Bayesian Word Sense Discrimination

Jenine Turner and Eugene Charniak

{jenine|ec}@cs.brown.edu

Brown University
Word Sense Disambiguation

Determine the correct meaning of an ambiguous word (such as *bank*) given

- The word’s context
- A set of possible meanings for the word
- Labeled training data
Word Sense Disambiguation

Determine the correct meaning of an ambiguous word (such as *bank*) given

- The word’s context
- A set of possible meanings for the word
- Labeled training data

Problem

- Obtaining labeled training data is expensive
Word Sense Disambiguation

Determine the correct meaning of an ambiguous word (such as *bank*) given

- The word’s context
- A set of possible meanings for the word
- Labeled training data

Problem

- Obtaining labeled training data is expensive

Solution?

- Unsupervised learning
In Word Sense **Discrimination** (unsupervised) you are still given
- The word’s context

But you are not given
- A set of possible meanings for the word
- Labeled training data
In Word Sense **Discrimination** (unsupervised) you are still given

- The word’s context

But you are not given

- A set of possible meanings for the word
- Labeled training data

Two types of unsupervised Word Sense Discrimination

- Data can include other outside labeled information
- Only data is from local context
Ambiguity

Wordnet lists 30 senses for the noun “line”

- A formation of people or things one behind another
- Text consisting of a row of words written across a page or computer screen
- Something (as a cord or rope) that is long and thin and flexible
Ambiguity

Some distinctions are more reasonable than others

- A formation of people or things one behind another
  - The line stretched clear around the corner
- A formation of people or things one beside another
  - The cast stood in line for the curtain call
Why Word Sense Disambiguation?

- Necessary for correct semantic interpretation
- Applications of sense disambiguation
  - Machine Translation
  - Question Answering
  - Information Retrieval
  - Language Modeling
Previous Work in Word Sense Discrimination

- Contexts drawn from Roget’s Thesaurus (Yarowsky, 1992)
- Bootstrapping from manually chosen seed collocations (Yarowsky, 1995)
- Choosing candidate seeds automatically (Eisner and Karakos, 2005)
- Expectation Maximization (EM) on context features (Schutze, 1998)
- Clustering similar contexts (Pedersen and Bruce, 1997),
- Clustering different nouns (Pantel and Lin 2002)
Previous methods

Some downsides

- EM prefers similar-sized groups
- Most methods need to be given the number of groups or a cap
- Some of the “unsupervised” methods need outside information
Our approach

We use a Bayesian generative model for unsupervised learning

- Finite model: number of senses given
- Infinite model: number of senses unknown
Our approach

We use a Bayesian generative model for unsupervised learning

- Finite model: number of senses given
- Infinite model: number of senses unknown

Advantages to this approach

- Model can handle data with senses of varying frequency
- The infinite model does not constrain the number of senses
Three different bag-of-words feature sets

Counts of context words for the ambiguous word

- All nearby words (1)
  - She made her way, still seemingly dancing to the tune, the huge crocodile-skin handbag on her arm swaying heavily in time, to the door down to the saloon.

- Words from a “stripped” version of the full parse (2)
  - she made her way still seemingly dancing tune huge crocodile skin handbag her arm swaying heavily time door down saloon

- The words from (2) with closed-class words taken out
  - made way still seemingly dancing tune huge crocodile skin handbag arm swaying heavily time door down saloon
The Finite Model
The Finite Model

- A probability distribution over possible senses ($w$)
The Finite Model

- A probability distribution over possible senses ($w$)
- For each possible sense $z$, a probability distribution over context words ($\theta_z$)
The Finite Model

- A probability distribution over possible senses ($w$)
- For each possible sense $z$, a probability distribution over context words ($\theta_z$)
- (The probability distributions are chosen from Dirichlet distributions with hyperparameters $\alpha$ and $\beta$)
The Finite Model

- A probability distribution over possible senses \((w)\)
- For each possible sense \(z\), a probability distribution over context words \((\theta_z)\)
- (The probability distributions are chosen from Dirichlet distributions with hyperparameters \(\alpha\) and \(\beta\))

The generative model describes how the observed data (i.e. the context words) are generated:
The Finite Model

- A probability distribution over possible senses ($w$)
- For each possible sense $z$, a probability distribution over context words ($\theta_z$)
- (The probability distributions are chosen from Dirichlet distributions with hyperparameters $\alpha$ and $\beta$)

The generative model describes how the observed data (i.e. the context words) are generated:

- For each instance of the ambiguous word choose a sense $z$ from $w$. 
The Finite Model

- A probability distribution over possible senses ($w$)
- For each possible sense $z$, a probability distribution over context words ($\theta_z$)
- (The probability distributions are chosen from Dirichlet distributions with hyperparameters $\alpha$ and $\beta$)

The generative model describes how the observed data (i.e. the context words) are generated:

- For each instance of the ambiguous word choose a sense $z$ from $w$.
- From sense $z$, generate $m$ context words from $\theta_z$
The Infinite Model

- A probability distribution over possible senses \((w)\)
- For each possible sense \(z\), a probability distribution over context words \((\theta_z)\)
- \((w \text{ is chosen from Dirichlet process, } \theta_z \text{ from Dirichlet distribution with hyperparameters } \alpha \text{ and } \beta)\)

The generative model describes how the observed data (i.e. the context words) are generated:

- For each instance of the ambiguous word choose its sense \(z\) from \(w\) or choose a new sense entirely
- From sense \(z\), choose \(m\) context words from \(\theta_z\)
What are we aiming for?

- Goal is to choose a sense for each ambiguous word that maximizes the joint probability
  \[ p(s_i...s_n|x, \alpha, \beta) \]
What are we aiming for?

- Goal is to choose a sense for each ambiguous word that maximizes the joint probability $p(s_i...s_n|x, \alpha, \beta)$
- We cannot compute this directly, so we sample using Gibbs Sampling
Gibbs Sampling

Let’s say our ambiguous word is *bank*
Let’s say our ambiguous word is *bank*

- First, assign every instance of *bank* to the same sense
Let's say our ambiguous word is *bank*

- First, assign every instance of *bank* to the same sense
- For some number of iterations:
  - For each instance of *bank*, remove it from its sense and choose a new sense
Gibbs Sampling

Let’s say our ambiguous word is *bank*

- First, assign every instance of *bank* to the same sense
- For some number of iterations:
  - For each instance of *bank*, remove it from its sense and choose a new sense

Over time, the sense assignments should converge to a sample from the joint distribution
Gibbs Sampling

Let’s say our ambiguous word is *bank*

- First, assign every instance of *bank* to the same sense
- For some number of iterations:
  - For each instance of *bank*, remove it from its sense and choose a new sense

Over time, the sense assignments should converge to a sample from the joint distribution

But how do we know which new sense to choose?
Sampling: Picking a new sense assignment

- Probability of any single sense assignment is a combination of two probabilities
Sampling: Picking a new sense assignment

- Probability of any single sense assignment is a combination of two probabilities
  - The probability of the sense, dependent on:
    - All the other sense assignments
    - $\alpha$
Sampling: Picking a new sense assignment

- Probability of any single sense assignment is a combination of two probabilities
  - The probability of the sense, dependent on:
    - All the other sense assignments
    - $\alpha$
  - The probability of each context word, dependent on:
    - The other sense assignments
    - All the other context words
    - $\beta$
Sampling: Probability of a sense

In both the finite and infinite models

- Probability of a sense is proportional to the current number of words with that sense assigned

In the infinite model

- A new sense is chosen with a probability dependent on $\alpha$
Sampling: Probability of the context words

Let’s say we have a context word *river*
Sampling: Probability of the context words

Let’s say we have a context word *river*

- The probability of *river* is dependent on
  - The sense assignment chosen for this *bank*
  - The other sense assignments
  - The other context words
Sampling: Probability of the context words

Let’s say we have a context word *river*

- The probability of *river* is dependent on
  - The sense assignment chosen for this *bank*
  - The other sense assignments
  - The other context words

- So the probability of *river* in the given sense assignment of this instance of *bank* is high if *river* occurs frequently in that sense compared to the other senses

- $\beta$ governs how sensitive the model is to noise
Recap

Find a distribution of sense assignments over all instances of *bank*
Recap

Find a distribution of sense assignments over all instances of *bank*

- Gibbs sampling
Recap

Find a distribution of sense assignments over all instances of *bank*

- Gibbs sampling
- Repeatedly choose a new group for each instance of *bank*
Recap

Find a distribution of sense assignments over all instances of \textit{bank}

- Gibbs sampling
- Repeatedly choose a new group for each instance of \textit{bank}
- Converges to a sample from the joint distribution
Evaluation

Difficult to evaluate, due to lack of manually tagged training data and lack of standardization

- Pseudo-ambiguous words
- Senseval
- Line corpus
Evaluation

Evaluation metric

- Overall accuracy
- Baseline: majority sense
Evaluation

Evaluation metric
- Overall accuracy
- Baseline: majority sense

Different methods in previous work
- Supervised: using sense-labeled training data
- Unsupervised: no sense-labeled training data
- Completely Unsupervised: no labeled data of any sort
14 nouns from Senseval1

Two sets of senses for each word
- The original set of senses
- A hand-chosen subset
14 nouns from Senseval1

Two sets of senses for each word

- The original set of senses
- A hand-chosen subset

Evaluation

- Experiments on both the finite and infinite versions
- Tried various values for $\alpha$ and $\beta$
- Accuracy score compared to majority score
Preliminary Results

The full set

- On both the finite and infinite versions, 8 words scored above baseline
- Infinite version tends to prefer 2 or 3 senses
Preliminary Results

The full set

- On both the finite and infinite versions, 8 words scored above baseline
- Infinite version tends to prefer 2 or 3 senses

The subset

- Surprisingly, the smaller set does not show better performance
Continuing Work

- Try different features
  - Dependency information
  - Co-occurring words
  - Take distance into consideration
- Topic modeling
- Choose hyperparameters automatically