Ensemble Learning of Named Entity Recognition Algorithms using Multilayer 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. We present the multilingual version of *FOX*, a knowledge extraction suite which supports this migration by providing named entity recognition based on ensemble learning for five languages. Our evaluation results show that our approach goes beyond the performance of existing named entity recognition systems on all five languages. In our best run, we outperform the state of the art by a gain of 32.38% F1-Score points on a Dutch dataset. More information and a demo can be found at http://fox.aksw.org as well as an extended version of the paper[1] descriping the evaluation in detail.

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

The recognition of named entities (Named Entity Recognition, short NER) in natural language texts plays a central role in knowledge extraction, i.e., the extraction of facts from texts in natural language. A plethora of approaches and frameworks (see, i.a., [7, 9, 11, 12, 17, 27, 33]) have hence been devised to address this task. The knowledge extraction suite *FOX* [31] integrates NER systems as well as named entity disambiguation approaches (NED). It is already an integral part of several applications [4, 14, 16, 18, 23, 28, 30, 35–37] and its demo service[2] receives more than 1 million calls per month from organizations around the world.

While the ensemble learning approach behind *FOX* has already been shown to work well for English [31], this approach was not deployed on other languages so far. In this paper, we present and evaluate the new version of the *FOX* application, which uses ensemble learning on five languages and outperforms the state of the art on the NER task.

The rest of this paper is structured as follows: We begin with a brief overview of the state of the art in NER and in the combination of NER systems. Then, in Section 3, we give a short overview of *FOX*'s inner workings. 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.

[1] http://github.com/dice-group/FOX/tree/master/evaluation/fox_long.pdf
[2] http://fox-demo.aksw.org

2 RELATED WORK

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 [20]. 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 [1, 38]. Following work then showed that these approaches did not perform well for NER tasks such as recognizing proper names [29]. Thus, rule-based approaches were introduced. These approaches rely on hand-crafted rules [5, 34] 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 [21]. When training examples are available, the methods of choice are borrowed from supervised machine-learning. Approaches such as Hidden Markov Models [39], Maximum Entropy Models [6] and Conditional Random Fields [13] have been applied to the NER task. Due to scarcity of large training corpora as necessitated by supervised machine-learning approaches, the semi-supervised [20, 25] and unsupervised machine-learning paradigms [10, 22] have also been used for extracting NER from text. [20] gives an exhaustive overview of approaches for the task.

This paper extends previous works ([23, 31, 32]) mainly by introducing a broadened language support and by performing a thorough evaluation of the extensions on multilingual datasets. Thus, the work in [31] (an ensemble learning approach for NER in English) is the closest related work to this paper.

3 OVERVIEW

FOX is an ensemble learning-based NER framework. For a given language, the framework integrates NER tools as follows: Given any input text t, *FOX* first forwards t to each of the n tools it integrates. The result of each tool T_i is a piece of annotated text t_i, in which each token is assigned either a particular class or a zero class (not part of the label of a named entity). Each token in t is then represented as a vector of length n which contains the classification assigned to it by each tool. This classification is forwarded to the multilayer perceptron (MLP), whose input layer contains one neuron for each possible combination of tool and class. The output layer of the network in the MLP contains exactly as many

classes as recognized by *FOX*. The trained neural network returns a classification for each token of t, which is the final classification assigned by *FOX* for the said token. In a final step, sequences of token which belong to the same class are merged to a single entity.

3.1 Evaluation Setup

We evaluated our approach using existing tools (which were not retrained) and a 10-fold cross validation to train our MLP. The MLP was implemented using the Weka library [15] and uses default options. We rely on the macro average F-measure to determine the performance of tools over the different entity types. Our evaluation was *token-based* like in [31], i.e., 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. We tested the significance of our experimental results using the Wilcoxon signed rank test [8] implemented in R [26]. We set the tests confidence interval to 95%.

We integrated five base classifiers (*Stanford* [11, 12, 17][3], *Illinois* [27][4], *OpenNLP* [2][5], *Balie* [19][6] and *Spotlight* [7][7]). 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 in the result tables in section 4. Throughout our experiments, 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.

We used the silver standard datasets *WikiDE*, *WikiEN*, *WikiES*, *WikiFR* and *WikiNL* provided by [24], which presents a multilingual state of the art semi-supervised learning approach that provides a multilingual annotated corpora by exploiting the text and structure of Wikipedia. In addition, we used the datasets *testa ES*, *testb ES*, *train ES* (Spanish) and *testa NL*, *testb NL*, *train NL* (Dutch), which are gold standard data sets from the CoNLL-2002 shared task[8]. 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 reused the datasets in our evaluation without the entity type *B-MISC* and *I-MISC*, as we aimed to classify persons, organizations and locations. For the German dataset *train DE*, we reused the full training dataset in [3]. The dataset is based on the GermEval 2014[9] dataset. We reuse this dataset without the entity type *B-OTH* and *I-OTH*.

4 RESULTS

The results of the 10-fold validation in terms of average F-measure ($F1\text{-}Score_T$) as well as the average precision (pre_T) achieved by the base classifiers and our approach over the 10 dataset sizes for each language for the silver standard datasets are shown in Table 1. An overview of the values of the $F1\text{-}Score_T$ and pre_T on the gold standard datasets are given in Table 2.

[3] http://nlp.stanford.edu:8080/ner/process
[4] http://cogcomp.cs.illinois.edu/page/demo_view/ner
[5] http://opennlp.apache.org/download.html
[6] http://balie.sourceforge.net
[7] http://spotlight.sztaki.hu/downloads
[8] http://www.cnts.ua.ac.be/conll2002/ner
[9] http://sites.google.com/site/germeval2014ner

4.1 Silver Standards Datasets

The highest value achieved by a classifier on each of the languages is marked in each row in Table 1. In our experiments, the combination of all base classifiers within the ensemble learning reaches the highest value on all languages.

The highest increase of the average $F1\text{-}Score_T$ is achieved on the *WikiNL* dataset for Dutch with an increase of +32.38% over the best base classifier (*OpenNLP* on this dataset). Datasets and increases for other languages include *WikiES* for Spanish with +29.45% increase (*Stanford* is 2nd-best), *WikiFR* for French +22.95% (vs. *Spotlight*), *WikiEN* for English with +3.01% increase (*Stanford* is 2nd-best) and +3.28% increase on *WikiDE* for German compared with *Stanford*. It is noteworthy that the *Stanford* tool performs significantly better on its supported languages (German, English, Spanish) than all other integrated base classifiers. For Dutch and French the best base classifier is *OpenNLP*. Our approach performs best on the English dataset (79.01% F1-score) and worst on the German (63.00%). Furthermore, our experiments reveal that training with 500 sentences suffices to train our MLP to achieve the F1-scores aforementioned.

Overall, *FOX* is significantly better (F1-score, Wilcoxon test, 95% confidence) than the single NER base classifiers on each of the five languages we evaluated against. In two cases (English and German), the averaged pre_T with the ensemble learning approach reaches the highest value as well.

4.2 Gold Standard Datasets

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.

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% $F1\text{-}Score_T$.

We also observe an increased performance on Spanish over all three datasets with the combination of all base classifiers and reach 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 in this scenario.

The increased performance is also seen in Dutch. Here, 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

Table 1: Averaged $F1\text{-}Score_T$ and averaged pre_T achieved on silver standards (in percentage).

dataset	*Balie*	*Illinois*	*OpenNLP*	*Spotlight*	*Stanford*	*FOX*
	$F1\text{-}Score_T/pre_T$	$F1\text{-}Score_T/pre_T$	$F1\text{-}Score_T/pre_T$	$F1\text{-}Score_T/pre_T$	$F1\text{-}Score_T/pre_T$	$F1\text{-}Score_T/pre_T$
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

Table 2: Averaged $F1\text{-}Score_T$ and averaged pre_T achieved on gold standard (in percentage).

dataset	*Balie*	*OpenNLP*	*Spotlight*	*Stanford*	*FOX*
	$F1\text{-}Score_T/pre_T$	$F1\text{-}Score_T/pre_T$	$F1\text{-}Score_T/pre_T$	$F1\text{-}Score_T/pre_T$	$F1\text{-}Score_T/pre_T$
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**

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

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 and showed empirically that the ensemble learning approach with a multilayer perceptron on the named entity recognition task improves the performance on nearly all datasets used in this paper, except for one case. 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. The results on the different dataset sizes reveal that training with 500 sentences suffices to train our MLP to achieve the F1-scores aforementioned.

In the evaluation process we carried out the experiments on all possible combinations of the named entity recognition base classifier. The combination of all base classifiers reached the highest performance on all datasets. We suggest that this combination works best for the task at hand.

We have integrated the results of this evaluation into the *FOX* framework[10], 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 NIF[11] and enriches the results with provenance information by using the PROV-O ontology[12] 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 version of *Agdistis* to support a better entity linking.

In the near future, we plan to integrate more 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 version, e.g. Dutch and for languages that are currently missing in the pipeline, e.g. Italian.

6 ACKNOWLEDGEMENT

This work has been supported by the H2020 project HOBBIT (no. 688227), the BMWI projects GEISER (no. 01MD16014E) and OPAL (no. 19F2028A), the EuroStars projects DIESEL (no. 01QE1512C) and QAMEL (no. 01QE1549C).

REFERENCES

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

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

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

[3] 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*.

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

[5] 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.

[6] 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.

[7] 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)*.

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

[9] 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.

[10] 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.

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

[12] 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.

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

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

[15] 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.

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

[17] 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.

[18] Diego Moussallem, Ricardo Usbeck, Michael Röder, and Axel-Cyrille Ngonga Ngomo. 2017. MAG: A Multilingual, Knowledge-base Agnostic and Deterministic Entity Linking Approach. In *K-CAP 2017: Knowledge Capture Conference*. ACM, 8.

[19] 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.

[20] 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.

[21] 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.

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

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

[24] 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.

[25] 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.

[26] 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.

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

[28] 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)*.

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

[30] 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.

[31] 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.

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

[33] 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.

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

[35] 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*.

[36] 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.

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

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

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