An unsupervised model for extracting MWEs, based on their collocational strength

1 Introduction

Multiword expressions (MWEs) are important linguistic structures that exist in a wide range of NLP tasks [12]. Being correctly extracted and identified, MWEs can considerably improve the accuracy of such tasks. In this work, we present a method of extraction of MWEs. We try to model collocational strength as a discriminant property of a subcategory of MWEs, i.e., Noun Compounds (NCs) - the most common type of English MWEs [1], based on the following hypothesis: *For a given compound, the head is more likely to co-occur with the modifier compare to its synonyms; and the modifier is more likely to co-occur with the head compare to its synonyms.* Preliminary experiments show that our method leads to a higher accuracy of extracting MWEs, compare to the state of the art methods. To the best of our knowledge, modeling collocational strength has never been proposed the way we present in this work, and similar models do not produce the same level of accuracy.

2 A Brief Literature Review

Attempts to extract MWEs are of different types. Some of them are rule-based and symbolic [13, 10]. Some rely on lexicons [7, 9]. Some are fully statistical [5, 15], and some are hybrid in the sense that they benefit from both statistical and linguistic information [14, 3]. To date, it has been shown that in practice, for extraction of MWEs, statistical models produce better results than rule-based models, in terms of precision.

Some extraction techniques build upon generic properties of MWEs, for instance non-compositionality in [16] and [6] which employs distributional similarity methods to estimate this property; or conventionalization [5]. There are also bilingual models which are mostly based on the assumption that a translation of a source language MWE exists in a target language, for instance models presented in [2] [8]. Another example is [4] which is a hybrid bilingual model that aims at extracting English-Chinese MWEs.

Finally, [11] presents a generic extraction model where the author develops a flexible platform that accepts different types of criteria (from statistical to deep linguistic) in order to extract and filter MWEs. However, in the this work, as the author claims, the quality of the extraction is highly dependent on the level of deep linguistic analysis, and thereby, statistical criterion’s role is less significant.

3 Method

We formulate collocational strength of a candidate MWE based on the following hypothesis ($H_0$): *In a given idiosyncratic Noun/Adjective-Noun pair, the head is more likely to appear with the present modifier compare to its synonyms, and reciprocally, the modifier is more likely to appear with the present head compare to its synonyms.* We formulate two models to test our hypothesis. The first model ($M_1$) which is described by Equation (1), implements a simplified form of our hypothesis ($H_3$): *In a given idiosyncratic pair, the head is more likely to appear with the present modifier compare to its synonyms:*

$$M_1 : P(w_2|w_1) > P(w_2|\text{Synset}(w_1))$$ (1)
where:

\[ P(w_2|w_1) = \frac{\#(w_1w_2)}{\#(w_1)} \]

and,

\[ P(w_2|\text{Synset}(w_1)) = \frac{\sum_{w'_1 \in \text{Synset}(w_1)} \#(w'_1w_2)}{\sum_{w'_1 \in \text{Synset}(w_1)} \#(w'_1)} \]

The second model \( M_2 \) on the other hand implements the original hypothesis \( H_a \):

\[ M_2 : P(w_1w_2|\text{Synset}(w_1)\text{Synset}(w_2)) > P(w_1|\text{Synset}(w_1)\text{Synset}(w_2)) \times P(w_2|\text{Synset}(w_1)\text{Synset}(w_2)) \]

which can be further expanded to:

\[ \frac{\#(w_1w_2)}{\sum_{w'_1 \in \text{Synset}(w_1)} \#(w'_1w_2)} > \frac{\sum_{w'_2 \in \text{Synset}(w_2)} \#(w_1w'_2)}{\sum_{w'_1 \in \text{Synset}(w_1)} \sum_{w'_2 \in \text{Synset}(w_2)} \#(w'_1w'_2)} \]

3.1 Evaluation

We create an evaluation set that consists of negative and positive MWE-NC examples. This set is partly created manually, and partly borrowed from the MWE lexicons presented in [1]. We then classify the elements of this set to MWE and non-MWE using our models; calculate the precision and recall of these models; and compare the classification results against the following two baselines: Multinomial Likelihood, and Mutual Information [5]. Afterward, we test our main hypothesis \( H_a \) against the following null hypothesis: *In an idiosyncratic pair, the head is not more likely to appear with the modifier compare to its synonyms, and the modifier is not more likely to appear with the head compare to its synonyms.*

References


