Automatic Detection of Idiomatic Language

Submitted as part of the requirements for:

CE902 Professional Practice and Research Methodology

**Name**: Jose Juan Zavala Iglesias

**Supervisor**: Aline Villavicencio

**Date**: 22 March 2019

**Abstract**. This sample document is not by any means complete and accurate. It merely illustrates some of the basic mechanisms by which Microsoft Word 2007 can be used to write technical reports. Guidelines about the structure and contents of each report will be distributed separately.

**Keywords**: Keywords help your reader to set the context to the report. Name the relevant disciplines and fields of research, such as: software design, computer networks, circuit theory.

Table of Contents

[1 Introduction 3](#_Toc2862690)

[2 Literature Review 3](#_Toc2862691)

[2.1 Supervised Methods 3](#_Toc2862692)

[2.2 Unsupervised Methods 3](#_Toc2862693)

[Acknowledgements 3](#_Toc2862694)

[References 4](#_Toc2862695)

[Appendix 4](#_Toc2862696)

# Introduction

// Mention relevance of idioms detection in translation and semantic parsing

// Idioms appear over time, so we need ever-growing corpora.

// Write examples of VNCs

Multiword Expressions (MWEs) are combinations of multiple words that exhibit some degree of idiomaticity, but not necessarily [1]. Verb-noun combinations (VNCs) are a common type of MWEs that consist of a verb with a noun in its direct object position [1].

# Literature Review

Much research on MWE identification focuses on the task of the identification of specific kinds on MWEs, such as English VNCs [1] [2] [3], while other authors focus on multilingual detection of MWEs [4]. Another great distinction is between the tasks of *idiom type* and *idiom token* classification; while *idiom type* classification is the task of identifying expression with possible idiomatic interpretations, *idiom token* classification focuses on distinguishing between idiomatic and literal usages of potentially idiomatic phrases [5].

This research proposal will focus on the task of detecting idiomatic usage of token-level (*idiom token* classification) English VNCs.

What distinguishes idiomatic and literal MWEs is the fact that an idiom has a different meaning from the simple composition of the meaning of its component words [3].

## Knowledge on Idioms

Past research focus on VNC analysis since they have been able to extract lexical and semantic consistencies across different idiomatic phrases in the English language. First is the observation that most idiomatic VNCs exhibit **lexico-syntactic fixedness** [1] [3]; i.e. the phrase *see stars* often presents idiomatic meaning when the verb has active voice, the determiner is null, and the noun is in plural form, as in *see stars* or *seeing stars*; while usages with a determiner (*see the stars*), singular noun form (*see a star)*, or passive voice (*stars where seen*) often have literal interpretation [1].

**Lexical-fixedness** of idiomatic phrases means that the substitution of a near synonym for a constituent does not preserve the idiomatic meaning of the expression [3] (i.e. *see stars* and *observe stars*). Even if some idioms allow lexical variations which generate closely related meanings, these are usually highly unpredictable substitutions that can’t be considered as a rule [3].

**Syntactic-fixedness** means that many idiomatic VNCs cannot undergo syntactic variations while retaining their idiomatic interpretation (i.e. the punch let him *seeing stars* / *seeing the star*); however, it is relevant to note that idiomatic VNCs differ with respect to their degree of tolerance to semantic operations (**syntactic flexibility**) [3].

Following on the concept of lexico-syntactic fixedness, a corpus-based study by [6] demonstrates that idiomatic phrases are not as fixed as literature assumed in the past, since “the corpus data in this chapter show that-in contrast to nonidiomatic combinations of words-idioms have strongly preferred canonical form, but at the same time the occurrence of idiom variation is too common to be ignored”. This sounds redundant, as it says that idiomatic VNC identification must be on the lookout for any form of VNC variation since they can all be idiomatic. However, it also stablishes the **canonical form**, which is the base form of the idiomatic VNC (i.e. *see stars*), as a startup point for their detection. Subsequent research works on the assumption that idiomatic VNCs are more likely to appear on canonical form that non-idiomatic phrases [3].

Phrasal idioms have also been found to involve a certain degree of semantic idiosyncrasy, which means that the idiom is hard to determine without special context or previous exposure even if the meaning of the component words is clear [3] [5]. Also, although it is traditionally believed that idioms are completely non-compositional, linguists and psycholinguists claim that they show some degree of semantic compositionality [3]. This suggests that many idioms have internal semantic structure, without ignoring the fact that they are non-compositional in a traditional sense, which opens the field for the introduction of terms such as **semantic decomposability** and/or **semantic analyzability** [3]. To say that an idiomatic VNC is semantically analysable means that the constituents contribute by their independent meanings to the idiomatic interpretation; so, the more semantically analyzable an idiom is, the easier it is to interpret the idiomatic meaning from its constituents [3].

## Supervised Methods

// Models: SVM (linear and polynomial kernels), K-NNs

Supervised model training for idiomatic phrase detection focuses on identifying if a given excerpt of a sentence is of idiomatic or literal meaning. This approach usually tackles token-level identification of VNCs as a supervised binary classification problem, classifying the use of a VNC as idiomatic of literal [1] [5].

Recent research makes use of classifier models such as k-Nearest Neighbours (k-NNs) SVM with linear and polynomial kernels, since these algorithms have been proven to work for binary classification problems [1] [5]. However, the VNC pre-processing and feature creation is the task in which most research focuses on. In an attempt to exploit the knowledge on lexical and syntactic patterns presented in Section 2.1, researchers have made use of unsupervised feature encoders to train the classifiers [1] [2] [5].

One such approach is that of **Skip-Thought Vectors** (Sent2Vec) [7], used by [5] and [1]. This model uses the continuity of text from books to train an encoder-decoder model that aims to reconstruct the surrounding sentences of an encoded passage, so sentences with similar semantic and syntactic properties are mapped to similar vector representations [7]. This results in an encoder that can product highly generic sentence representations [7]. Sent2Vec was first used for idiom detection by [5] on the assumption that in a real-world application, target phrases won’t have access to a surrounding context; which motivated the exploration of distributed compositional semantic models to produce reliable estimates of idiom token classification [5]. Utility found in the Sent2Vec model is that it is possible to infer properties of the surrounding context only from the input sentence [5] [7], which allows the classifier to learn lexical and syntactic patterns without complex methods. [5] uses the resulting encodings to train three SVM classifiers with the VNC-Tokens Dataset [8]: Linear Kernel with C=1.0, Polynomial Kernel of degree = 2 and C = 1000, and Linear Kernel trained using Stochastic Gradient Descent with a learning rate of 0.0001. Results on the classifiers (Table 1) show an improvement on the baseline set by the authors, which used entire context extracted from several paragraphs.

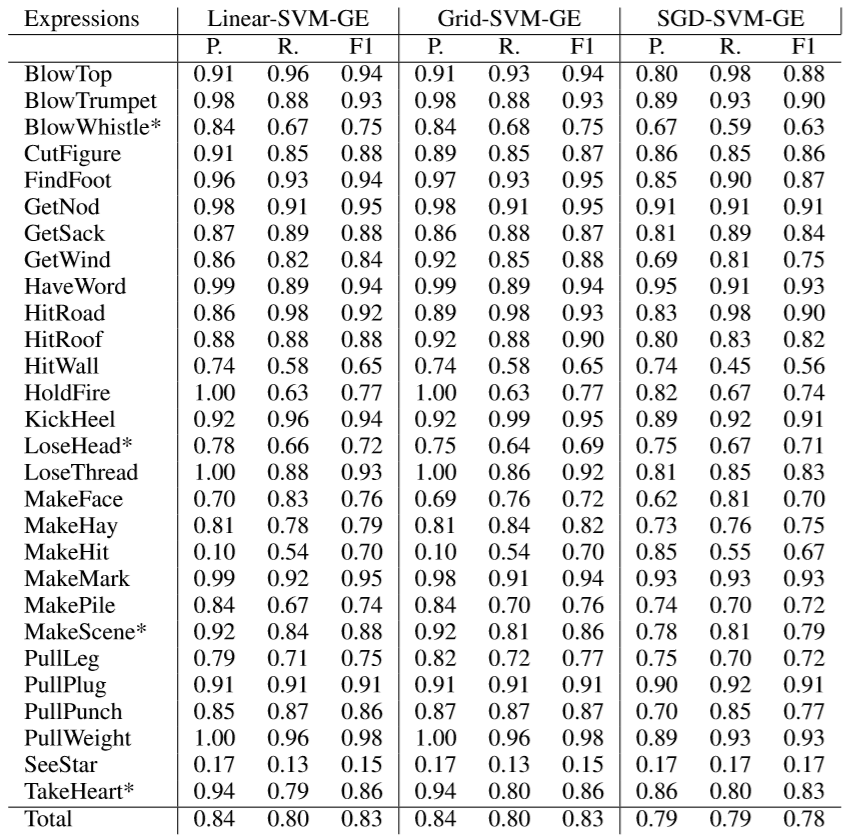


Table - Precision (P.), Recall (R.), and F1-Score (F1) results on Generic Classifiers by [5]

## Unsupervised Methods

// Talk about unsupervised methods such as [4] [3] [2]

# Acknowledgements

// Should I include acknowledgements?

# References

|  |  |
| --- | --- |
| [1] | M. King and P. Cook, “Leveraging distributed representations and lexico-syntactic fixedness for token-level prediction of the idiomaticity of English verb-noun combinations,” in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, Melbourne, 2018. |
| [2] | W. Gharbieh, V. C. Bhavsar and P. Cook, “A Word Embedding Approach to Identifying Verb-Noun Idiomatic Combinations,” in *Proceedings of the 12th Workshop on Multiword Expressions*, Berlin, 2016. |
| [3] | A. Fazly, P. Cook and S. Stevenson, “Unsupervised Type and Token Identification of Idiomatic Expressions,” *Computational Linguistics,* vol. 35, no. 1, pp. 61-103, 2009. |
| [4] | A. Savary, C. Ramish, S. R. Cordeiro, F. Sangati, V. Vincze, B. QasemiZadeh, M. Candito, F. Cap, V. Giouli, I. Stoyanova and A. Doucet, “The PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions,” in *Proceedings of the 13th Workshop on Multiword Expressions*, Valencia, 2017. |
| [5] | G. D. Salton, R. J. Ross and J. D. Kelleher, “Idiom Token Classification using Sentential Distributed Semantics,” in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, Berlin, 2016. |
| [6] | S. Z. Riehemann, “A Constructional Approach to Idioms and Word Formation,” Stanford University, Stanford, 2001. |

# Appendix

// Should I include an appendix?