Using Gaussian Mixture Models to Detect Figurative Language in Context

Linlin Li and Caroline Sporleder
Saarland University, Postfach 15 11 50
66041 Saarbrücken, Germany
{linlin, csporled}@coli.uni-saarland.de

Abstract

We present a Gaussian Mixture model for detecting different types of figurative language in context. We show that this model performs well when the parameters are estimated in an unsupervised fashion using EM. Performance can be improved further by estimating the parameters from a small annotated data set.

1 Introduction

Figurative language employs words in a way that deviates from their normal meaning. It includes idiomatic usage, metaphor, metonymy or other types of creative language. Being able to detect figurative language is important for a number of NLP applications, e.g., machine translation.

Simply checking the input against an idiom dictionary does not solve the problem. While some expressions (e.g., *trip the light fantastic*) are always used idiomatically, many expressions (e.g., *spill the beans*), can take on a literal meaning as well. Whether such expression is used idiomatically or not has to be inferred from the discourse context. Likewise, simple dictionary look-up would not work for truly creative, one-off usages; these can neither be found in a dictionary nor can they be detected by standard idiom extraction methods, which apply statistical measures to accumulated corpus evidence for an expression to assess its 'idiomaticity'. An example of a fairly creative usage can be found in (1), which is a variation of the idiom *put a sock in*.

(1) **Take the sock out of your mouth** and create a brand-new relationship with your mom.

We propose a method for detecting figurative language in context. Because we use context information rather than corpus statistics, our approach works also for truly creative usages.

2 Related Work

Most studies on the detection of idioms and other types of figurative language focus on one of three aspects: type-based extraction (detect idioms on the type level), token-based classification (given a potentially idiomatic phrase in context, decide whether it is used idiomatically), token-based detection (detect figurative expressions in running text).

Type-based extractions exploit the fact that idioms have many properties which differentiate them from other expressions, e.g., they often exhibit a degree of syntactic and lexical fixedness. These properties can be used to identify potential idioms, for instance, by employing measures of association strength between the elements of an expression (Lin, 1999).

Type-based approaches are unsuitable for expressions which can be used both figuratively and literally. These have to be disambiguated in context. *Token-based classification* aims to do this. A number of token-based approaches have been proposed: supervised (Katz and Giesbrecht, 2006), weakly supervised (Birke and Sarkar, 2006), and unsupervised (Fazly et al., 2009; Sporleder and Li, 2009).

Finally, *token-based detection* can be viewed as a two stage task which is the combination of type-based extraction and token-based classification. There has been relatively little work on this so far. One exception are Fazly et al. (2009) who detect idiom types by using statistical methods that model the general idiomaticity of an expression and then combine this with a simple second-stage process that detects whether the target expression is used figuratively in a given context, based on whether the expression occurs in canonical form or not.

However, modeling token-based detection as a

combination of type-based extraction and token-based classification has some drawbacks. First, type-based approaches typically compute statistics from multiple occurrences of a target expression, hence they cannot be applied to novel usages. Second, these methods were developed to detect figuratively used multi-word expressions (MWEs) and do not work for figuratively used individual words, like *sparrow* in example (2). Ideally, one would like to have a generic model that can detect any type of figurative usage in a given context. The model we propose in this paper is one step in this direction.

(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.

3 Using Gaussian Mixture Model to Detect Figurative Language

We address the problem by using Gaussian Mixture Models (GMMs). We assume that the literal (l) and non-literal (n) data are generated by two different Gaussians (*literal* and *nonliteral* Gaussian). The token-based detection task is done by comparing which Gaussian has the higher probability of generating a specific instance.

The Gaussian mixture model is defined as:

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

Where, c is the category of the Gaussian, μ_c is the mean, Σ_c is the covariance matrix, and w_c is the Gaussian weight.

Our method is based on the insight that figurative language exhibits less semantic cohesive ties with the context than literal language (Sporleder and Li, 2009). We use Normalized Google Distance to model semantic relatedness (Cilibrasi and Vitanyi, 2007) and represent the data by five types of semantic relatedness features $x = (x1, x2, x3, x4, x5)$:

$x1$ is the average relatedness between the target expression and context words,

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

where w_i is a component word of the target expression (T); c_j is one of the context words (C); $|T|$ is the total number of words in the target expression, and $|C|$ is the total number of words in the context. The term $\frac{2}{|T| \times |C|}$ is the normalization factor, which is the total number of relatedness pairs between target component words and context words.

$x2$ is the average semantic relatedness in the context of the target expression,

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

$x3$ is the difference between the average semantic relatedness between the target expression and the context words and the average semantic relatedness of the context (i.e., $x3 = x1 - x2$). It is an indicator of how strongly the target expression is semantically related to the discourse context.

$x4$ is the feature used by Sporleder and Li (2009) for predicting literal or idiomatic use in the cohesion graph based method,

$$x4 = \begin{cases} 1 & \text{if } x3 < 0 \\ 0 & \text{else} \end{cases}$$

$x5$ is a high dimensional vector which represents the top relatedness scores between the component words of the target expression and the context,

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

where the function $max(k, A)$ is defined to choose the k^{th} highest element from the set A.[1]

The detection task is done by a Bayes decision rule, which chooses the category by maximizing the probability of fitting the data into the different Gaussian components:

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

4 Evaluating the GMM Approach

4.1 Data

We evaluate our method on two data sets. The first set (*idiom set*) is taken from Sporleder and Li (2009) and consists of 3964 idiom occurrences (17 idiom types) which were manually labeled as 'literal' or 'figurative'. The second data set (*V+NP set*), consists of a randomly selected sample of 500 V+NP constructions from the Gigaword corpus, which were manually labeled.

To determine how well our model deals with different types of figurative usage, we distinguish four phenomena: *Phrase-level figurative* means that the

[1] We set k to be 100 in our experiment.

whole phrase is used figuratively. We further divide this class into expressions which are potentially ambiguous between literal and figurative usage (**nsa**), e.g., *spill the beans*, and those that are unambiguously figurative irrespective of the context (**nsu**), e.g., *trip the light fantastic*. The latter can, theoretically, be detected by dictionary look-up, the former cannot. The label *token-level figurative* (**nw**) is used when part of the phrase is used figuratively (e.g., *sparrow* in (2)). Often it is difficult to determine whether a word is still used in a 'literal' sense or whether it is already used figuratively. Since we are interested in improving the performance of NLP applications such as MT, we take a pragmatic approach and classify usages as 'figurative' if they are not lexicalized, i.e., if the specific sense is not listed in a dictionary.[2] For example, we would classify *summit* in the 'meeting' sense as 'literal' (**l**). In our data set, 7.3% of the instances were annotated as 'nsa', 1.9% as 'nsu', 9.2% as 'nw' and 81.5% as 'l'. A randomly selected sample (100 instances) was annotated independently by a second annotator. The kappa score (Cohen, 1960) is 0.84, which suggest that the annotations are reliable.

4.2 GMM Estimated by EM

We used the MatLab package provided by Calinon (2009) for estimating the GMM model. The GMM is trained by the EM algorithm. The priors of Gaussian components, means and covariance of each components, are initialized by the k-means clustering algorithm (Hartigan, 1975).

To determine whether the GMM is able to perform token-based idiom classification, we applied it to the idiom data set. The results (see Table 1) show that the GMM can distinguish usages quite well and gains equally good results as Sporleder and Li's cohesion graph method (*Co-Graph*). In addition, this method can deal with unobserved occurrences of non-literal language.

Table 2 shows the results on the second data set. The baseline predicts 'idiomatic' and 'literal' according to a biased probability which is based on the true distribution in the annotated set. *GMM* shows the performance on the whole V+NP set. We also split the test set into three different subsets to determine how the GMM performs on distinguishing literal usage from the different types of figurative usage: *GMM*$\{nsu, l\}$, *GMM*$\{nsa, l\}$, *GMM*$\{nw, l\}$.

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	

Table 1: Results on the idiom data set, n(on-literal) is the union of the predefined three sub-classes (nsu, nsa, nw), l(iteral), Acc(uracy), Pre(cision), Rec(all), F-S(core)

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	

Table 2: Results on the V+NP data set, Gaussian component parameters estimated by EM

The unsupervised GMM model beats the baseline and achieves good results on the V+NP data set. It also outperforms the Co-Graph approach, which suggests that the statistical model, GMM, is more likely to boost the performance by capturing statistical properties of the data for more difficult cases (*idioms* vs. *general figurative usages*), compared with the Co-Graph approach.

In conclusion, the model is not only able to classify idiomatic expressions but also to detect new figurative expressions. However, the performance on the second data set is worse compared with running the same model on the idiom data set. This is because the V+NP data set contains more difficult examples, e.g., expressions which are only partially figurative (e.g., (2)). One would expect the literal part of the expression to exhibit cohesive ties with the context, hence the cohesion based features may fail to detect this type of figurative usage. Consequently the performance of the GMM is lower for figuratively used words ('nw') than for idioms ('nsa', 'nsu'). However, even for 'nw' cases the model still obtains a relatively high accuracy.

[2] We used `http://www.askoxford.com`.

4.3 GMM estimated from Annotated Data

In a second experiment, we tested how well the GMM performs when utilizing the annotated idiom data set to estimate the two Gaussian components instead of using EM. We give equal weights to the two Gaussian components and predict the label on the V+NP data set by fixing the mixture model which is estimated from the training set (GMM+f). This method further improves the performance compared to the unsupervised approach (Table 3).

We also experimented with setting a threshold and abstaining from making a prediction when the probability of an instance belonging to the Gaussian is below the threshold (GMM+f+s). Table 3 shows the performance when only evaluating on the subset for which a classification was made. It can be seen that the accuracy and the overall performance on the literal class improve, but the precision for the non-literal class remains relatively low, i.e., many literal instances are still misclassified as 'non-literal'. One reason for this may be that there are a few instances containing named entities, which exhibit weak cohesive ties with the context even if though they are used literally. Using a named-entity tagger before applying the GMM might solve the problem.

Model	C	Pre.	Rec.	F-S.	Acc.
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	

Table 3: Results on the V+NP data set, Gaussian component parameters estimated by annotated data

Finally, Table 4 shows the result when using different idioms to generate the nonliteral Gaussian. The literal Gaussian can be generated from the automatically obtained nonliteral examples by Li and Sporleder (2009). We found the estimation of the GMM is not sensitive to idioms; our model is robust and can use any existing idiom data to discover new figurative expressions. Furthermore, Table 4 also shows that the GMM does not need a large amount of annotated data for parameter estimation. A few hundred instances are sufficient.

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	

Table 4: Results on the V+NP dataset, Gaussian component parameters estimated on different idioms

5 Conclusion

We described a GMM based approach for detecting figurative expressions. This method not only works for distinguishing literal and non-literal usages of a potential idiomatic expression in a discourse context, but also discovers new figurative expressions.

The components of the GMM can be effectively estimated using the EM algorithm. The performance can be further improved when employing an annotated data set for parameter estimation. Our results show that the estimation of Gaussian components are not idiom-dependent. Furthermore, a small annotated data set is enough to obtain good results.

Acknowledgments

This work was funded by the DFG within the Cluster of Excellence "Multimodal Computing and Interaction". Thanks to Benjamin Roth for discussions and comments.

References

J. Birke, A. Sarkar. 2006. A clustering approach for the nearly unsupervised recognition of nonliteral language. In *Proceedings of EACL-06*.

S. Calinon. 2009. *Robot Programming by Demonstration: A Probabilistic Approach*. EPFL/CRC Press.

R. L. Cilibrasi, P. M. B. Vitanyi. 2007. The Google similarity distance. *IEEE Trans. on Knowl. and Data Eng.*, 19(3):370–383.

J. Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurements*, 20:37–46.

A. Fazly, P. Cook, S. Stevenson. 2009. Unsupervised type and token identification of idiomatic expressions. *Computational Linguistics*, 35(1):61–103.

J. A. Hartigan. 1975. *Clustering Algorithm*. Wiley.

G. Katz, E. Giesbrecht. 2006. Automatic identification of non-compositional multi-word expressions using latent semantic analysis. In *Proceedings of the ACL06 Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*.

L. Li, C. Sporleder. 2009. Contextual idiom detection without labelled data. In *Proceedings of EMNLP-09*.

D. Lin. 1999. Automatic identification of non-compositional phrases. In *Proceedings of ACL-99*.

C. Sporleder, L. Li. 2009. Unsupervised recognition of literal and non-literal use of idiomatic expressions. In *Proceedings of EACL-09*.