simple exhaustive search for the best partition of French translations and indicator values would take exponential time
- The Flip-Flop algorithm is a linear time algorithm based on Brieman et al.'s (1984) splitting theorem
  - Run the algorithm for all possible indicators and choose the indicator with the highest mutual information
  - Once the indicator and partition of its values is determined, disambiguation is simple:
    - For each ambiguous word, determine the value x_i of the indicator
    - If x_i is in Q_1, assign sense 1; if x_i is in Q_2, assign sense 2
- Brown et al. (1991) obtained a 20% improvement in MT system using this approach (translations used as senses)

Dictionary-Based Disambiguation

- If we have no information about the senses of specific instances of words, we can fall back on a general characterization of the senses provided by a lexicon
- We will be looking at three different methods:
  - Disambiguation with sense definitions in a dictionary (Lesk, 1986)
  - Thesaurus-based disambiguation (Walker, 1987 & Yarowsky, 1992)
  - Disambiguation based on translations in a second-language corpus (Dagan and Itai, 1994): not discussed in class
- Also, we will learn about how a careful examination of the distributional properties of senses can lead to significant improvements in disambiguation
  - Ambiguous words tend to be used with only one sense in a given discourse with a given collocate

Sense Definition Disambiguation

- (Lesk, 1986) uses the simple idea that a word's dictionary definitions are likely to be good indicators for the senses they define
- For example, the words in definitions associated with the word *cone* (seed bearing cone versus ice cream containing cone) can be matched to the words in the definitions of all of the words in the context of the word
  - Let $D_1, D_2, \ldots., D_K$ be the definitions of the senses $s_1, s_2, \ldots., s_K$ of an ambiguous word w, each represented as a bag of words in the definition
  - Let E_{v_j} be the dictionary definition(s) for word v_j occurring in context c of w, represented as a bag of words; if $s_{j1}, s_{j2}, \ldots, s_{jL}$ are the senses of v_j, then $E_{v_j} = \cup_{jt} D_{jt}$

Sense Definition Disambiguation Algorithm

- Disambiguate the ambiguous word by choosing the sub-definition of the ambiguous word that has the greatest overlap with the words occurring in its context. Overlap can be measured by counting common words or other types of similarity measures

```
Given context $c$
for all senses $s_k$ of $w$ do
    score($s_k$) = overlap($D_k$, $\cup_{v_j \text{ in } c} E_{v_j}$)
end
Choose $s'$=argmax$_{s_k}$ score ($s_k$)
```

Sense Definition Disambiguation (Cont.)

- By itself, this method is insufficient to achieve highly accurate word sense disambiguation; Lesk obtained accuracies between 50% and 70% on a sample of ambiguous words
- There are possible optimizations that can be applied to improve the algorithm:
  - Run several iterations of the algorithm on a text, and instead of using a union of all words E_{v_j} occurring in the definition for v_j, use only the contextually appropriate definitions based on a prior iteration
  - Expand each word in context c with synonyms from a thesaurus

Thesaurus-Based Disambiguation

- This approach exploits the semantic categorization provided by a thesaurus (e.g., Roget’s) or lexicon with subject categories (e.g., Longman’s)
- The basic idea is that semantic categories of the words in a context determine the semantic category of the context as a whole. This category, in turn, determines which word senses are used
- Two approaches:
  - (Walker, 87)
  - (Yarowski, 92)

Thesaurus-Based Disambiguation (Cont.)

- Assumption: given a word is assigned multiple subject codes in a thesaurus, maybe each is associated with a different sense of the word

$$S(s_k \mid c) = \sum_{v_j \text{ in } c} \delta(t(s_k), v_j)$$

$$\delta(t(s_k), v_j) = 1 \text{ if } t(s_k) \text{ is one of the subject codes of } v_j$$

- Decision Rule

$$\text{Decide } \hat{s} \text{ if } \hat{s} = \operatorname{argmax}_{s_k} S(s_k \mid c)$$

Thesaurus-Based Approaches

- (Walker, 87): each word is assigned one or more subject codes in a dictionary corresponding to its different meanings
  - If more than one subject code is found, then assume that each code corresponds to a different word sense
  - Let $t(s_k)$ be the subject code for sense s_k of word *w* in context *c*
  - Then *w* can be disambiguated by counting the number of words from the context *c* for which the thesaurus lists $t(s_k)$ as a possible subject code. We select the sense that has the subject code with the highest count
- Black(1988) achieved only moderate success on 5 ambiguous words with this approach (~ 50% accuracies)

Context Categorization

- Assumption: given a word is assigned multiple subject codes in a thesaurus, maybe each is associated with a different sense of the word

$$S(s_k \mid c) = \sum_{v_j \text{ in } c} \delta(t(s_k), v_j)$$

$$\delta(t(s_k), v_j) = 1 \text{ if } t(s_k) \text{ is one of the subject codes of } v_j$$

- Decision Rule

$$\text{Decide } \hat{s} \text{ if } \hat{s} = \text{argmax}_{s_k} S(s_k \mid c)$$

Walker's Algorithms

```
Given context $c$
for all senses $s_k$ of $w$ do
    score($s_k$) = $\Sigma_{v_j \text{ in } c}\, \delta(\mathrm{t}(s_k), v_j)$
end
Choose $s' = \mathrm{argmax}_{s_k}$ score ($s_k$)
```

- Note that $\delta(\mathrm{t}(s_k), v_j)=1$ iff $\mathrm{t}(s_k)$ is one of the subject codes for v_j and 0 otherwise. The score is the number of words compatible with the subject code of s_k
- One problem with this algorithm is that a general categorization of words into topics may be inappropriate in a particular domain (e.g., *mouse* as a mammal or electronic device in the context of computer manual)
- Another problem is coverage, e.g., names like *Navratilova* suggests the topic of sports and yet appear in no lexicon

Yarowsky’s Algorithm

Categorize contexts based on categorization of words

```
for all contexts $c_i$ in the corpus do
    for all thesaurus categories $t_l$ do
        $\text{score}(c_i,t_l) = \log (P(c_i| t_l)/P(c_i)) \times P(t_l)$
    end
end
```

$t(c_i) = \{t_l \mid \text{score} (c_i,t_l) > \alpha\}$

Categorize words based on categorization of contexts

```
for all words $v_j$ in the vocabulary do
    $V_j = \{c \mid v_j \text{ in } c\}$
end
```

Yarowsky’s Algorithm (Cont.)

```
for all topics $t_l$ do
    $T_l = \{c \mid t_l \in \mathrm{t}(c)\}$
end
for all words $v_j$, all topics $t_l$ do
    $P(v_j| t_l) = |V_j \cap T_l| / \Sigma_\mathrm{j} |V_j \cap T_l|$
end
for all topics $t_l$ do
    $\mathrm{P}(t_l) = \Sigma_\mathrm{j} |V_j \cap T_l| / \Sigma_\mathrm{l} \Sigma_\mathrm{j} |V_j \cap T_l|$
End

comment: disambiguation
for all senses $s_k$ of $w$ occurring in $c$ do
    $\mathrm{score}(s_k) = \log \mathrm{P}(\mathrm{t}(s_k)) + \Sigma_{v_j \text{ in } c} \log \mathrm{P}(v_j | \mathrm{t}(s_k))$
end
choose $s' = \mathrm{argmax}_{s_k}$ score $(s_k)$
```

Results with Yarowsky's Algorithm

- The method achieves a high accuracy when thesaurus categories and senses align well with topics (e.g., *bass, star*), but when a sense spreads over topics (e.g., *interest*), the algorithm fails
- Topic independent distinctions between senses are problematic– when *interest* means advantage, it is not topic specific. In this case, it makes sense that topic-based classification would not work well

Categorization-Based Disambiguation

- Assumption: disambiguation can be done by replacing a word with its corresponding topic, find relevant topics:

$$t(c_i) = P(t_l \mid S(c_i, t_l) = \log \frac{P(c_i \mid t_l)}{P(c_i)} * P(t_l) > \alpha)$$

- ML estimation of features and topics in context: for proportion of contexts of topic *t* that contains the word *v*

$$V_j = \{c \mid v_j \text{ in } c\}, T_l = \{c \mid t_l \in t(c)\}$$

$$P(v_j \mid t_l) = \frac{\mathrm{C}(V_j \cap T_l)}{\sum_j \mathrm{C}(V_j \cap T_l)}, P(t_l) = \frac{\sum_j \mathrm{C}(V_j \cap T_l)}{\sum_l \sum_j \mathrm{C}(V_j \cap T_l)}$$

- Decision Rule (Figure 7.5 for examples)

$$\text{Decide } \hat{s} \text{ if } \hat{s} = \operatorname{argmax}_{s_k} [\textstyle\sum_{v_j \text{ in } c} \log P(v_j \mid t(s_k)) + \log P(t(s_k))]$$

Supervised Disambiguation: Summary

- Tagging information to design sense models
  - Context and context features
  - Thesaurus- and dictionary-based definitions
  - Translation from a second-language corpus
  - One sense per discourse
  - One sense per collocation
- Performance evaluation
  - Upper bound: human performance, 95%
  - Lower bound: 70-90%?
- Unsupervised disambiguation
  - Word sense clustering through EM learning algorithm

Unsupervised Disambiguation

- It may be useful to disambiguate among different word senses in cases where there are no available lexical resources
  - in a specialized domain (e.g., linguistics)
  - could be quite important for information retrieval in a domain
- Of course, it is impossible to do sense tagging in a situation where there is no labeled data; however, it is possible to carry out sense discrimination in a completely unsupervised manner
- Without supporting tools such as dictionaries and thesauri and in the absence of labeled text, we can simply cluster the contexts of an ambiguous word into a number of groups and discriminate between these groups without labeling them

Context-Group Discrimination

- The probabilistic model is the same Bayesian model as the one used by Gale et al.’s Bayes classifier, except that each P(v_j|s_k) is estimated using the EM algorithm (Schutze, 1998)
  - Start with a random initialization of the parameters of P(v_j|s_k)
  - Compute for each context c_i of *w*, the probability P(c_j|s_k) generated by s_k i.e. estimate the conditional probability of each word v_j occurring in *w*’s context being used with sense s_k
  - Use this preliminary categorization of contexts as our training data and then re-estimate P(v_j|s_k) to maximize the likelihood of the data given the model
  - EM is guaranteed to increase the log likelihood of the model given the data at each step; therefore, the algorithm stops when the likelihood does not increase significantly

Schutze's Algorithm

- Once parameters are estimated, we can disambiguate contexts *w* by computing the probability of each of the senses based on the words v_j occurring in context with the Naïve Bayes decision rule (Schutze, 1998):
  - Decide *s'* if $s' = \text{argmax}_{s_k} [\log P(s_k) + \Sigma_{v_j \text{ in } c} \log P(v_j | s_k)]$
- The granularity of senses of a word can be chosen by running the algorithm over a range of values
  - The larger the number of senses the better it will be able to explain the data
  - Relative increase in likelihood may help to distinguish important senses from random variations
  - Could make # of senses dependent on the amount of training data
  - Can get finer grained distinctions than in supervised approaches
- Works better for senses with topic dependency

So What Is a Word Sense Really?

- It might seem reasonable to define word senses as the mental representations of different word meanings
  - Not much is known about mental representations because it is hard to design experiments to get at what that is
  - Humans can categorize word usage using introspection, but is that reasonable? Also agreement tends to be low
  - Humans could label word senses using dictionary definitions, but this works best for skewed distributions where one sense is predominant. Also, definitions can often be vague
  - Words with the highest frequencies have the highest disagreement rate, so selecting words based on frequency would bias results

So What is a Word Sense Really? (Cont.)

- It may be that it is common for humans to have a simultaneous activation of different senses when comprehending words in text or discourse (leading to high levels of disagreement)
- These coactivations may be cases of **systematic polysemy**, where lexico-semantic rules apply to the class of words to systematically change or extend their meaning. For example, *competition* can refer to *the act of X* or *the people doing X*
- Propernouns also create problems, e.g., *Brown*, *Army*, etc.
- Could consider only course-grained distinctions among word senses (like those that show up across languages). Clustering approaches to word sense disambiguation adopt this strategy

Word Sense Disambiguation Evaluation

- If the disambiguation task is embedded in a task like translation, then it is easy to evaluate in the context of that application. This leads to application-oriented notions of sense
- Direct evaluation of disambiguation accuracy is more difficult in an application-independent sense. It would be easier if there were standard evaluation sets (Senseval project is addressing this need)
- There is a need for researchers to evaluate their algorithms on a representative sample of ambiguous words

Designing Disambiguation Test

- Assumption: conditional independence between noun and verb attachments
  - Example 8.18: “*He put the book on World War II on the table*.”
  - *P(*VA***p****,* NA***p*** *| v, n) = P(*VA***p***| v) * P(*NA***p*** *| n)*
  - *Log probability ratio test*

$$LPR(v,n,p) = \log \frac{P(\text{Attach}(p) = v \mid v,n)}{P(\text{Attach}(p) = n \mid v,n)} = \log \frac{P(\text{VA}_p = 1 \mid v) * P(\text{NA}_p = 0 \mid n)}{P(\text{NA}_p = 1 \mid n)}$$

- Estimation of attachment probabilities

$$P(\text{VA}_p = 1 \mid v) = \frac{C(v,p)}{C(v)}, P(\text{NA}_p = 1 \mid v) = \frac{C(n,p)}{C(n)}$$

- But how do we handle tagging of “ground truth”?

Factors Influencing the Notion of Sense

- The type of information used in disambiguation affects the notion of sense used:
  - Co-occurrence (bag-of-words model): topical sense
  - Relational information (e.g., subject, object)
  - Other grammatical information (e.g., part-of-speech)
  - Collocations (one sense per collocation)
  - Discourse (one sense per discourse segment): How much context is needed to determine sense?
  - Combinations of the above
- Different types of information may be more useful for different parts of speech (e.g., verb meaning is affected by its complements, but nouns are more affected by wider context)

Attachment Disambiguation

- Some Definitions
  - Example 8.14: "*The children eat the cake with a spoon*."
  - PP: propositional phrase: studied extensively in literature
  - VA***p*** : is there a PP headed by *p* and following the verb *v* which attaches to *v* (VA***p*** =1) or not (VA***p*** =0)
  - NA***p*** : is there a PP headed by *p* and following the noun *n* which attaches to *n* (NA***p*** =1) or not (NA***p*** =0)
- Log probability ratio test to resolve the ambiguity
  - Step 1: compute *P(p|v)* and *P(p|n)*
  - Step 2: hypothesis testing – assume a null hypothesis that the PP is attached to the verb, and an alternative hypothesis that the PP is attchaed to the noon, and accept the null hypothesis if

$$LPR(v,n,p) = \log \frac{P(p\,|\,v)}{P(p\,|\,n)} > \tau$$

Summary

- Today’s Class
  - Word sense disambiguation
- Next Class
  - Project summary due on 2/24, let’s start our discussion
  - Project plan finalize on 3/3 (presentation on 4/16 ???)
  - Lab3 assigned on 2/12 and due on 2/26
  - Midterm on 3/12 (???)
  - Final at 8am on 4/27 (shall we try a take-home ???)
- Reading Assignments
  - Manning and Schutze, Chapters 7 & 8