Ensemble Learning of Named Entity Recognition Algorithms using Multi Layer Perceptron for the Multilingual Web of Data

René Speck
Data Science Group, University of Leipzig
Augustusplatz 10
Leipzig, Germany 04109
speck@informatik.uni-leipzig.de

Axel-Cyrille Ngonga Ngomo
Data Science Group, University of Paderborn
Pohlweg 51
Paderborn, Germany 33098
axel.ngonga@upb.de

ABSTRACT

Implementing the multilingual Semantic Web vision requires transforming unstructured data in multiple languages from the Document Web into structured data for the multilingual Web of Data, the current implementation of the Semantic Web. Named entity recognition is heavily involved in this transformation process and it has been shown that ensemble learning on the named entity recognition task improves its performance on English with a reduced error rate of 40%. Contributing towards the multilingual Semantic Web vision, we carry out an evaluation of ensemble learning with a multilayer perceptron for multilingual named entity recognition for Dutch, English, French, German and Spanish. The evaluation results show that our approach improves the performance of named entity recognition task on all five researched languages. In our best run, we increased the performance on a Dutch dataset by 32.38% F1-Score. We integrated the results of our evaluation in the open source framework *FOX* and thus provide a state-of-the-art system for transforming unstructured data in machineprocessable data for the multilingual Semantic Web.

CCS CONCEPTS

• **Information systems** → **Information extraction**; • **Computing methodologies** → *cross validation*; *Ensemble methods*; *Supervised learning by classification*;

KEYWORDS

Named Entity Recognition, Ensemble Learning, Multilingual, Semantic Web

1 INTRODUCTION

Named Entity Recognition (NER) has been received great attention in the last decades. It is the identification of proper names in natural language (NL) text such as "`Leipzig`", "`Leipzig University`" and "`Gottfried Wilhelm Leibniz`" as well as the typing of this names with predefined types such as *Location*, *Organization* and *Person*. Over in the past few years, the problem of recognizing named entities in natural language texts has been addressed by several approaches and frameworks [9, 11, 13, 14, 19, 29]. Moreover, NER is a pre-processing step for deeper linguistic processing such as coreference resolution, syntactic and semantic parsing. Named Entity Disambiguation (NED) links this proper names to a knowledge base, such as DBpdiea, Wikidata or Yago. NER and NED are closely related and NER serves as processing step for NED as well [20, 40].

Currently, NER together with NED plays a central roll in many application primarily in the Web of Data community. The state-of-the-art knowledge extraction suite *FOX* [35] integrates among others NER systems as well as NED approaches and contributes in several applications with different domains in relation to the Web of Data [5, 16, 18, 20, 25, 32, 34, 39–41].

The input data of such applications is usually unstructured and can be multilingual. Thus, this data needs transformed into machine-processable data in order to analysed by machines. One possible use case of our approach is related to machine translation systems, that usually have problems with proper names. This problem can be solved with NER and NED by replacing the proper names with the target language equivalent from a knowledge base supporting multiple languages. The new version of *FOX* proposed in this paper supports a state-of-the-art NER and NED framework for multiple languages. With its uniform interface it can easily applied in this use case.

Implementing the vision of the multilingual Web of Data thus requires transforming multilingual unstructured data from the Document Web into structured data to the Web of Data with highly accurate NER and NED approaches. Providing multilingual NER and NED for the Web of Data is still a current challenging topic [11]. In this paper we address this research gap by providing an evaluation on ensemble learning of multilingual NER by using a multilayer perceptron. We include our evaluation results in the *FOX* framework and extend the framework with the new *Agdistis* version[20] in order provide knowledge extraction for multiple languages that produces machineprocessable Linked Data in a standardised format.

The goal of our evaluation is to answer the following research questions:

(1) Does for all the researched languages ensemble learning of NER algorithms with multilayer perceptron achieves higher F1-Score then the NER tools within the system?
(2) How does the dataset size to learn a multilayer perceptron model influences the F1-Score of the perceptron model?
(3) Which combination of the base NER tools within the multilayer perceptron reaches the highest performance?

The rest of this paper is structured as follows: We begin with an overview of the state of the art in NER, in the combination of NER systems and in benchmarking platforms. Then, in Section 3, we explain our evaluation pipeline and its setup together with the datasets used in this paper. In Section 4, we compare the results achieved by our evaluation on the silver and gold standard datasets. Finally, we discuss the insights provided by our evaluation and possible extensions of our approach in Section 5.2.

2 RELATED WORK

In this section we firstly give a brief introduction on work done by the natural language processing (NLP) community on NER systems. Secondly we explain approaches with the aim to improve the performance of the NER task by combining multiple NER systems. Thirdly we mention related work on NER and NED benchmarking systems.

2.1 NER Systems

NER tools and frameworks implement a broad spectrum of approaches, which can be subdivided into three main categories: dictionary-based, rule-based and machine-learning approaches [22]. The first systems for NER implemented dictionary-based approaches, which relied on a list of named entities (NEs) and tried to identify these in text [2, 44]. Following work then showed that these approaches did not perform well for NER tasks such as recognizing proper names [33]. Thus, rule-based approaches were introduced. These approaches rely on hand-crafted rules [6, 38] to recognize NEs. Most rule-based approaches combine dictionary and rule-based algorithms to extend the list of known entities. Nowadays, hand-crafted rules for recognizing NEs are usually implemented when no training examples are available for the domain or language to process [23]. When training examples are available, the methods of choice are borrowed from supervised machine-learning. Approaches such as Hidden Markov Models [46], Maximum Entropy Models [8] and Conditional Random Fields [15] have been applied to the NER task. Due to scarcity of large training corpora as necessitated by supervised machine-learning approaches, the semi-supervised [22, 27] and unsupervised machine-learning paradigms [12, 24] have also been used for extracting NER from text. [22] gives an exhaustive overview of approaches for the task.

2.2 NER Systems Combinations

This paper extends previous works ([25, 35, 36]) mainly by introducing a broadened language support and by performing a thorough evaluation of the extensions on multilingual datasets.

Thus, the work in [35] is the closest related work to this paper, its based on an earlier observation introduced and evaluated in [25]. The authors in [25] developed the Semantic Content Management Systems (SCMS) framework. The frameworks goals are the extraction of knowledge from unstructured data in any CMS and the integration of the extracted knowledge into the same CMS. In order to do this, it integrates a highly accurate knowledge extraction pipeline in its *extraction and storage layer*. The pipelines NER and machine-learning components of the *extraction layer* are extended and thoroughly evaluated in [35]. The evaluation results are integrated in the open source framework *FOX* [36]. *FOX* integrates several state of the art NER and ensemble learning algorithms as well as it implements a pipeline based on the 10-fold cross validation technique. The authors present an evaluation on ensemble learning for NER with the framework on five English datasets. The results show an improved performance of NER based on ensemble learning with a reduced error rate of 40% and suggest that Multilayer Perceptron and AdaBoostM1 with J48 as base classifier work best for the task at hand for English language in [35]. Another reason to choose the framework for our extension is that *FOX*participated in the Open Knowledge Extraction (OKE)) Challenge 2017 and won three out of four tasks[37], thus the framework showed promising performance. We extend the work with multiple languages and carry out an evaluation of the extensions.

The different to this work is the lack of the supported languages. We extend the *FOX* framework with NER systems to support multiple languages and we evaluate the extensions on newly integrated datasets for the languages. Furthermore, this paper researches the influence of the dataset size on the results. Our choice to reuse this framework is also based on the basis of the support of NED in the framework. It integrates *Agdistis* [40], a highly accurate NED approach, to link the NEs to DBpedia. Therewith, *FOX* produces its output as machineprocessable Linked Data in an RDF document serialisation in Turtle or RDF/XML, so that the results are reusable by the Web of Data community. Moreover, its open source and free usable.

More work on improving the performance on NER with machine-learning is present in *Nerd-ML*[43]. The work is based on [30] and produces Linked Data [31] as well. The different to this work is the domain of the tools and data researched by the authors. There approach is based on Micropost datasets and they show that the combination of several NER systems for Micropost with additional features outperform off-the-shelf approaches and a customised state of the art approach. The results indicate that an hybrid system may be better equipped to deal with the task.

In [45], a system was presented that combines with the ensemble learning stacking and voting classifiers which were trained with several languages, for language-independent NER.

2.3 NER Benchmarks

Over the last years, several benchmarks for NER have been proposed. For example, [7] presents a benchmark for NER and NED approaches. Especially, the authors define the named entity annotation task.

Gerbil [42] is another NER benchmarking platform and an extension of [7]. *Gerbil* provides comparable results to tool developers so as to allow them to easily discover the strengths and weaknesses of their implementations with respect to the state of the art. With the permanent experiment URIs provided by our framework, it ensures the reproducibility and archiving of evaluation results. Moreover, the framework generates data in machineprocessable format, allowing for the efficient querying and post-processing of evaluation results. Finally, the tool diagnostics provided by GERBIL allows deriving insights pertaining to the areas in which tools should be further refined, thus allowing developers to create an informed agenda for extensions and end users to detect the right tools for their purposes. GERBIL aims to become a focal point for the state of the art, driving the research agenda of the community by presenting comparable objective evaluation results. *Gerbil* defines many benchmarking and annotation tasks. To the best of our knowledge, *Gerbil* [42] supports English language in its current state only. Consequently, the platform mismatches our needs and we reuse and extend the evaluation components within *FOX* for our evaluation.

3 EVALUATION

In this section we explain the applied evaluation pipeline with the performance measures it uses. We then describe the pipelines setup comprising the integrated systems and datasets.

3.1 Pipeline

Our evaluation pipeline consists of a multilayer perceptron along with several diverse NER approaches serving as base classifiers for the ensemble learning algorithm. The pipelines goal is to combine the results of the base classifiers to improve the performance on the NER task. Hence, we apply the k-fold cross validation technique in our evaluation pipeline.

3.1.1 K-Fold Cross Validation. The k-fold cross validations preprocessing step splits a given dataset in k distinct equal sized subsets, each subset consists of an input text in NL and NEs. The cross validation process uses k-1 subsets as training data and the remaining subset as test data on the ensemble learning algorithm. Then, the cross validation passes four steps, A) the input text of the training data is send to the NER base classifiers to request the NEs in the text, B) the base classifiers NEs together with the training data are send to the ensemble learning algorithm to train a prediction model and store it if needed, C) the test datasets input text is send to the NER base classifiers to request the NEs in the text and D) the test datasets input text together with the NEs from the base NER classifiers and the trained model are used by the ensemble learning algorithm in order to predict NEs for the test dataset.

The whole process performs k times, so that each of the k subsets is tested once. The results of each test dataset from the folds together produce a single estimation over the whole dataset.

3.1.2 Measures. Equation 1 and 2 formalize precision pre_t and recall rec_t, the performance measures we compute on the test datasets for each entity type $t \in T$. They consist of the number of true positives TP_t, the number of true negatives TN_t, the number of false positives FP_t and the number of false negatives FN_t. We applied the one-against-all approach [1] to convert the multi-class confusion matrix of each dataset into a binary confusion matrix to compute the performance measures.

We macro average the performances over the entity types with Equation 3, the key performance indicators (KPIs) in our pipeline.

$$pre_t = \frac{TP_t}{TP_t + FP_t}; rec_t = \frac{TP_t}{TP_t + FN_t} \tag{1}$$

$$F1\text{-}Score_t = 2 \cdot \frac{pre_t \cdot rec_t}{pre_t + rec_t} \tag{2}$$

$$F1\text{-}Score_T = \frac{\sum_{t \in T} F1\text{-}Score_t}{|T|}; pre_T = \frac{\sum_{t \in T} pre_t}{|T|} \tag{3}$$

The conditions the result of an annotator has to fulfil to be a correct result is defined as *weak annotation matching* in [42].

In the *weak annotation matching*[1], we regarded partial matches of multi-word units as being partially correct. For example, our evaluation dataset considered "`Franziska Barbara Ley`" as being an instance of `Person`. If a tool generated "`Franziska`" as being a `Person` and omitted "`Barbara Ley`", it was assigned 1 true positive and 2 false negatives.

3.2 Setup

Our evaluation pipeline consists of five base classifiers, each supports one or more of the five languages we take into account in this paper (German, English, Spanish, French and Dutch). The exact language support of each tool can be seen later in the result tables in Section 4.

3.2.1 Integrated Systems. The integrated NER base classifiers in our evaluation are *Stanford* [13, 14, 19][2], *Illinois* [29][3], *OpenNLP* [3][4], *Balie* [21][5] and *Spotlight* [9][6]. The classifiers have been integrated in the *FOX* framework [36] for English already but we extended the framework and the tool integration to support multiple languages. Given that some of these tools at hand do not allow accessing their confidence score without any major alteration of their code, we considered the output of the tools to be binary (i.e., either 1 or 0).

The ensemble learning is carried out with a *Multilayer Perceptron (MLP)* as suggested in [35]. The MLP is implemented in the Weka library [17] and uses default options in our evaluation pipeline. Which means, in case we apply five NER base systems to our pipeline together with the entity types plus a type indicating its not an entity `NA`. Then, the MLP input features are 20, five NER base systems times four types, for the input layer. The hidden layer size is the half of the number of features plus the number of types, in our example 12. The output layer size is equal to the type size, 4.

For the evaluation process, we reuse the 10-fold cross validation integrated within the *FOX* framework and implemented in Weka as well. We test the results of our experiments of its significance with the Wilcoxon signed rank test [10] implemented in R [28]. We set the tests confidence interval to 95%.

We fist applied 10-fold cross validation to the base classifiers and then to each possible combination with the base classifiers. Finally, we choose the combination with the highest averaged $F1\text{-}Score_T$ that is significantly better than all base classifiers.

We only considered the performance on the entity types *Location*, *Organization* and *Person*. To this end, we mapped the entity types of each of the datasets and tools to these three types.

3.3 Datasets

Our evaluation pipeline integrates several datasets for multiple languages. Our first experiments are carried out on silver standard datasets in five different languages and ten different sizes. We test our approach on seven gold standard datasets in three different languages as well.

3.3.1 Silver Standard Datasets. *WikiDE*, *WikiEN*, *WikiES*, *WikiFR* and *WikiNL* are datasets in German, English, Spanish, French and Dutch by [26], a multilingual state-of-the art semi-supervised learning approach that provides a multilingual annotated corpora by exploiting the text and structure of Wikipedia. The silver-standard annotations outperform traditional training on a manually-annotated

[1] Equivalent to *token-based evaluation* in [35].

[2] http://nlp.stanford.edu:8080/ner/process

[3] http://cogcomp.cs.illinois.edu/page/demo_view/ner

[4] http://opennlp.apache.org/download.html

[5] http://balie.sourceforge.net/

[6] http://spotlight.sztaki.hu/downloads/

collection of Wikipedia articles. We reuse the silver-standard datasets in our evaluation with entities of the type *PER*, *LOC*, *ORG*. Since the datasets are large we research the influence of the dataset size. Therefore, we run 10 experiments on each dataset, starting with 500 sentences in the first experiment, followed by twice the number of sentences for the next experiment, until 5000 sentences in the last experiment.

3.3.2 Gold Standard Datasets. The datasets for Spanish *testa ES*, *testb ES*, *train ES* and for Dutch *testa NL*, *testb NL*, *train NL* are gold standard datas from the CoNLL-2002 shared task[7]. Where the first dataset of each language is the test a, the second the test b and the last the training dataset from the shared task. We reuse the dataset in our evaluation without the entity type *B-MISC* and *I-MISC*.

The Spanish data is a collection of news wire articles made available by the Spanish EFE News Agency. The articles are from May 2000.

The Dutch data consist of four editions of the Belgian newspaper "De Morgen" of 2000 (June 2, July 1, August 1 and September 1).

For the German dataset *train DE*, we reused the full training dataset in [4]. The dataset is based on the GermEval 2014[8] dataset. We reuse the datasets without the entity type *B-OTH* and *I-OTH*. dp

4 RESULTS

This section presents first the results of our experiments on the silver standard datasets and second on the gold standard datasets.

4.1 Silver Standard Datasets

Figures 1 to 5 depict the $F1\text{-}Score_T$ measured on the silver standard datasets for the five researched languages (German, English, Spanish, French and Dutch) in this paper. Each of the 5 figures shows the $F1\text{-}Score_T$ on 10 different dataset sizes on all the base classifiers which support the current language as well as the best combination of the classifiers, the ensemble learning. In 49 out of the 50 datasets, with different sizes and languages, the ensemble learning approach reaches a higher $F1\text{-}Score_T$ than the base classifiers. In one case, on the German dataset with the size of 1500 sentences, *Stanford* reaches a minimal higher value (62.02%) than the ensemble learning (61.57%).

The averaged $F1\text{-}Score_T$ and the averaged pre_T over the 10 dataset sizes for each language are depicted in Table 1. The rows of the table provide the performance on the researched five languages and the columns the performance of the base classifiers along with the ensemble learning approach named *FOX* in the table. The highest value of a classifier on a language is marked in each row. In our experiments, the combination of all base classifiers within the ensemble learning, in comparison with all other possible combinations, reaches the highest value on all languages. These values are shown in the last column of Table 1.

The best run regarding to the rise of the averaged $F1\text{-}Score_T$, we observe on the *WikiNL* dataset for Dutch with an increased value of +32.38% compared to the best base classifier *OpenNLP* on this dataset. And on *WikiES* for Spanish +29.45% compared to *Stanford*,

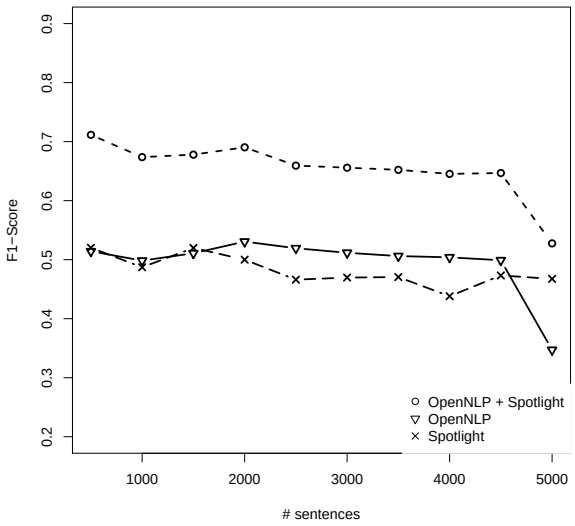

Figure 1: $F1\text{-}Score_T$ **on the Dutch dataset *WikiNL*.**

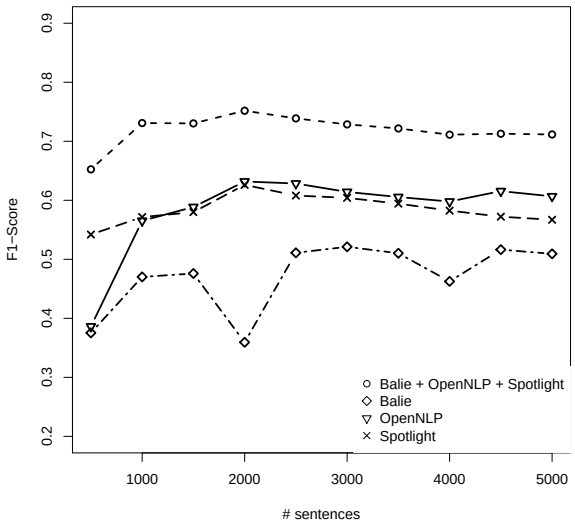

Figure 2: $F1\text{-}Score_T$ **on the French dataset *WikiFR*.**

Table 1: Averaged $F1\text{-}Score_T$ **and averaged** pre_T **in percentage.**

dataset	*Balie*	*Illinois*	*OpenNLP*	*Spotlight*	*Stanford*	*FOX*
DE	35.91/50.88	-	-	34.06/**79.17**	61.33/74.20	**63.00**/74.46
EN	56.23/64.87	70.57/70.14	46.30/58.53	57.22/74.21	76.70/78.61	**79.01**/**81.33**
ES	38.71/63.02	-	35.80/45.57	30.75/34.42	49.88/50.13	**64.57**/**74.58**
FR	47.12/71.53	-	58.40/86.01	58.48/**87.97**	-	**71.90**/82.95
NL	-	-	49.41/**79.96**	48.12/75.12	-	**65.41**/79.91

[7] http://www.cnts.ua.ac.be/conll2002/ner
[8] http://sites.google.com/site/germeval2014ner

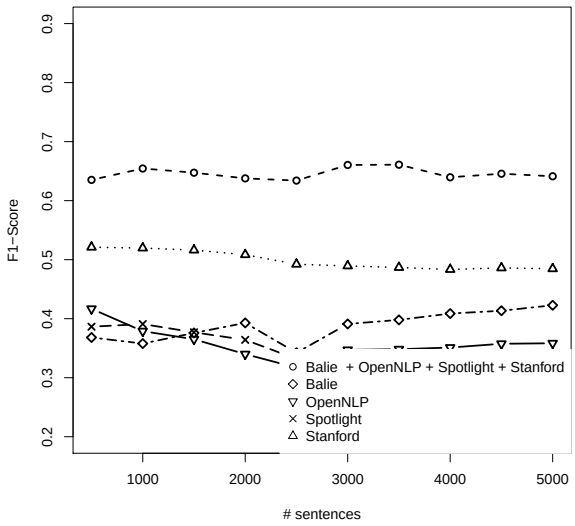

Figure 3: $F1\text{-}Score_T$ **on the Spanish dataset *WikiES*.**

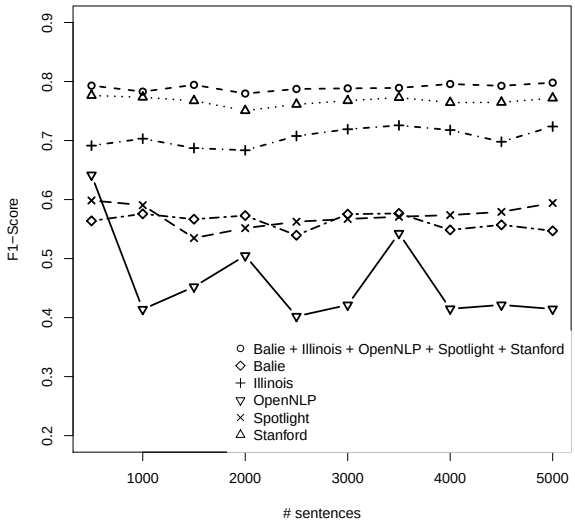

Figure 4: $F1\text{-}Score_T$ **on the English dataset *WikiEN*.**

on *WikiFR* for French +22.95% compared to *Spotlight*, on *WikiEN* for English +3.012% compared with *Stanford* and on *WikiDE* for German +3.279% compared with *Stanford* as well.

It is noteworthy, that the *Stanford* tool performs significantly better on its supported languages (German, English, Spanish) than all other integrated base classifiers, see Figures 3 to 5. For Dutch and French its *OpenNLP* see Figures 1 and 2.

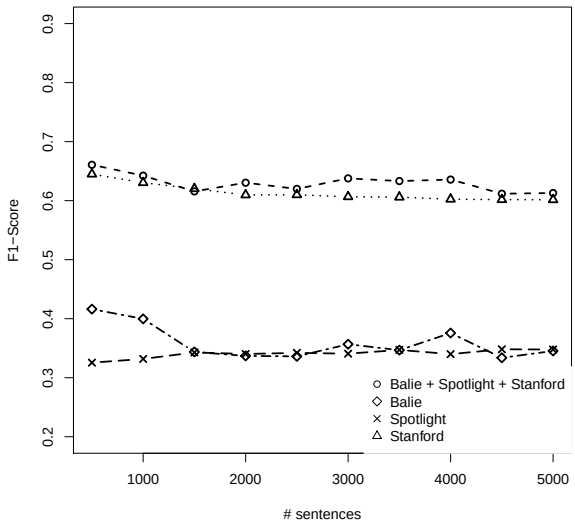

Figure 5: $F1\text{-}Score_T$ **on the German dataset *WikiDE*.**

The highest value over all languages we observe on the English dataset (79.01%) and the lowest on the German dataset (63.00%) with our approach. Further, the measured values on the datasets with different sizes indicate, that the dataset sizes, larger than 500 sentences, have no crucial influence on the ensemble learning approach in our case, since the values barely change.

Overall, the ensemble learning approach reaches, based on the significance test, a better averaged $F1\text{-}Score_T$ than a single NER base classifier on all five languages over the different dataset sizes. In two cases, for English and German, also the averaged pre_T with the ensemble learning approach reach the highest value.

4.2 Gold Standard Datasets

Figures 6 to 8 depict the $F1\text{-}Score_T$ on each gold standard dataset for the three languages German, Spanish and Dutch. Each figure shows the performance of the base classifiers as well as the the ensemble learning approach.

An overview of the values of the $F1\text{-}Score_T$ and pre_T on the gold standard datasets is given in Table 2. The rows of the table provide the performance on the researched datasets and the columns the performance of the base classifiers along with the ensemble learning approach named *FOX* in the table. The highest value of a classifier on a language is marked in each row.

Figure 6 shows the performance results on the *train DE* dataset for German. We observe a performance boost by +31.96% on $F1\text{-}Score_T$ with the combination of all base classifiers. In comparison, the *Stanford* system performs best as single base classifier with 45.97% $F1\text{-}Score_T$ and the combination of all NER base classifiers with ensemble learning reaches 60.66% $F1\text{-}Score_T$.

Figure 7 shows results on the *testa ES*, *testb ES* and *train ES* dataset for Spanish. We also observe an increased performance on all three datasets with the combination of all base classifiers and reach

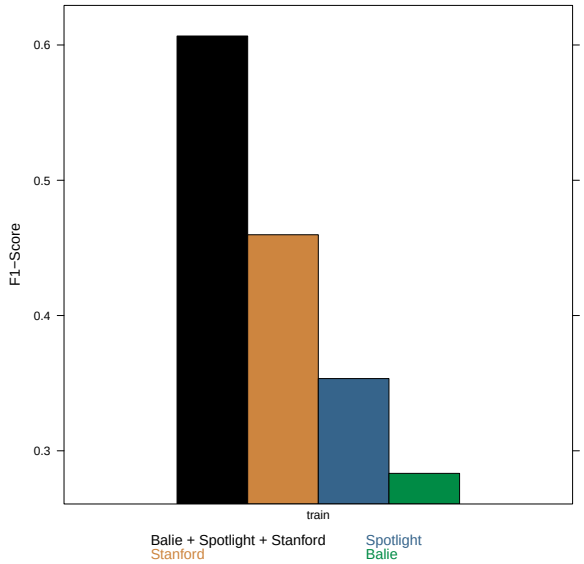

Figure 6: $F1\text{-}Score_T$ **on the *train DE* dataset.**

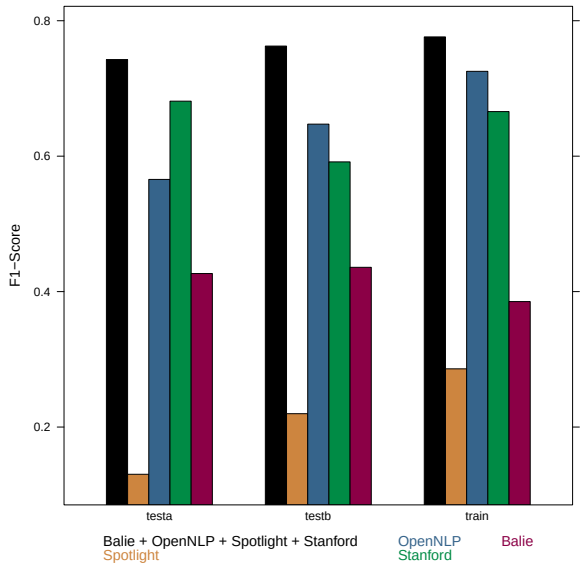

Figure 7: $F1\text{-}Score_T$ **on the *testa ES*, *testb ES* and *train ES* datasets.**

74.26%, 76.26% and 77.61% $F1\text{-}Score_T$. Here, the *Stanford* system performs best as single base classifier on the *testa* dataset (68.12%), but *OpenNLP* on the *testb* (64.73%) and *train* dataset (72.53%). On this datasets we increased the performance of $F1\text{-}Score_T$ by +9.014%, +17.81%, +7.004%. *Spotlight* performs worst, most likely founded in the mismatch of the domain of the datasets and the tools domain.

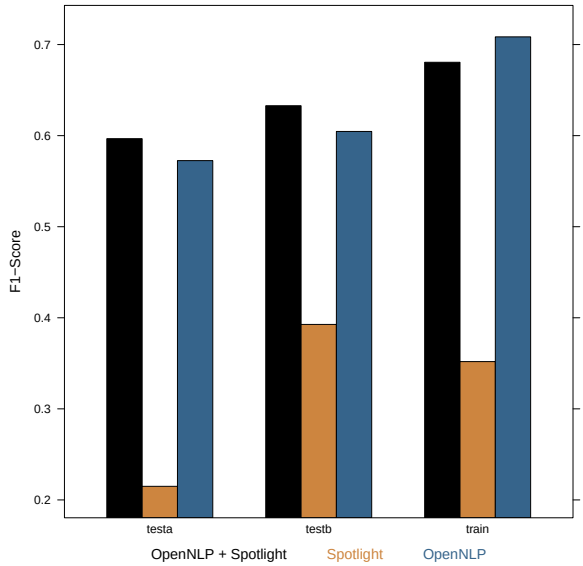

Figure 8: $F1\text{-}Score_T$ **on the *testa NL*, *testb NL* and *train NL* datasets.**

Table 2: Averaged $F1\text{-}Score_T$ **and averaged** pre_T **in percentage.**

dataset	*Balie*	*OpenNLP*	*Spotlight*	*Stanford*	*FOX*
testa ES	42.67/61.00	56.57/70.34	13.03/17.54	68.12/65.20	**74.26**/**74.70**
testb ES	43.59/65.09	64.73/73.17	21.97/50.04	59.16/68.98	**76.26**/**76.53**
train ES	38.53/58.66	72.53/72.65	28.60/28.30	66.58/63.96	**77.61**/**77.35**
testa NL	-	57.26/79.09	21.49/66.24	-	**59.67**/**82.02**
testb NL	-	60.46/**77.75**	39.27/71.92	-	**63.28**/71.57
train NL	-	**70.85**/74.11	35.19/64.45	-	68.06/**79.28**
train DE	28.33/37.70	-	35.34/76.00	45.97/53.69	**60.66**/**78.22**

The datasets consist of news wire articles but the *Spotlight* system is build on Wikipedia.

Figure 8 shows the results on the *testa NL*, *testb NL* and *train NL* datasets for Dutch. We observe an increased performance on two datasets (*testa NL* and *testb NL*) with the combination of all base classifiers. On these datasets we reach 59.57% and 63.28% $F1\text{-}Score_T$. On the *train NL* the ensemble learning approach reaches just 68.06% but *OpenNLP* reaches a slightly better performance with 70.85% $F1\text{-}Score_T$, which reduces the performance by -4.10%. On the other side we observe an increased pre_T by +6.97% from 74.11% to 79.28% on this dataset. The reason for this rogue result can be founded in the low number of base classifiers. In our evaluation, just two NER base classifiers support the Dutch language. *Spotlight* performs poor on this datasets for the same reason as on the Spanish gold standard datasets.

Overall, our approach improved the performance of the $F1\text{-}Score_T$ measure on six out of seven datasets compared to the base classifiers in our evaluation pipeline on the gold standard datasets. Moreover, our approach improved the performance of the pre_t measure on

also six gold standard datasets. On five datasets, the ensemble learning for NER performs better on the $F1\text{-}Score_T$ as well as on the pre_t measure.

5 CONCLUSION AND FUTURE WORK

5.1 Conclusion

In this paper, we presented an ensemble learning approach for multilingual named entity recognition for improving the performance on the named entity recognition task. We presented the underlying pipeline with its components including the setup and datasets. We evaluated the ensemble learning approach for multilingual named entity recognition. We showed empirically that the ensemble learning approach on the named entity recognition task improves at least one of our key performance indicators on all datasets. Thus we reached our goal and answer our first question in this work with yes.

The results on the different dataset sizes suggest that the dataset size influences barely or by no means the ensemble learning approach in our evaluation. This answers our second question.

In the evaluation process we carried out the experiments on all possible combinations of the named entity recognition base classifier. Except for one case, the combination of all base classifiers reached the highest performance on all datasets. The one exception is most likely reasoned in the lack of the numbers of base classifiers for the ensemble learning algorithm but finding the reason to this is out of the scope of this paper and possible future work. This answers our last research question. On nearly all datasets except for one reached the combination of all named entity recognition base classifiers the highest performance. We suggest that this combination works best for the task at hand.

We have integrated the results of this evaluation into the *FOX* framework[9], which is open source, freely available and ready to use via a RESTful web service by the community. Thus, we push forward the version of the multilingual Web of Data with a multilingual state of the art system. Moreover, *FOX* provides the results in the NIF[10] and enriches the results with provenance information by using the PROV-O ontology[11] as well as it links the results with the integrated NED tool *Agdistis* to the DBpedia knowledge base. We extended the framework with the new versions of *Agdistis* to support a better entity linking.

5.2 Future Work

In the near future, we plan to integrating more state of the art NER tools in the frameworks pipeline with the aim to improve the performance particularly for languages with just a few tools integrated in the current state, e.g. Dutch and for languages that are currently missing in the pipeline, e.g. Italian.

We also plan to extend Gerbil with our components and datasets to benchmark our system and allow the community to reproduce the evaluation results.

More future work on the framework includes benchmarking of the ensemble learning approach compared to the base named entity recognition systems included in the pipeline in terms of runtime. For the benchmarking we suggest to use the benchmarking platform *Gerbil* to make the evaluation results machineprocessable, easier reusable and reproducible.

Another research direction is the study of the impact of other features included in the ensemble learning process, e.g., features like POS-tags or a specific number of predecessor and successor words. Which of these features improve the performance best.

Further research questions rise up by configuring the multilayer perceptron. Questions about, how to choose the optimal hidden layer size, the best momentum or the optimal learning rate. Future work to answer this questions could include an improved performance on the task.

[9] http://fox.aksw.org
[10] http://persistence.uni-leipzig.org/nlp2rdf/
[11] https://www.w3.org/TR/prov-o/

REFERENCES

[1] Erin L. Allwein, Robert E. Schapire, and Yoram Singer. 2001. Reducing Multiclass to Binary: A Unifying Approach for Margin Classifiers. *J. Mach. Learn. Res.* 1 (Sept. 2001), 113–141.

[2] R. Amsler. 1989. Research Towards the Development of a Lexical Knowledge Base for Natural Language Processing. *SIGIR Forum* 23 (1989), 1–2.

[3] Jason Baldridge. 2005. The OpenNLP project. *URL: http://opennlp.apache.org/index.html, (accessed 17 May 2017)* (2005).

[4] Darina Benikova, Seid Muhie, Yimam Prabhakaran, and Santhanam Chris Biemann. 2015. C.: GermaNER: Free Open German Named Entity Recognition Tool. In *In: Proc. GSCL-2015.*

[5] Lorenz Bühmann, Ricardo Usbeck, and Axel-Cyrille Ngonga Ngomo. 2015. ASSESS — Automatic Self-Assessment Using Linked Data. In *International Semantic Web Conference (ISWC).*

[6] Sam Coates-Stephens. 1992. The Analysis and Acquisition of Proper Names for the Understanding of Free Text. *Computers and the Humanities* 26 (1992), 441–456. Issue 5. 10.1007/BF00136985.

[7] Marco Cornolti, Paolo Ferragina, and Massimiliano Ciaramita. 2013. A framework for benchmarking entity-annotation systems. In *Proceedings of the 22nd international conference on World Wide Web.* International World Wide Web Conferences Steering Committee, 249–260.

[8] James R. Curran and Stephen Clark. 2003. Language independent NER using a maximum entropy tagger. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 - Volume 4.* 164–167.

[9] Joachim Daiber, Max Jakob, Chris Hokamp, and Pablo N. Mendes. 2013. Improving Efficiency and Accuracy in Multilingual Entity Extraction. In *Proceedings of the 9th International Conference on Semantic Systems (I-Semantics).*

[10] Janez Demšar. 2006. Statistical Comparisons of Classifiers over Multiple Data Sets. *J. Mach. Learn. Res.* 7 (Dec. 2006), 1–30.

[11] Maud Ehrmann, Guillaume Jacquet, and Ralf Steinberger. 2017. JRC-Names: Multilingual entity name variants and titles as Linked Data. *Semantic Web* 8, 2 (2017), 283–295.

[12] Oren Etzioni, Michael Cafarella, Doug Downey, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S. Weld, and Alexander Yates. 2005. Unsupervised named-entity extraction from the web: an experimental study. *Artif. Intell.* 165 (June 2005), 91–134. Issue 1.

[13] Manaal Faruqui and Sebastian Padó. 2010. Training and Evaluating a German Named Entity Recognizer with Semantic Generalization. In *Proceedings of KONVENS 2010.* Saarbrücken, Germany.

[14] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics.* Association for Computational Linguistics, 363–370.

[15] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by Gibbs sampling. In *ACL.* 363–370.

[16] Daniel Gerber and Axel-Cyrille Ngonga Ngomo. 2014. From RDF to Natural Language and Back. In *Towards the Multilingual Semantic Web.* Springer.

[17] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA Data Mining Software: An Update. *SIGKDD Explor. Newsl.* 11, 1 (Nov. 2009), 10–18.

[18] Ali Khalili, Sören Auer, and Axel-Cyrille Ngonga Ngomo. 2014. conTEXT – Lightweight Text Analytics using Linked Data. In *Extended Semantic Web Conference (ESWC 2014).*

[19] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations.* 55–60.

[20] Diego Moussallem, Ricardo Usbeck, Michael Röder, and Axel-Cyrille Ngonga Ngomo. 2017. MAG: A Multilingual, Knowledge-based Agnostic and Deterministic Entity Linking Approach. (2017). arXiv:1707.05288

[21] David Nadeau. 2005. *Balie—baseline information extraction: Multilingual information extraction from text with machine learning and natural language techniques.* Technical Report. Technical report, University of Ottawa.

[22] David Nadeau. 2007. *Semi-supervised Named Entity Recognition: Learning to Recognize 100 Entity Types with Little Supervision.* Ph.D. Dissertation. Ottawa, Ont., Canada, Canada. AAINR49385.

[23] David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Linguisticae Investigationes* 30, 1 (January 2007), 3–26. Publisher: John Benjamins Publishing Company.

[24] David Nadeau, Peter Turney, and Stan Matwin. 2006. Unsupervised Named-Entity Recognition: Generating Gazetteers and Resolving Ambiguity. *Advances in Artificial Intelligence*, 266–277.

[25] Axel-Cyrille Ngonga Ngomo, Norman Heino, Klaus Lyko, René Speck, and Martin Kaltenböck. 2011. SCMS - Semantifying Content Management Systems. In *ISWC 2011*.

[26] Joel Nothman, Nicky Ringland, Will Radford, Tara Murphy, and James R. Curran. 2012. Learning multilingual named entity recognition from Wikipedia. *Artificial Intelligence* 194 (2012), 151–175.

[27] Marius Pasca, Dekang Lin, Jeffrey Bigham, Andrei Lifchits, and Alpa Jain. 2006. Organizing and searching the world wide web of facts - step one: the one-million fact extraction challenge. In *proceedings of the 21st national conference on Artificial intelligence - Volume 2*. AAAI Press, 1400–1405.

[28] R Development Core Team. 2008. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

[29] L. Ratinov and D. Roth. 2009. Design Challenges and Misconceptions in Named Entity Recognition. In *CoNLL*.

[30] Giuseppe Rizzo and Raphaël Troncy. 2012. NERD: A Framework for Unifying Named Entity Recognition and Disambiguation Extraction Tools. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL '12)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 73–76.

[31] Giuseppe Rizzo, Raphaël Troncy, Sebastian Hellmann, and Martin Brümmer. 2012. NERD meets NIF: Lifting NLP extraction results to the linked data cloud. In *LDOW, 5th Workshop on Linked Data on the Web, April 16, 2012, Lyon, France*. Lyon, FRANCE.

[32] Michael Röder, Ricardo Usbeck, René Speck, and Axel-Cyrille Ngonga Ngomo. 2015. CETUS – A Baseline Approach to Type Extraction. In *1st Open Knowledge Extraction Challenge @ 12th European Semantic Web Conference (ESWC 2015)*.

[33] G. Sampson. 1989. How Fully Does a Machine-usable Dictionary Cover English Text. *Literary and Linguistic Computing* 4, 1 (1989).

[34] Mohamed Ahmed Sherif, Axel-Cyrille Ngonga Ngomo, and Jens Lehmann. 2015. Automating RDF Dataset Transformation and Enrichment. In *12th Extended Semantic Web Conference, Portorož, Slovenia, 31st May - 4th June 2015*. Springer.

[35] René Speck and Axel-Cyrille Ngonga Ngomo. 2014. Ensemble Learning for Named Entity Recognition. In *The Semantic Web – ISWC 2014*. Lecture Notes in Computer Science, Vol. 8796. Springer International Publishing, 519–534.

[36] René Speck and Axel-Cyrille Ngonga Ngomo. 2014. Named Entity Recognition using FOX. In *International Semantic Web Conference 2014 (ISWC2014), Demos & Posters*.

[37] René Speck, Michael Röder, Sergio Oramas, Luis Espinosa-Anke, and Axel-Cyrille Ngonga Ngomo. 2017. Open Knowledge Extraction Challenge 2017. In *Semantic Web Challenges: Fourth SemWebEval Challenge at ESWC 2017 (Communications in Computer and Information Science)*. Springer International Publishing.

[38] Christine Thielen. 1995. An Approach to Proper Name Tagging for German. In *In Proceedings of the EACL-95 SIGDAT Workshop*.

[39] Ricardo Usbeck, Axel-Cyrille Ngonga Ngomo, Lorenz Bühmann, and Christina Unger. 2015. HAWK - Hybrid Question Answering over Linked Data. In *12th Extended Semantic Web Conference, Portorož, Slovenia, 31st May - 4th June 2015*.

[40] Ricardo Usbeck, Axel-Cyrille Ngonga Ngomo, Michael Röder, Daniel Gerber, SandroAthaide Coelho, Sören Auer, and Andreas Both. 2014. AGDISTIS - Graph-Based Disambiguation of Named Entities Using Linked Data. In *The Semantic Web – ISWC 2014*, Peter Mika, Tania Tudorache, Abraham Bernstein, Chris Welty, Craig Knoblock, Denny Vrandečić, Paul Groth, Natasha Noy, Krzysztof Janowicz, and Carole Goble (Eds.). Lecture Notes in Computer Science, Vol. 8796. Springer International Publishing, 457–471.

[41] Ricardo Usbeck, Michael Röder, Peter Haase, Artem Kozlov, Muhammad Saleem, and Axel-Cyrille Ngonga Ngomo. 2016. Requirements to Modern Semantic Search Engines. In *KESW*.

[42] Ricardo Usbeck, Michael Röder, Axel-Cyrille Ngonga Ngomo, Ciro Baron, Andreas Both, Martin Brümmer, Diego Ceccarelli, Marco Cornolti, Didier Cherix, Bernd Eickmann, Paolo Ferragina, Christiane Lemke, Andrea Moro, Roberto Navigli, Francesco Piccinno, Giuseppe Rizzo, Harald Sack, René Speck, Raphaël Troncy, Jörg Waitelonis, and Lars Wesemann. 2015. GERBIL – General Entity Annotation Benchmark Framework. In *24th WWW conference*.

[43] Marieke Van Erp, Giuseppe Rizzo, and Raphaël Troncy. 2013. Learning with the Web: Spotting named entities on the intersection of NERD and machine learning. In *WWW 2013, 3rd International Workshop on Making Sense of Microposts (#MSM'13), Concept Extraction Challenge, May 13, 2013, Rio de Janeiro, Brazil*.

[44] D. Walker and R. Amsler. 1986. The Use of Machine-readable Dictionaries in Sublanguage Analysis. *Analysing Language in Restricted Domains* (1986).

[45] Dekai Wu, Grace Ngai, and Marine Carpuat. 2003. A Stacked, Voted, Stacked Model for Named Entity Recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4 (CONLL '03)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 200–203.

[46] GuoDong Zhou and Jian Su. 2002. Named entity recognition using an HMM-based chunk tagger. In *Proceedings of ACL*. 473–480.