Never-Ending Multiword Expressions Learning

Alexandre C. Rondon
Federal Univ. of São Carlos
São Carlos, Brazil
alex.crondon@gmail.com

Helena de Medeiros Caseli
Federal Univ. of São Carlos
São Carlos, Brazil
helenacaseli@dc.ufscar.br

Carlos Ramisch
Aix Marseille Université
Marseille, France
carlos.ramisch@lif.univ-mrs.fr

Abstract

This paper introduces NEMWEL, a system that performs Never-Ending MultiWord Expressions Learning. Instead of using a static corpus and classifier, NEMWEL applies supervised learning on automatically crawled news texts. Moreover, it uses its own results to periodically retrain the classifier, bootstrapping on its own results. In addition to a detailed description of the system's architecture and its modules, we report the results of a manual evaluation. It shows that NEMWEL is capable of learning new expressions over time with improved precision.

1 Introduction

Multiword expressions (MWEs) are combinations of two or more lexemes which present some lexical, syntactic, semantic, pragmatic or statistical idiosyncrasies with respect to regular combinations (Baldwin and Kim, 2010). Examples include idioms (*saw logs* as *to snore*), phrasal verbs (*pull over*, *give up*), noun compounds (*machine learning*, *support vector machine*) and complex function words (*as well as*, *with respect to*).

In human languages, such constructions are very frequent, as native speakers rarely realize how often they employ them (Sag et al., 2002; Jackendoff, 1997b). However, they are not frequent in NLP resources such as lexicons and grammars, and this represents a bottleneck for building robust and accurate NLP applications.

Since the construction of such resources is onerous and demands highly qualified linguistic expertise, *automatic MWE lexicon extraction* is an attractive alternative which has been one of the most active topics in the MWE research community. Proposed methods are often based on supervised and unsupervised learning of MWE lists from textual corpora (Evert and Krenn, 2005; Pecina, 2008). In spite of the availability of very large corpora like the Gigaword or WaC (Baroni et al., 2009), these methods are still limited by the coverage of the texts in the source corpus.

This paper presents NEMWEL, a machine learning system able to learn MWEs following the never-ending approach (Mitchell et al., 2015). NEMWEL automatically extracts MWE candidates from a corpus periodically crawled from a Brazilian online news portal. Then, based on supervised training, NEMWEL classifies the candidates and promotes some of them to the status of "true MWEs", which are used to retrain the classifier. This process is repeated endlessly, taking into consideration the true MWEs learned in previous steps. By doing so, NEMWEL tries to resemble the way human beings learn.

We have developed a prototype that implements this idea. To the best of our knowledge, this is the first attempt to build MWE lexicons using a never-ending learning approach. We have manually evaluated the extracted MWEs and we show that the precision of the learner seems to increase with time.

Proceedings of NAACL-HLT 2015, pages 45–53,
Denver, Colorado, May 31 – June 5, 2015.

The remainder of this paper is structured as follows: we discuss related work on MWE extraction (Section 2) and never-ending learning methods (Section 3). Then, we present the architecture and detail the modules in NEMWEL (Section 4). Finaly, we present the results of automatic and manual evaluation in Brazilian Portuguese (Section 5) and ideas for future work (Section 6).

2 MWE Extraction

Automatic unsupervised MWE learning from corpora has been proposed based on pairwise association measures (Church and Hanks, 1990; Smadja, 1993; Pedersen et al., 2011), string matching (Duan et al., 2006), extraction patterns based on expert linguistic knowledge and automatic analysis (Justeson and Katz, 1995; Seretan and Wehrli, 2009) or a combination of these methods (Araujo et al., 2011).

Supervised machine learning methods have also been used for MWE lexicon learning.[1] Pecina (2008) proposes a logistic regression classifier which uses as features a set of 84 different lexical association measures. Ramisch et al. (2008) use decision trees for classifying MWEs based on standard association measures as well, but they add variation entropy. In terms of classifiers, many alternatives have been tested like bayesian networks (Dubremetz and Nivre, 2014) and support vector machines (Farahmand and Martins, 2014). Zilio et al. (2011) use a stable set of features, but compare several classification algorithms implemented in Weka. Furthermore, in-context MWE tagging has been performed using sequence learning models like conditional random fields (Constant and Sigogne, 2011) and structured perceptron (Schneider et al., 2014).[2]

Many alternative sources and methods have been tested for MWE extraction, like parallel texts (Caseli et al., 2010; Tsvetkov and Wintner, 2010), bilingual lexicons (Salehi and Cook, 2013), Wikipedia interlingual links (Attia et al., 2010), WordNet synonyms (Pearce, 2001) and distributional neighbors (Reddy et al., 2011). The web has also been considered as a source for MWE learning, often using page hit counts from search engines (Lapata and Keller, 2005; Kim and Nakov, 2011). However, in related work, candidates are not extracted from web texts, but from traditional corpora.

Differently from previous corpus-based or web-based learning approaches, our goal is not to build one static MWE lexicon. Instead, we propose to build a system that continuously learns new expressions from the web. It populates and enriches the lexicon with new MWEs every day. Our proposal is to employ bootstrapping on a traditional supervised machine learning setting, enriched with new features and dynamically crawled corpora. At any given time, a snapshot of the database will include the current MWE lexicon, which can be exported, evaluated and used to retrain the classifier. To the best of our knowledge, this is the first time never-ending learning is applied to MWE lexicon discovery.

3 Never-Ending Learning

In traditional machine learning, an algorithm is usually applied to learn a model from a fixed amount of labeled training data. Although effective in many applications, this way of learning is very limited and also far from the way that human beings learn. Never-ending learning is an approach that tries to resemble the way humans learn, taking into account different sources of information and using previous experience to guide subsequent learning (Mitchell et al., 2015). It can be classified as a bootstrapping algorithm. It requires a small set of annotated items, used to initialize the model, and then it uses its own results to retrain the classifier in future iterations.

The main system developed following the never-ending learning approach is the Never-Ending Language Learner (NELL) of Carlson et al. (2010). NELL is the learning system of the Read the Web project[3] and it is running 24 hours/day since 2010. NELL's goals are (1)

[1] Usually, such methods require a list of candidate expressions annotated as true or false MWEs.

[2] Such models require corpora where sentences are annotated with the MWE sequences they contain.

[3] `http://rtw.ml.cmu.edu/rtw/`

to read the web extracting beliefs (true facts) that populate a knowledge base and (2) to learn better day by day. To do so, NELL is able to perform different learning tasks (category classification, relation classification, etc.) and combine different learning functions to make decisions and improve its learning methods (Mitchell et al., 2015).

In this paper we describe the Never-Ending MultiWord Expressions Learner (NEMWEL). Different from NELL, NEMWEL is in its first year of life and is intended only to learn MWEs. But, following the main never-ending learning premise, NEMWEL uses its previously learned knowledge to better learn new MWEs.

According to Jackendoff (1997a), there are as many MWEs in a lexicon as single words. For Sag et al. (2002) this is an underestimation and the real number of MWEs grows with language evolution. These findings corroborate our idea that a never-ending learning system is a good solution to tackle the MWE extraction problem.

4 The Never-Ending MWE Learner

The NEMWEL was developed in Java and is divided into four modules – crawler, extractor, processor and promoter – explained in the next subsections. These four modules are applied in sequence and repeatedly in each iteration of NEMWEL.

4.1 Crawler

The first module, the Crawler, is responsible for collecting texts from the web to build a corpus. In our current prototype, in each iteration, 40 different articles from the G1 news portal[4] are downloaded randomly, cleaned by removing HMTL markup and boilerplate content, and concatenated in one unique file. Figure 1 shows an excerpt of a text from one iteration of the Crawler module.

4.2 Extractor

After collecting and cleaning the texts, the Extractor annotates the tokens in each text with its surface form, part-of-speech tag and lemma. To do so, we used the TreeTagger (Schmid, 1994) with a model trained for Portuguese[5]. Tagging the corpus is required because we evaluate our learner using nominal MWEs, thus we need to be able to identify nouns and their complements. The TreeTagger was chosen because it is free, easy to use and fast, enabling us to quickly process large amounts of crawled texts. The same excerpt of Figure 1 processed by the Extractor is shown in Figure 2.

Mais de 100 famílias de baixa renda ocuparam casas de um **conjunto habitacional**, em Paulínia (SP), na madrugada desta quarta-feira (19).
*More than 100 low-income families occupied houses of a **housing development** in Paulinia (SP) in the early hours of this Wednesday (19).*

Figure 1: Excerpt of a text crawled from the news portal. Original text (in Brazilian Portuguese) and its English translation (manually prepared for this paper).

The sequences of tagged tokens in the crawled texts are processed by the *mwetoolkit* (Ramisch, 2015), which is the core of our Extractor and Processor modules. In the Extractor, a list of MWE candidates is obtained by matching a multilevel regular-expression pattern (Figure 3) against the tagged corpus. Figure 4 shows an example of MWE candidate extracted from our example sentence, using the pattern of Figure 3. The pattern is based on intuitive noun phrase descriptions, but it also captures more candidates, that are not necessarily nominal compounds. Further filters must be applied to remove regular noun phrases and keep only nominal MWEs.

4.3 Processor

In this module, the mwetoolkit calculates some association measures that will be used by the Promoter in the next step. These measures are calculated based on the number of occurrences of the MWE candidate and of the words that

[4] `http://g1.globo.com`

[5] `http://gramatica.usc.es/~gamallo/tagger\_intro.htm`

```
Mais            ADV     mais
de              PRP     de
100             CARD    @card@
...
casas           NOM     casa
de              PRP     de
um              DET     um
conjunto        NOM     conjunto
habitacional    ADJ     habitacional
,               VIRG    ,
em              PRP     em
Paulínia        NOM     paulínia
(               QUOTE   (
SP              NOM     SP
)               QUOTE   )
,               VIRG    ,
na              PRP     em
madrugada       NOM     madrugada
desta           PRP     de
quarta-feira    NOM     quarta-feira
(               QUOTE   (
19              CARD    @card@
)               QUOTE   )
.               SENT    .
```

Figure 2: The excerpt from Figure 1 after part-of-speech tagging by TreeTagger.

```
<patterns>
  <pat>
    <w pos="NOM"/>
    <pat repeat="{1,3}"/>
      <either>
        <pat>
          <w pos="PRP*" lemma="de"/>
          <w pos="NOM"/>
        </pat>
        <pat>
          <w pos="ADJ"/>
        </pat>
      </either>
    </pat>
  </pat>
</patterns>
```

Figure 3: List of part-of-speech sequences describing nominal multiword expressions in Brazilian Portuguese. They correspond to a noun followed by 1 to 3 complements, which can be either an adjective or a prepositional phrase introduced by *de*.

```
<cand candid="684">
  <ngram>
    <w lemma="conjunto">
      <freq name="g1" value="10"/>
      <freq name="plnbr" value="3005"/>
    </w>
    <w lemma="habitacional">
      <freq name="g1" value="3"/>
      <freq name="plnbr" value="359"/>
    </w>
    <freq name="g1" value="3"/>
    <freq name"plnbr" value="86"/>
  </ngram>
  ...
</cand>
```

Figure 4: MWE candidate extracted from the sentence of Figure 1 using the pattern of Figure 3.

compose it. In our experiments, these numbers of occurrences were calculated using the G1 corpus and also the PLN-BR corpus[6], which contains around 29 million words of news articles from the *Folha de São Paulo* newspaper, from 1994 to 2004. The use of the larger, static corpus may help because it provides more accurate association measures as features. For instance, in Figure 4, we can see that G1 returns 3 occurrences for *conjunto habitacional*, and 10 and 3 occurrences for the individual words. It is known that association measures are sensitive to low-frequency data, so it is probably a good idea to complement this with a measure calculated on PLN-BR, where the frequencies are of 86 occurrences for the expression, 3006 occurrences for the first words and 359 occurrences for the second word.

4.3.1 Features

The next module, the Promoter, uses supervised training performed using the 17 features defined below.

- **Association measures** – measure of the strength of the association between the frequency of an n-gram and the frequency of each word that forms the n-gram. In our experiments, four measures were used: normalized frequency, Student's t score, point-

[6] http://www.nilc.icmc.usp.br/plnbr

wise mutual information and Dice's coefficient. All of these measures were calculated by the mwetoolkit in the two corpora: G1 and PLN-BR. Thus, in total, we have eight features based on association measures.

- **Translatability** – measure based on the non-translatability property of true MWEs. First, we estimate the probability of a content word w[7] to be translated into a word x in English (`en`) and then back to Portuguese (`pt`), using a bilingual weighted lexicon:

$$T(w) = \sum_x P_{pt \to en}(w, x) \times P_{en \to pt}(x, w)$$

Two new features were proposed based on this probability:

$$\texttt{translatability\_mult} = \prod_{i=1}^{n} T(w_i)$$

$$\texttt{translatability\_mean} = \frac{1}{n} \sum_{i=1}^{n} T(w_i)$$

Figure 5 shows an example of these features for the candidate expression *taxa de juros* (*interest rate*).

- **POS context** – the part of speech of the three previous and the three next tokens around the MWE candidate. We also use the concatenated parts of speech of the words that form the MWE candidate. When there are more than one possible contexts, the most frequent one is chosen. Thus, seven features are based on the POS context, three in each direction and the POS sequence of the target candidate.

The new features proposed in this paper, based on translatability, are based on linguistic tests that show that MWEs have limited variability and thus, in most cases, cannot be translated word by word. It is calculated using two probabilistic bilingual dictionaries generated by NATools[8] from the FAPESP parallel corpus corpus[9]. This corpus contains

[7] In our experiments, content words are nouns and adjectives.

[8] `http://corpora.di.uminho.pt/natools`

[9] `http://www.nilc.icmc.usp.br/nilc/tools/FapespCorpora.htm`

$$\begin{aligned}
T(\text{taxa}) &= P_{pt \to en}(\text{taxa}, \text{rate}) \times \\
&\quad P_{en \to pt}(\text{rate}, \text{taxa}) + \\
&\quad P_{pt \to en}(\text{taxa}, \text{level}) \times \\
&\quad P_{en \to pt}(\text{level}, \text{taxa}) + \\
&\quad P_{pt \to en}(\text{taxa}, \text{interest}) \times \\
&\quad P_{en \to pt}(\text{interest}, \text{taxa}) \\
&= 0.583 \times 0.537 + 0.251 \times 0.096 + \\
&\quad 0.008 \times 0 \\
&= 0.3372 \\
T(\text{juros}) &= P_{pt \to en}(\text{juros}, \text{interest}) \times \\
&\quad P_{en \to pt}(\text{interest}, \text{juros}) + \\
&\quad P_{pt \to en}(\text{juros}, \text{rates}) \times \\
&\quad P_{en \to pt}(\text{rates}, \text{juros}) + \\
&= 0.628 \times 0.032 + 0.372 \times 0.114 \\
&= 0.0625 \\
&\texttt{translatability_mult} \\
&= T(\text{taxa}) \times T(\text{juros}) \\
&= 0.0211 \\
&\texttt{translatability_mean} \\
&= \tfrac{1}{2} T(\text{taxa}) + T(\text{juros}) \\
&= 0.1998
\end{aligned}$$

Figure 5: Example of the two features based on translatability of the MWE candidate *taxa de juros* (*interest rate*).

a set of sentence-aligned Portuguese-English and English-Portuguese articles about research projects. From this corpus, NATools outputs, for each source word, a list of up to 10 best translations accompanied by its probability.

To the best of our knowledge, this is the first time that translatability is implemented for MWE automatic extraction using automatically built bilingual lexicons. Related methods are based on non weighted, standard bilingual lexicons like PanLex or Wikipedia titles (Salehi and Cook, 2013; Attia et al., 2010).

4.4 Promoter

The last module, the Promoter, analyses the MWE candidates and promotes to *beliefs* the ones with the best scores. Beliefs are candidates that were classified as true MWEs in a previous iteration of the learner.

The Promoter applies a classification model trained using Weka (Hall et al., 2009) as a wrapper and LibSVM (Chang and Lin, 2011) as the

core. The result is a support vector machine that distinguishes true MWEs from ordinary noun phrases. As training data, it uses previously annotated instances. The Promoter is generated based on examples that were already classified, either manually, for the Promoter-0, or manually+automatically, for the Promoters built in subsequent iterations.

SVM was the chosen classifier because it has presented good performance on diverse NLP tasks such as text categorization (Sassano, 2003), sentiment analysis (Mullen and Collier, 2004) and named entity recognition (Li et al., 2008), as well as standard corpus-based MWE extraction (Farahmand and Martins, 2014).

5 Evaluation

An initial training corpus was generated from texts of the G1 news portal. From this corpus, NEMWEL extracted 1,100 candidate MWEs which were manually annotated by two native speakers of Brazilian Portuguese: 600 candidates for each one with an intersection of 100 candidates. The annotation interface showed the candidate and the sentences from the G1 corpus from which the candidate was extracted (see Figure 6). The annotators had to perform a binary choice as to whether the candidate was a true MWE ("Sim") or not ("Não"). Each annotator cross-checked the other one's items. This last cross-checking step was crucial because, even though some guidelines were provided, some cases were hard to decide and required discussion. From this first annotation, 19% of the candidates were evaluated as true MWEs. The kappa agreement (Cohen, 1960) was 0.85, which indicates a very good agreement.

The annotated set was used to train our Promoter-0 as explained in section 4.4. NEMWEL, then, run for 15 iterations and, at each 5 iterations (a generation), a new Promoter was trained using the beliefs and false MWEs classified in the previous iterations.[10] After these 15 iterations, a new sample of 1,200 MWE candidates was manually evaluated by the two native speakers, but with no overlap between the annotators. To allow the analysis of the learning curve over time, this sample contained 400 candidates extracted in each generation, from which each annotator judged half, that is, 600 candidates per annotator, 200 for each generation.

	Iterations			
	1-5	6-10	11-15	All
Precision	24.6%	32.2%	34.3%	30.5%
Recall	55.6%	65.5%	52.3%	57.0%
F1	34.1%	43.2%	41.4%	39.7%
Accuracy	85.5%	87.5%	83.8%	85.6%

Table 1: Results of NEMWEL's evaluation after 15 iterations and 3 generations of new Promoters.

From the 1,200 candidates, 15.6% were classified as true MWE. The results are shown in Table 1 in terms of precision, recall, F-measure and accuracy calculated regarding true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN):

- Precision = $\frac{TP}{TP+FP}$
- Recall = $\frac{TP}{TP+FN}$
- F1 = $2 \times \frac{P \times R}{P+R}$
- Accuracy = $\frac{TP+TN}{TP+FP+TN+FN}$

As we can notice from Table 1, the precision rises 10 percentage points from the first to the last iteration, indicating that NEMWEL is capable of improving its learning performance, as expected for a never-ending learning system. The decay in recall from 65.5% to 52.3% from the second to the third generation seems to be related to overfitting. Another possible explanation for this decay is that only the candidate MWEs annotated as true by both annotators were taking into account. Furthermore, since the dataset is unbalanced, the classifier

[10] Thus, in our experiments, three Promoters were generated: (1) Promoter-0, trained only with manually annotated data, run from iteration 1 to 5 (first generation); (2) Promoter-1, trained with manually annotated data and the true/false MWEs learned in the first generation, run from iteration 6 to 10; and (3) Promoter-2, trained with manually annotated data and the true/false MWEs learned in the first two generations, run from iteration 11 to 15.

Figure 6: Interface for manual annotation of MWE candidates.

may tend to classify new candidates always as non MWEs. New experiments will be carried out to investigate this decay. Table 2 shows some examples of MWE candidates extracted by NEMWEL.

6 Conclusions

From the results presented in this paper, it is possible to conclude that the never-ending learning approach can be applied to the automatic extraction of MWEs. Although with just a few iterations (15), it was already possible to see that NEMWEL is able to improve its learning based on previously learned knowledge, with an increase of 10 percentage points in precision.

The next steps of this work include running NEMWEL for a long period, ideally 24 hours per day, continuously. It is also our intention to expand NEMWEL to be able to learn other MWEs, from other sources and for different languages, such as English, maybe following a multilingual extraction process. Finally, some new features can be added such as the one that tests the substitutability of a MWE candidate, i.e., the non-replacement of words that form the MWE candidate by synonyms. NEMWEL's source code and search interface will be available soon at: `http://www.lalic.dc.ufscar.br/never-ending/`.

MWE candidate	NEWMEL	Reference
horário comercial *business hours*	F	T
dona de casa *housewife*	F	T
dor de cabeça *headache*	F	T
fogo de artifício *firework*	T	T
empate técnico *technical draw*	T	T
terminal de ônibus *bus terminal*	T	T
estado do Rio *state of Rio*	F	F
ano passado *last year*	F	F
local de exame *test site*	F	F
redução de custo *cost reduction*	T	F
banco traseiro *rear seat*	T	F
processo de seleção *selection process*	T	F

Table 2: Examples of true MWE candidates extracted by NEMWEL, respectively: false negatives, true positives, true negatives and false positives.

Acknowledgments

We would like to thank São Paulo Research Foundation (FAPESP) for the grants #2013/11811-0 and #2013/50757-0 (AIM-WEST project). This work is also part of CAMELEON (CAPES-COFECUB #707/11) and RITA (CAPES #047/14) projects.

References

Vitor De Araujo, Carlos Ramisch, and Aline Villavicencio. 2011. Fast and flexible MWE candidate generation with the mwetoolkit. In Kordoni et al. (Kordoni et al., 2011), pages 134–136.

Mohammed Attia, Antonio Toral, Lamia Tounsi, Pavel Pecina, and Josef van Genabith. 2010. Automatic extraction of Arabic multiword expressions. In Éric Laporte, Preslav Nakov, Carlos Ramisch, and Aline Villavicencio, editors, *Proc. of the COLING Workshop on MWEs: from Theory to Applications (MWE 2010)*, pages 18–26, Beijing, China, Aug. ACL.

Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of Natural Language Processing*, pages 267–292. CRC Press, Taylor and Francis Group, Boca Raton, FL, USA, 2 edition.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Lang. Res. & Eval.*, 43(3):209–226, Sep.

Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R. Hruschka Jr., and Tom M. Mitchell. 2010. Toward an architecture for never-ending language learning. In *Proceedings of the Twenty-Fourth Conference on Artificial Intelligence (AAAI 2010)*.

Helena de Medeiros Caseli, Carlos Ramisch, Maria das Graças Volpe Nunes, and Aline Villavicencio. 2010. Alignment-based extraction of multiword expressions. *Lang. Res. & Eval. Special Issue on Multiword expression: hard going or plain sailing*, 44(1-2):59–77, Apr.

C. C. Chang and C. J. Lin. 2011. LIBSVM: A Library for Support Vector Machines. In *ACM Transactions on Intelligent Systems and Technology*.

Kenneth Church and Patrick Hanks. 1990. Word association norms mutual information, and lexicography. *Comp. Ling.*, 16(1):22–29.

J. Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Matthieu Constant and Anthony Sigogne. 2011. MWU-aware part-of-speech tagging with a CRF model and lexical resources. In Kordoni et al. (Kordoni et al., 2011), pages 49–56.

Jianyong Duan, Ruzhan Lu, Weilin Wu, Yi Hu, and Yan Tian. 2006. A bio-inspired approach for multi-word expression extraction. In James Curran, editor, *Proc. of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 176–182, Sidney, Australia, Jul. ACL.

Marie Dubremetz and Joakim Nivre. 2014. Extraction of nominal multiword expressions in french. In *Proceedings of the 10th Workshop on Multiword Expressions (MWE)*, pages 72–76, Gothenburg, Sweden, April. Association for Computational Linguistics.

Stefan Evert and Brigitte Krenn. 2005. Using small random samples for the manual evaluation of statistical association measures. *Comp. Speech & Lang. Special issue on MWEs*, 19(4):450–466.

Meghdad Farahmand and Ronaldo Martins. 2014. A supervised model for extraction of multiword expressions, based on statistical context features. In *Proceedings of the 10th Workshop on Multiword Expressions (MWE)*, pages 10–16, Gothenburg, Sweden, April. Association for Computational Linguistics.

Nicole Grégoire, Stefan Evert, and Brigitte Krenn, editors. 2008. *Proc. of the LREC Workshop Towards a Shared Task for MWEs (MWE 2008)*, Marrakech, Morocco, Jun.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The weka data mining software: An update. In *SIGKDD Explorations*, volume 11.

Ray Jackendoff. 1997a. *The Architecture of the Language Faculty*. Number 28 in Linguistic Inquiry Monographs. MIT Press, Cambridge, MA, USA. 262 p.

Ray Jackendoff. 1997b. Twistin' the night away. *Language*, 73:534–559.

John S. Justeson and Slava M. Katz. 1995. Technical terminology: some linguistic properties and an algorithm for identification in text. *Nat. Lang. Eng.*, 1(1):9–27.

Su Nam Kim and Preslav Nakov. 2011. Large-scale noun compound interpretation using bootstrapping and the web as a corpus. In Regina Barzilay and Mark Johnson, editors, *Proc. of the 2011 EMNLP (EMNLP 2011)*, pages 648–658, Edinburgh, Scotland, UK, Jul. ACL.

Valia Kordoni, Carlos Ramisch, and Aline Villavicencio, editors. 2011. *Proc. of the ACL Workshop on MWEs: from Parsing and Generation to the Real World (MWE 2011)*, Portland, OR, USA, Jun. ACL.

Mirella Lapata and Frank Keller. 2005. Web-based models for natural language processing. *ACM Trans. Speech and Lang. Process. (TSLP)*, 2(1):1–31.

D. Li, G. Savova, and K. Kipper-Schuler. 2008. Conditional random fields and support vector machines for disorder named entity recognition in clinical texts. In *Proceedings of the workshop on current trends in biomedical natural language processing*, pages 94–95, Columbus, Ohio.

T. Mitchell, W. Cohen, E. Hruschka, P. Talukdar, J. Betteridge, A. Carlson, B. Dalvi, M. Gardner, B. Kisiel, J. Krishnamurthy, N. Lao, K. Mazaitis, T. Mohamed, N. Nakashole, E. Platanios, A. Ritter, M. Samadi, B. Settles, R. Wang, D. Wijaya, A. Gupta, X. Chen, A. Saparov, M. Greaves, and J. Welling. 2015. Never-ending learning. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI-15)*.

T. Mullen and N. Collier. 2004. Sentiment analysis using support vector machines with diverse information sources. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 412–418.

Darren Pearce. 2001. Synonymy in collocation extraction. In *WordNet and Other Lexical Resources: Applications, Extensions and Customizations (NAACL 2001 Workshop)*, pages 41–46, Jun.

Pavel Pecina. 2008. Reference data for Czech collocation extraction. In Grégoire et al. (Grégoire et al., 2008), pages 11–14.

Ted Pedersen, Satanjeev Banerjee, Bridget McInnes, Saiyam Kohli, Mahesh Joshi, and Ying Liu. 2011. The n-gram statistics package (text::NSP) : A flexible tool for identifying n-grams, collocations, and word associations. In Kordoni et al. (Kordoni et al., 2011), pages 131–133.

Carlos Ramisch, Paulo Schreiner, Marco Idiart, and Aline Villavicencio. 2008. An evaluation of methods for the extraction of multiword expressions. In Grégoire et al. (Grégoire et al., 2008), pages 50–53.

Carlos Ramisch. 2015. *Multiword Expressions Acquisition: A Generic and Open Framework*, volume XIV of *Theory and Applications of Natural Language Processing*. Springer.

Siva Reddy, Diana McCarthy, and Suresh Manandhar. 2011. An empirical study on compositionality in compound nouns. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 210–218, Chiang Mai, Thailand, November. Asian Federation of Natural Language Processing.

Ivan Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Proc. of the 3rd CICLing (CICLing-2002)*, volume 2276/2010 of *LNCS*, pages 1–15, Mexico City, Mexico, Feb. Springer.

Bahar Salehi and Paul Cook. 2013. Predicting the compositionality of multiword expressions using translations in multiple languages. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 266–275, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

M. Sassano. 2003. Virtual Examples for Text Classification with Support Vector Machines. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 208–215.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, pages 44–49.

Nathan Schneider, Emily Danchik, Chris Dyer, and Noah A. Smith. 2014. Discriminative lexical semantic segmentation with gaps: running the MWE gamut. *Transactions of the Association for Computational Linguistics*, 2:193–206, Apr.

Violeta Seretan and Eric Wehrli. 2009. Multilingual collocation extraction with a syntactic parser. *Lang. Res. & Eval. Special Issue on Multilingual Language Resources and Interoperability*, 43(1):71–85, Mar.

Frank A. Smadja. 1993. Retrieving collocations from text: Xtract. *Comp. Ling.*, 19(1):143–177.

Yulia Tsvetkov and Shuly Wintner. 2010. Extraction of multi-word expressions from small parallel corpora. In Chu-Ren Huang and Dan Jurafsky, editors, *Proc. of the 23rd COLING (COLING 2010) — Posters*, pages 1256–1264, Beijing, China, Aug. The Coling 2010 Organizing Committee.

Leonardo Zilio, Luiz Svoboda, Luiz Henrique Longhi Rossi, and Rafael Martins Feitosa. 2011. Automatic extraction and evaluation of mwe. In *STIL 2011 - Cuiabá, MT, Brasil.*