Extraction of Nominal Multiword Expressions in French

Marie Dubremetz and **Joakim Nivre**
Uppsala university
Department of Linguistics and Philology
Uppsala, Sweden

Abstract

Multiword expressions (MWEs) can be extracted automatically from large corpora using association measures, and tools like mwetoolkit allow researchers to generate training data for MWE extraction given a tagged corpus and a lexicon. We use mwetoolkit on a sample of the French Europarl corpus together with the French lexicon Dela, and use Weka to train classifiers for MWE extraction on the generated training data. A manual evaluation shows that the classifiers achieve 60–75% precision and that about half of the MWEs found are novel and not listed in the lexicon. We also investigate the impact of the patterns used to generate the training data and find that this can affect the trade-off between precision and novelty.

1 Introduction

In alphabetic languages, words are delimited by spaces. Some words can combine to create a new unit of meaning that we call a multiword expression (MWE). However, MWEs such as *kick the bucket* must be distinguished from free combinations of words such as *kick the ball*. A sequence of several words is an MWE if “at least one of its syntactic, distributional or semantic properties cannot be deduced from the properties of its component” (Silberztein and L.A.D.L., 1990). So how can we extract them?

Statistical association measures have long been used for MWE extraction (Pecina, 2010), and by training supervised classifiers that use association measures as features we can further improve the quality of the extraction process. However, supervised machine learning requires annotated data, which creates a bottleneck in the absence of large corpora annotated for MWEs. In order to circumvent this bottleneck, mwetoolkit (Ramisch et al., 2010b) generates training instances by first extracting candidates that fit a certain part-of-speech pattern, such as Noun-Noun or Noun-Adjective, and then marking the candidates as positive or negative instances depending on whether they can be found in a given lexicon or not. Such a training set will presumably not contain any false positives (that is, candidates marked as positive instances that are not real MWEs), but depending on the coverage of the lexicon there will be a smaller or larger proportion of false negatives. The question is what quality can be obtained using such a noisy training set. To the best of our knowledge, we cannot find the answer for French in literature. Indeed, Ramisch et al. (2012) compares the performance of mwetoolkit with another toolkit on English and French corpora, but they never use the data generated by mwetoolkit to train a model. In contrast, Zilio et al. (2011) make a study involving training a model but use it only on English and use extra lexical resources to complement the machine learning method, so their study does not focus just on classifier evaluation.

This paper presents the first evaluation of mwetoolkit on French together with two resources very commonly used by the French NLP community: the tagger TreeTagger (Schmid, 1994) and the dictionary Dela.[1] Training and test data are taken from the French Europarl corpus (Koehn, 2005) and classifiers are trained using the Weka machine learning toolkit (Hall et al., 2009). The primary goal is to evaluate what level of precision can be achieved for nominal MWEs, using a manual evaluation of MWEs extracted, and to what extent the MWEs extracted are novel and can be used to enrich the lexicon. In addition, we will investigate what effect the choice of part-of-speech patterns used to generate the training data has on precision and novelty. Our results indicate that classifiers

[1] http://www-igm.univ-mlv.fr/~unitex/index.php?page=5&html=bibliography.html

Proceedings of the 10th Workshop on Multiword Expressions (MWE 2014), pages 72–76,
Gothenburg, Sweden, 26-27 April 2014. ©2014 Association for Computational Linguistics

achieve precision in the 60–75% range and that about half of the MWEs found are novel ones. In addition, it seems that the choice of patterns used to generate the training data can affect the tradeoff between precision and novelty.

2 Related Work

2.1 Extraction Techniques

There is no unique definition of MWEs (Ramisch, 2012). In the literature on the subject, we notice that manual MWE extraction often requires several annotators native of the studied language. Nevertheless, some techniques exist for selecting automatically candidates that are more likely to be the true ones. Candidates can be validated against an external resource, such as a lexicon. It is possible also to check the frequency of candidates in another corpus like the web. Villaviciencio (2005), for example, uses number of hits on Google for validating the likelihood of particle verbs.

However, as Ramisch (2012) states in his introduction, MWE is an institutionalised phenomenon. This means that an MWE is frequently used and is part of the vocabulary of a speaker as well as the simple words. It means also that MWEs have specific statistical properties that have been studied. The results of those studies are statistical measures such as dice score, maximum likelihood estimate, pointwise mutual information, T-score. As Islam et al. (2012) remark in a study of Google Ngram, those measures of association are language independent. And it is demonstrated 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 extraction and validation of MWE is not considered the most effective method. The tendency these last years is to combine association measures with linguistic features (Ramisch et al., 2010a; Pecina, 2008; Tsvetkov and Wintner, 2011).

2.2 Mwetoolkit

Among the tools developed for extracting MWEs, mwetoolkit is one of the most recent. Developed by Ramisch et al. (2010b) it aims not only at extracting candidates for potential MWEs, but also at extracting their association measures. Provided that a lexicon of MWEs is available and provided a preprocessed corpus, mwetoolkit makes it possible to train a machine learning system with the association measures as features with a minimum of implementation.

Ramisch et al. (2010b) provide experiments on Portuguese, English and Greek. Zilio et al. (2011) provide experiments with this tool as well. In the latter study, after having trained a machine on bigram MWEs, they try to extract full n-gram expressions from the Europarl corpus. They then reuse the model obtained on bigrams for extraction of full n-gram MWEs. Finally, they apply a second filter for getting back the false negatives by checking every MWE annotated as False by the algorithm against a online dictionary. This method gets a very good precision (over 87%) and recall (over 84%). However, we do not really know if this result is mostly due to the coverage of the dictionary online. What is the contribution of machine learning in itself? Another question raised by this study is the ability of a machine trained on one kind of pattern (e.g., Noun-Adjective) to extract correctly another kind of MWE pattern (e.g., Noun-Noun). That is the reason why we will run three experiments close to the one of Zilio et al. (2011) but were the only changing parameter is the pattern that we train our classifiers on.

3 Generating Training Data

3.1 Choice of Patterns

In contrast to Zilio et al. (2011) we run our experiment on French. The choice of a different language requires an adaptation of the patterns. French indeed, as a latin language, does not show the same characteristic patterns as English. We know that there is a strong recurrence of the pattern Noun-Adjective in bigram MWEs in our lexicon (Silberztein and L.A.D.L., 1990, p.82), and the next most frequent pattern is Noun-Noun. Therefore we extract only candidates that correspond to these patterns. And, since we have two patterns, we will run two extra experiments where our models will be trained only on one of the patterns. In this way, we will discover how sensitive the method is to the choice of pattern.

3.2 Corpus

As Ramisch et al. (2012) we work on the French Europarl corpus. We took the three first million words of Europarl and divided it into three equal parts (one million words each) for running our experiments. The first part will be devoted at 80% to training and 20% to development test set, when training classifiers on Noun-Adjective or Noun-Noun patterns, or both. We use the second million as a secondary development set that is not used in this study. The third million is used as a final test set and we will present results on this set.

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) to obtain the tags and the lemmas.

3.4 Extracting Data and Features

The mwetoolkit takes as input a preprocessed corpus plus a lexicon and gives two main outputs: an arff file which is a format adapted to the machine learning framework Weka, and an XML file. At the end of the process we obtain, for each candidate, a binary classification as an MWE (True) or not (False) depending on whether it is contained in the lexicon. For each candidate, we also obtain the following features: maximum likelihood estimate, pointwise mutual information, T-score, dice coefficient, log-likelihood ratio. The machine learning task is then to predict the class (True or False) given the features of a candidate.

3.5 Choice of a Lexicon in French

The evaluation part of mwetoolkit is furnished with an internal English lexicon as a gold standard for evaluating bigram MWEs, but for French it is necessary to provide an external resource. We used as our gold standard the French dictionary Dela (Silberztein and L.A.D.L., 1990), the MWE part of which is called Delac. It is a general purpose dictionary for NLP and it includes 100,000 MWE expressions, which is a reasonable size for leading an experiment on the Europarl corpus. Also the technical documentation of the Delac (Silberztein and L.A.D.L., 1990, p.72) says that this dictionary has been constructed by linguists with reference to several dictionaries. So it is a manually built resource that contains MWEs only referenced in official lexicographical books.

3.6 Processing

Thanks to mwetoolkit we extracted all the bigrams that correspond to the patterns Noun-Adjective (NA), Noun-Noun (NN) and to both Noun-Adjective and Noun-Noun (NANN) in our three data sets and let mwetoolkit make an automatic annotation by checking the presence of the MWE candidates in the Delac. Note that the automatic annotation was used only for training. The final evaluation was done manually.

4 Training Classifiers

For finding the best model we think that we have to favour the recall of the positive candidates. Indeed, when an MWE candidate is annotated as True, it means that it is listed in the Dela, which means that it is an officially listed MWE. However, if an MWE is not in the Dela, it does not mean that the candidate does not fulfil all the criteria for being an 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.

4.1 Training on NA

We tested several algorithms offered by Weka as well as the training options suggested by Zilio et al. (2011). We also tried to remove some features and to keep only the most informative ones (MLE, T-score and log-likelihood according to information gain ratio) but we noticed each time a loss in the recall. At the end with all the features kept and for the purpose of evaluating NA MWE candidates the best classification algorithm was the Bayesian network.

4.2 Training on NN

When training a model on NN MWEs, our aim was to keep as much as possible the same condition for our three experiments. However, the NN training set has definitely not the same properties as the NA and NANN ones. The NN training set is twenty-four times smaller than NA training set. Most of the algorithms offered by Weka therefore ended up with a dummy systematic classification to the majority class False. The only exceptions were ibk, ib1, hyperpipes, random trees and random forest. We kept random forest because it gave the best recall with a very good precision. We tried several options and obtained the optimum results with 8 trees each constructed while considering 3 random features, one seed, and unlimited depth of trees. As well as for NA we kept all features.

4.3 Training on NA+NN

For the training on NANN candidates we tried the same models as for NN and for NA candidates.

The best result was obtained with the same algorithm as for NA: Bayesian network.

5 Evaluation

The data automatically annotated by mwetoolkit could be used for training, but to properly evaluate the precision of MWE extraction on new data and not penalize the system for 'false positives' that are due to lack of coverage of the lexicon, we needed to perform a manual annotation. To do so, we randomly picked 100 candidates annotated as True by each model (regardless if they were in the Delac or not). We then annotated all such candidates as True if they were found in Delac (without further inspection) and otherwise classified them manually following the definition of Silberztein and L.A.D.L. (1990) and the intuition of a native French speaker. The results are in Table 1.

Extracting NANN	NA model	NN model	NANN model
In Delac	40 ±9.4	18 ±7.2	28 ±8.6
Not in Delac	34 ±9.0	41 ±9.2	38 ±9.3
Precision	74 ±8.4	59 ±9.2	66 ±9.0

Table 1: Performance of three different models on the same corpus of Noun-Adjective and Noun-Noun candidates. Percentages with 95% confidence intervals, sample size = 100.

As we see in Table 1, the experiment reveals a precision ranging from almost 60% up to 74%. The results of our comparative manual annotation indicate that the model trained on NN candidates has the capacity to find more MWEs not listed in our lexicon (41 out of 59) even if it is the least precise model. On the other hand, we notice that the model based on Noun-Adjective patterns is more precise but at the same time extracts fewer MWEs that are not already in the lexicon (34 out of 74). Our mixed model confirms these two tendencies with a performance in between (38 new MWEs out of 66). Thus, the method appears to be sensitive to the patterns used for training.

We notice during evaluation different kinds of MWEs that are successfully extracted by models but that are not listed in the Delac. Most of them are the MWEs specific to Europarl (e.g., 'dimension communautaire', 'législation européenne'[2]). Another category are those MWEs that became popular in the French language after the years 2000's and therefore could not be included in the Delac, released in 1997. Indeed by reading the first paragraph of the French version of Europarl we notice that the texts have been written after 1999. Of course, they are not the majority of the successfully extracted MWEs but we still manage to find up to 3 of them in a sample of 100 that we checked ('développement durable', 'radiophonie numérique', 'site internet'[3]). Furthermore the corpus in itself is already more than ten years old, so in a text of 2014 we can expect to find even more of them. Finally, there are MWEs that are not in French (e.g., 'Partido popular'), these, however, did not appear systematically in our samples.

It is tricky to learn statistical properties of MWEs when, actually, we do not have all the information necessary for extracting the MWEs in the corpus. Indeed, for this purpose the corpus should ideally be read and annotated by humans. However, we still managed to train models with decent performance, even if it is likely that a lot of candidates pre-annotated as False in the training data were probably perfect MWEs. This means that the Delac has covered enough MWEs for the features to not appear as completely meaningless and arbitrary. The final precision would never be as good as it is, if the coverage had been not sufficient enough. This shows that the method of automatic annotation offered by mwetoolkit is reliable given a lexicon as large as Delac.

6 Conclusion

We wanted to know if the method of automatic extraction and evaluation offered by mwetoolkit could have a decent precision in French. We annotated automatically part of the Europarl corpus given the lexical resource Dela as a gold standard and generated in this way annotated training sets. Classifiers trained on this data using Weka achieved a maximum precision of 74%, with about half of the extracted MWEs being novel compared to the lexicon. In addition, we found that the final precision and novelty scores were sensitive to the choice of patterns used to generate the training data.

[2] 'community scale', 'European legislation'

[3] 'sustainable development', 'digital radio','website'

References

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. 2009. The WEKA data mining software: an update. *SIGKDD Exploration Newsletter*, 11(1):10–18.

Aminul Islam, Evangelos E Milios, and Vlado Keselj. 2012. Comparing Word Relatedness Measures Based on Google n-grams. In *COLING, International Conference on Computational Linguistics (Posters)*, pages 495–506, Mumbai, India.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *The Tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand.

Pavel Pecina. 2008. A Machine Learning Approach to Multiword Expression Extraction. In *Proceedings of the LREC 2008 Workshop Towards a Shared Task for Multiword Expressions*, pages 54–57, Marrakech, Morocco.

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

Carlos Ramisch, Helena de Medeiros Caseli, Aline Villavicencio, André Machado, and Maria José Finatto. 2010a. A Hybrid Approach for Multiword Expression Identification. In *Proceedings of the 9th International Conference on Computational Processing of Portuguese Language (PROPOR)*, pages 65–74.

Carlos Ramisch, Aline Villavicencio, and Christian Boitet. 2010b. Multiword Expressions in the wild? The mwetoolkit comes in handy. In *COLING, International Conference on Computational Linguistics (Demos)*, pages 57–60.

Carlos Ramisch, Vitor De Araujo, and Aline Villavicencio. 2012. A Broad Evaluation of Techniques for Automatic Acquisition of Multiword Expressions. In *Proceedings of ACL 2012 Student Research Workshop*, pages 1–6, Jeju Island, Korea. Association for Computational Linguistics.

Carlos Ramisch. 2012. Une plate-forme générique et ouverte pour l'acquisition des expressions polylexicales. In *Actes de la 14e Rencontres des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues*, pages 137–149, Grenoble, France.

Helmut Schmid. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of International Conference on New Methods in Language Processing*, pages 44–49, Manchester, Great Britain.

Max Silberztein and L.A.D.L. 1990. Le dictionnaire électronique des mots composés. *Langue française*, 87(1):71–83.

Yulia Tsvetkov and Shuly Wintner. 2011. Identification of Multi-word Expressions by Combining Multiple Linguistic Information Sources. In *Empirical Methods in Natural Language Processing*, pages 836–845.

Aline Villavicencio. 2005. The availability of verb–particle constructions in lexical resources: How much is enough? *Computer Speech & Language*, 19(4):415–432.

Leonardo Zilio, Luiz Svoboda, Luiz Henrique Longhi Rossi, and Rafael Martins Feitosa. 2011. Automatic extraction and evaluation of MWE. In *8th Brazilian Symposium in Information and Human Language Technology*, pages 214–218, Cuiabá, Brazil.