Modeling the Statistical Idiosyncrasy of Multiword Expressions

Abstract

The focus of this work is statistical idiosyncrasy (or collocational weight) as a discriminant property of multiword expressions. We formalize and model this property, compile a 2-class dataset of MWE and non-MWE examples, and evaluate our models on this dataset. We present a possible empirical implementation of collocational weight and study its effects on identification and extraction of MWEs. Our models prove to be more effective than baselines in identifying noun/adjective-noun MWEs.

1 Introduction

Multiword Expressions (MWEs) are sequences of words that show some levels of idiosyncrasy. For instance they can be semantically idiosyncratic (i.e., their meaning can not be readily inferred from the meaning of their components), syntactically idiosyncratic (their syntax can not be extracted from the syntax of their components), statistically idiosyncratic (their components tend to co-occur more often than expected by chance), or have other forms of idiosyncrasy. MWEs comprise several types and sub-types. Although it is not always clear where to draw the line between various types of MWEs, the two broadest categories are lexicalized MWEs and institutionalized MWEs (Sag et al., 2002). The main property of lexicalized MWEs is syntactic or semantic idiosyncrasy and the main property of institutionalized MWEs is statistical idiosyncrasy. Semantic idiosyncrasy is closely related to the concept of non-compositionality. It is important to note that a MWE is often idiosyncratic in more than one way (Baldwin and Kim, 2010). This means lexicalized MWEs can be statistically idiosyncratic, and institutionalized MWEs can be semantically idiosyncratic. Institutionalized MWEs are closely related to collocations\(^1\). They can be compositional (seat belt) or non-compositional (hard drive), but statistically they co-occur more often than expected by chance.

Efficient extraction and identification of MWEs can positively influence some important Natural Language Processing (NLP) tasks such as parsing (Nivre and Nilsson, 2004), and Statistical Machine Translation (Ren et al., 2009). Identification and extraction of MWEs are therefore important research questions in the area of NLP.

In this work we refer to statistical idiosyncrasy as collocational weight and present a method of modeling this property for two-word noun-noun and adjective-noun compounds. Comparative evaluation reveals better performance of proposed models compared to that of the baselines.

In previous work, it has often been suggested that collocations can be identified by their non-substitutability. This means we can not replace a collocation’s components with their near synonyms (Manning and Schütze, 1999). For instance we can not say brief film instead of short film. Pearce (2001) defines collocations as pairs of words where “one of the words significantly prefers a particular lexical realization of the concept the other represents.” To the

\(^{1}\)Although the major property of collocations is known to be statistical idiosyncrasy, in many works, semantically idiosyncratic multiword expressions have also been regarded as collocation.
best of our knowledge, however, non-substitutability (with near synonyms) or in other words collocational weight has never been empirically tested. In this work, we present two models that partially, and fully, model collocational weight, and investigate its effects on extraction of MWEs.

2 Related work

Extraction of MWEs has been widely researched from different perspectives. Various models from rule-based to statistical have been employed to address this problem.

Examples of rule-based models are Seretan (2011) and Jacquemin et al. (1997), who base their extraction on linguistic rules and formalism in order to identify and filter MWE candidates. Other examples are Baldwin (2005), who aims at extracting verb particle constructions based on their linguistic properties using a chunker and dependency grammar, and Goldman et al. (2001), who present a MWE extraction model which is based on the output of a symbolic parser. A similar work to ours is Pearce (2001) who uses WordNet in order to produce anti-collocations from synonyms of the components of a MWE candidate, and decides about “MWEhood” based on these anti-collocations.

There are also approaches that are mainly based on statistics and statistical models. For instance Pecina (2010), Evert (2005), Lapata and Lascarides (2003), Smadja et al. (1996), and the early work Xtract (Smadja, 1993). Farahmand and Martins (2014) present a method of extracting MWEs based on their statistical contextual properties and Hermann et al. (2012) employ distributional semantics to model non-compositionality and use it as a way of identifying noun compounds.

Other approaches are hybrid in the sense that they benefit from both statistical and linguistic information (Seretan and Wehrli, 2006; Baldwin and Villavicencio, 2002; Piao and McEnery, 2001; Dias, 2003).

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 (Smith, 2014; de Caseli et al., 2010; Ren et al., 2009; Nerima et al., 2003). Another example of bilingual models is Moirón and Tiedemann (2006) which measures MWEs candidates’ idiosyncrasy based on translational entropy. Lastly, Ramisch (2012) implements a flexible platform that accepts both statistical and deep linguistic criteria in order to extract and filter MWEs.

3 Method

Following previous work by Manning and Schütze (1999), and Pearce (2001), we define collocational weight -a discriminant property of mainly institutionalized but also lexical MWEs, for noun/adjective-noun pairs according to the following hypotheses:

**Simplified Hypothesis** For a given two-word compound, the head word is more likely to co-occur with the modifier than with synonyms of the modifier.

**Main Hypothesis** For a given two-word compound, the head word is more likely to co-occur with the modifier than with synonyms of the modifier, and the modifier is more likely to co-occur with the head than with synonyms of the head.

We formalize these hypotheses in the form of $M_1$ and $M_2$ models which implement the simplified and main hypotheses and are described by equations (1) and (2), respectively.

$$M_1 : P(w_2|w_1) > \alpha P(w_2|\text{Syns}(w_1))$$

where:

$$P(w_2|w_1) = \frac{\#(w_1 w_2)}{\#(w_1)}$$

and

$$P(w_2|\text{Syns}(w_1)) = \frac{\sum_{w_1' \in \text{Syns}(w_1)} \#(w_1' w_2)}{\sum_{w_1' \in \text{Syns}(w_1)} \#(w_1' + \mathcal{L})}$$

$w_1 w_2$ represents a compound. $\text{Syns}(w)$ represents a set of synonyms of $w$, and in order to obtain such a set we use WordNet’s $\text{synset}(\cdot)$ function. $\mathcal{L}$ is the smoothing factor, which is set to 0.1, and $\alpha$ is a parameter that we altered between $[1−30]$. $\mathcal{L}$ and $\alpha$ are also present in $M_2$ and are assigned the same values as in $M_1$. 
\[ M_2 : P(w_2|w_1) > \alpha P(w_2|\text{Syns}(w_1)) \quad (2) \]
\[ \& \& P(w_1|w_2) > \alpha P(w_1|\text{Syns}(w_2)) \]

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

and
\[
P(w_2|\text{Syns}(w_1)) = \frac{\sum_{w'_1 \in \text{Syns}(w_1)} #(w'_1w_2)}{\sum_{w'_1 \in \text{Syns}(w_1)} #(w'_1) + \mathcal{L}}
\]
\[
P(w_1|\text{Syns}(w_2)) = \frac{\sum_{w'_2 \in \text{Syns}(w_2)} #(w_1w'_2)}{\sum_{w'_2 \in \text{Syns}(w_2)} #(w'_2) + \mathcal{L}}
\]

\section{Experiments}

In order to test our hypotheses, we implement the two models described above and two baselines, and run a comparative evaluation. We divide our data into two subsets: development and test sets. The evaluation is carried out in two phases. In the first phase we perform model selection and find the optimal parameters for various models on the development set. In the second phase we evaluate the selected models with optimal parameters on the test set, which remains unseen by the models up to this phase.

\subsection{Data}

Although there are several datasets for English MWEs, e.g., datasets introduced by Baldwin and Kim (2010), and Reddy et al. (2011), to the best of our knowledge there is no dataset with annotations for both MWE and non-MWE classes. This was required for us to evaluate our models. We therefore compiled our own dataset. We extracted a set of noun/adjective-noun pairs from POS-tagged Wikipedia and annotated them as MWE and non-MWE. These pairs were extracted randomly from across English Wikipedia. The annotations were based on the unanimous decision of two linguists. The pairs which were either semantically or statistically idiosyncratic (collocational), or both, were annotated as MWE. The pairs which were neither semantically nor syntactically nor statistically idiosyncratic were annotated as non-MWE. This resulted in a set of \( \approx 1200 \) annotated examples. For the sake of more reliable results, we kept only the pairs whose both head and modifier had more than three synonyms according to WordNet. This resulted in a set that comprises 257 MWE and 216 non-MWE examples. We divide this set into development (2/3) and test (1/3) sets, which contain the same proportion of MWE and non-MWE classes. Table 1 describes our dataset in more detail.

\begin{table}[h]
\centering
\begin{tabular}{|l|c|c|}
\hline
Set & MWE & non-MWE \\
\hline
original set & 257 & 216 \\
dev. set & 171 & 143 \\
test set & 86 & 73 \\
\hline
\end{tabular}
\caption{Dataset statistics.}
\end{table}

\subsection{Evaluation}

We implement the following two baselines: (1) Multinomial likelihood (Evert, 2005), which calculates the probability of the observed contingency table for a given pair under the null hypothesis of independence. (2) Mutual information (Church and Hanks, 1990), which calculates the mutual dependency of words of a co-occurrence, and has been proved efficient in identification and extraction of MWEs (Pecina, 2010; Evert, 2005). With respect to the range of scores, we set and alter a threshold for multinomial likelihood (\( M.N.L \) hereafter) and mutual information (\( M.I. \) hereafter). Pairs that obtain a score above the threshold are considered MWE, and pairs that obtain a score below the threshold are considered non-MWE. Figure 1 illustrates the precision-recall curve for our models and the baselines on the development set.

The two baseline models i.e., \( M.N.L. \) and \( M.I. \) reach a high precision only at the cost of a dramatic loss in recall. They behave quite similarly, however, \( M.I. \) performs slightly better with respect to maximum precision. \( M_2 \) reaches a high precision and recall, however, its precision declines rather quickly when recall increases. \( M_1 \) shows a more steady be-
 Recall

 Precision

 M1

 M2

 M.N.L.

 M.I.

 Behaviour in the sense that reaching a higher recall doesn’t significantly impact its precision. Figure 2 shows how $F_1$ score changes for various models when tuning parameters in order to go from high precision to high recall. $M_1$ and $M_2$ constantly have a higher $F_1$ score, where $M.I.$ and $M.N.L.$ start off with a low score and reach a score which is comparable with that of the other models.

![Figure 1: Precision-recall curve for various models.](image1.png)

Out of the four tested models, with respect to $F_1$ scores, we select $M_1$, $M_2$, and $M.I.$ for further experiments. We set the relevant parameters to optimal values (obtained by looking at the highest $F_1$ score) and run the next experiments on the test set, which has remained unseen by the models up to this point. Table 2 shows the result of these experiments. The performance of all three models on the test set is consistent with their performance on the development set. $M_2$ reaches the highest precision and $F_1$ score. $M_1$ has the highest recall, and $M.I.$ has a high recall and a reasonable but not very high precision.

![Figure 2: $F_1$ score for various models.](image2.png)

<table>
<thead>
<tr>
<th>Model</th>
<th>precision</th>
<th>recall</th>
<th>$F_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$M_1$</td>
<td>0.65</td>
<td>0.96</td>
<td>0.78</td>
</tr>
<tr>
<td>$M_2$</td>
<td>0.79</td>
<td>0.86</td>
<td>0.82</td>
</tr>
<tr>
<td>$M.I.$</td>
<td>0.56</td>
<td>0.91</td>
<td>0.70</td>
</tr>
</tbody>
</table>

Table 2: Evaluation results in terms of precision, recall and $F_1$ score for the three selected models.

5 Conclusions

We showed that statistical idiosyncrasy can play a significant role in identification and extraction of MWEs. We showed that this property can be used efficiently to extract compounds which constitute the largest subset of English MWEs. We referred to statistical idiosyncrasy as collocational weight and formalized this property and implemented two corresponding models. We empirically tested the performance of these models against two baselines and showed that one of our models constantly performs better than the baselines in terms of $F_1$ score.

References


Pavel Pecina. 2010. Lexical association measures and collocation extraction. Language resources and evaluation, 44(1-2):137–158.


