UNAL: Discriminating between Literal and Figurative Phrasal Usage Using Distributional Statistics and POS tags

Sergio Jimenez, Claudia Becerra
Universidad Nacional de Colombia
Ciudad Universitaria,
edificio 453, oficina 114
Bogotá, Colombia
sgjimenezv@unal.edu.co
cjbecerrac@unal.edu.co

Alexander Gelbukh
CIC-IPN
Av. Juan Dios Bátiz, Av. Mendizábal,
Col. Nueva Industrial Vallejo
CP 07738, DF, México
gelbukh@gelbukh.com

Abstract

In this paper we describe the system used to participate in the sub task 5b in the Phrasal Semantics challenge (task 5) in SemEval 2013. This sub task consists in discriminating literal and figurative usage of phrases with compositional and non-compositional meanings in context. The proposed approach is based on part-of-speech tags, stylistic features and distributional statistics gathered from the same development-training-test text collection. The system obtained a relative improvement in accuracy against the most-frequent-class baseline of 49.8% in the "unseen contexts" (*LexSample*) setting and 8.5% in "unseen phrases" (*AllWords*).

1 Introduction

The Phrasal Semantics task-5b in SemEval 2013 consisted in the discrimination of literal of figurative usage of phrases in context (Korkontzelos et al., 2013). For instance, the occurrence in a text of the phrase "a piece of cake" can be used whether to refer to something that is pretty easy or to an actual piece of cake. The motivation for this task is that such discrimination could improve the quality and performance of other tasks like machine translation and information retrieval.

This problem has been studied in the past. Lin (1999) observed that the distributional characteristics of the literal and figurative usage are different. Katz and Giesbrecht (2006) showed that the similarities among contexts are correlated with their literal or figurative usage. Birke and Sarkar (2006) clustered literal and figurative contexts using a word-sense-disambiguation approach. Fazly et al. (2009) showed that literal and figurative usages are related to particular syntactical forms. Sporleder and Li (2009) showed that for a particular phrase the contexts of its literal usages are more cohesive than those of its figurative usages. Inspired by these works and in a new observation, we proposed a set or features based on cohesiveness, syntax and stylometry (Section 2), which are used to train a machine learning classifier.

The cohesiveness between a phrase an its context can be measured aggregating the relatedness of the context words against the target phrase. This cohesiveness should be high for phrases used literally. Conversely, figurative usages can occur in a large variety of contexts implying low cohesiveness. For instance, the cohesiveness of the phrase "a piece of cake" against context words such as "coffee", "birthday" and "bakery" should be high. The distributional measures used to obtain the needed relatedness scores and the proposed measures of cohesiveness are presented in subsection 2.1.

Moreover, we observed a stylistic trend in the training data set. That is, figurative usage tends to occur later in the document in comparison with the literal usage. Consequently, a small set of features that exploits this particular observation is proposed in subsection 2.2.

Fazly et al. (2009) showed that idiomatic phrases composed of a verb and a noun (e.g. "break a leg") differ from their literal usages in the use of some syntactic structures. For instance, idiomatic phrases are less flexible in the use of determiners, pluraliza-

*Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 114–117, Atlanta, Georgia, June 14-15, 2013.

tion and passivization. In order to capture that notion in a simple way, a set of features form a part-of-speech tagger was included in the feature set (see subsection 2.3).

In Section, additional details of the proposed system are provided jointly with the obtained official results. Finally, in sections 4 and 5 a brief discussion of the results and some concluding remarks are presented.

2 Features

Each instance of the training and test sets consist of a short document d where one or more occurrences of its target phase p_d are annotated. For each particular phrase p, several instances are provided corresponding to literal or figurative usages. In this section, the set of features that was extracted from each instance to provide a vectorial representation is presented.

2.1 Cohesiveness Features

Let's start with some definitions borrowed from the information retrieval field: D is a collection of documents, $\mathrm{df}(w)$ is the number of documents in D where the word w occurs (document frequency), $\mathrm{df}(w \wedge p_d)$ is the number of documents where w and a target phrase p_d co-occur, $\mathrm{tf}(w,d)$ is the number of occurrences of w in a document $d \in D$ (term frequency), and $\mathrm{idf}(w) = \log_2 \frac{\mathrm{df}(w)}{|D|}$ is the inverse document frequency of w (Jones, 2004).

A simple distributional measure of relatedness between w and p can be obtained with the following ratio:

$$\mathrm{R}(w,p) = \frac{\mathrm{df}(w \wedge p_d)}{\mathrm{df}(w)} \qquad (1)$$

Pointwise mutual information (PMI) (Church and Hanks, 1990) is another distributional measure that can be used for measuring the relatedness of w and p. The probabilities needed for its calculation can be obtained by maximum likelihood estimation (MLE): $P(w) \approx \frac{\mathrm{df}(w)}{|D|}$, $P(p_d) \approx \frac{\mathrm{df}(p_d)}{|D|}$ and $P(w \wedge p_d) \approx \frac{\mathrm{df}(w \wedge p_d)}{|D|}$.

Thus, PMI is given by this expression:

$$\mathrm{PMI}(w,p_d) = \log_2 \left(\frac{P(w \wedge p_d)}{P(w) \cdot P(p_d)} \right) \qquad (2)$$

F1:	$\sum_{w \in d'} \mathrm{R}(w,p_d)$
F2:	$\sum_{w \in d'} \mathrm{tf}(w,d)$
F3:	$\sum_{w \in d'} \mathrm{idf}(w)$
F4:	$\sum_{w \in d'} \mathrm{PMI}(w,p_d)$
F5:	$\sum_{w \in d'} \mathrm{NPMI}(w,p_d)$
F6:	$\sum_{w \in d'} (\mathrm{tf(w,d)} \cdot \mathrm{R}(w,p_d))$
F7:	$\sum_{w \in d'} (\mathrm{idf}(w) \cdot \mathrm{R}(w,p_d))$
F8	$\sum_{w \in d'} (\mathrm{R}(w,p_d) \cdot \mathrm{PMI}(w,p_d))$
F9:	$\sum_{w \in d'} (\mathrm{R}(w,p_d) \cdot \mathrm{NPMI}(w,p_d))$
F10:	$\sum_{w \in d'} (\mathrm{tf}(w,d) \cdot \mathrm{idf(w)})$
F11:	$\sum_{w \in d'} (\mathrm{tf}(w,p_d) \cdot \mathrm{PMI}(w,p_d))$
F12:	$\sum_{w \in d'} (\mathrm{tf}(w,p_d) \cdot \mathrm{NPMI}(w,p_d))$
F13:	$\sum_{w \in d'} (\mathrm{idf}(w) \cdot \mathrm{PMI}(w,p_d))$
F14:	$\sum_{w \in d'} (\mathrm{idf}(w) \cdot \mathrm{NPMI}(w,p_d))$
F15:	$\sum_{w \in d'} (\mathrm{PMI}(w,p_d) \cdot \mathrm{NPMI}(w,p_d))$
F16:	$\sum_{w \in d'} (\mathrm{tf}(w,d) \cdot \mathrm{idf(w)} \cdot \mathrm{R(w,p_d)})$
F17:	$\sum_{w \in d'} (\mathrm{tf}(w,d) \cdot \mathrm{R}(w,p_d) \cdot \mathrm{PMI}(w,p_d))$
F18:	$\sum_{w \in d'} (\mathrm{tf}(w,d) \cdot \mathrm{R}(w,p_d) \cdot \mathrm{NPMI}(w,p_d))$
F19:	$\sum_{w \in d'} (\mathrm{tf}(w,d) \cdot \mathrm{idf(w)} \cdot \mathrm{PMI(w,p_d)})$
F20:	$\sum_{w \in d'} (\mathrm{tf}(w,d) \cdot \mathrm{idf(w)} \cdot \mathrm{NPMI(w,p_d)})$

Table 1: Cohesiveness features

Furthermore, the scores obtained through eq. 2 can be normalized in the interval [+2,0] with the following expression:

$$\mathrm{NPMI}(w,p_d) = \frac{\mathrm{PMI}(w,p_d)}{-\log_2(P(w \wedge p_d))} + 1 \qquad (3)$$

A measure of the cohesiveness between a document d against its target phrase p_d, can be obtained by aggregating the pairwise relatedness scores between all the words in d and p_d. For instance, using eq. 1 that measure is $\sum_{w \in d'} \mathrm{R}(w,p_d)$, where d' is the set of different words in d. The equations 1, 2 and 3 can be used as weights associated to each word, which can also be combined among them and with *tf* and *idf* weights. Such weight combinations produce measures that can be used as cohesiveness features for a document. The set of 20 features obtained using this approach is shown in Table 1.

2.2 Stylistic Features

The set of stylistic features related to the document length, vocabulary size and relative position of the occurrence of the target phrase in a document is shown in Table 2.

F21:	Relative position of p_d in d
F22:	Document length in characters
F23:	Document length in tokens
F24:	Number of different words

Table 2: Stylistic features

2.3 Syntactic Features

The features F25 to F67 correspond to the set of 43 part-of-speech tags of the NLTK English POS tagger (Loper and Bird, 2002). Each feature contains the frequency of occurrence of each POS-tag in a document d.

3 Experimental Setup and Results

The data provided for this task consists of two data sets *LexSample* and *AllWords*, which are divided into development, training and test sets. Nevertheless, we considered a single training set aggregating the development and training parts from both data sets for a total of 3,230 instances. Each training instance has a class label whether "literally" or "figuratively" depending on the usage or the target phrase. Similarly, the aggregated test set contains 1,112 instances, but with unknown values in the class attribute.

Firstly, the syntactic features for each text were obtained using the POS tagger included in the NLTK v.2.0.4 (Loper and Bird, 2002). Secondly, all texts were preprocessed by tokenizing, lowecasing, stopword removing, punctuation removing and stemming using the Porter's algorithm (1980). This preprocessed version of the texts was used to obtain the remaining cohesiveness and stylistic features. The resulting vectorial data set was used to produce the predictions labeled "UNAL.RUN1" through a Logistic classifier (Cessie and Houwelingen, 1992). The implementation used for this classifier was the included in WEKA v.3.6.9 (Hall et al., 2009). The accuracies obtained by the different feature groups in the training set using 10-fold cross validation are shown in Table 3. The last column shows the percentage of relative improvement of different feature sets combinations from the most frequent class baseline to our best system using all features.

The predictions labeled "UNAL.RUN2" were obtained with the same vectorial data set but adding as a nominal feature the target phrase of each instance. The official results obtained by both submitted runs are shown in Table 4. Note that official results in the test set are reported separately for the data sets *LexSample* and *AllWords*. The *LexSample* test set contains instances whose target phrases were seen in the training set (i.e. unseen contexts). Unlike *LexSample*, *AllWords* contains instances whose target phrases were unseen in the training set (i.e. unseen phrases).

Features	Accuracy	% improv.
All features	**0.7272**	100.0%
Cohesiveness+Syntactic	0.7034	87.1%
Cohesiveness	0.6833	76.2%
Syntactic	0.6229	43.5%
Stylistic	0.5492	3.5%
Baseline MFC	0.5427	0.0%

Table 3: Results by group of features in the training set using 10-fold cross validation

System	*LexSample*	*AllWords*	Both
UNAL.RUN1	0.7222	**0.6680**	0.6970
UNAL.RUN2	**0.7542**	0.6448	**0.7032**
Baseline MFC	0.5034	0.6158	0.5558
Best SemEval'13	0.7795	0.6680	0.7276
# test instances	594	518	1,112

Table 4: Official results in the test set (accuracy)

4 Discussion

As it was expected, the results obtained in the "unseen context" setting were consistently better than in "unseen phrases". This result suggests that the discrimination of literal and figurative usage heavily depends on particular idiomatic phrases. This can also be confirmed by the best accuracy obtained by RUN2 compared with RUN1 in *LexSample*. Clearly, the classifier used in RUN2 exploited the identification of the phrase to leverage a priori information about the phrase such as the most frequent usage.

Another factor that could undermine the results in the "unseen phrases" setting is the low number of instances per phrase in the *AllWords* test set, roughly a third in comparison with *LexSample*. Given that the effectiveness of the cohesiveness features depends

on the number of documents where the idiomatic phrase occurs, the predictions for this test set relied mainly on the less effective features, namely syntactic and stylistic features (see Table 3). However, this problem could be alleviated obtaining the distributional statistics from a large corpus with enough occurrences of the unseen phrases.

Besides it is important to note, that in spite of the low individual contribution of the stylistic features to the overall accuracy (3.5%), when these are combined with the remaining features they provide an improvement of 12.9% (see Table 3).

5 Conclusions

We participated in the Phrasal Semantics sub task 5b in SemEval 2013. Our system proved the effectiveness of the use of cohesiveness, stylistic and syntactic features for discriminating literal from figurative usage of idiomatic phrases. The most-frequent-class baseline was overcame by 49.8% in the "unseen contexts" setting (*LexSample*) and 8.5% in "unseen phrases" (*AllWords*).

Acknowledgments

This research was funded in part by the Systems and Industrial Engineering Department, the Office of Student Welfare of the National University of Colombia, Bogotá, and through a grant from the Colombian Department for Science, Technology and Innovation, Colciencias, proj. 1101-521-28465 with funding from "El Patrimonio Autónomo Fondo Nacional de Financiamiento para la Ciencia, la Tecnología y la Innovación, Francisco José de Caldas." The third author recognizes the support from Mexican Government (SNI, COFAA-IPN, SIP 20131702, CONACYT 50206-H) and CONACYT–DST India (proj. 122030 "Answer Validation through Textual Entailment").

References

Julia Birke and Anoop Sarkar. 2006. A clustering approach for nearly unsupervised recognition of nonliteral language. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy.

S. Le Cessie and J. C. Van Houwelingen. 1992. Ridge estimators in logistic regression. *Applied Statistics*, 41(1):191.

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Comput. Linguist.*, 16(1):22–29, March.

Afsaneh Fazly, Paul Cook, and Suzanne Stevenson. 2009. Unsupervised type and token identification of idiomatic expressions. *Comput. Linguist.*, 35(1):61–103, March.

Mark Hall, Frank Eibe, Geoffrey Holmes, and Bernhard Pfahringer. 2009. The WEKA data mining software: An update. *SIGKDD Explorations*, 11(1):10–18.

Karen Spärck Jones. 2004. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 60(5):493–502, October.

Graham Katz and Eugenie Giesbrecht. 2006. Automatic identification of non-compositional multi-word expressions using latent semantic analysis. In *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, MWE '06, pages 12–19, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ioannis Korkontzelos, Torsten Zesch, Fabio Massimo Zanzotto, and Chris Biemann. 2013. SemEval-2013 task 5: Evaluating phrasal semantics. In *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013)*.

Dekang Lin. 1999. Automatic identification of non-compositional phrases. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, ACL '99, page 317–324, Stroudsburg, PA, USA. Association for Computational Linguistics.

Edward Loper and Steven Bird. 2002. NLTK: the natural language toolkit. In *Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics. Philadelphia*. Association for Computational Linguistics.

Martin Porter. 1980. An algorithm for suffix stripping. *Program*, 3(14):130–137, October.

Caroline Sporleder and Linlin Li. 2009. Unsupervised recognition of literal and non-literal use of idiomatic expressions. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, page 754–762, Stroudsburg, PA, USA. Association for Computational Linguistics.