Using Gaussian Mixture Models to Detect Figurative Language in Context

Linlin Li and Caroline Sporleder

Cluster of Excellence, MMCI
Saarland University, Germany

NAACL-HLT, 2010
Outline

1. Introduction
2. Using Gaussian Mixture Model to Detect Figurative Language
3. Evaluating the GMM Approach
4. Conclusion
What is figurative language and why is it a problem?

**Unambiguous Idiom**
The 19th century windjammers like Cutty Sark were able to maintain progress *by and large* even in bad wind conditions.

**Ambiguous Idiom**
The government agent *spilled the beans* on the secret dossier. When Peter reached for the salt he knocked over the can and *spilled the beans* all over the table.

**General Creative Usage**
Take the sock out of your mouth, and create a brand new relationship with your mom.
Machine Translation (Babel Fish)

Example

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

**Idea**

Literal and non-literal data are generated by two different Gaussians, **literal** and **non-literal** Gaussian

**Model**

\[ p(x) = \sum_{c \in \{l,n\}} w_c \times N(x|\mu_c, \Sigma_c) \]

- \( c \): the category of the Gaussian
- \( \mu_c \): mean
- \( \Sigma_c \): covariance matrix
- \( w_c \): Gaussian weight
Figurative Language Detection

Idea

Which Gaussian has the higher probability of generating the instance?

Decision Rule

\[ c(x) = \arg \max_{i \in \{l,n\}} \{ w_i \times N(x | \mu_i, \Sigma_i) \} \]

1. \( w_i \times N(x | \mu_i, \Sigma_i) \): fit the data to different Gaussians
2. \( \arg \max_{i \in \{l,n\}} \): choose the Gaussian that maximizes the probability of generating the specific instance
Feature Design

Aim

- Phrase independent features
- Generalize across different figurative usages

Features

- Semantic cohesion features
- Use normalized Google distance (Cilibrasi and Vitanyi, 2007), to model semantic cohesion
Semantic Cohesion Features (5 types)

- $x_1$: the average relatedness between the target expression and context words
  \[ x_1 = \frac{2}{|T| \times |C|} \sum_{(w_i, c_j) \in T \times C} \text{relatedness}(w_i, c_j) \]

- $x_2$: the average semantic relatedness of the context words
  \[ x_2 = \frac{1}{\binom{|C|}{2}} \sum_{(c_i, c_j) \in C \times C, i \neq j} \text{relatedness}(c_i, c_j) \]

- $x_3$: $x_1 - x_2$

- $x_4$: prediction of the co-graph (Sporleder and Li, 2009)

- $x_5$: the top $n$ relatedness scores ($n = 100$)
  \[ x_5(k) = \max_{(w_i, c_j) \in T \times C} (k, \{\text{relatedness}(w_i, c_j)\}) \]
Cohesion Features
An Example

Literal Case
- beans
- can
- table
- reach
- knock

Nonliteral Case
- beans
- secret
- dossier
- govern
- agent

Features:
- target word connectivity ($x_1$)
Cohesion Features
An Example

Literal Case
- beans
- can
- reach
- table
- knock

Nonliteral Case
- beans
- secret
- govern
- dossier
- agent

Features:
- average discourse connectivity ($x_2$)
Cohesion Features
An Example

Literal Case

- Beans
- Can
- Table
- Reach
- Knock

Nonliteral Case

- Beans
- Secret
- Govern
- Dossier
- Agent

Features:

- Cohesion graph
  \((x_1 - x_2)\)
Cohesion Features
An Example

**Literal Case**
- beans
- can
- table
- reach
- knock

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

Features:
- top connected words ($x_5$)
Cohesion Features
An Example

**Literal Case**
- **beans**
- **can**
- **reach**
- **table**
- **knock**

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

**Features:**
- target word connectivity ($x_1$)
- average discourse connectivity ($x_2$)
- cohesion graph ($x_1 - x_2$)
- top connected words ($x_5$)
Data

Datesets:

- Idiom dataset
  - 3964 idiom occurrences (17 types)
  - manually labeled as literal or figurative

- Random V+NP dataset
  - Randomly selected sample of 500 V+NP constructions from the idiom corpus (subset from the Gigaword corpus)
Annotation

Different types of figurative usage

- **nas**: ambiguous phrase-level figurative (7.3%)
  - spill the beans
- **nsu**: unambiguous phrase-level figurative (1.9%)
  - trip the light fantastic
- **nw**: token-level figurative (9.2%)
  - During the Iraq war, he was a *sparrow*; he didn’t condone the bloodshed but wasn’t bothered enough to go out and protest.
- **l**: literal (81.5%)
  - steer the industry (word senses)
Two Experimental Settings

- GMM estimated by **EM**
  - Priors of Gaussian components, means and covariance of each components, are initialized by the k-means clustering algorithm (Hartigan, 1975)
- GMM estimated from **annotated data**
## GMM Estimated by EM

### Idiom Dataset

<table>
<thead>
<tr>
<th>Model</th>
<th>C</th>
<th>Pre.</th>
<th>Rec.</th>
<th>F-S.</th>
<th>Acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Co-Graph</td>
<td>n</td>
<td>90.55</td>
<td>80.66</td>
<td>85.32</td>
<td>78.38</td>
</tr>
<tr>
<td></td>
<td>l</td>
<td>50.04</td>
<td>69.72</td>
<td>58.26</td>
<td></td>
</tr>
<tr>
<td>GMM</td>
<td>n</td>
<td>90.69</td>
<td>80.66</td>
<td>85.38</td>
<td>78.39</td>
</tr>
<tr>
<td></td>
<td>l</td>
<td>50.17</td>
<td>70.15</td>
<td>58.50</td>
<td></td>
</tr>
</tbody>
</table>
## GMM Estimated by EM
### V+NP Dataset

<table>
<thead>
<tr>
<th>Model</th>
<th>C</th>
<th>Pre.</th>
<th>Rec.</th>
<th>F-S.</th>
<th>Acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>n</td>
<td>21.79</td>
<td>22.67</td>
<td>22.22</td>
<td>71.87</td>
</tr>
<tr>
<td></td>
<td>l</td>
<td>83.19</td>
<td>82.47</td>
<td>82.83</td>
<td></td>
</tr>
<tr>
<td>Co-Graph</td>
<td>n</td>
<td>37.29</td>
<td>84.62</td>
<td>51.76</td>
<td>70.92</td>
</tr>
<tr>
<td></td>
<td>l</td>
<td>95.12</td>
<td>67.83</td>
<td>79.19</td>
<td></td>
</tr>
<tr>
<td>GMM</td>
<td>n</td>
<td>40.71</td>
<td>73.08</td>
<td>52.29</td>
<td>75.41</td>
</tr>
<tr>
<td></td>
<td>l</td>
<td>92.58</td>
<td>75.94</td>
<td>83.44</td>
<td></td>
</tr>
<tr>
<td>GMM{nsu,l}</td>
<td>n</td>
<td>8.79</td>
<td>1.00</td>
<td>16.16</td>
<td>76.49</td>
</tr>
<tr>
<td></td>
<td>l</td>
<td>1.00</td>
<td>75.94</td>
<td>86.33</td>
<td></td>
</tr>
<tr>
<td>GMM{nsa,l}</td>
<td>n</td>
<td>22.43</td>
<td>77.42</td>
<td>34.78</td>
<td>76.06</td>
</tr>
<tr>
<td></td>
<td>l</td>
<td>97.40</td>
<td>75.94</td>
<td>85.34</td>
<td></td>
</tr>
<tr>
<td>GMM{nw,l}</td>
<td>n</td>
<td>23.15</td>
<td>64.10</td>
<td>34.01</td>
<td>74.74</td>
</tr>
<tr>
<td></td>
<td>l</td>
<td>94.93</td>
<td>75.94</td>
<td>84.38</td>
<td></td>
</tr>
</tbody>
</table>
### GMM Estimated from Annotated Data

#### V+NP Dataset

<table>
<thead>
<tr>
<th>Model</th>
<th>C</th>
<th>Pre.</th>
<th>Rec.</th>
<th>F-S.</th>
<th>Acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>GMM</td>
<td>n</td>
<td>40.71</td>
<td>73.08</td>
<td>52.29</td>
<td>75.41</td>
</tr>
<tr>
<td></td>
<td>l</td>
<td>92.58</td>
<td>75.94</td>
<td>83.44</td>
<td></td>
</tr>
<tr>
<td>GMM+f</td>
<td>n</td>
<td>42.22</td>
<td>73.08</td>
<td>53.52</td>
<td>76.60</td>
</tr>
<tr>
<td></td>
<td>l</td>
<td>92.71</td>
<td>77.39</td>
<td>84.36</td>
<td></td>
</tr>
<tr>
<td>GMM+f+s</td>
<td>n</td>
<td>41.38</td>
<td>54.55</td>
<td>47.06</td>
<td>83.44</td>
</tr>
<tr>
<td></td>
<td>l</td>
<td>92.54</td>
<td>87.94</td>
<td>90.18</td>
<td></td>
</tr>
</tbody>
</table>

- **f**: fix the Gaussian components, estimate from the annotated idiom data
- **s**: select most confident examples, abstain from making a prediction when the probability of belonging to a certain Gaussian is below the selected threshold
Conclusion

- Distinguish potential idiomatic expressions, and discover new figurative expressions
- Due to the homogeneity of nonliteral language, features can be designed in a cross-expression manner
- The components of GMM can be effectively estimated using EM in an unsupervised way
- The performance can be further improved when employing an annotated data set for parameter estimation
## GMM Estimated from different Idiom Data

### V+NP Dataset

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>bite one’s tongue (166)</td>
<td>n</td>
<td>40.79</td>
<td>79.49</td>
<td>53.91</td>
<td>74.94</td>
</tr>
<tr>
<td></td>
<td>l</td>
<td>94.10</td>
<td>73.91</td>
<td>82.79</td>
<td></td>
</tr>
<tr>
<td>break the ice (541)</td>
<td>n</td>
<td>39.05</td>
<td>52.56</td>
<td>44.81</td>
<td>76.12</td>
</tr>
<tr>
<td></td>
<td>l</td>
<td>88.36</td>
<td>81.45</td>
<td>84.77</td>
<td></td>
</tr>
<tr>
<td>pass the buck (262)</td>
<td>n</td>
<td>41.01</td>
<td>73.08</td>
<td>52.53</td>
<td>75.65</td>
</tr>
<tr>
<td></td>
<td>l</td>
<td>92.61</td>
<td>76.23</td>
<td>83.62</td>
<td></td>
</tr>
<tr>
<td>play with fire (566)</td>
<td>n</td>
<td>39.29</td>
<td>84.62</td>
<td>53.66</td>
<td>73.05</td>
</tr>
<tr>
<td></td>
<td>l</td>
<td>95.29</td>
<td>70.43</td>
<td>81.00</td>
<td></td>
</tr>
</tbody>
</table>

- None of the difference is statistically significant