French Multi Word Expressions: Using Data on Different Patterns for Extraction and Validation.

Marie Dubremetz
Uppsala universitet,
Institutionen för lingvistik och filologi
- Box 635 -
751 26 Uppsala, Suede

Abstract

Multi Word Expressions (MWE) are an important problem in NLP. Many researchers use association measures for collecting and evaluating MWE candidates. In this paper we propose to check if it is legitimate to use those measures when data are only collected on one pattern of MWE (e.g. Noun-Adjective) for evaluating candidates belonging to another pattern (e.g. Noun-Noun). For this purpose, we run tests on the French Europarl corpus. Using association measures extracted from Noun-Adjective patterns as features, we train a model that we evaluate on instances of Noun-Noun candidates. We notice with this method that the model will still evaluate correctly a quarter of the candidates. However the result tend to be lower.

1 Introduction

This paper deals with Multi Word Expressions (MWE) we will focus on the problem of their extraction for a lexicographical purpose. This task is important for improving lexical resources used for tasks such as tokenization, parsing or translation. After presenting the previous researches, we will present our method based on a tool called the MWEtoolkit and on the use of a lexicon, we then will present our result and discuss it.

2 Previous Researches

2.1 Techniques of Extractions

The automatic validation of MWE extracted from corpora is a laborious and very repetitive task, the main reason is because even if we know some grammatical pattern relevant to MWE structure, the elements extracted automatically tend to be noisy. Some techniques exist for selecting candidates that are more likely to be the true ones. Candidates can be validated against an external resource, such as lexicons. It is possible also to check the frequency of the word expression in an other corpora like the web. Villavicencio [2005], for example, uses number of hits on Google for validating the likelihood of particle verbs.

However as Ramisch [2012] remembers in his introduction, MWE is an institutionalised phenomenon. This means that a MWE is frequently used and is part of the vocabulary of a speaker as well as the simple words. It means also that MWE have specific statistical properties that have been studied. The result of those studies are statistical measures such as DICE score, maximum likelihood Estimator, pointwise mutual information, T-score. As Islam et al. [2012] remember in a study of Google Ngram, those measures of association are language independent. And it is proven by Pecina [2008] that combining different collocation measures using standard statistical classification methods improves over using a single collocation measure. However nowadays using only lexical association
measures for extracting and validating MWE is not considered as the most efficient methods. The tendency those last years is to combine measure associations with linguistics features [Ramisch et al., 2010a, Pecina, 2008, Tsvetkov and Wintner, 2011].

2.2 MWEtoolkit
Among the tools developed for extracting MWE, MWEtoolkit is one of the most recent. Developed by Ramisch et al. [2010b] it aims not only at extracting candidates for potential multi word expressions but also at extracting their association measures. Provided that a lexicon of MWE is available and provided a preprocessed corpus, it makes possible to train a machine learning system with the association measures as a feature with a few amount of implementation to do. Ramisch et al. [2010b] provide experiments on Portuguese, English and Greek. To the best of our knowledge only Zilio et al. [2011] provide experiments with this tool as well. In this research, after having trained a machine on bigrams MWE, he tries to extract full n-gram expressions on the Europarl corpus. He reuses then the threshold obtained on bigrams for validating full n-grams expressions. After that, he applies a second filter for getting back the false negatives by checking every MWE annotated as ‘false’ by his algorithm against a dictionary online. With this method he obtains a very good precision (over 87%) and recall (over 84%). However, we don’t really know if this result is mostly due to the coverage of the dictionary online. What is the contribution of machine learning in itself ? An other question raised by this research, is the ability of a machine trained on one kind of pattern (e.g. Adjectiv-Noun) to validate correctly an other kind of MWE pattern (e.g. Preposition-Noun). That is the reason why we will run a experiment close to the one of Zilio et al. [2011] but without getting back the false negatives with an external dictionary.

3 Our Method
Therefore the aim of our study is to check whether or not a machine trained on one particular pattern can serve for extraction MWE that fit an other pattern.

3.1 Choose of a pattern
At the opposite of Zilio et al. [2011] we run our experiment on French. The choice of a different language requires to adapt the patterns. French indeed as a latin language doesn’t show the same characteristic patterns as English. We know that there is a strong recurrence of the pattern Noun-adjective in bigrams MWE in the lexicons [Silberztein and L.A.D.L., 1990, p.82], the next most frequent pattern is Noun-Noun. Therefore thanks to candidates Noun-Adjectives extracted and validated we will make an algorithm of machine learning that will extract Noun-Noun MWE.

3.2 Corpus
We ran our experiment on a language that, to the best of our knowledge, has never been used with the MWEtoolkit in previous researches : French. We work on the Europarl corpus. We extracted three parts of equal size (one million words each) for running our experiments, so we had parts for training, developing and validating our model. For avoiding any bias in the corpus selection we took the three first millions words of Europarl and sliced it into three equal parts.

3.3 Preprocessing
For preprocessing we used the same processes as described in Zilio et al. [2011]. First we ran the sentence splitter and the tokenizer provided with the Europarl corpus. Then we ran TreeTagger [Schmid, 1994] for obtaining the tags and the lemmas.

3.4 Tool for extraction
The MWEToolkit takes as an input a preprocessed corpus and give, among other outputs, an arff file which is a format adapted to the machine learning framework Weka. At the end of the process we obtain for each candidates the following features: Maximum Likelihood Estimator, Pointwise Mutual Information, T-score, Dice’s coefficient, Loglikelihood, Validation Result. All those features are numerical features
except for the last one that is a binary value (True or False). And it is this last value that will permit to do supervised machine learning.

### 3.5 Choose of a Lexicon in French

For doing supervised Machine learning, we needed to validate MWE. This information is obtained thanks to a list of MWE. In the English version of MWE toolkit a list of collocations is already provided and implemented as a gold standard. In French it is necessary to provide an external resource. So our gold standard will be the french dictionary Dela [Silberztein and L.A.D.L., 1990], the MWE part of this dictionary is called Delac. It is a generalist dictionary for NLP and it includes one hundred thousands MWE expressions so it is a good lexicon to associate to the corpus Europarl. Also the technical documentation of the Delac [Silberztein and L.A.D.L., 1990, p.72] tells that this dictionary has been constituted by linguists with reference to several dictionaries. So it is a manually constituted basis that contains MWE only referenced in official lexicographic books.

In order to use the lexicon some preprocessing were required for being read into the XML format of MWE toolkit, we tagged and lemmatized each entry of the Delac with the same lemmatizer as the corpus.

### 3.6 Processing

Thanks to the MWE toolkit we extracted all the bigrams that corresponded to the patterns Noun-Adjective and Noun-Noun in two separated files for each of our three samples of one million words. Then the MWEtoolkit made an automatic annotation by checking the presence of the MWE candidates in the Delac.

### 3.7 Machine Learning

We then use the first sample of extracted and annotated candidates as Adjective-Noun for training and testing a model (80% for training 20% as development test). For finding the best model we thought that we had to privilege the recall of the positive candidates. Indeed, when a MWE candidates is annotated as true it means that it is listed in the Dela, which means that it is an officially listed MWE. However if a MWE is not in the Dela, it doesn’t mean that the candidates doesn’t fulfil all the criteria for being a MWE. For this reason obtaining a good recall is much more difficult than getting a good precision, but it is also the most important if we stay on a lexicographical purpose. We tested several algorithms offered by Weka as well as the training options suggested by Zilio et al. [2011]. However with all the features kept and for this purpose the best classification algorithm was the Bayesian Network.

### 4 Result of the machine learning tentative of automatic Noun-Noun MWE extraction.

#### 4.1 Result of training on Noun-Adjective for validating Noun-Adjective candidates

We then have tested our model the Noun-Adjective patterns extracted on the second million words Europarl sample as a validation set, we obtained the result presented in table 1

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>F-mesure</th>
</tr>
</thead>
<tbody>
<tr>
<td>True</td>
<td>0.367</td>
<td>0.316</td>
<td>0.34</td>
</tr>
</tbody>
</table>

Table 1: recall and precision for detecting Noun-Adjective patterns with a model trained on Noun Adjectiv patterns

We notice Table 1 that our model can find about one third of the Noun-Adjective MWE present in the Delac. The precision is not very high however we must remind that the candidates are extracted without any filter, so the MWE are a minority (1091/7465).

#### 4.2 Result of training on Noun-Adjective for validating Noun-Noun candidates

We then took the same model and, without any retraining, re-evaluated it on all the candidates extracted on our third sample (the third extract of one million words, not used for Noun-Adjective training and evaluation).
The recall lost about twenty per cent compared to the previous evaluation. However the biggest lost is on the precision that lost more than fifty per cent.

5 Analyse of the results

5.1 The Delac Coverage: A Problem That Explain Only Partly the Difference of Performance

The difference of recall is not extreme but still fall from a third to a quarter of the MWE that are correctly classified as 'True', we lost more than twenty per cent recall plus fifty per cent precision. One could think that this is due to a lack from the Delac its self: in Silberztein and L.A.D.L. [1990] a table shows that about 79 000 Noun-Adjective forms has been collected for 3100 Noun-Noun MWE. It represents twenty-five times less. This can seem disproportional however our corpus in itself reminds us that the pattern Noun-Adjective is much more frequent than the Noun-Noun pattern in French. We noticed that we had for each sample between twenty-three and twenty-four times more candidates as Noun-Adjective. So there is a difference of coverage but if we look at the proportions and if we assume that the patterns Noun-Noun concentrate the same proportion of MWE as the pattern Noun-Adjective it could explain a difference of twenty-three divided by twenty-five which means only eight per cent.

5.2 Evaluation

We noticed that the amount of MWE extracted is fewer for the pattern Noun-Noun (see table 3 and 4).

This difference allows us to look into details the candidates misclassified by the model. We noticed, indeed, different categories of errors that we detail in the next part.

5.2.1 The False Positives

The misclassified candidates has been annotated by a french nativ speaker. About the 64 false positives we noticed that:

- 18 were due to candidates that could be filtered or were due to a processing error. For example a big majority of them were actually involving days of the week (“jeudi soir”) and titles (“Mme+Name Entity”) which is something recurrent in the genre of the Europarl corpus. Sometimes we could notice as well mistakes in the tokenization (“avons-nous besoin”)
- 7 seem perfect candidates for being MWE, and could have been in the Delac. They are indeed belonging to the general vocabulary, express a particular concept, or are synonymous of MWE already present in the Delac. We can consider them as forgotten by our lexicon. These are: “élément clé” (note that on the opposite “position clé”, “facteur clé” and “question clé” belong to the Delac but were classified as false in the corpus by our model), “zone non-fumeur”, “temps record”, “rôle clé”, “père fondateur”, “salaire minimum”, “congé maternité”.
- 5 are foreign MWE and so can’t be part of the Delac: “partido popular”, “for money
• 5 were MWE that would have been impossible to find in the Delac for the simple reason that they were concept that became very popular in France during the years 2000 and the Delac has not been updated since 1997. The French corpora starts after the year 1999 so we can assume that most of the Europarl corpus in French has been written after Delac’s last release. These 5 MWE are: “tolérance zéro”, “cyber crime”, “site internet”, “site web”, “immigration zéro”.

• 10 are definitely not Noun-Noun multiword expressions in French. e.g. “bien jusqu”, “pays tiers”, “prochaine année”.

For the 19 that last we decided to not statue on them because, as we have only one annotator native speaker but not specialist on all the domains that concern Europarl, their case would merit more annotators for judging whether they correspond to a MWE or not. Indeed, 15 of them seem very specialised terminology “valeur seuil”, “service logistique”, “processus euroméditerranéen” that belong to more specific domains such as economy, politics or legislative. That could explain why they have not their seat in the Delac that is a generalistic lexicon.

5.2.2 The False Negatives

About the thirty seven MWE that were classified illegitimately as false, only four of them appeared two times, all the others appeared only one time in the corpus. This is something that can explain a big part of the recall lost if not the precision: their is proportionally more candidates in Noun-Noun that will figure only one time in the corpus (about 80% of them) whereas on the same amount of text only 72% of the Noun-Adjective appeared only one time. The informations on an instance that appeared only one time are to few for evaluating its statistical properties.

6  Discussion

As we saw with our evaluation part, it is tricky to learn statistical properties of MWE when actually we have not all the informations necessary for extracting all the MWE in the corpus. Indeed, for this purpose the corpus need to be read and annotated by humans. However we still managed to train a model even if it is likely that a lot of candidates pre-annotated as false were probably perfect MWE. This means that the Delac has covered enough MWE for the features to not appear as completely meaningless and arbitrary.

7  Conclusion

We wanted to check the performance of machine learning for validating MWE candidates on two different types of patterns. For answering this question we trained a Bayesian Network algorithm using association measures as features on candidates Noun-Adjective extracted in a sample of the Europarl corpus. We used the Delac as our gold standard for validating the candidates. Therefore we have checked the performance of our algorithm on Noun-Noun extracted as well from the Europarl corpus. To the question can a system trained on one type of MWE (Noun-Adjective) classify correctly MWE belonging to another pattern (Noun-Noun)? we will answer yes even if the performance tend to be lower. We found two differences that could explain the lower performance: first a slightly different treatment in the number of Noun-Noun collected in our lexicon, they were proportionally less listed than the Noun-Adjective, secondly, and this is the most important, the property of the pattern Noun-Noun in itself to be more rare in French. Furthermore we can notice after annotation that at least 20% of the MWE considered as false positives can definitely belong to MWE (for our lexicon they were either anachronisms either forbidden). This would mean that the machine learning algorithm is really pertinent when it comes to fulfil lexicons.
8 Future works

After those results we noticed that the biggest problem in our method came from the enormous difference in the number of apparitions between our patterns. Maybe it would be necessary to evaluate on a corpus twenty five time bigger for the Noun-Noun candidate extraction than for the Noun-Adjective extraction. However if we keep the same size, what could be the performance of a system trained on the Noun-Noun patterns? Would it be worse since the Noun-Noun pattern offers less candidates to train on? Would it be more pertinent to run the same test on patterns that have about the same proportion in French language? What would be the results on patterns that doesn’t share any common tags such as Verb-Preposition and Adjective-Noun?

References


Leonardo Zilio, Luiz Svoboda, Luiz Henrique Loughi Rossi, and Rafael Martins Feitosa. Automatic extraction and evalu-