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

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
- μ_c: mean
- Σ_c: covariance matrix
- w_c: Gaussian weight

Figurative Language Detection

Idea

Which Gaussian has the higher probablity 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 probablity 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)

- $x1$: the average relatedness between the target expression and context words

$$x1 = \frac{2}{|T| \times |C|} \sum_{(w_i, c_j) \in T \times C} relatedness(w_i, c_j)$$

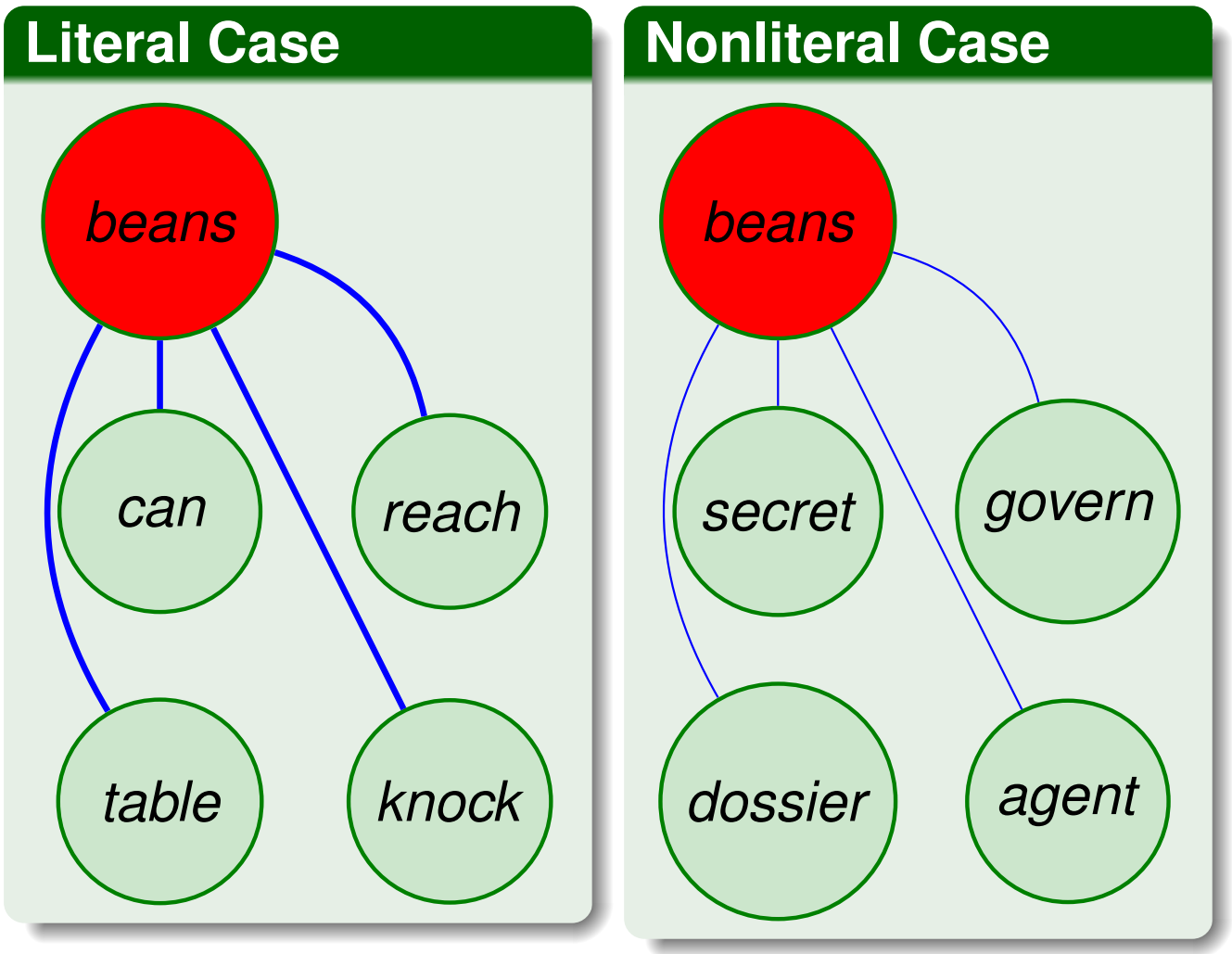
- $x2$: the average semantic relatedness of the context words

$$x2 = \frac{1}{\binom{|C|}{2}} \sum_{(c_i, c_j) \in C \times C, i \neq j} relatedness(c_i, c_j)$$

- $x3$: $x1 - x2$
- $x4$: prediction of the co-graph (Sporleder and Li, 2009)
- $x5$: the top n relatedness scores ($n = 100$)

$$x5(k) = \max_{(w_i, c_j) \in T \times C} (k, \{relatedness(w_i, c_j)\})$$

Cohesion Features
An Example

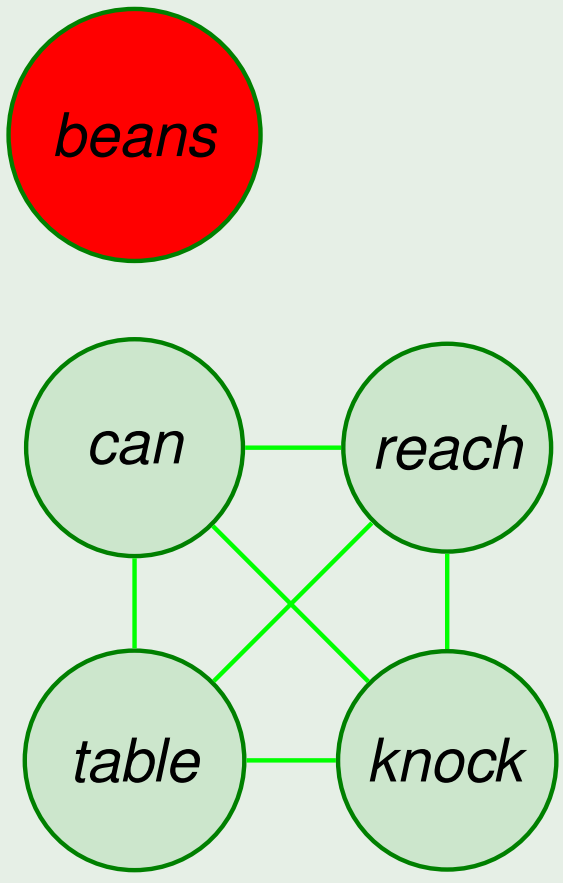


Features:

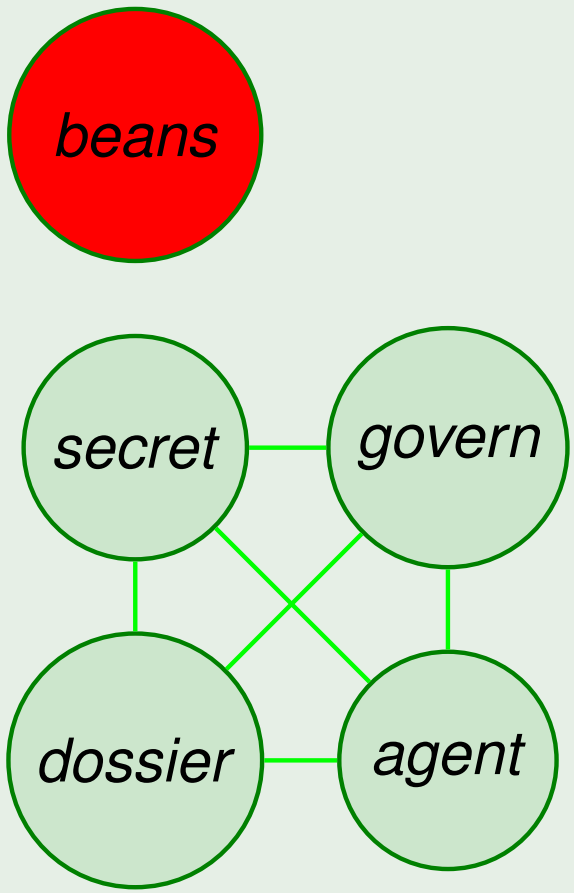
- target word connectivity (x_1)

Cohesion Features
An Example

Literal Case



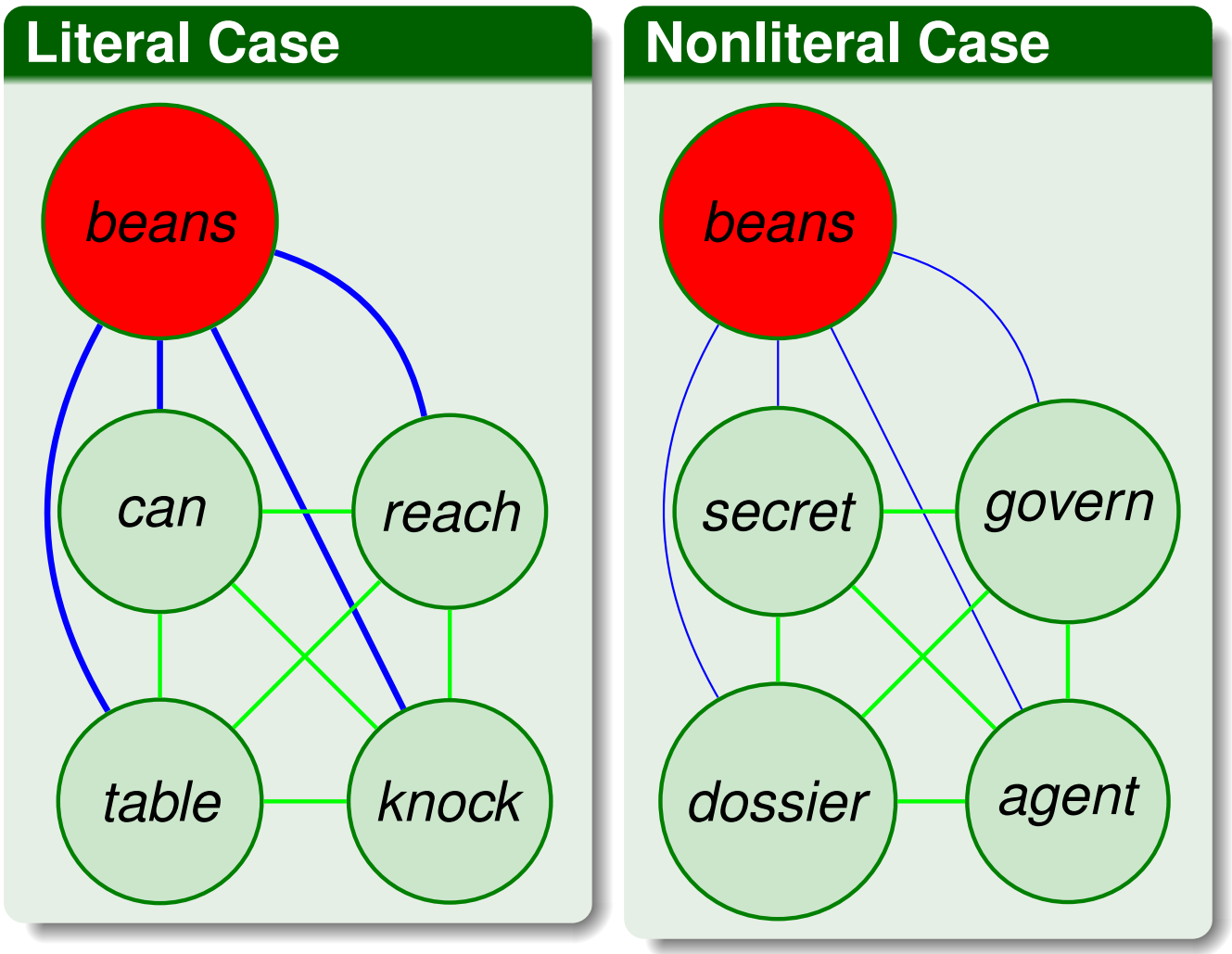
Nonliteral Case



Features:

- average discourse connectivity (x_2)

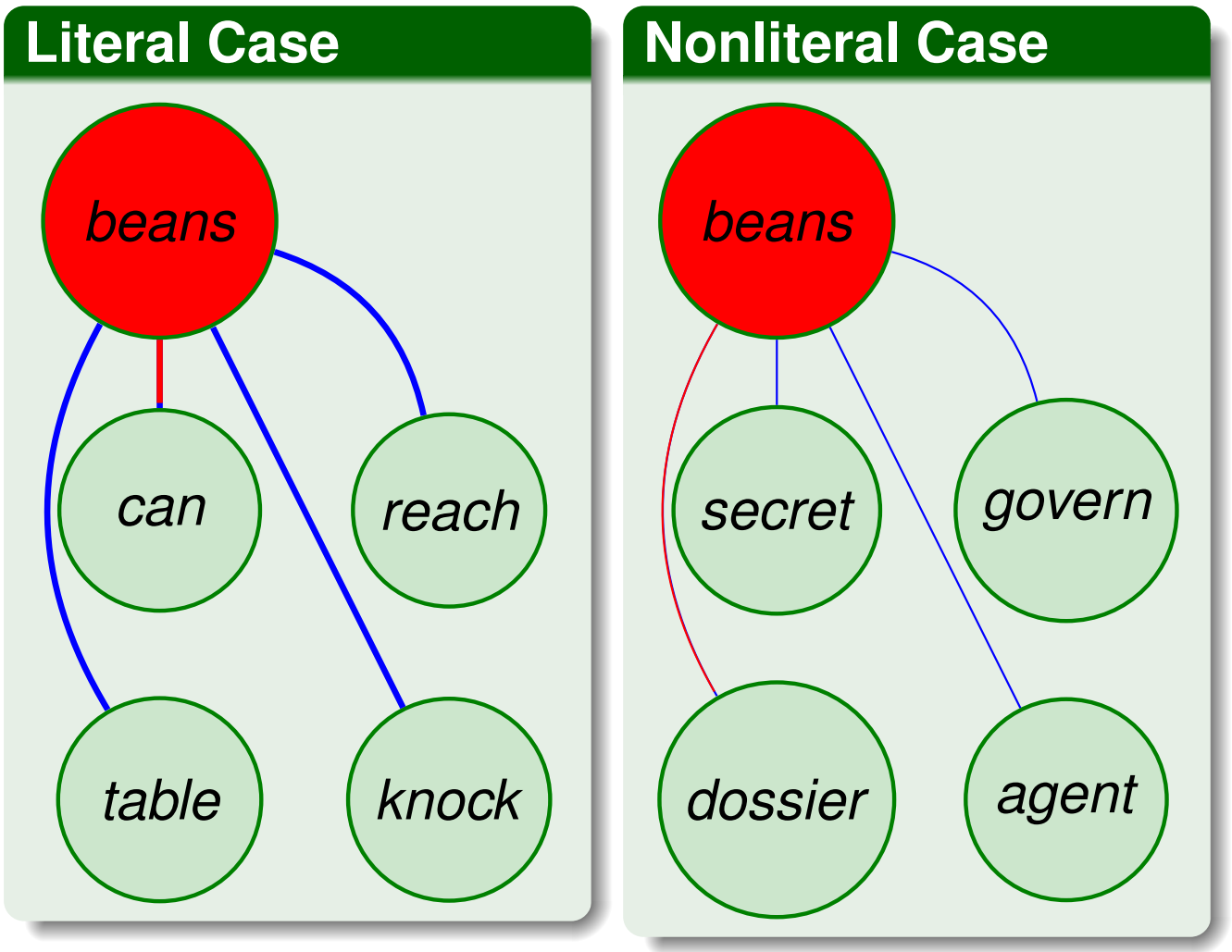
Cohesion Features
An Example



Features:

- cohesion graph ($x_1 - x_2$)

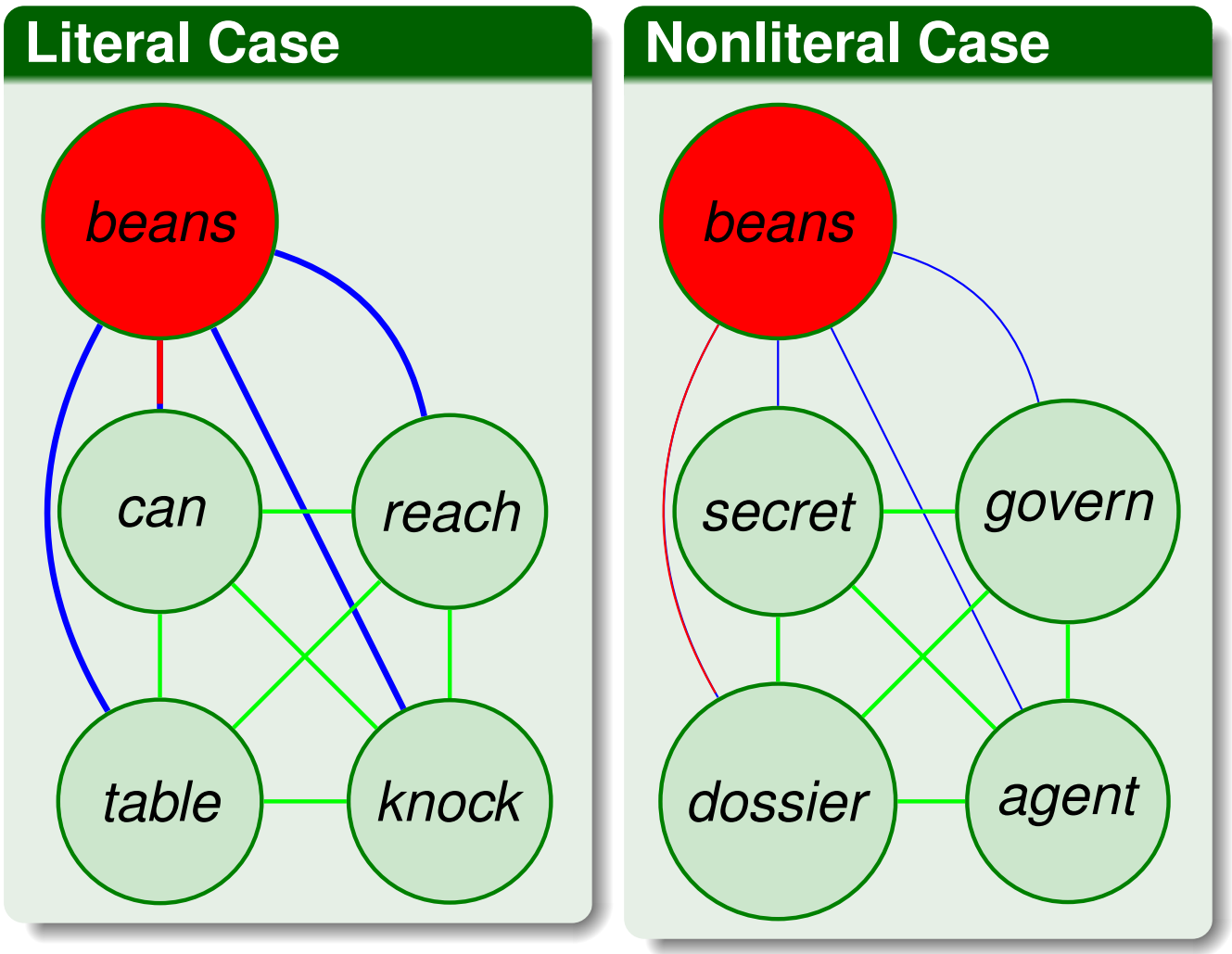
Cohesion Features
An Example



Features:

- top connected words (x_5)

Cohesion Features
An Example



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

Model	C	Pre.	Rec.	F-S.	Acc.
Co-Graph	n	90.55	80.66	85.32	78.38
	l	50.04	69.72	58.26	
GMM	n	90.69	80.66	85.38	78.39
	l	50.17	70.15	58.50	

GMM Estimated by EM
V+NP Dataset

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

GMM Estimated from Annotated Data
V+NP Dataset

Model	C	Pre.	Rec.	F-S.	Acc.
GMM	n	40.71	73.08	52.29	75.41
	l	92.58	75.94	83.44	
GMM+f	n	42.22	73.08	53.52	76.60
	l	92.71	77.39	84.36	
GMM+f+s	n	41.38	54.55	47.06	83.44
	l	92.54	87.94	90.18	

- **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

Train (size)	C	Pre.	Rec.	F-S.	Acc.
bite one's tongue (166)	n	40.79	79.49	53.91	74.94
	l	94.10	73.91	82.79	
break the ice (541)	n	39.05	52.56	44.81	76.12
	l	88.36	81.45	84.77	
pass the buck (262)	n	41.01	73.08	52.53	75.65
	l	92.61	76.23	83.62	
play with fire (566)	n	39.29	84.62	53.66	73.05
	l	95.29	70.43	81.00	

- None of the difference is statistically significant