Out-of-context Fine-grained Multi-word Entity Classification

Exploring token, character n-gram and NN-based models for Multilingual Entity Classification

Guillaume Jacquet
European Commission
Joint Research Centre
Ispra, Italy
guillaume.jacquet@ec.europa.eu

Jakub Piskorski
European Commission
Joint Research Centre
Ispra, Italy
jakub.piskorski@ec.europa.eu

Sophie Chesney
Queen Mary University of London
London, UK
s.chesney@qmul.ac.uk

ABSTRACT

In this paper, we present a number of experiments on the construction of fine-grained and out-of-context multi-word entity classification models. These models exploit a large BabelNet-derived multilingual Named Entity corpus of 49 languages from 7 different scripts, which is also presented in this work. In particular, we compare SVM-based character and token *n*-gram models with neural network-based ones and also explore language-specific variants against multilingual models. The various models have been evaluated on additional external Named Entity resources to gain further insight into the quality and re-usability of the trained models.

The language-independent character *n*-gram SVM-based model outperforms the corresponding token *n*-gram SVM-based model for a large majority of tested languages and obtained a 95.7% average precision. When applied to a number of external resources, we did see a slight drop in performance, but still achieved an average precision of 84.7% in all experiments, demonstrating the applicability of the proposed model in a range of contexts. Finally, the experiments applying a neural network model show comparable results for language-specific and language-independent approaches.

CCS CONCEPTS

• **Computing methodologies** → **Information extraction**; **Neural networks**; *Language resources*; *Classification and regression trees*; • **Information systems** → *Clustering and classification*;

KEYWORDS

named-entity classification, multi-word entity, highly multilingual, character n-grams, SVM, Neural Network

ACM Reference format:
Guillaume Jacquet, Jakub Piskorski, and Sophie Chesney. 2019. Out-of-context Fine-grained Multi-word Entity Classification. In *Proceedings of The 34th ACM/SIGAPP Symposium on Applied Computing, Limassol, Cyprus, April 8–12, 2019 (SAC '19),* 10 pages.
https://doi.org/10.1145/3297280.3297379

1 INTRODUCTION

Named Entities (NE) constitute a significant part of natural language text found on the web, and are important bearers of information. Their detection is of paramount importance in the development of many downstream natural language processing (NLP) applications, such as Information Extraction [4, 38], Knowledge Base Construction [15], Question Answering [29] and Machine Translation [27]. The Named Entity Recognition and Classification (NERC) task addresses the problem of the identification (detection) and classification of predefined types of named entities [32], such as organizations (e.g., '*European Medicines Agency*'), persons (e.g., '*Silvio Berlusconi*'), toponyms (e.g., '*Red Sea*'), products (e.g., '*Subaru Forester XT*') and temporal and numerical expressions, etc.

The majority of the early research in the context of NERC focused on: (a) the development of language-specific solutions, usually only covering major languages [32], (b) the exploitation of rule-based [20, 22] and supervised machine learning (ML) [7, 31] methods or combinations thereof, and (c) the recognition of limited and coarse-grained entity types only [10, 14, 48]. Recently, NERC research, in particular in the context of processing texts on the web, has focused on: (a) fine-grained entity recognition [30], (b) the development of scalable multilingual solutions [2], and (c) the exploitation of large-scale encyclopedic and semantic resources for distant supervision [35].

Our work contributes to the main trends sketched out above. In particular, we exploit BabelNet [33], a large-scale multilingual encyclopedic dictionary and semantic network for constructing fine-grained and out-of-context multi-word multilingual entity classification models. Contrary to recent research on fine-grained named entity classification [30, 43] that heavily exploit the contextual information with which the entities appear, our experiments focus on learning models that exclusively exploit internal NE features. Various scenarios exist in which contextual information for NEs is either: (a) not available at all, e.g., in the case of historically accumulated names from the web with no link to the original documents, (b) scarce and very limited, e.g., in the case of short social media messages and conversations, (c) not straightforward to exploit, e.g., in case of classifying potential names in html metadata, web pages containing mostly tables [50], query logs [23], etc. Therefore, we believe, exploring methods for classification of names that rely solely on internal NE features is well motivated in the context of the analysis of web documents of various kinds.

For the reasons outlined above, our experiments focused solely on entity classification in an out-of-context environment. As a consequence, the number of accessible features is limited, the comparison with other classical NE classification system is difficult and the

SAC '19, April 8–12, 2019, Limassol, Cyprus

ACM ISBN 978-1-4503-5933-7/19/04.
https://doi.org/10.1145/3297280.3297379

evaluation of classification of ambiguous entities is harder. We will show in the following sections how we address these challenges.

The work presented in this paper builds on top of the work presented in [9], which reports on developing an approach for the classification of multi-word NEs in 43 languages based only on the tokens they contain.It was shown that a single language-independent SVM classifier is successful in classifying multi-word NEs into 13 types, suggesting that the small languages in the corpus benefit from the concatenated training data. The aforementioned research constitutes our starting point to explore whether a character *n*-gram model can maximise cross-lingual overlap already seen in [9]'s work in a similar out-of-context setting, without using any external linguistic features (e.g. WordNet or word embeddings). However, contrarily to [9]'s work which was excluding ambiguous entities inside each language, we consider and evaluate them. As a matter of fact, we formulate the task at hand as multi-label name entity classification.

Furthermore, in this work, we conducted a number of experiments comparing the character-level *n*-gram based models with neural network-based ones, as well as language-specific models against multilingual ones. All experiments were carried out on an extended version of the BabelNet-derived NE resource described in [9] . In particular, the tagset has been extended with person names and 6 additional languages with non-latin scripts have been added (Chinese, Japanese, Hindi, Farsi, Hebrew and Bengali). In addition, in order to gain insight into the quality and usability of the trained models on unseen data, we evaluated them on a series of external resources disjoint from the BabelNet-derived NE resource.

The main contributions of our work is summarised as follows:

- we explore how character-level multilingual *n*-gram and neural network-based models perform for out-of-context NE classification of multi-word NEs in comparison with the token-based multilingual model reported in [9],
- we provide results of evaluation on external resources to gain better insight into the re-usability of the presented approaches, and
- we create a large-scale multilingual NE dataset (49 languages) with 14 fine-grained NE types and make it available to researchers as a silver standard

The rest of the paper is structured as follows. First, Section 2 provides an overview of related work. Next, Section 3 describes the BabelNet-derived NE resource used for training, and the various external resources used for evaluation purposes. Subsequently, Section 4 presents the various classification models explored in our work, while Section 5 provides evaluation results. Finally, Section 6 concludes the work and gives an outlook on future work.

2 RELATED WORK

An overview of historical work on NERC is given in [32]. A vast proportion of recent work in the context of NERC focuses on exploiting various supervised and unsupervised ML techniques [12, 18, 39, 40]. However, most of the work in the area of NERC focuses on systems that only classify entities into a few coarse-grained categories, i.e., persons, organisations, locations and miscellaneous. Recent work in this area has moved from coarse types towards more fine-grained classifications of semantic subtypes.

The idea of more fine-grained NE types was first introduced in [41] and [28], both proposing more than 100 fine-grained NE categories. Some other prior work also reports on concrete systems and approaches that identify fine-grained entity types, e.g. [21] proposed an unsupervised approach based on lexical entailment to annotate persons and locations with 21 fined-grained types.

More recently, [30] proposed a system for Fine-grained Entity Classification (FETC), which used overlapping 112 NE tags and exploits linear classifier perceptron multi-class NE classification. [53] deployed SVM-based classifiers to tag NEs using a set of 505 labels. [43] present an attentive neural model for the classification entities into 112 types, achieving state-of-the-art performance on the FIGER (GOLD) dataset introduced in [30]. [25] address hierarchical class structure in the fine-grained entity typing task, focusing on the issue of type ambiguity, while [3] utilise context-dependent training data to jointly learn entity mentions and their context in order to eliminate the use of hand-crafted features. [52] show that fine-grained NE classification can be improved through joint multi-level representations of entities on three complementary levels, namely, character, token and entity, while [51] reports on the benefits of jointly tackling the task of fine-grained NE tagging and relation extraction.

While most of the aforementioned work heavily relies on contextual information for fine-grained NE tagging and evaluates the various approaches and systems on monolingual corpora, our work focuses on exploring methods for fine-grained entity classification using exclusively entity internal features and the development of highly multilingual solutions that do not rely on sophisticated language-specific linguistic analysis.

In particular, we build on the findings of the work presented in [9] reporting on developing an approach for classification of multi-word NEs in 49 languages based only on the token information they contain and through exploitation of BabelNet [33], a highly multilingual language resource. Since gold standard linguistically annotated resources are scarce for the vast majority of languages and are costly to produce, distant supervision has gained lot of attention recently. For instance, [35] classifies Wikipedia articles into NE types, utilising the links between in-text entities and their corresponding Wikipedia pages, developing in this manner a silver-standard annotated corpus for 9 languages for training NERC systems. An overview of various distant supervision-based approaches to Information Extraction (including NERC) that rely on corpora and background knowledge bases such as DBpedia [5], Freebase [8], and YAGO [45] is presented in [6].

3 RESOURCES

3.1 BabelNet-derived NE Resource

For training the entity classification models, we have exploited an extended version of the silver-standard NE resource described in [9] which was semi-automatically extracted from BabelNet (BN) [33]. The resource contains ca. 4.7 million multi-word NEs in 49 languages, divided into 14 fine-grained classes (organisations, locations[1], persons, products and events) that could be considered as a

[1] LOC-FA category encompasses sport, recreation, cultural and other urban and non-urban facilities such as railroads, etc., whereas the category LOC-OT covers mentions of landforms, e.g., mountains, islands, water bodies, etc.

simplified version of NE class hierarchy introduced by Sekine [42]. Quantitative data about the aforementioned resource is provided in Table 1. An entity in the resource consists on average of 2.94 tokens across all languages. The breakdown of resource figures with respect to language specific data is given in Table 2.

ORGANISATION		
Subtype (Encoding)	**Example**	**#entries**
political-public (org-pp)	*Republican Party*	143 002
commercial (org-co)	*Google Inc.*	212 646
sport (org-sp)	*HSV Hamburg*	217 202
educ-research (org-er)	*ET Zurich*	168 343
LOCATION		
Subtype (Encoding)	**Example**	**#entries**
facility (loc-fa)	*Louvre Museum*	557 591
other (loc-ot)	*Lago Maggiore*	276 894
PERSON		
Subtype (Encoding)	**Example**	**#entries**
person (per)	*Umberto Eco*	2 583 531
PRODUCT		
Subtype (Encoding)	**Example**	**#entries**
electronics (pro-el)	*Iphone 6*	16 974
weapon (pro-we)	*US109L shotgun*	14 301
vehicle (pro-ve)	*Honda Pilot*	26 534
art (pro-ar)	*Indiana Jones*	259 098
EVENT		
Subtype (Encoding)	**Example**	**#entries**
incident (evt-in)	*Kennedy Assassination*	113 009
natural (evt-na)	*Haiti 2010 earthquake*	10 636
occasion (evt-oc)	*EMNLP 2017 Conference*	68 723

Table 1: Data on the BN-derived resource.

The aforementioned NE resource was created in the following manner. First, we used the BabelNet API[2] to select therefrom all NE-related synsets. Since the NE-related BabelNet synsets are not tagged with a specific NE tag, the NE type was inferred by using the hypernym information provided in BabelNet (i.e. using WordNet hypernyms and Wikipedia categories). More precisely, based on hypernym frequency information for the entire set of NEs contained in BabelNet, for each NE type a list of hypernyms was manually created[3]. These lists were subsequently used to extract NEs of each particular type for 49 languages. A given NE-related synset was extracted if there was at least one hypernym for the main sense of the synset in the list of hypernyms. For instance, the list of hypernyms for extracting names of commercial organisations (org-co) included terms like *company, bank*. From the set of extracted NEs, we have removed all single-token entities and potentially problematic ones, e.g. entities consisting of two tokens, one of which is a single character. Circa 2.73% of the entries in the dataset are ambiguous, where the observed maximum number of classes assigned to an entry is 4 (e.g. *St George* which has the types loc-fa, loc-ot, pro-ar and per). More than 88% of ambiguous NEs have per as one of the NE types assigned which can be partly explained by the high representation of per class in the BN-derived resource. Table 3 provides information on the most frequent types of ambiguities observed. The main difference between the BN-derived NE resource vis-a-vis the version described in [9] is: (a) introduction of per category, (b) non usage of the 'negative hypernyms' mentioned in [9] for disambiguation purposes and (c) introduction of 6 non-latin languages.

Lg	**#entr.**	**Lg**	**#entr.**	**Lg**	**#entr.**	**Lg**	**#entr.**	**Lg**	**#entr.**
EN	958248	NL	159144	RO	48660	MK	28924	BN	5564
ES	352896	IT	138600	EL	47196	LV	24024	IS	4784
FR	334040	PT	111668	AR	47148	SL	20456	NN	4692
DE	323228	FA	105452	NO	40916	SK	16224	LB	4272
RU	271340	CA	79100	HE	34040	HI	14068	MT	3136
JA	264092	CS	56584	TR	32296	HR	13756	BS	2676
SV	213532	BG	51788	BE	32060	EU	13020	FO	344
ZH	209840	FI	51100	LT	32032	SQ	8380	RM	176
PL	167832	SR	50660	DA	31804	CY	7836	LAD	168
UK	163804	HU	48684	ET	31648	GA	6556		

Table 2: Number of entries per languages in the BN-derived resource.

A qualitative evaluation of a subset of this NE resource (i.e. ca. 1.5 million of the entries, excluding person names) reported in [9] showed that human annotations on randomly selected subsets of 200 entities for 5 languages yielded precision and recall figures ranging from 87.6% to 92.5% and 85.0% and 90.5% respectively.

Ambiguity	**%entr.**	Ambiguity	**%entr.**
per/pro-ar	42.7	evt-in/per	1.7
org-sp/per	35.0	loc-fa/org-co	1.3
org-co/per	5.4	org-co/pro-ve	0.9
evt-oc/per	2.9	org-er/per	0.9
loc-fa/loc-ot	1.9	other	7.3

Table 3: Most prevalent entity type ambiguities.

3.2 External Testing Resources

We completed the evaluation of the entity classification models by testing their performance on existing external NE resources. For this purpose we exploited resources in 13 different languages, gathered from the 'general' or 'news' domains from the publicly available NE resources described in [16][4].

In particular, we selected a set of five manually annotated, freely available corpora in five different languages (English, German, Italian, Spanish, Dutch) from the CoNLL2002[5] and 2003[6] evaluation campaigns. Additionally, we used a Hungarian NE corpus [46][7], and a set of Slavic NE corpora (which will be referred to with BSNLP 2017[8]), containing texts in seven different languages (Croatian,

[2] http://babelnet.org/guide

[3] Each such list contained not more than 100 entries

[4] http://damien.nouvels.net/resourcesen/

[5] http://www.cnts.ua.ac.be/conll2002/ner/

[6] http://www.cnts.ua.ac.be/conll2003/ner/

[7] http://www.inf.u-szeged.hu/rgai/nlp

[8] BSNLP - Balto-Slavic Natural Language Processing

Czech, Polish, Russian, Ukrainian, Slovakian and Slovene) [37][9]. A substantial part of the latter corpus contains inflected variants of named entities, i.e., named-entity mentions being non-base forms constitute from ca. 37% to ca. 58% of all names in a given corpus (depending on the language and topic). For all of these corpora, we selected only multi-word entities, whose statistics are listed in Table 6.

Because each annotated corpus has its own entity type hierarchy, some more fine-grained than others, we harmonised the annotated expressions with three types: PERSON, ORGANISATION and MISC, where MISC should correspond to EVENT and PRODUCT for our classifier. Of course, this is much more coarse-grained than the classifier is capable of, but the fine-grained quality is already measured in the initial evaluation, and the goal of this additional evaluation being instead to test the classifier in a context of distant supervision.

4 CLASSIFICATION METHODS

We compare a re-implementation of the language-dependent and language-independent token-based SVM classifier presented in [9] with two newly implemented classifiers based on character *n*-grams: one SVM classifier and one based on a Neural Network (NN) model.

4.1 Token-based Model: SVM_TOKEN

The best-performing classifier in [9] was a distantly-supervised SVM with a TF.IDF-weighted bag-of-words data representation, re-implemented here. The only features used in the classification task are the tokens from the entities themselves. The vectorisation is implemented with L2 normalisation, in order to normalise for the number of expressions in each class, and sublinear TF calculations (which log-scales the TF counts). Given that NEC is a multi-class classification task, One-Versus-One (OvO) classification was used, where a binary classifier was trained for each pair of classes and the classification for a given entity is the most highly-voted class. The classifiers were implemented using Scikit-learn [36].

4.2 Character *N*-gram Model: SVM_CHAR

In text classification tasks, working beyond the token level is particularly advantageous in a multilingual context: preprocessing tools, such as lemmatisers, are seldom available for under-resourced languages and can therefore limit work on them. Using character *n*-gram models reduces the amount of preprocessing necessary and thus avoids the need for such tools. It has been shown that language-independent text classification can be successfully approached using character *n*-gram models [13, 24].

Our comparative model is also implemented using Scikit-learn and follows the exact implementation of the above model (TF.IDF vectorisation with L2 normalisation and sublinear TF counts), but with character *n*-gram feature extraction in place of the token-based bag-of-words approach used above. We use an *n*-gram range between 3 and 5-grams for all languages except for Chinese (ZH) and Japanese (JA) where the range is between 1 and 3-grams. All other aspects of the implementation are as in [9].

4.3 Neural Network-based model: NN_CHAR

The last chunk of experiments focused on exploiting Neural Networks (NN). We used a multilayer perceptron described as a sequential NN model with three densely-connected NN layers. The first layer is collecting character *n*-gram TF.IDF values as features (TF.IDF vectorisation with L2 normalisation and sublinear TF counts like the two previous models), the second layer contains 256 hidden units, the third layer contains 64 hidden units and the final layer contains 14 hidden units, corresponding to the 14 categories. The model is based on the Adam optimizer [26] with learning rate= 0.001, epsilon = 1e-08 and the learning rate decay over each update = 0. The experiments presented in section 5.4 are based on 200 epochs. The implementation of this model uses the python library Keras [11] which can be run based on both Theano [1] and Tensorflow [47] frameworks.

5 EVALUATIONS

Given that our research focuses on named entity classification without considering the context and the fact that a significant proportion of the entities in the evaluation resource is ambiguous (2.73%) implies that the NE classification task at hand is defined as multi-label classification task.

Since both SVM and NN approaches can return, instead of one type per entity, a ranked set of types with their corresponding confidence score we have adapted three metrics for multi-label classification described in [49], namely, *One-Error*, *Coverage* and *Average Precision*. For the formal description of these metrics, we will denote with $\Lambda = \lambda_1, \dots, \lambda_k$ the finite set of NE types, $E = e_1, \dots, e_n$ set of entities in the corpus, $T(e) \subseteq \Lambda$ a set of true types assigned to e, and $r_e(\lambda)$ the rank of type λ in system response for entity e.

One-Error (OE) measures how many times the top predicted type (i.e. the one the classifier is most confident with) is not in the set of true types of the entity:

$$OE = \frac{1}{|E|} \cdot \sum_{i=1}^{|E|} \delta(\arg\min_{\lambda \in \Lambda} r_{e_i}(\lambda)) \tag{1}$$

where

$$\delta(\lambda) = \begin{cases} 1, & \text{if } \lambda \notin T(e_i) \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

The smaller the value of *OE* the better the performance.

Coverage (C) measures how far on average one needs to go down the ranked list of types returned by the system in order to cover all true types for the given entity, and is formally defined as follows.

$$C = \frac{1}{|E|} \cdot \sum_{i=1}^{|E|} \frac{|T(e_i)|}{\max\limits_{\lambda \in T(e_i)} r_{e_i}(\lambda)} \tag{3}$$

The higher the value of C the better the performance, where $C = 1$ corresponds to 'perfect' classification.

Average Precision (AP) is somewhat similar to the *Coverage* and measures the average fraction of types in the response that are ranked higher than a given label $\lambda \in T(e_i)$ which are also in $T(e_i)$.

[9] http://bsnlp-2017.cs.helsinki.fi/shared_task_results.html

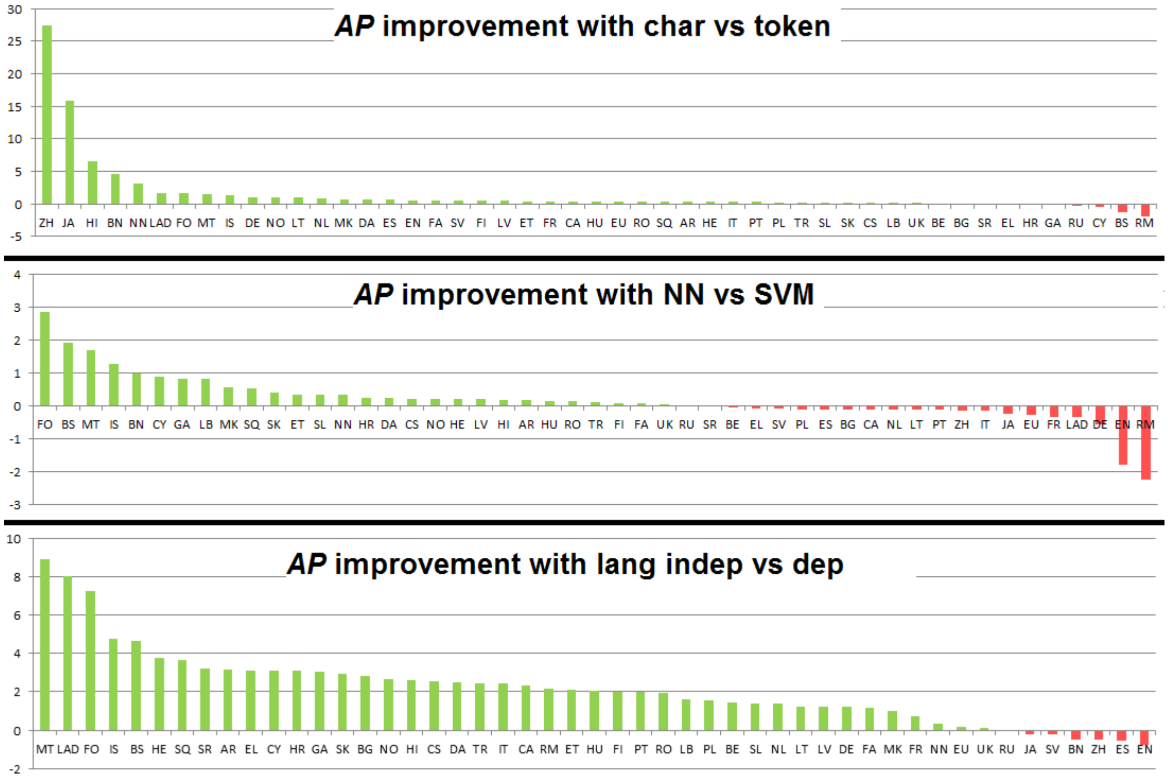


Figure 1: Comparison of the Average Precision per language according to different parameters: token vs char (SVM_dep_token vs SVM_dep_char), language-dependent vs language-independent (SVM_dep_char vs SVM_indep_char) and SVM vs NN (SVM_dep_char vs NN_dep_char). In the three figures, the results are sorted from left to right according to the gain in Average Precision score.

$$AP = \frac{1}{|E|} \cdot \sum_{i=1}^{|E|} \frac{1}{|T(e_i)|} \cdot \sum_{\lambda \in T(e_i)} \frac{|\lambda' \in T(e_i) : r_{e_i}(\lambda') \leq r_{e_i}(\lambda)|}{r_{e_i}(\lambda)} \qquad (4)$$

The higher the value of *AP* the better the performance, where *AP* = 1 corresponds to 'perfect' classification.

The description of the outcome of the evaluations is divided into three parts.

First, in Section 5.2, we compare the performance of the language-independent SVM_TOKEN and SVM_CHAR classifiers, with 10-fold ShuffleSplit cross-validation over the BN-derived resource for 49 languages. Following the methodology presented in [9, p. 17], the language-independent training set is the concatenation of the training set from all languages in the BN-derived resource after it is split for testing and training (25%/75%). Overlaps between the language-independent training set and any language-specific test sets are removed from the language-independent training set.

Subsequently, in Section 5.3, we report on the results of applying the best-performing classifier, i.e., the language-independent SVM_CHAR, to several external resources described in Section 3.2.

Finally, we report in Section 5.4 on the performance of the NN_CHAR language-dependent and independent classifiers.

5.1 Percentage Thresholds

Usually, 'confidence thresholds' control a trade-off between precision and recall. By excluding the least confident classifications, precision can be improved by upwards of 5% when averaged across all languages in the data set, but with a loss in recall [9]. Analgously, we test the different models with *OE*, *C* and *AP* metrics at five percentile thresholds (0%, 5%, 10%, 15% and 20%).

5.2 SVM models

Table 4 shows the performance of all SVM and NN-based classifiers on the BN-derived resource over five percentile thresholds. We can see a constant and steady improvement with the SVM_CHAR over the SVM_TOKEN model for all percentile levels, for both language-dependent and independent configurations, reaching an Averaged Precision (AP) of 97.8% over all 49 languages at the highest exclusion rate. Furthermore, we see that One-Error score is retained clearly better for SVM_CHAR. Over the 49 languages, the AP measure is improved for 39 and worse for 4 (see Figure 1). Chinese (ZH) and Japanese (JA) show the largest improvement of 28 and 16 points respectively above SVM_TOKEN. In [9] we reported on a marked improvement with the introduction of language-independent training data, particularly for the smaller languages in the dataset. For example, a boost from 50% to 76% F1 for Faroese (FO), and similar

Excluded percentile	Lang. dep. OE	Lang. dep. C	Lang. dep. AP	Lang. indep. OE	Lang. indep. C	Lang. indep. AP
SVM_token						
0	11.1%	92.3%	92.5%	8.6%	94.6%	94.7%
5	9.9%	93.1%	93.4%	7.3%	95.4%	95.5%
10	8.4%	94.0%	94.3%	6.7%	95.8%	95.9%
15	7.4%	94.7%	95.0%	6.1%	96.1%	96.2%
20	7.1%	95.0%	95.2%	5.8%	96.3%	96.4%
SVM_char						
0	7.6%	94.9%	95.1%	**7.1%**	**95.5%**	**95.7%**
5	6.1%	95.8%	96.0%	**5.8%**	**96.3%**	**96.5%**
10	4.9%	96.7%	96.8%	5.1%	**96.8%**	**96.9%**
15	3.9%	97.3%	97.4%	4.6%	97.1%	97.2%
20	3.3%	97.7%	97.8%	4.2%	97.4%	97.5%
NN_char						
0	8.0%	94.4%	94.6%	7.9%	94.6%	94.8%
5	6.2%	95.5%	95.8%	5.9%	95.9%	96.1%
10	4.8%	96.4%	96.6%	**4.6%**	96.7%	**96.9%**
15	**3.8%**	**97.4%**	**97.5%**	3.9%	97.2%	97.4%
20	**3.1%**	**97.9%**	**98.0%**	3.3%	97.5%	97.7%

Table 4: Average results for One-Error (OE), Coverage (C) and Average Precision (AP) across the 49 languages tested: Comparison of language-dependent vs independent models token vs character n-gram-based models,and SVM vs NN-based models for the 5 tested percentile thresholds.

languages	SVM_char with Lang. independent OE	C	AP
EN (biggest)	14.3%	90.7%	90.9%
RO (best)	1.3%	99.1%	99.2%
LAD (smallest & worst)	21.3%	86.1%	86.3%
AR (non-latin)	1.95%	98.6%	98.7%
ZH (1-3 char ngram)	7.7%	94.8%	95.0%

Table 5: Results obtained for some 'key' languages, with language-independent SVM_char model.

improvements for Ladino (LAD) and Luxembourgish (LB) (p. 18) were observed. We could observe further improvements in all of these languages with the language-independent SVM_char classifier, as illustrated in Figure 1, i.e., AP measure is improved for 42 languages and became worse for only 6. Figure 1 also shows that one obtains better results for scarcely-resourced languages with the language-independent model. The 7 best improvements are for languages with less than 40K entries, and the two languages with the most significant performance deterioration are the ones with the highest number of entries, English (EN) and Spanish (ES). Finally, this is not directly visible from Figure 1, but through comparing results obtained with language-dependent SVM_token and language-independent SVM_char models, we noticed a strong improvements for the cluster of Scandinavian languages (Faroese (FO), Norwegian Nynorsk (NN), Norwegian (NO), Swedish (SV) and Danish (DA)). This may be due to the fact that the character-level classifier is capturing closely-related cross-over terms that differ marginally orthographically between languages or occur in compounds. For example, the term 'miljö/miljø/miljø' ('environment' in SV/DA/NO) appears as part of compound nouns such as 'danmarks mijløundersøgelser' (DA) or 'miljøpartiet de grønne' (NO). A token-based model will not be able to utilize this linguistic similarity, but the character-level one can.

Figure 2 provides the One-Error confusion matrix for language-independent SVM_char run on the test data for one fold of the 10-fold ShuffleSplit cross-validation with 0% excluded percentile. One can observe that the level of misclassifications within each main entity type is relatively low for events (max. 0.5%), somewhat higher for organisations (max. 2.0%) and products (max. 2.1%). Generally, the most significant errors are misclassifications of sport organisation (ORG-SP) and art products (pro-ar) as person (PER) which corresponds to the most frequent type ambiguities listed in Table 3. pro-ar false positives occur as art product names frequently contain location, organisation and event terms. Furthermore, product names, disregarding their particular type, are frequently misclassified as persons (2.9% to 16.1% depending on product type). This is not surprising and results from the fact that often first names (e.g. *'Mercedes'*) or surnames are often used as part of product names. Finally, vehicles (pro-ve) and electronics (pro-el) are also sometimes misclassified as org-co (e.g. with the entities *'Nike One'*, *'General Motors Corsa'*, *'Dell Netbooks'*).

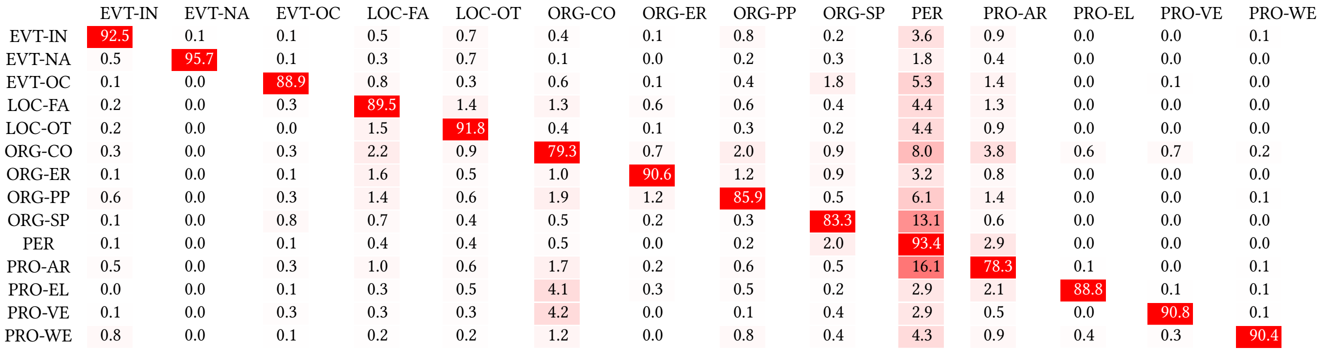
5.3 Results on External Resources

The AP scores achieved by the language-independent SVM_char model for the 7 external resources are shown in Table 6. On average across all resources, we observe an AP of 84.7% across the three categories, per, org and misc. Although this is below the average achieved on the BN-derived resource with the same classifier, we see a strong performance on par with the previous results, particularly for Hungarian and English.

Additionally, across all languages in the external resources tested, we observe high AP scores for the per category, which reflects the performance seen with the same classifier on the BN-derived resource, detailed in 5.2. This further suggests that this classifier's strong performance in the initial experiments is not isolated, since a relatively good performance is achievable in a variety of external settings.

However, it is important to emphasize at this stage that the harmonisation of external resources was difficult since the annotation guidelines varied across corpora, as did the definition of NE types. Hence, we drastically reduced the range of evaluated categories. The permeability between categories like org and loc-fa (facility) between the external data sets and the original BN-derived training data explains in part the comparatively low AP: e.g. museums or airports were often classified as loc-fa, in line with the BN types, but gold-annotated as org in the external resources. Furthermore, the resulting misc category is not especially concise and is therefore difficult to evaluate, at least from automatic metrics alone.

Nevertheless, the experiments on the external resources aim to evaluate the classifier on external data, to demonstrate its applicability in more varied contexts. Despite annotation inconsistency, the performance results summarized in Table 6 show that the classifier achieves particularly high AP scores in the org category for Hungarian, a highly inflected language, and reasonably high scores for those contained within the BSNLP2017 corpus. This is a positive result for a classifier which uses such simple features,



	EVT-IN	EVT-NA	EVT-OC	LOC-FA	LOC-OT	ORG-CO	ORG-ER	ORG-PP	ORG-SP	PER	PRO-AR	PRO-EL	PRO-VE	PRO-WE
EVT-IN	92.5	0.1	0.1	0.5	0.7	0.4	0.1	0.8	0.2	3.6	0.9	0.0	0.0	0.1
EVT-NA	0.5	95.7	0.1	0.3	0.7	0.1	0.0	0.2	0.3	1.8	0.4	0.0	0.0	0.0
EVT-OC	0.1	0.0	88.9	0.8	0.3	0.6	0.1	0.4	1.8	5.3	1.4	0.0	0.1	0.0
LOC-FA	0.2	0.0	0.3	89.5	1.4	1.3	0.6	0.6	0.4	4.4	1.3	0.0	0.0	0.0
LOC-OT	0.2	0.0	0.0	1.5	91.8	0.4	0.1	0.3	0.2	4.4	0.9	0.0	0.0	0.0
ORG-CO	0.3	0.0	0.3	2.2	0.9	79.3	0.7	2.0	0.9	8.0	3.8	0.6	0.7	0.2
ORG-ER	0.1	0.0	0.1	1.6	0.5	1.0	90.6	1.2	0.9	3.2	0.8	0.0	0.0	0.0
ORG-PP	0.6	0.0	0.3	1.4	0.6	1.9	1.2	85.9	0.5	6.1	1.4	0.0	0.0	0.1
ORG-SP	0.1	0.0	0.8	0.7	0.4	0.5	0.2	0.3	83.3	13.1	0.6	0.0	0.0	0.0
PER	0.1	0.0	0.1	0.4	0.4	0.5	0.0	0.2	2.0	93.4	2.9	0.0	0.0	0.0
PRO-AR	0.5	0.0	0.3	1.0	0.6	1.7	0.2	0.6	0.5	16.1	78.3	0.1	0.0	0.1
PRO-EL	0.0	0.0	0.1	0.3	0.5	4.1	0.3	0.5	0.2	2.9	2.1	88.8	0.1	0.1
PRO-VE	0.1	0.0	0.3	0.3	0.3	4.2	0.0	0.1	0.4	2.9	0.5	0.0	90.8	0.1
PRO-WE	0.8	0.0	0.1	0.2	0.2	1.2	0.0	0.8	0.4	4.3	0.9	0.4	0.3	90.4

Figure 2: Confusion matrix for language-independent SVM_CHAR (character *n*-gram) model.

without specifically taking into account any morphological analysis for these highly inflective languages. It should be emphasized too that vast fraction of the NEs in the BSNLP2017 corpus are inflected variants of NEs (ranging from 30 to 55% depending on the language).

Corpora	ORG	PER	MISC	Average	Support
BSNLP2017 (Slavic)	80.4%	93.1%	49.7%	79.4%	1832
CoNLL2002 (Italian)	74.1%	97.4%	42.3%	84.7%	2696
CoNLL2002 (Dutch)	62.4%	93.1%	58.2%	75.1%	2614
CoNLL2002 (Spanish)	75.1%	96.3%	55.0%	80.4%	4271
CoNLL2003 (German)	76.2%	95.2%	47.6%	80.4%	2452
CoNLL2003 (English)	88.9%	99.5%	50.8%	92.1%	4389
HNE corpus (Hungarian)	92.1%	78.3%	19.0%	86.8%	2699
Average	**81.5%**	**96.3%**	**52.9%**	**84.7%**	**20953**

Table 6: Average precision scores obtained on external resources with language-independent SVM_CHAR classifier and percentile threshold = 0.

5.4 Neural Network-based models

NN-based classifier (henceforth NN_CHAR) was evaluated with language-dependent and language-independent training sets using character *n*-gram TF.IDF values as features.

Table 4 shows that on average for the 49 languages, the NN_CHAR classifier performs well, with a 94.6% Average Precision. Furthermore, the NN_CHAR classifier slightly outperforms the SVM_CHAR by 0.1 to 0.2 point of AP when the exclusion percentile thresholds is higher than 10% and actually also outperforms the language-independent SVM_CHAR model. Comparing the One-Error confusion matrices in Figures 3 and 4 gives an additional view of this improvement, even when the percentile threshold is 0%. Namely, it shows that with the SVM_CHAR model, most of the erroneous classifications are related to the PER category, defaulting to assigning the PER category to uncertain entities. This is likely due to the fact that this is the largest category, hence the most probable. In contrast, with the NN_CHAR model, the erroneous classifications seems to be spread more evenly across all categories, reducing the quality of the most probable category but providing better results to almost all the other ones. The high ratio of all products misclassified as ORG-CO noticeable in both Figure 3 and 4 reflects the type confusion already noticed in the case of language-independent SVM_CHAR confusion matrix (cf. Table 2) .

Finally, Table 4 shows the results obtained with the language-independent NN_CHAR model trained on the BN-resource for 49 languages. The performance is comparable to the results obtained with the language-dependent NN_CHAR models, where we were expecting some improvement vis-a-vis the results observed for the SVM_CHAR model. When comparing language-dependent and language-independent NN_CHAR models, we observed that the result improvements varied a lot from one language to another, even when considering only languages with latin script: from 5 point of *AP* improvement for Maltese (MT) to 7 point deterioration for Gaeilge (GA). We also observed a higher sensibility of language scripts with an *AP* deterioration for all the non-latin script languages.

6 CONCLUSIONS AND FUTURE WORK

This paper reported on numerous experiments on out-of-context entity classification models based on character n-gram SVM and neural network architectures. In particular, our work focused on exploring solutions for fine-grained multi-word entity classification without considering any contextual information, and relying solely on entity internal features, whose computation does not require much linguistic sophistication. We created a large NE resource with 14 fined grained NE types that covers 49 languages and which has been derived from BabelNet. This silver-standard resource was used for training classification models and for evaluation using a 10-fold ShuffleSplit cross-validation. The trained models have been evaluated on a pool of additional external resources to get a better insight on their performance on unseen data. The main outcomes of the reported research can be summarised as follows:

- a language-independent out-of-context character *n*-gram SVM approach to multi-word entity classification outperformed a token-vectorised SVM model presented earlier in [9], achieving a 95.7% Average Precision (*AP*) score over 14 categories across all languages, where one could observe an improvement for 39 out of 49 languages for *AP*,
- a particular boost in performance of the character n-gram model vis-a-vis the token n-gram model could be observed in the context of less-resourced languages (e.g. Faroese (FO),

	EVT-IN	EVT-NA	EVT-OC	LOC-FA	LOC-OT	ORG-CO	ORG-ER	ORG-PP	ORG-SP	PER	PRO-AR	PRO-EL	PRO-VE	PRO-WE
EVT-IN	86.6	0.1	0.1	0.9	0.8	0.7	0.1	1.3	0.2	7.4	1.6	0.0	0.1	0.1
EVT-NA	1.8	89.4	0.1	0.7	1.1	0.3	0.1	0.3	0.1	5.2	0.9	0.0	0.0	0.0
EVT-OC	0.2	0.0	78.7	2.1	0.4	1.5	0.2	0.8	2.8	11.1	2.2	0.0	0.1	0.0
LOC-FA	0.1	0.0	0.2	91.0	1.1	1.0	0.5	0.4	0.2	4.7	0.8	0.0	0.0	0.0
LOC-OT	0.3	0.0	0.0	2.4	86.7	0.5	0.1	0.2	0.2	8.7	0.8	0.0	0.0	0.0
ORG-CO	0.3	0.0	0.2	3.2	0.7	71.7	0.9	1.8	0.7	15.8	3.5	0.3	0.7	0.1
ORG-ER	0.0	0.0	0.1	2.2	0.4	1.3	89.4	1.2	0.4	4.1	0.7	0.0	0.0	0.0
ORG-PP	0.7	0.0	0.2	2.4	0.6	2.9	1.7	77.6	0.4	11.6	1.8	0.0	0.0	0.1
ORG-SP	0.1	0.0	1.1	1.0	0.3	0.9	0.4	0.3	65.1	30.3	0.5	0.0	0.0	0.0
PER	0.1	0.0	0.1	0.2	0.1	0.3	0.0	0.1	0.7	97.3	1.1	0.0	0.0	0.0
PRO-AR	0.6	0.0	0.3	1.8	0.8	2.2	0.3	0.6	0.2	34.2	59.0	0.0	0.0	0.0
PRO-EL	0.2	0.0	0.2	1.2	0.7	13.5	0.5	0.9	0.6	17.6	6.0	57.8	0.4	0.4
PRO-VE	0.3	0.0	0.4	0.8	0.2	8.9	0.1	0.3	1.0	17.1	1.0	0.1	69.2	0.4
PRO-WE	1.6	0.0	0.2	0.6	0.6	5.6	0.1	1.3	0.8	22.1	2.3	0.5	0.8	63.4

Figure 3: Confusion matrix for language-dependent SVM_char model run on 49 languages of BabelNet-derived NE resource.

	EVT-IN	EVT-NA	EVT-OC	LOC-FA	LOC-OT	ORG-CO	ORG-ER	ORG-PP	ORG-SP	PER	PRO-AR	PRO-EL	PRO-VE	PRO-WE
EVT-IN	86.9	0.2	0.1	1.2	1.0	1.0	0.2	1.4	0.3	5.6	1.9	0.0	0.1	0.2
EVT-NA	2.1	90.3	0.3	0.6	1.1	0.3	0.3	0.3	0.6	2.3	1.5	0.1	0.1	0.1
EVT-OC	0.3	0.0	79.5	2.0	0.5	1.8	0.3	1.0	3.0	8.6	2.7	0.1	0.2	0.0
LOC-FA	0.1	0.0	0.2	90.9	1.3	1.2	0.6	0.4	0.4	3.8	1.0	0.0	0.0	0.0
LOC-OT	0.3	0.1	0.1	2.9	86.7	0.8	0.2	0.3	0.3	7.0	1.2	0.0	0.0	0.0
ORG-CO	0.4	0.0	0.3	3.7	0.9	70.8	1.1	2.0	1.0	13.3	4.5	0.5	1.1	0.2
ORG-ER	0.1	0.0	0.1	2.4	0.5	1.6	89.4	1.3	0.5	3.3	0.8	0.0	0.0	0.0
ORG-PP	0.8	0.0	0.4	2.6	0.9	3.4	2.0	77.4	0.5	9.7	2.0	0.1	0.1	0.1
ORG-SP	0.1	0.0	1.1	1.4	0.6	1.2	0.6	0.4	66.0	27.7	0.8	0.0	0.1	0.1
PER	0.1	0.0	0.1	0.3	0.2	0.5	0.1	0.2	1.0	95.7	1.7	0.0	0.0	0.0
PRO-AR	0.8	0.0	0.4	2.1	1.1	2.8	0.4	0.9	0.4	30.9	59.9	0.1	0.1	0.1
PRO-EL	0.5	0.0	0.1	1.3	0.9	16.3	0.5	1.2	0.8	10.6	8.0	57.7	0.9	1.2
PRO-VE	0.7	0.0	0.4	0.8	0.4	10.7	0.2	0.3	1.5	11.9	1.4	0.4	70.4	1.1
PRO-WE	2.3	0.0	0.4	0.9	0.9	5.7	0.1	1.6	1.5	14.9	2.9	1.3	2.4	65.1

Figure 4: Confusion matrix for language-dependent NN_char run on 49 languages of BabelNet-derived NE resource.

Norwegian (NO), Norwegian (NN)), which illustrates the advantage of using character-based models for maximal performance in language-independent contexts,

- when applied to a number of external resources, a slight drop in performance of the character n-gram language-independent model was observed, although an *AP* of 84.7% in all experiments demonstrates the applicability of the proposed model in a range of contexts, not to mention relatively good performance for highly inflected languages, e.g. Hungarian and Slavic languages,
- promising performance of a simple neural network classifier which obtained comparable results vis-a-vis SVM-based classifiers both with language-dependent and independent training settings

All relevant resources for reproducing the presented experiments, including, the BN-derived NE resource and the trained models are available to the research community[10].

To our best knowledge, no similar evaluation of n-gram based models which exploit internal entity features for such a wide range of languages (49) was reported in the past. The presented models and the corresponding results reported in this paper constitute sort of more sophisticated baselines for multilingual fine-grained NE classification tasks and point of departure for future work.

In future work, we envisage to further test SVM_char on corpora from different domains, such as social media data and other non-news sources, as well as more fine-grained NE-tagged resources. Furthermore, given the strong performance of the character *n*-gram model, it would be interesting to compare this with other sub-word-based models, e.g. skipgrams [17], as well as a direct comparison with some of the recent neural fine-grained entity typing systems presented in Section 2. Finally, one could consider exploiting additional internal features of entities, e.g. their internal structure in recursively embedded NEs [19, 34] that might potentially further improve the classification performance, although for many languages, e.g. Slavic ones, the computation thereof is not trivial. Another interesting area of research could focus on evaluating whether the inclusion of inflected forms in the training data impacts the performance of the models explored in the presented work.

ACKNOWLEDGMENTS

The authors are highly indebted to the team behind the JEODPP platform [44] for their support and providing access to their GPUs which were indispensable for carrying out the reported research.

[10] https://cidportal.jrc.ec.europa.eu/ftp/jrc-opendata/LANGUAGE-TECHNOLOGY/BabelNet_derived_entity_resource/

REFERENCES

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. (2015). https://www.tensorflow.org/ Software available from tensorflow.org.

[2] Rami Al-Rfou', Vivek Kulkarni, Bryan Perozzi, and Steven Skiena. 2015. POLYGLOT-NER: Massive Multilingual Named Entity Recognition.. In *Proceedings of the 2015 SIAM International Conference on Data Mining (SDM 2015)*, Suresh Venkatasubramanian and Jieping Ye (Eds.). SIAM, 586–594. http://dblp.uni-trier.de/db/conf/sdm/sdm2015.html#Al-RfouKPS15

[3] Ashish Anand, Amit Awekar, et al. 2017. Fine-Grained Entity Type Classification by Jointly Learning Representations and Label Embeddings. *arXiv preprint arXiv:1702.06709* (2017).

[4] Douglas E. Appelt. 1999. Introduction to Information Extraction. *AI Communications* 12, 3 (Aug. 1999), 161–172. http://dl.acm.org/citation.cfm?id=1216155.1216161

[5] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. DBpedia: A Nucleus for a Web of Open Data. In *Proceedings of the 6th International The Semantic Web and 2nd Asian Conference on Asian Semantic Web Conference (ISWC'07/ASWC'07)*. Springer-Verlag, Berlin, Heidelberg, 722–735. http://dl.acm.org/citation.cfm?id=1785162.1785216

[6] Isabelle Augenstein. 2016. Web Relation Extraction with Distant Supervision. PhD thesis. http://etheses.whiterose.ac.uk/13247/. (2016).

[7] Daniel M. Bikel, Scott Miller, Richard Schwartz, and Ralph Weischedel. 1997. Nymble: A High-performance Learning Name-finder. In *Proceedings of the Fifth Conference on Applied Natural Language Processing (ANLC '97)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 194–201. https://doi.org/10.3115/974557.974586

[8] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: A Collaboratively Created Graph Database for Structuring Human Knowledge. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data (SIGMOD '08)*. ACM, New York, NY, USA, 1247–1250. https://doi.org/10.1145/1376616.1376746

[9] Sophie Chesney, Guillaume Jacquet, Ralf Steinberger, and Jakub Piskorski. 2017. Multi-word Entity Classification in a Highly Multilingual Environment. *Proceedings of EACL 2017 Multi-Word Expressions Workshop* (2017).

[10] Nancy A. Chinchor. 1998. Proceedings of the Seventh Message Understanding Conference (MUC-7) Named Entity Task Definition. In *Proceedings of the Seventh Message Understanding Conference (MUC-7)*. Fairfax, VA, 21 pages. version 3.5, http://www.itl.nist.gov/iaui/894.02/related_projects/muc/.

[11] François Chollet et al. 2015. Keras. https://github.com/fchollet/keras. (2015).

[12] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural Language Processing (Almost) from Scratch. *J. Mach. Learn. Res.* 12 (Nov. 2011), 2493–2537. http://dl.acm.org/citation.cfm?id=1953048.2078186

[13] Marc Damashek. 1995. Gauging similarity with n-grams: Language-independent categorization of text. *Science* 267, 5199 (1995), 843.

[14] George R Doddington, Alexis Mitchell, Mark A Przybocki, Lance A Ramshaw, Stephanie Strassel, and Ralph M Weischedel. 2004. The Automatic Content Extraction (ACE) Program-Tasks, Data, and Evaluation.. In *Proceedings of 2004 Language Resources and Evaluation Conference (LREC 2004)*, Vol. 2. 1.

[15] Xin Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, and Wei Zhang. 2014. Knowledge Vault: A Web-scale Approach to Probabilistic Knowledge Fusion. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '14)*. ACM, New York, NY, USA, 601–610. https://doi.org/10.1145/2623330.2623623

[16] Maud Ehrmann, Damien Nouvel, and Sophie Rosset. 2016. Named Entity Resources - Overview and Outlook. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)* (23-28), Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association (ELRA), Paris, France.

[17] Ahmed K. Elmagarmid, Panagiotis G. Ipeirotis, and Vassilios S. Verykios. 2007. Duplicate Record Detection: A Survey. *IEEE Trans. on Knowl. and Data Eng.* 19, 1 (Jan. 2007), 1–16.

[18] Micha Elsner, Eugene Charniak, and Mark Johnson. 2009. Structured Generative Models for Unsupervised Named-entity Clustering. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL '09)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 164–172. http://dl.acm.org/citation.cfm?id=1620754.1620778

[19] Jenny Rose Finkel and Christopher D. Manning. 2009. Joint Parsing and Named Entity Recognition. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL '09)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 326–334. http://dl.acm.org/citation.cfm?id=1620754.1620802

[20] Sofía N. Galicia-Haro, Alexander Gelbukh, and Igor A. Bolshakov. 2004. Recognition of Named Entities in Spanish Texts. In *Proceedings of the 3rd Mexican International Conference on Artificial Intelligence (MICAI 2004)*. 420–429.

[21] Claudio Giuliano and Alfio Gliozzo. 2008. Instance-based Ontology Population Exploiting Named-entity Substitution. In *Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1 (COLING)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 265–272. http://dl.acm.org/citation.cfm?id=1599081.1599115

[22] Ralph Grishman. 1995. The NYU System for MUC-6 or Where's the Syntax?. In *Proceedings of the 6th Conference on Message Understanding (MUC6 '95)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 167–175. https://doi.org/10.3115/1072399.1072415

[23] Jiafeng Guo, Gu Xu, Xueqi Cheng, and Hang Li. 2009. Named Entity Recognition in Query. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '09)*. ACM, New York, NY, USA, 267–274. https://doi.org/10.1145/1571941.1571989

[24] Ioannis Kanaris, Konstantinos Kanaris, Ioannis Houvardas, and Efstathios Stamatatos. 2007. Words versus character n-grams for anti-spam filtering. *International Journal on Artificial Intelligence Tools* 16, 06 (2007), 1047–1067.

[25] Sanjeev Karn, Ulli Waltinger, and Hinrich Schütze. 2017. End-to-End Trainable Attentive Decoder for Hierarchical Entity Classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Association for Computational Linguistics, 752–758. http://aclweb.org/anthology/E17-2119

[26] Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *CoRR* abs/1412.6980 (2014). arXiv:1412.6980 http://arxiv.org/abs/1412.6980

[27] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions (ACL '07)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 177–180. http://dl.acm.org/citation.cfm?id=1557769.1557821

[28] Changki Lee, Yi-Gyu Hwang, Hyo-Jung Oh, Soojong Lim, Jeong Heo, Chung-Hee Lee, Hyeon-Jin Kim, Ji-Hyun Wang, and Myung-Gil Jang. 2006. Fine-Grained Named Entity Recognition Using Conditional Random Fields for Question Answering. In *Proceedings of the 3rd Asia Information Retrieval Symposium (AIRS 2006) (Lecture Notes in Computer Science)*, Vol. 4182. Springer, 581–587.

[29] Thomas Lin, Mausam, and Oren Etzioni. 2012. No Noun Phrase Left Behind: Detecting and Typing Unlinkable Entities. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL '12)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 893–903. http://dl.acm.org/citation.cfm?id=2390948.2391045

[30] Xiao Ling and Daniel S. Weld. 2012. Fine-grained Entity Recognition. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence (AAAI'12)*. AAAI Press, 94–100. http://dl.acm.org/citation.cfm?id=2900728.2900742

[31] Andrew McCallum and Wei Li. 2003. Early Results for Named Entity Recognition with Conditional Random Fields, Feature Induction and Web-enhanced Lexicons. 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, 188–191. https://doi.org/10.3115/1119176.1119206

[32] David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes* 30, 1 (2007), 3–26.

[33] Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network. *Artificial Intelligence* 193 (2012), 217–250.

[34] Sofía N.Galicia-Haroand and Alexander Gelbukh. 2007. Complex named entities in Spanish texts: Structures and properties. *Lingvisticae Investigationes* 2007 (2007), 69–94. Issue 1.

[35] Joel Nothman, Nicky Ringland, Will Radford, Tara Murphy, and James R. Curran. 2013. Learning Multilingual Named Entity Recognition from Wikipedia. *Artif. Intell.* 194 (Jan. 2013), 151–175. https://doi.org/10.1016/j.artint.2012.03.006

[36] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research.* 12 (Nov. 2011), 2825–2830. http://dl.acm.org/citation.cfm?id=1953048.2078195

[37] Jakub Piskorski, Lidia Pivovarova, Jan Šnajder, Josef Steinberger, and Roman Yangarber. 2017. The First Cross-Lingual Challenge on Recognition, Normalization, and Matching of Named Entities in Slavic Languages. In *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing (BSNLP 2017)*. Association for Computational Linguistics, Valencia, Spain, 76–85.

[38] Jakub Piskorski and Roman Yangarber. 2013. Information Extraction: Past, Present and Future. In *Multi-source, Multilingual Information Extraction and Summarization*, Thierry Poibeau, Horacio Saggion, Jakub Piskorski, and Roman Yangarber (Eds.). Springer Berlin Heidelberg, 23–49.

[39] Lev Ratinov and Dan Roth. 2009. Design Challenges and Misconceptions in Named Entity Recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL '09)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 147–155. http://dl.acm.org/citation.cfm?id=1596374.1596399

[40] Alan Ritter, Sam Clark, Mausam, and Oren Etzioni. 2011. Named Entity Recognition in Tweets: An Experimental Study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '11)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 1524–1534. http://dl.acm.org/citation.cfm?id=2145432.2145595

[41] Satoshi Sekine and Chikashi Nobata. 2004. Definition, Dictionaries and Tagger for Extended Named Entity Hierarchy.. In *Proccedings of Language Resources and Evaluation Conference 2012 (LREC 2004)*. European Language Resources Association.

[42] S. Sekine, K. Sudo, and C. Nobata. 2002. Extended Named Entity Hierarchy. In *Proceedings of 3rd International Conference on Language Resources and Evaluation (LREC'02)*, M. Gonzáles Rodríguez and C. Paz Suárez Araujo (Eds.). Canary Islands, Spain, 1818–1824.

[43] Sonse Shimaoka, Pontus Stenetorp, Kentaro Inui, and Sebastian Riedel. 2017. Neural Architectures for Fine-grained Entity Type Classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. Association for Computational Linguistics, 1271–1280. http://aclweb.org/anthology/E17-1119

[44] P. Soille, A. Burger, D. De Marchi, P. Kempeneers, D. Rodriguez, V. Syrris, and V. Vasilev. 2018. A versatile data-intensive computing platform for information retrieval from big geospatial data. *Future Generation Computer Systems* 81 (2018), 30–40. https://doi.org/10.1016/j.future.2017.11.007

[45] Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2008. YAGO: A Large Ontology from Wikipedia and WordNet. *Web Semantics: Science, Services and Agents on the World Wide Web* 6, 3 (2008), 203 – 217. https://doi.org/10.1016/j.websem.2008.06.001

[46] György Szarvas, Richárd Farkas, László Felföldi, András Kocsor, and János Csirik. 2006. A highly accurate Named Entity corpus for Hungarian. In *Proceedings of International Conference on Language Resources and Evaluation (LREC 2006)*.

[47] Theano Development Team. 2016. Theano: A Python framework for fast computation of mathematical expressions. *arXiv e-prints* abs/1605.02688 (May 2016). http://arxiv.org/abs/1605.02688

[48] Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 Shared Task: Language-independent 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, 142–147. https://doi.org/10.3115/1119176.1119195

[49] Grigorios Tsoumakas, Ioannis Katakis, and Ioannis Vlahavas. 2010. Mining multi-label data. In *In Data Mining and Knowledge Discovery Handbook*. 667–685.

[50] Wern Wong, David Martinez, and Lawrence Cavedon. 2009. Extraction of Named Entities from Tables in Gene Mutation Literature. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing (BioNLP '09)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 46–54. http://dl.acm.org/citation.cfm?id=1572364.1572371

[51] Yadollah Yaghoobzadeh, Heike Adel, and Hinrich Schütze. 2017. Noise Mitigation for Neural Entity Typing and Relation Extraction. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. Association for Computational Linguistics, 1183–1194. http://aclweb.org/anthology/E17-1111

[52] Yadollah Yaghoobzadeh and Hinrich Schütze. 2017. Multi-level Representations for Fine-Grained Typing of Knowledge Base Entities. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. Association for Computational Linguistics, 578–589. http://aclweb.org/anthology/E17-1055

[53] Mohamed Amir Yosef, Sandro Bauer, Johannes Hoffart, Marc Spaniol, and Gerhard Weikum. 2012. HYENA: Hierarchical Type Classification for Entity Names. In *24th Intl. Conference on Computational Linguistics (Coling 2012) (Proceedings of the 24th Intl. Conference on Computational Linguistics (Coling 2012), Mumbai, India, December 8-15, 2012, pp. 1361-1370)*. Mumbai, India. https://hal.archives-ouvertes.fr/hal-01122707