Extending FrameNet to Machine Learning Domain

Piotr Jakubowski[1], Agnieszka Ławrynowicz[1]

Institute of Computing Science, Poznan University of Technology, Poland
{pjakubowski,alawrynowicz}@cs.put.poznan.pl

Abstract. In recent years, several ontological resources have been proposed to model machine learning domain. However, they do not provide a direct link to linguistic data. In this paper, we propose a linguistic resource, a set of several semantic frames with associated annotated initial corpus in machine learning domain, we coined MLFrameNet. We have bootstrapped the process of (manual) frame creation by text mining on the set of 1293 articles from the Machine Learning Journal from about 100 volumes of the journal. It allowed us to find frequent occurences of words and bigrams serving as candidates for lexical units and frame elements. We bridge the gap between linguistics analysis and formal ontologies by typing the frame elements with semantic types from the DMOP domain ontology. The resulting resource is aimed to facilitate tasks such as knowledge extraction, question answering, summarization etc. in machine learning domain.

1 Introduction

For arguably any scientific domain, there exists big amount of textual content that includes probably interesting information buried in linguistic structures. Each of the domains has aspects that are typical only for it. For example in the field of machine learning there are sentences dealing with various measures, numerical data or comparisons. A method for automatic extraction of such specific information could facilitate exploration of text corpus, for instance when we are looking for information about accuracy or popularity of a concrete algorithm among all articles on machine learning.

From the other side there are ontological resources that model domain knowledge using formal, logic-based languages such as OWL[1]. We aim to leverage those for facilitating tasks such as knowledge extraction, question answering, summarization etc. in machine learning domain.

We propose therefore to fill the gap between linguistic analysis and formal semantics by combining *frame semantics* [4] with mapping to a machine learning specific ontology. To this end, we extend FrameNet [10] – a lexicon for English based on *frame semantics* – to the machine learning domain. In this paper, we

[1] https://www.w3.org/TR/owl-features/

present an initial version of this extension, we coined *MLFrameNet*, consisting of several *semantic frames* that cover a part of the machine learning domain.

The rest of the paper is organized like follows. In Section 2 we discuss related works including a short introduction to FrameNet, other extensions of FrameNet and machine learning ontologies. in Section 3 we describe the process of developing the extension which includes collecting a corpus of ML domain-specific articles and is based on automatic extraction of *lexical units (LU)* from the corpus; the lexical units can help to identify parts of a semantic frame. In Section 5 we provide a discussion, and Section 6 concludes the paper.

2 Preliminaries and Related Works

2.1 FrameNet

Frame semantics developed by Fillmore [5] is a theory of linguistic meaning. It describes the following elements that characterize events, relations or entities and the participants in it: frame, frame elements, lexical units. The main concept is a *frame*. It is a conceptual structure modeling a prototypical situation. *Frame Elements (FEs)* are a part of the frame that represents the *roles* played during the situation realization by its participants. The other part of a semantic frame are *Lexical Units (LUs)*. They are predicates that linguistically express the situation represented by the frame. We can say that the frame is evoked in texts through the occurence of its lexical unit(s).

Each semantic frame usually contains more than one LU and may come into relationship, such as hyponymy, with other frames.

The standard approach for creating semantic frames described by Fillmore [6] is based on five main steps: i) characterizing situations in particular domain which could be modeled as a semantic frame, ii) describing Frame Elements, iii) selecting lexical units that can evoke a frame, iv) annotating sample sentences from large corpus of texts, and finally v) generating lexical entries for frames, which are derived for each LU from annotations, and describe how FEs are realized in syntactic structures.

The FrameNet project [10] is constructing a lexical database of English based on frame semantics, containing 1,020 frames (release 1.5).

2.2 Extensions of FrameNet

There have been several extensions of FrameNet to specific domains including biomedical domain (BioFrameNet [2]), legal domain [13] and sport (Kicktionary [11]). In all of these cases, the authors pointed that each specific domain is characterized by specific challenges related to creating semantic frames. One major decision concerns whether it is necessary to create a new frame or we can use one of those existing in FrameNet and extend it.Another design aspect deals with typing of frame elements with available controlled vocabularies and/or ontologies. For instance, the structure of Kicktionary, a multi-lingual extension

of FrameNet for football domain, allows to connect it to the concrete football ontology [1]. Even better developed BioFrameNet extension has its structure connected to biomedical ontologies [2].

2.3 Machine Learning Ontologies

There have been proposed a few ML ontologies or vocabularies such as DMOP [7], OntoDM [9], Exposé [12] and MEX vocabulary [3]. A common proposed standard schema unifying these efforts, ML Schema, is only on the way being developed by the W3C Machine Learning Schema Community Group[2]. Despite of the existence of the ontological resources and vocabularies which formalize the ML domain, a linguistic resource linking those to textual data is missing. Therefore we propose to fill this gap by MLFrameNet and to link it to an existing ML ontology.

3 Frame Construction Pipeline – Our Approach

We propose a pipeline in order to extract information needed for creating semantic frames on machine learning that consists of five steps (Figure 1).

At first we crawled websites from `http://www.springer.com` to extract data for creating a text corpus based on the *Machine Learning Journal* articles. All articles were stored in text files without any preprocessing like stemming or removing stopwords. The reason for this is that whole sentences were later used for creating semantic frames. In the second step, we applied statistical approach based on calculating histogram for articles to find out, which words or phrases (e.g., bigrams) occur most frequently. This is the major part of our method and it aims to find candidates for lexical units or frame elements for new frames based on text mining. We envisage that those candidates could play a role of lexical units or instantiations of frame elements. Usage of them should simplify the process of new semantic frames creation. In the third step, we gather the sentences that contain the found expressions. In the fourth step, we created the frames manually, leveraging the candidates for the frame parts and sentences containing them. In the final step, after creating frame drafts that could fit existing FrameNet structure, we connected the frame elements to terms from the DMOP ontology that covers machine learning domain.

3.1 Corpus

The data for this research comes from *Machine Learning Journal* and covers 1293 articles from 101 volumes of that journal stored in filesystem as text files with metadata stored in a database. Importantly: *Springer grants text and data-mining rights to subscribed content, provided the purpose is non-commercial research*[3]. We used an open source framework written in Python for crawling web

[2] `https://www.w3.org/community/ml-schema/`

[3] Sentence from the licence `http://www.springer.com/gp/rights-permissions/springer-s-text-and-data-mining-policy/29056`

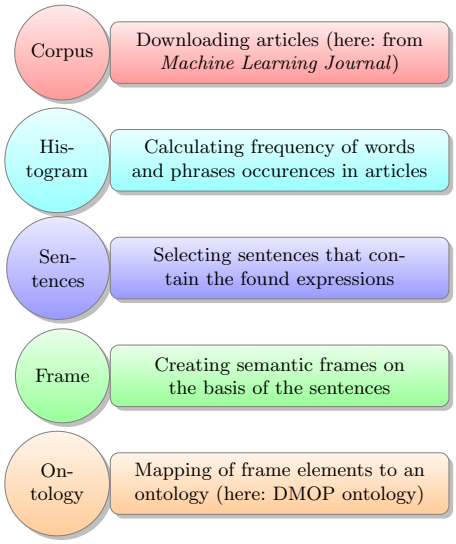

Fig. 1. The pipeline of the method for creating semantic frames in ML domain

pages and downloading articles. Preliminary preprocessing of stored content was made by Python library NLTK[4].

3.2 Data Mining Optimization Ontology

The *Data Mining OPtimization Ontology (DMOP)* [7] has been developed with the primary purpose of the automation of algorithm and model selection via *semantic meta-mining* that is an ontology-based approach to meta-learning of complete data mining processes in view of extracting patterns associated with performance. DMOP contains detailed descriptions of data mining tasks (e.g., learning, feature selection, model application), data, algorithms, hypotheses (models or patterns), and workflows. In response to many non-trivial modeling problems that were encountered due to the complexity of the data mining domain details, the ontology is highly axiomatized and modeled with the use of the OWL 2 DL[5] profile. DMOP was evaluated for semantic meta-mining in several problems and used in building the Intelligent Discovery Assistant a plugin to the popular data mining tool RapidMiner. We use DMOP to provide the semantic types for the frame elements.

[4] `http://www.nltk.org`
[5] `https://www.w3.org/TR/owl2-overview/`

Table 1. The most common bigrams from *Machine Learning Journal* articles

Bigram	Number of occurences	Bigram	Number of occurences
machine learning	718	bayes net	192
data set	489	experimental results	189
learning algorithm	377	training examples	182
training set	364	loss function	177
training data	325	upper bound	177
active learning	277	data points	174
feature selection	259	feature space	171
reinforcement learning	224	sample complexity	159
value function	217	learning methods	153
time series	201	decision trees	152
natural language	192	lower bound	143

3.3 Methods

In this section we will describe in more detail the execution of the subsequent steps of our pipeline.

During searching for candidates for lexical units or frame elements we tried three different histograms. At first we found simple words which occur most frequently in our corpus. We restricted the number of results to 521 words that occur more than 300 times. In the second approach, instead of words we searched for bigrams (phrases consisting of two words) and restricted the results to those which occur more than 32 times in the corpus, what resulted in 490 bigrams. Finally, we checked the quality of the results using tf-idf numerical statistic - for each of 1294 articles we chosen ten words with the highest tf-idf measure.

The most interesting results pertain to bigrams that occur most frequently in the corpus. The most frequent bigrams are presented in Table 1.
We use them as elements of semantic frames, e.g. as lexical units or instances of a frame element. The clue of our method was to select sentences containing the found expressions. Those sentences could be very likely occurences of semantic frames in the domain of machine learning. Additionally, we were looking for sentences in which our bigrams were parts of a noun expression or a verb expression (lexical units and frame elements are often such parts of speech).

4 MLFrameNet

On the basis of sentences extracted during the process described in the previous section, we manually developed several semantic frames. Each of the sentences contains at least one of the most common word or bigrams in the corpus. They are very often the part of a frame element or a lexical unit.
By now we have developed eight frames that cover the basics of the machine learning domain. The names of those frames are: *Algorithm*, *Data*, *Model*, *Task*, *Measure*, *Error*, *TaskSolution* and *Experiment*.

Below, we present the frames in a FrameNet style. The proposed lexical units are underlined, frame elements are in brackets (with adequate number superscripted in the definition of situation) and phrases extracted from the histogram are in **bold**.

Task:

- Definition of situation: This is a frame for representing ML task[1], and optionally an algorithm[2] for solving it.
- Frame Elements: (1) ML task; (2) ML algorithm
- Lexical Units: *supervised, unsupervised, reinforcement learning, classification, regression, clustering, density estimation, dimensionality reduction*
- An example of annotated sentence:
  [**Supervised learning** ML task] can be used to build class probability estimates.

Algorithm:

- Definition of situation: This frame represents classes of ML Algorithms[1], their instances[2], tasks[3] they address, data[4] they specify, the type of hypothesis[5] they produce, ML software (environment)[6] where they are implemented and the optimization problem they try to solve[7].
- Frame Elements: (1) ML algorithm type (2) instance; (3) ML task; (4) data; (5) hypothesis; (6) software; (7) optimization problem
- Lexical Units: *algorithm, learning algorithm, method, learning method*
- An example of annotated sentence:
  [Expectation Maximization instance] is the standard [semi-supervised **learning algorithm** ML algorithm type] for [generative models hypothesis].

Data:

- Definition of situation: This frame represents data[1], the quantity or dimensions[2] associated with given data (e.g, a number of datasets, number of features), identifies the origin[3] of data, its characteristic [4], its name[5] (e.g., of a particular dataset).

– Frame Elements: (1) data (2) quantity; (3) origin; (4) characteristic; (5) name.

– Lexical Units: *data, data set, training set, training data, training examples, examples, data point, test set, test data, label ranking, preference information, background knowledge, prior knowledge, missing values, ground truth, unlabeled data, data stream, positive examples, data streams, class labels, gene expression, real data, missing data, synthetic data, labeled data, high dimensional, negative examples, training samples, multi-label data, training instances, instances, real-world data, data values, labeled examples, feature vector, feature set, validation set, observed data, relational data, large data, time points, sample*

– An example of annotated sentence:
We note that the [extreme sparsity $_{\text{characteristic}}$] of this [**data set** $_{\text{data}}$] makes the prediction problem extremely difficult.

Model:

– Definition of situation: This frame represents ML models[1], identifies ML algorithms[2] that produce the models, and model's characteristics[3].
– Frame Elements: (1) model (2) ML algorithm; (3) characteristic.
– Lexical Units: *model, models, hypothesis, hypotheses, cluster, clusterings, rules, patterns, bayes net, decision tree, graphical model, joint distribution, neural network, generative model, bayesian network*
– An example of annotated sentence: [RIDOR $_{\text{ML algorithm}}$] creates a set of [**rules** $_{\text{model}}$], but does not keep track of the number of training instances covered by a rule.

Measure:

– Definition of situation: This frame represents information about specific measure[2] (and its value[5]) used to estimate the performance of a specific ML algorithm[1] on some dataset[4] in a specific way[6]. The ML algorithm solves ML task[3].

– Frame Elements: (1) ML algorithm/model; (2) measure; (3) ML task (4) dataset (5) measure value (6) measure method

– Lexical Units: *result, measure, estimate, performance, better, worse, precision, recall, accuracy, lift, ROC, confusion matrix, cost function*

– An example of annotated sentence:
Additional experiments based on ten runs of [10-fold **cross validations** $_{\text{measure method}}$] on [40 **data sets** $_{\text{dataset}}$] further support the effectiveness

of the [reciprocal-sigmoid model $_{\text{ML Algorithm/model}}$], where its [**classification accuracy** $_{\text{measure}}$] is seen to be comparable to several top classifiers in the literature.

Error:

- Definition of situation: This frame describes type of error[1] that coud be used for specific ML algorithm[2], that solves ML task[3]. The error value[4] can be calculated for specific data[5].

- Frame Elements: (1) error type; (2) ML task; (3) error value (4) ML algorithm (5) dataset

- Lexical Units: *error, measure, minimize, maximize, validation set error, prediction error, expected error, error rate, error loss, generalization error, training error, approximation error*

- An example of annotated sentence:
  We present an efficient [**algorithm** $_{\text{ML algorithm}}$] for [computing the optimal two-dimensional region $_{\text{ML task}}$] that minimizes the [mean squared **error** $_{\text{error type}}$] of an objective numeric attribute in a given database.

Task_Solution:

- Definition of situation: This is a frame for representing relations between ML task[1] and method[2] that solves it. The solution method could be wider described[3]. The method or collateral problems are probably described in reference article[4].

- Frame Elements: (1) ML task (2) solution type; (3) solution description (4) authors/references

- Lexical Units: *solve, solving, model, assume, perform*

- An example of annotated sentence:
  Indeed, [MCTS $_{\text{solution type}}$] has been recently used by [Gaudel and Sebag (2010) $_{\text{authors/references}}$] in their [FUSE (Feature Uct SElection) $_{\text{solution type}}$] system to perform [**feature selection** $_{\text{ML task}}$].

Experiment:

- Definition of situation: This is a frame for representing relations between ML experiment[1] and data[2] used in the expriment, an ML algorithms/models applied[3], measure[4] used to assess the results of an experiment or possibly an error [5] calculated based on the experiment results, measure or error value[6] and indication of possible loss or gain[7] in a comparison.

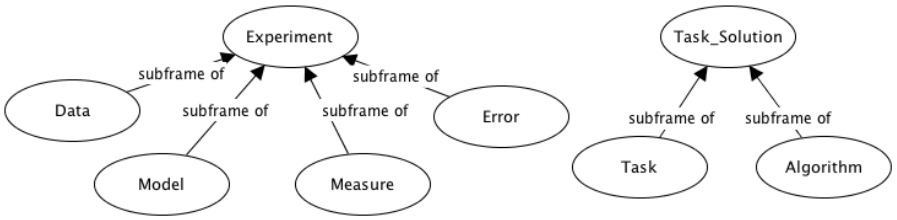

Fig. 2. The 'subframe of' relations between the frames.

- Frame Elements: (1) ML experiment (2) data; (3) ML algorithm/model; (4) measure; (5) error; (6) measure or error value; (7) loss or gain indication.

- Lexical Units: *experiment, investigation, empirical investigation, study, run, evaluation*

- An example of annotated sentence:
  [**Experiments** ML experiment] on a [large OCR data set $_{\text{data}}$] have shown [CB1 ML algorithm/model] to [significantly increase loss or gain indication] [generalization accuracy measure] over [SSE or CE optimization ML algorithm/model], [from 97.86% and 98.10% measure or error value], respectively, to [99.11% measure or error value] .

The Table 2 presents a set of mappings of frame elements to DMOP terms. DMOP was selected from among the available machine learning domain ontologies, since it links to the foundational ontology *Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE)* [8]. Due to this alignement, we have found it more relevant for applications related to computational linguistics than the other available ontologies. We only presented the exising mappings, omitting the frame elements for which no precise mapping exists yet. Sometimes it is due to the ontological ambiguity of the common language (discussed in the next Section). The other times, the DMOP ontology does not contain an adequate vocabulary term as for instance the author of an algorithm (such information as scientific papers describing particular algorithms are placed in DMOP in the annotations).

The 'subframe of' relations between frames are illustrated in Figure 2. They highlight the nature of the developed frames. Some of the frames, Task, Algorithm, Data, Model, Measure, Error, represent objects (corresponding to nouns), while the others, Task_Solution and Experiment, represent more complex situation in the former case or an event in the latter case (which is also reflected by their LUs that are mostly verbs).

The MLFrameNet data is being made available at `https://semantic.cs.put.poznan.pl/wiki/aristoteles/`.

Table 2. The mappings of the frame elements to DMOP terms.

Frame Element	DMOP term
Algorithm.ML algorithm type	dmop:DM-Algorithm
Algorithm.instance	dmop:DM-Algorithm
Algorithm.ML task	dmop:DM-Task
Algorithm.data	dmop:DM-Data
Algorithm.hypothesis	dmop:DM-Hypothesis
Algorithm.software	dmop:DM-Software
Algorithm.optimization problem	dmop:OptimizationProblem
Data.data	dmop:DM-Data
Data.characteristic	dmop:DataCharacteristic
Model.model	dmop:DM-Hypothesis
Model.ML algorithm	dmop:InductionAlgorithm
Model.characteristic	dmop:HypothesisCharacteristic
Measure.measure	dmop:HypothesisEvaluationMeasure
Measure.ML task	dmop:DM-Task
Measure.dataset	dmop:DM-Data
Error.error type	dmop:HypothesisEvaluationFunction
Error.ML task	dmop:DM-Task
Error.ML algorithm	dmop:DM-Algorithm
Error.dataset	dmop:DM-Data
Task_Solution.ML task	dmop:DM-Task
Experiment.experiment	dmop:DM-Experiment
Experiment.data	dmop:DM-Data
Experiment. measure	dmop:HypothesisEvaluationMeasure
Experiment.error	dmop:HypothesisEvaluationFunction

5 Discussion

The creation of the most frequent occurences of words and bigrams was very helpful in creating semantic frames since it introduced filtering such that there was no need to analyze the whole corpus of articles.

After the process of making frames, we investigated some inconvenience in our approach and things that we could do better.
First of them is that sometimes it turns out that we want to know the context of particular sentence to build a valuable frame from it or to extract more frame elements. For example for the sentence " *This problem could be solved by logistic regression.*" we can assume that in the previous few sentences there occurs the information about the name of the problem. Our method does not solve this issue, as the sentence is not bound to the previous sentence.

During the process of creating semantic frames for machine learning it occurs that in such restricted domain the amount of lexical units is much smaller than for general FrameNet. This situation cause that a number of frames can be evoked by the same lexical units.

An interesting modeling problem that we have encountered is an interchangeable usage of the concepts of an algorithm and a model (the algorithm produces) in machine learning texts while describing the performance of the algorithms and models. Ontologically, it is the model that is being used to produce the performance measurement and not the algorithm that produced the model. In a common language, however, it often appears that the term algorithm is that associated with producing the performance. Since those terms played many times this particular role interchangeably in the sentences, we have modeled such frame elements as 'Measure.ML algorithm/model'. However, it poses problems for semantic typing as clearly algorithm and model are disjoint in the DMOP ontology.

Due to the licence issues we are only able to publish a corpus of annotated sentences where there is only maximum one sentence per each Machine Learning Journal non-open access article. There is no such restriction in case of the open access articles. It is noteworthy, that this restriction does not prevent text mining of the journal articles for scientific purposes such as our automatic statistical analysis of most frequent words which is allowed.

6 Conclusions and future work

In this paper, we have proposed an initial extension to the FrameNet resource for the machine learning domain: MLFrameNet. We have discussed our approach to the problem of creating semantic frames for this specific technical domain of machine learning. So far, our main objective was to create a valuable resource for machine learning domain in the FrameNet style that could also serve as a *seed resource* for further automatic methods. Thus we have been less concentrated on the pipeline itself that will be a topic of the future work. Nevertheless, our attempts have shown that statistical analysis of domain-specific corpus of text is an effective way of finding appropriate vocabulary, that can be treated as a part of semantic frames. Gradually we will be building new semantic frames in this domain.

In the future work, we will conduct an external evaluation with use of one of available crowdsourcing platforms for evaluating resources that we have created so far. Especially, we plan to perform a crowdsourcing experiment in which contributors will decide whether a sample sentence is properly annotated. We want to tackle the problem of taking into account the context of the sentence and investigate the implications of that multiple frames can be evoked by the same lexical units. We also plan to extend our corpus by new annotations that may be published without publishing the original sentences or new texts. Moreover, we want to search for new candidates for frame elements automatically. That approach could be built on the basis of parts of speech or parts of sentences, for example through finding similarities between existing, manually annotated, sentences and new examples. We plan to use the created MLFrameNet resource for relation extraction from the scientific articles, in order to populate data mining ontologies (DMOP) and schemas (ML Schema) and create Linked Data describing machine learning experiments described in scientific articles.

Acknowledgments This research has been supported by the National Science Centre, Poland, within the grant number 2014/13/D/ST6/02076.

References

1. Buitelaar, P., Eigner, T., Gulrajani, G., Schutz, A., Siegel, M., Weber, N., Cimiano, P., Ladwig, G., Mantel, M., Zhu, H.: Generating and visualizing a soccer knowledge base. In: Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics: Posters & Demonstrations. pp. 123–126. Association for Computational Linguistics (2006)
2. Dolbey, A., Ellsworth, M., Scheffczyk, J.: Bioframenet: A domain-specific framenet extension with links to biomedical ontologies. In: In Proceedings of the Biomedical Ontology in Action Workshop at KR-MED. pp. 87–94 (2006)
3. Esteves, D., Moussallem, D., Neto, C.B., Soru, T., Usbeck, R., Ackermann, M., Lehmann, J.: MEX vocabulary: a lightweight interchange format for machine learning experiments. In: Proceedings of the 11th International Conference on Semantic Systems, SEMANTICS 2015, Vienna, Austria, September 15-17, 2015. pp. 169–176 (2015), http://doi.acm.org/10.1145/2814864.2814883
4. Fillmore, C.J.: Frame semantics and the nature of language. Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech 280(1), 20–32 (1976)
5. Fillmore, C.J.: Frames and the semantics of understanding. Quaderni di semantica 6(2), 222–254 (1985)
6. Fillmore, C.J., Baker, C.: A frames approach to semantic analysis. The Oxford handbook of linguistic analysis pp. 313–339 (2010)
7. Keet, C.M., Lawrynowicz, A., d'Amato, C., Kalousis, A., Nguyen, P., Palma, R., Stevens, R., Hilario, M.: The data mining optimization ontology. J. Web Sem. 32, 43–53 (2015), http://dx.doi.org/10.1016/j.websem.2015.01.001
8. Masolo, C., Borgo, S., Gangemi, A., Guarino, N., Oltramari, A.: Ontology library. WonderWeb Deliverable D18 (ver. 1.0, 31-12-2003). (2003), http://wonderweb.semanticweb.org
9. Panov, P., Soldatova, L.N., Dzeroski, S.: Ontology of core data mining entities. Data Min. Knowl. Discov. 28(5-6), 1222–1265 (2014), http://dx.doi.org/10.1007/s10618-014-0363-0
10. Ruppenhofer, J., Ellsworth, M., Petruck, M.R., Johnson, C.R., Scheffczyk, J.: FrameNet II: Extended Theory and Practice. International Computer Science Institute, Berkeley, California (2006), distributed with the FrameNet data
11. Schmidt, T.: The Kicktionary: Combining corpus linguistics and lexical semantics for a multilingual football dictionary. na (2008)
12. Vanschoren, J., Blockeel, H., Pfahringer, B., Holmes, G.: Experiment databases - A new way to share, organize and learn from experiments. Machine Learning 87(2), 127–158 (2012), http://dx.doi.org/10.1007/s10994-011-5277-0
13. Venturi, G., Lenci, A., Montemagn, S., Vecchi, E.M., Sagri, M.T., Tiscornia, D.: Towards a FrameNet resource for the legal domain. In: Proceedings of the Third Workshop on Legal Ontologies and Artificial Intelligence Techniques. Barcelona, Spain (June 2009), http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-465/