Computational Modeling of Lexical Ambiguity

Linlin Li

Cluster of Excellence (MMCI), Saarland University

{linlin}@coli.uni-sb.de

Saarbrücken, Germany
Oct 10, 2012
Word Meaning

Full Moon?
Word Meaning

Full Moon?
Full Moon?
Motivation: Examples

Words
- I hardly ever need to water the plant that grows in my yard because of the leak in the drains.
- Germany’s coalition government has announced a reversal of policy that will see all the country’s nuclear power plants phased out by 2022.

Idioms
- Dissanayake said that Kumaratunga was “playing with fire” after she accused military’s top brass of interfering in the peace process.
- Grilling outdoors is much more than just another dry-heat cooking method. It’s the chance to play with fire, satisfying a primal urge to stir around in coals.
A Machine Translation Example (Babel Fish)

- The government agent *spilled the beans* on the secret dossier.
- Der Regierungsbeauftragte *verschüttete die Bohnen* auf dem geheimen Dossier.

Natural language processing systems need to deal with lexical ambiguity...
“You can't retire. You know too much. You might talk.”

“The beans are all over the table.”

Figure: ‘literal’ vs ‘nonliteral’ senses of spill the beans.

Figure: Senses for bank, represented by sense keys from WordNet 2.1
Figure: Sense Category is not predefined.
Part I: Modeling Idiomatic Expressions (Binary Senses)

spill the beans?

"You can’t retire. You know too much. You might talk."
Part I: Modeling Idiomatic Expressions (Binary Senses)

spill the beans?

"You can't retire. You know too much. You might talk."
We played $v_1$ a couple of party $v_2$ games $v_3$ to break $v_4$ the ice $v_5$.

Graph-based Classifier ($\Delta c > 0 \Rightarrow$ literal):

$$\Delta c = c(G) - c(G')$$

$$(G : \{v_1, v_2, v_3, v_4, v_5\}, G' : \{v_1, v_2, v_3\})$$
A Bootstrapping Model to Detect Idioms

Unlabeled Corpus

Cohesion Graph

Methods Agree & Confidence > T

SVM

Labeled Set

Training Set

Evaluation

Auto Labeled Literal

Boosting Literal Class

Unlabeled Set
Features (1): Lexical Cohesion

- The average relatedness between the target expression and context words (target word connectivity):
  \[ f_1 = \frac{2}{|T| \times |C|} \sum_{(w_i, c_j) \in T \times C} relatedness(w_i, c_j) \]

- The average semantic relatedness of the context words (discourse connectivity):
  \[ f_2 = \frac{1}{\binom{|C|}{2}} \sum_{(c_i, c_j) \in C \times C, i \neq j} relatedness(c_i, c_j) \]

- The difference between the target expression relatedness and contextual relatedness \( f_3 \)

- Prediction of the cohesion graph \( f_4 \)

- The top \( n \) relatedness scores, the \( k^{th} \) highest score is defined as:
  \[ f_5(k) = \max_{(w_i, c_j) \in T \times C} (k, \{relatedness(w_i, c_j)\}) \]

where, \( T \) is the target expression and \( C \) is the context.
Features (2): An Example of Lexical Cohesion

- I always end up **spilling** the **beans** **all** over the **floor** and looking foolish. (l)
- The **government** **agent** **spilled** the **beans** on the **secret dossier**. (n)

Literal Case

- **beans**
- **spill**
- **all**
- **over**
- **floor**

Nonliteral Case

- **beans**
- **govern**
- **agent**
- **secret**
- **dossier**

- Target Word Connectivity $f_1$
I always end up **spilling** the **beans** **all over** the **floor** and looking foolish. (l)
The **government agent** **spilled** the **beans** on the **secret dossier**. (n)

**Literal Case**
- **beans**
- **spill**
- **all**
- **over**
- **floor**

**Nonliteral Case**
- **beans**
- **gov**
- **agent**
- **secret**
- **dossier**

**Avg. Discourse Connectivity** $f_2$
I always end up **spilling** the beans **all** over the floor and looking foolish. (l)
The government agent **spilled** the beans on the secret dossier. (n)

**Literal Case**

- beans
- spill
- all
- over
- floor

**Nonliteral Case**

- beans
- govern
- agent
- secret
- dossier

- Cohesion Graph $f_1, f_2$
Features (2): An Example of Lexical Cohesion

- I always end up **spilling** the beans **all over** the floor and looking foolish. (l)
- The government agent **spilled** the beans **on** the secret dossier. (n)

**Literal Case**

- **beans**
- **spill**
- **all**
- **over**
- **floor**

**Nonliteral Case**

- **beans**
- **govern**
- **agent**
- **secret**
- **dossier**

- Top Relatedness Score $f_5$
Features (2): An Example of Lexical Cohesion

- I always end up **spilling** the beans **all over** the floor and looking foolish. (l)
- The government agent **spilled** the beans on the secret dossier. (n)

<table>
<thead>
<tr>
<th>Literal Case</th>
<th>Nonliteral Case</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>beans</em></td>
<td><em>govern</em></td>
</tr>
<tr>
<td><em>spill</em></td>
<td><em>secret</em></td>
</tr>
<tr>
<td><em>all</em></td>
<td><em>agent</em></td>
</tr>
<tr>
<td><em>over</em></td>
<td></td>
</tr>
<tr>
<td><em>floor</em></td>
<td><em>dossier</em></td>
</tr>
</tbody>
</table>

- Target Word Connectivity $f_1$
- Avg. Discourse Connectivity $f_2$
- Cohesion Graph $f_1$, $f_2$
- Top Relatedness Score $f_5$
Features (3): Syntax and Other Linguistic Indicators

Global Lexical Context
- salient words
- related words

Local Lexical Context
- n-gram in a window of ±5w

Named Entities
- 'Russia', 'Bill Clinton', 'Secretary of State'

Indicative Terms
- 'literally', 'proverbially'

Scare Quotes

Syntax
- coordinated verb
  e.g., He may break the ice and fall through.
- head verb dependency relation
- modal verbs
- subjects
- verb subcat
- modifiers

Li’s PhD Defense
Experimental Data

- taken from Sporleder and Li (2009)
- 17 idioms (mainly V+NP and V+PP) with literal and non-literal senses
- all occurrences extracted from the Gigaword corpus (3964 instances)
- five paragraph context
- manually labelled as ‘literal’ (862) or ‘non-literal’ (3102)
### Results and Comparison

<table>
<thead>
<tr>
<th>model</th>
<th>acc.</th>
<th>prec /</th>
<th>rec /</th>
<th>f-score /</th>
</tr>
</thead>
<tbody>
<tr>
<td>base&lt;sub&gt;maj&lt;/sub&gt;</td>
<td>78.25</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>co-graph</td>
<td>78.38</td>
<td>50.04</td>
<td>69.72</td>
<td>58.26</td>
</tr>
<tr>
<td>combined</td>
<td>86.30</td>
<td>83.86</td>
<td>45.82</td>
<td>59.26</td>
</tr>
<tr>
<td>combined+boost</td>
<td>86.13</td>
<td>70.26</td>
<td>62.76</td>
<td>66.30</td>
</tr>
<tr>
<td>combined+it*</td>
<td>86.68</td>
<td>85.68</td>
<td>46.52</td>
<td>60.30</td>
</tr>
<tr>
<td>combined+boost+it*</td>
<td>87.03</td>
<td>71.86</td>
<td>66.36</td>
<td>69.00</td>
</tr>
<tr>
<td>super. 10CV</td>
<td>90.34</td>
<td>85.80</td>
<td>66.60</td>
<td>75.00</td>
</tr>
</tbody>
</table>

- **base<sub>maj</sub>**: majority baseline, i.e., ‘non-literal’
- **co-graph**: state-of-the-art
- **combined**: combine two base classifiers, one iteration
- **combined+boost**: boosting literal class
- **combined+boost+it***: bootstrapping, iterative training and boosting literal instances in each iteration
- **super. 10CV**: 10-fold cross validation, supervised
Part II: Modeling Word Senses (Multi Senses)

bank?

Li’s PhD Defense
Computational Modeling of Lexical Ambiguity
An Idea for Word Sense Disambiguation

context(c) → model

Target? → sense paraphrase₁

Target? → sense paraphrase₂

Target? → sense paraphraseᵢ

Target? → sense paraphraseᵣ
An Idea for Word Sense Disambiguation

context($c$)

Target?

model

$p(s|c)$

sense paraphrase;
Background

- PLSA (Hofmann, 1999)
  \[ p(w|d) = \sum_z p(w|z)p(z|d) \]

- Bayesian version, LDA (Blei et al., 2003)

A Sense Disambiguation Topic Model

- The probability of a sense given a context is proportional to the sense prior times the dot product of the context topic distributions and the sense topic distribution:
  \[ p(ds|dc) \propto p(s) \sum_z p(z|dc)p(z|ds) \]
Coarse-grained WSD

- SemEval-2007 Task-07 benchmark dataset (Navigli et al., 2009)
- Sense categories were obtained by clustering senses from WordNet 2.1 sense inventory (Navigli, 2006).

Fine-grained WSD

- SemEval-2007 Task-17 dataset (Pradhan et al., 2009)
- The sense inventory is from WordNet 2.1.
The most-frequent-sense baseline curse

- Word sense distribution is highly skewed (Zipf’s law)
- The most frequent one sense predominantly appears (McCarthy, 2009)
- SemEval07 coarse-grained WSD task: The most-frequent-sense baseline beats the best unsupervised systems.

<table>
<thead>
<tr>
<th>System</th>
<th>Noun</th>
<th>Verb</th>
<th>Adj</th>
<th>Adv</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>UPV-WSD</td>
<td>79.33</td>
<td>72.76</td>
<td>84.53</td>
<td>81.52</td>
<td>78.63</td>
</tr>
<tr>
<td>BL_{mfs}</td>
<td>77.44</td>
<td>75.30</td>
<td>84.25</td>
<td>87.50</td>
<td>78.99</td>
</tr>
</tbody>
</table>

- SemEval07 fine-grained WSD task: the most-frequent-sense baseline outperforms the best unsupervised system.

<table>
<thead>
<tr>
<th>System</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>RACAI</td>
<td>52.7</td>
</tr>
<tr>
<td>BL_{mfs}</td>
<td>55.91</td>
</tr>
</tbody>
</table>
## Results and Comparison (2)

### Coarse-grained WSD

<table>
<thead>
<tr>
<th>System</th>
<th>Noun</th>
<th>Verb</th>
<th>Adj</th>
<th>Adv</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>BL&lt;sub&gt;mfs&lt;/sub&gt;</td>
<td>77.44</td>
<td>75.30</td>
<td>84.25</td>
<td>87.50</td>
<td>78.99</td>
</tr>
<tr>
<td>UPV-WSD</td>
<td>79.33</td>
<td>72.76</td>
<td>84.53</td>
<td>81.52</td>
<td>78.63</td>
</tr>
<tr>
<td>MI+ref</td>
<td>79.96</td>
<td>75.47</td>
<td>83.98</td>
<td>86.06</td>
<td>79.99</td>
</tr>
</tbody>
</table>

- Comparison to the state-of-the-art system UPV-WSD (Buscaldi and Rosso, 2007)
- Comparison to the most-frequent-sense baseline BL<sub>mfs</sub>

### Fine-grained WSD

<table>
<thead>
<tr>
<th>System</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>BL&lt;sub&gt;mfs&lt;/sub&gt;</td>
<td>55.91</td>
</tr>
<tr>
<td>RACAI</td>
<td>52.7</td>
</tr>
<tr>
<td>MI+ref</td>
<td>56.99</td>
</tr>
</tbody>
</table>

- Comparison to the state-of-the-art system RACAI (Ion and Tufis, 2007)
- Comparison to the most-frequent-sense baseline BL<sub>mfs</sub>
Part III: Evaluation Study

Word Sense Induction: a new task that does not rely on pre-defined sense category

- Graded representation of word meaning better reflect linguistic realities than rigid sense inventories (Cruse, 2000)
- Applications such as information retrieval and machine translation have been shown to benefit from induced sense (Veronis, 2004; Vickrey et al., 2005)

Without pre-defined sense inventory, how do we evaluate?

- State-of-the-art evaluation approach: Normalized Mutual Information (V-Measure)
  \[ V(c, k) = \frac{2I(c, k)}{H(c) + H(k)} \]

- Entropy is estimated by the Maximum Likelihood Estimator (MLE)
  \[ \hat{H}_{MLE}(p_N) = -\sum_{i=1}^{m} p_{N,i} \log p_{N,i} \]
### Rankings from SemEval10 Shared Task

<table>
<thead>
<tr>
<th>System</th>
<th>Cluster nr.</th>
<th>V-Measure</th>
<th>ranks</th>
</tr>
</thead>
<tbody>
<tr>
<td>1cl1inst</td>
<td>89.1</td>
<td>31.6</td>
<td>1</td>
</tr>
<tr>
<td>Hermit</td>
<td>10.8</td>
<td>16.2</td>
<td>2</td>
</tr>
<tr>
<td>KSU KDD</td>
<td>17.5</td>
<td>15.7</td>
<td>3</td>
</tr>
<tr>
<td>UoY</td>
<td>11.5</td>
<td>15.7</td>
<td>4</td>
</tr>
<tr>
<td>KCDC-PT</td>
<td>1.5</td>
<td>1.9</td>
<td>26</td>
</tr>
<tr>
<td>Duluth-Mix-Uni-Gap</td>
<td>1.4</td>
<td>1.4</td>
<td>27</td>
</tr>
<tr>
<td>Duluth-WSI-SVD-Gap</td>
<td>1.0</td>
<td>0.0</td>
<td>28</td>
</tr>
<tr>
<td>MFS</td>
<td>1.0</td>
<td>0.0</td>
<td>29</td>
</tr>
</tbody>
</table>

- The simple baseline one-cluster-per-instance (1cl1inst) gains the best performance.
- The top ranked systems tend to have more clusters whereas the bottom ranked systems tend to have less ones.
- V-Measure seems to be co-related to cluster numbers as higher cluster number tends to achieve better score.
Given $N$ instances in the sample set

- The class entropy is $\hat{H}(c)$;

- The cluster entropy is $\hat{H}(k) = -\sum_{i=1}^{N} \frac{1}{N} \log \frac{1}{N} = \log N$;

- Similarly, the joint entropy is $\hat{H}(c, k) = \log N$;

- The V-Measure is estimated as:

$$\hat{V}(c, k) = \frac{2\hat{H}(c)}{\log N + \hat{H}(c)}$$

When the sample size goes to infinity, $N \to +\infty$:

$$\lim_{N \to +\infty} \hat{V}(c, k) = \lim_{N \to +\infty} \frac{2\hat{H}(c)}{\log N + \hat{H}(c)} = 0$$

V-Measure overrates the 1cl1inst baseline on a finite sample set
Analysis (2): V-Measure is biased

Finding:
- V-Measure is positively biased

The cause of the V-Measure bias
- The ML entropy estimator used in V-Measure is biased
- The sample size ($N$) is too small compared to classes for joint entropy estimation $H(c,k)$; It is very often that:

$$\frac{N}{|c| \times |k|} < 1$$

Alternative entropy estimators: add bias correction factor to the standard ML estimator
- Miller-Madow’s Estimator (MM) (Miller, 1995)
- Jackknifed Estimator (JK) (Quenouille, 1956; Tukey, 1958)
- “Best Upper Bounds” Estimator (BUB) (Paninski, 2003)
Figure: The estimated and true entropy of Zipf’s law ($s=2$), the number of classes is set to be $m = 10$
ML estimator has higher discrepancy to the more precise estimators when cluster number is big

Figure: Discrepancy in entropy estimators (V-Measure) as function of the predicted number of clusters. The dots represent different systems of SemEval10.
Ranking Correlation of Different Estimators

(a) JK v.s. ML

(b) BUB v.s. ML

(c) BUB v.s. JK
Evaluation Summary

Observation:
- The state-of-the-art V-Measure is in favor of fine-grained output

Findings:
- V-Measure is positively biased
- The cause of this bias: entropy estimator is biased on a finite sample set

Proposal:
- More precise estimators (JK and BUB) should replace ML for future WSI evaluation
- We further suggest limiting the number of clusters based on test set size would alleviate the entropy bias problem
Main Contributions

Progress in the state-of-the-art idiom detection
- A novel bootstrapping model
- Multi-dimensional linguistic features such as lexical cohesion, syntax and some semantic features

Progress in the state-of-the-art fine-/coarse-grained WSD
- A novel topic model
- Explore sense priors

New findings on WSI evaluation
- State-of-the-art evaluation V-Measure is biased
- Proposal:
  - More precise estimators (JK and BUB) should replace standard ML estimator
  - WSI task should limit the number of clusters allowed for system output based on the size of the evaluation set
Outlooks

- Prior knowledge in statistical modeling
- Meaning of words/phrases can change over time
- More factors influencing lexical semantics
  - Trendy lexical choice, e.g., forum/review corpora
  - Granularity mismatch in multilingual corpora, e.g., how to translate a lexical entry when an equivalent term in the target language is missing
Linlin Li, Benjamin Roth and Caroline Sporleder.
Topic Models for Word Sense Disambiguation and Token-based Idiom Detection.
ACL, 2010.

Linlin Li and Caroline Sporleder.
Using Gaussian Mixture Models to Detect Figurative Language in Context.
NAACL, Short Papers, 2010.

Linlin Li and Caroline Sporleder.
Linguistic Cues for Distinguishing Literal and Non-Literal Usage.
CoLing, 2010.

Linlin Li and Caroline Sporleder.
Classifier combination for Contextual idiom detection without labelled data.
EMNLP, 2009.

Caroline Sporleder and Linlin Li.
Unsupervised Recognition of Literal and Non-Literal Use of Idiomatic Expressions.
EACL, 2009.
Contexts and senses paraphrases are both treated as documents.

\[ s = \arg \max_{ds_i} p(ds_i|dc) \]
Contexts and senses paraphrases are both treated as documents.

\[ s = \arg \max_{d_{si}} p(d_{si} | d_c) \]

Assume \( d_s \) is conditionally independent of \( d_c \), given \( z \).

\[ p(d_s | d_c) = \sum_z p(z | d_c) p(d_s | z) \]
Contexts and senses paraphrases are both treated as documents.

\[ s = \arg \max_{ds_i} p(ds_i|dc) \]

Assume \( ds \) is conditionally independent of \( dc \), given \( z \).

\[
p(ds|dc) = \sum_z p(z|dc)p(ds|z)
\]

No direct estimation of \( p(ds|z) \).

\[
p(ds|dc) = p(ds) \sum_z \frac{p(z|dc)p(z|ds)}{p(z)}
\]
\[ p(ds|dc) = p(ds) \sum_z \frac{p(z|dc)p(z|ds)}{p(z)} \]

Use prior sense information \( p(s) \) to estimate \( p(ds) \).

The sense distribution in real corpus is often highly skewed (McCarthy, 2009). \( p(s) \) can be taken from existing resource (e.g., sense frequency given in WordNet).

Assume topic distribution is uniform.

\[ p(ds|dc) \propto p(s) \sum_z \frac{p(z|dc)p(z|ds)}{p(z)} \]

Li’s PhD Defense
Computational Modeling of Lexical Ambiguity 33/31
\[ p(ds|dc) = p(ds) \sum_z \frac{p(z|dc)p(z|ds)}{p(z)} \]

- Use prior sense information \( p(s) \) to estimate \( p(ds) \).

\[ \hat{p}(ds) = p(s) \]

- The sense distribution in real corpus is often highly skewed (McCarthy, 2009).
- \( p(s) \) can be taken from existing resource (e.g., sense frequency given in WordNet).
\[ p(ds|dc) = p(ds) \sum_z \frac{p(z|dc)p(z|ds)}{p(z)} \]

- Use prior sense information \( p(s) \) to estimate \( p(ds) \).
  \[ \hat{p}(ds) = p(s) \]

- The sense distribution in real corpus is often highly skewed (McCarthy, 2009).
- \( p(s) \) can be taken from existing resource (e.g., sense frequency given in WordNet).
- Assume topic distribution is uniform.
  \[ p(ds|dc) \propto p(s) \sum_z p(z|dc)p(z|ds) \]
In case no prior sense information is available?

\[ p(ds \mid dc) \propto p(s) \sum_z p(z \mid dc) p(z \mid ds) \]

- Vector-space model on inferred topic frequency statistics \( v(z \mid d) \).
- Maximizing the cosine value of two document vectors \( \cos(ds, dc) \).

Model II:

\[ \text{arg max}_{ds_i} \cos(v(z \mid dc), v(z \mid ds_i)) \]
Sometimes, a sense paraphrase is characterized only by one typical, strongly connected word.

- Take the **maximum** of the probability of observing a word in the sense paraphrase.

  - “rock the boat” → \{“break the norm”, “cause trouble”\}
  - \(p(\text{“break the norm, cause trouble”} | dc)\), very strong requirement
  - \(p(\text{“norm”} | dc) \text{ OR } p(\text{“trouble”} | dc) \Rightarrow \text{idiomatic sense}\)

**Model III:**

\[
\arg \max_{d_{s_j}} \left\{ \max_{w_i \in d_{s_j}} \sum_z p(w_i | z)p(z | dc) \right\}
\]
The test set and sense paraphrase set are relatively small.

- Estimate topics from a very large corpus (a Wikipedia dump), with broad thematic diversity and vocabulary coverage.

- Represent sense paraphrase documents and context documents by topics $p(z|dc)$, $p(z|ds)$. 
Sense Paraphrases

WSD Tasks

- The WordNet word forms, glosses and example sentences of
  - the sense synset
  - the reference synsets (excluding hypernym)

Idiom Task

- Paraphrases the nonliteral meaning from several online idiom dictionaries.
  - e.g., rock the boat → {"break the norm", "cause trouble"}
- For the literal sense, we use 2-3 manually selected words.
  - e.g., break the ice → {"ice", "water", "snow"}
The system statistically significantly outperforms the majority baseline.

The system also outperforms two state-of-the-art systems: the cohesion-graph based approach (Li and Sporleder, 2009), and the bootstrapping approach (Li and Sporleder, 2009).

<table>
<thead>
<tr>
<th>System</th>
<th>Prec₁</th>
<th>Rec₁</th>
<th>F₁</th>
<th>Acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basemaj</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>78.25</td>
</tr>
<tr>
<td>co-graph</td>
<td>50.04</td>
<td>69.72</td>
<td>58.26</td>
<td>78.38</td>
</tr>
<tr>
<td>boot.</td>
<td>71.86</td>
<td>66.36</td>
<td>69.00</td>
<td>87.03</td>
</tr>
<tr>
<td>Model III</td>
<td>67.05</td>
<td>81.07</td>
<td>73.40</td>
<td>87.24</td>
</tr>
</tbody>
</table>
We propose three models for sense disambiguation tasks by incorporating a hidden variable which is estimated from a Wikipedia dump.

- Model I directly optimizes the conditional probability of a sense paraphrase.
- Model II is a vector space model on topic frequencies.
- Model III maximizes the conditional probability of a particular word in the paraphrase.

The proposed models outperform state-of-the-art systems.
Future Work

The limitations:

- The sense granularity problem.
- Evaluation is isolated from real applications.

Interesting directions to explore in the future...

- Application based word sense disambiguation evaluation.
- Multilingual word sense disambiguation.
- Word sense disambiguation for machine translation (application + multilingual).
Context as features

Idiomatic Usage v.s. Non-idiomatic Usage

- Dad had to break\textsubscript{v4} the ice\textsubscript{v5} on the chicken\textsubscript{v1} troughs\textsubscript{v2} so that they could get water\textsubscript{v3}.
- We played\textsubscript{v1} a couple of party\textsubscript{v2} games\textsubscript{v3} to break\textsubscript{v4} the ice\textsubscript{v5}.

We use conditional probabilities to model those context word dependencies.
Estimation of Topics

Gibbs Sampling (Griffiths and Steyvers, 2004)

- **Topic Assignment Step:**
  
  \[ p(z_j|w_i, d) = p(w_i|z_j) \times p(z_j|d) \]  
  \[ (1) \]

- **Distribution Estimation Step:**

  \[ p(z_j|d) = \frac{\sum_{w_k \in d} f(w_k, z_j) + \alpha}{\sum_{z_n} \sum_{w_k \in d} f(w_k, z_n) + T \alpha} \]  
  \[ (2) \]

  \[ p(w_i|z_j) = \frac{f(w_i, z_j) + \beta}{\sum_{w_k} f(w_k, z_j) + W \beta} \]  
  \[ (3) \]

\( T \) is the number of topics; \( W \) is size of vocabulary; \( \alpha \) and \( \beta \) are hyperparameters (prior).