JRC TMA-CC
Slavic Named Entity Recognition and Linking
Participation in the BSNLP-2019 shared task

Guillaume Jacquet Jakub Piskorski Hristo Tanev Ralf Steinberger
Joint Research Centre
European Commission
Ispra, Italy
{fname.lname}@ec.europa.eu

Abstract

We report on the participation of the JRC *Text Mining and Analysis Competence Centre* (TMA-CC) in the BSNLP-2019 Shared Task, which focuses on named-entity recognition, lemmatisation and cross-lingual linking. We propose a hybrid system combining a rule-based approach and light ML techniques. We use multilingual lexical resources such as JRC-NAMES and BABELNET together with a named entity guesser to recognise names. In a second step, we combine known names with wild cards to increase recognition recall by also capturing inflection variants. In a third step, we increase precision by filtering these name candidates with automatically learnt inflection patterns derived from name occurrences in large news article collections. Our major requirement is to achieve high precision. We achieved an average of 65% F-measure with 93% precision on the four languages.

1 Introduction

Multilingual Named Entity Recognition (NER) and the grounding of names to real-world entities is an essential component of the JRC TMA-CC's[1] large-scale, multi-annual and highly multilingual media monitoring effort called *Europe Media Monitor* - EMM[2] (Steinberger et al., 2017).

EMM has been analysing online news articles since 2003, reaching a current average of 320K articles per day from about 12K news sources in up to 70 languages. EMM clusters related news, categorises them into thousands of categories, detects breaking news and tracks topics over short periods of time. For a subset of about two dozen languages, EMM recognises and disambiguates entity mentions. The EMM-NER component constitutes the backbone of our submissions to the BSNLP-2019 Shared Task (Piskorski et al., 2019).

2 Approach

We submitted four system instance results, all of which are based on our in-house NER system *NERONE* (Ehrmann et al., 2017; Steinberger et al., 2015), which we describe first.

NERONE identifies and disambiguates mentions of persons, organisations, locations, events and products by first looking up *known names* and by then guessing new names. The list of *known names* contains about 1.2 million names. 600 000 unique entities have an average of 2 variants, the biggest number of variants for one entity being 6 200. The guessing of new names is based on large lexical resources (1.5 million entries) and ca 200 language-agnostic recognition patterns using the finite-state formalism described in (Piskorski, 2007). NERONE continuously updates the list of *known names*. Newly guessed names can become part of the list of *known names* if they are considered reliable enough. Reliability is mostly based on the frequency of the newly guessed name, the number of languages where it appears, the number of sources where it appears. Once eligible it is automatically added as a new known name or merged as a new variant of an existing name, including across languages and scripts (Steinberger et al., 2011). On average, 150 new variants and new names are automatically added daily to the list of *known names*. This list of known names (JRC-NAMES), is distributed publicly, together with the name variants, the titles, the language and date when it was found (Ehrmann et al., 2017). Based on previous work focused on multi-word entities (Jacquet et al., 2019), we furthermore added 2.1 million names and variants of the relevant entity

[1] https://ec.europa.eu/jrc/en/text-mining-and-analysis
[2] http://emm.newsbrief.eu

Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing, pages 100–104,
Florence, Italy, 2 August 2019. ©2019 Association for Computational Linguistics

categories from BabelNet (Navigli and Ponzetto, 2012). In the disambiguation steps, names that are part of a larger name are ignored (e.g. *John F Kennedy Airport*) and location names are disregarded if a homographic entity name of another category exists (e.g. Мартин (*Martin*) which could be both a small city in Slovakia and a person name).

In the remaining part of this Section we describe the four approaches explored, all of which are built on top of *NERONE*, which is known to have high precision, but low recall. We modified it to extend the recall, knowing that the precision will fall (*NERONE* with wildcards), then tried different levels of filtering to optimise the balance between precision and recall.

It is important to emphasise at this point that the four NER approaches presented in this paper are JRC's contribution (as one of the co-organisers of the Shared Task) to the provision of 'good' baseline systems to compare against.

2.1 JRC-TMA-CC-1: *NERONE*

The JRC-TMA-CC-1 variant uses *NERONE* as described before. We only did a slight adaptation for the location recognition. As our list of known location names (LOC), derived from GeoNames[3] is very short for some languages, we merged the LOC lists for the Cyrillic script languages Russian, Bulgarian, Bosnian, Macedonian and Serbian and we did the same for the Latin script west Slavic languages Polish, Czech and Slovak. It corresponds to the update of 200 000 entries among the existing 1.3 million location name resource.

2.2 JRC-TMA-CC-2: *NERONE* + wildcards

In addition to the system used in JRC-TMA-CC-1, we added wildcards to each name part of all entity types except for the GeoNames-derived LOC lists. The objective is to increase Recall by also capturing morphological variants of the known names. During morphological inflection, suffixes can be added to the base form of the name (e.g. *Andrej Babiš* inflected as *Andrejem Babišem*), but it also happens that final letters get replaced (suffix replacement, e.g. *Garbině Muguruzaová* inflected as *Garbiňe Muguruzaovou*). We therefore removed the last two letters of each name part and added a wildcard (*Garbině Muguruzaová* would become *Garbi% Muguruzao%*). To avoid overgenerating wildcard patterns, we did not remove letters from name parts that are three letters or shorter and we only removed one letter in four-letter words. Note that we use the term 'suffix' not in the morphological sense, but simply to denote the final letters of a name string.

2.3 JRC-TMA-CC-3: *NERONE* + wildcards and suffix filtering

Due to the vast number of different names, of which some can also be a string subset of longer names, the wildcards do occasionally overgenerate, i.e. capture names that are not variants, but names in their own right (e.g. *Josef Mill* would create the wildcard pattern *Jos% Mil%* which would wrongly match *Josefa Miller* as a possible inflection of *Josef Mill*). Submission JRC-TMA-CC-3 is based on the previous method, but here we aim to reduce such false positives (increase Precision) by filtering the names matched with the wildcards against a list of the more frequent suffix replacement rules.

To create such suffix replacement rules, we first searched in an average of 2 million news articles per language[4] for all our known names with the wildcard described in JRC-TMA-CC-2 to gather possible inflections of names, resulting in variant frequency lists for each name (see Table 1 for examples of collected variants). We then applied the following algorithm:

1. We hypothesise that the main form according to BabelNet and JRC Names is the main form. We have found a good empirical evidence this is true.

2. Tokenise all the names

3. For each token from the main variant Tm find the corresponding token from one of the derivations Td.

4. Find the common parts between the token Tm and Td. For example (cf. first case in Table 1), the common part between *Kotleby* and *Kotleba* is *Kotleb*.

5. Find the difference between the two forms and produce a list of candidate suffix rules,

[3] https://www.geonames.org/

[4] EMM collects daily meta-data from thousands of news articles, including article URLs. Exploiting these URLs, we collected (for the still active URLs) one year of articles for each of the four analysed languages.

Known name	potential variant list	freq
Mariana Kotleby	*Mariana Kotleby*	82
	Marian Kotleba	64
	Mariana Kotlebu	23
	Marianem Kotlebou	22
Garbin Muguruza	*Garbiňe Muguruzaovou*	92
	Garbiňe Muguruzaová	44
	Garbiňe Muguruzaové	22
	Garbině Muguruzaovou	8
	Garbiňe Muguruzaovou	7
Andrej Babiš	*Andrej Babiš*	29934
	Andreje Babiše	20470
	Andrej Babi	5935
	Andreje Babie	4271
	Andrejem Babišem	3979
Harvey Weinstein	*Harveyho Weinsteina*	278
	Harvey Weinstein	162
	Harveymu Weinsteinovi	20
	Harvey Weinsteinem	10
	Harvey Weinsteinovi	10
Energetický a průmyslový holding	*Energetický a průmyslový holding*	169
	Energetického a průmyslového holdingu	155
	Energetickému a průmyslovému holdingu	14
	Energetickým a průmyslovým holdingem	6
	Energetickém a průmyslovém holdingu	5

Table 1: Example of variant lists extracted from news.

endings	inflections	ratio
-uza	uzaová	0.4000
	uzaovou	0.3000
	uzaové	0.2000
-za	z	0.1125
	ze	0.1029
	zem	0.0386
-a	u	0.0696
	em	0.0657
	y	0.0602
-rej	reje	0.2656
	rejem	0.0938
	reji	0.0938
-ey	eyho	0.1366
	eym	0.0478
	y	0.0260
-cky	cký	0.2083
	ckého	0.1667
	ckém	0.1667

Table 2: Example of inflection probabilities

in this last case the rules will look like $y \rightarrow a$; $by \rightarrow ba$; $eby \rightarrow eba$.

6. In the case when the first token is completely contained in the second one, like *Marian* and *Mariana*, we extract a rule by taking the last two letters from the main form and the last corresponding ending from the derivative form $an \rightarrow ana$.

7. The inflection rules are gathered and we calculate various statistics. For example, the conditional probability that the first part of the rule is transformed into the second part of the rule. The statistics were collected from the list of word variants

Table 2 shows some examples of inflection rules obtained with this algorithm. This list was then used to filter acceptable inflections according to the initial base form: only those suffix replacement rules that had a probability higher than 0.01 were considered valid suffixes. If a name inflection found belonged to the eliminated low-frequency suffix replacement rules, it was not considered.

2.4 JRC-TMA-CC-4: *NERONE* + wildcards and less strict suffix filtering

This variant is identical to JRC-TMA-CC-3 with a lower threshold for filtering set to 0.001.

3 Results

While the Shared Task was subdivided into three subtasks, namely, Entity Recognition, Normalisation and Linking, our contribution focused less on the normalisation subtask and more on recognition, with a priority on precision scores and on cross-lingual entity-linking. Table 3 shows the results obtained by the four systems we submitted. The scores reported only refer to F-measure scores. For each evaluation category and each language, the bold score corresponds to the highest obtained F-measure. As a first observation, according to the description of the four systems, we were expecting the JRC-TMA-CC-1 system to obtain high precision but low recall, the JRC-TMA-CC-2 system to obtain high recall but low precision, and JRC-TMA-CC-3 and JRC-TMA-CC-4 to filter the too noisy recognition from JRC-TMA-CC-2 and deliver good precision/recall balance, therefore better F-measure. This is what one could observe when evaluating on the training set for the four languages. On the test set, one can observe the same phenomenon for Polish and Bulgarian, both for the relaxed partial and strict recognition, however, it applies to a smaller extent for Russian on the *Nord-Stream* topic and Czech on the *Ryanair* topic. By checking the error logs, these differences appear to be due to mis-recognition of key entities for these specific topics. Additionally to the F-measure scores reported in Table 3, the high precision scores we obtained for all languages are worth mentioning . We obtained best precision ranking compare to the other shared task participants for all four languages. As an average for both topics, our JRC-TMA-CC-4 system obtained for Czech, Russian, Bulgarian and Polish a precision of, respectively, 94.4%, 90.2%, 95.4%, 93.7% (at a price of lower recall). This precision is also quite well-distributed across entity types. For

AVERAGE ON BOTH TOPICS			*Language*			
Phase	*Metric*	system	cs	ru	bg	pl
Recognition	Relaxed Partial	JRC-TMA-CC-1	**62.0**	73.6	73.2	56.13
		JRC-TMA-CC-2	61.6	72.0	72.4	54.8
		JRC-TMA-CC-3	58.0	**73.7**	73.8	50.9
		JRC-TMA-CC-4	58.0	73.5	**74.2**	**57.4**
	Relaxed Exact	JRC-TMA-CC-1	**55.6**	**68.7**	67.3	48.6
		JRC-TMA-CC-2	54.3	66.2	66.7	45.5
		JRC-TMA-CC-3	55.3	68.2	67.6	48.4
		JRC-TMA-CC-4	55.3	68.0	**67.9**	**49.6**
	Strict	JRC-TMA-CC-1	47.6	59.9	63.9	41.8
		JRC-TMA-CC-2	49.9	60.4	64.4	44.0
		JRC-TMA-CC-3	**50.0**	**60.6**	64.6	44.6
		JRC-TMA-CC-4	**50.0**	60.5	**65.2**	**45.9**
Entity linking	Single language	JRC-TMA-CC-1	29.8	41.8	51.8	21.9
		JRC-TMA-CC-2	**35.9**	**42.9**	51.5	**30.3**
		JRC-TMA-CC-3	33.5	41.9	51.5	25.8
		JRC-TMA-CC-4	33.9	41.8	**52.4**	28.2
	Cross-lingual	JRC-TMA-CC-1	24.7			
		JRC-TMA-CC-2	**29.7**			
		JRC-TMA-CC-3	26.4			
		JRC-TMA-CC-4	27.3			

Table 3: Evaluation results (F-scores) across all scenarios and languages on the test data.

PER, LOC, ORG, PRO and EVT, we respectively obtained 92.4%, 95.9%, 89.2%, 96.0% and 83.3%. The fact that we were able to improve our existing system with quite a simple adaptation is promising and encourages us to push further this process of name ending/inflection filtering. Concerning the entity-linking evaluation, Table 3 shows results for each single language and, more importantly, for cross-lingual linking. Despite the low recall of our four systems compare to other teams, our F-measure scores are ranked 2nd for both single language and cross-lingual linking. We will have to analyse the error logs in more detail to investigate possible improvements. Also, we observe that in almost all languages and topics, the best results are obtained by the JRC-TMA-CC-2 system, which is most likely correlated to a high recall.

4 Related Work

NER systems are often the first step in event detection, question answering, information retrieval, co-reference resolution, topic modelling, etc. The first NER task was organised by (Grishman and Sundheim, 1996) in the Sixth Message Understanding Conference. Early NER systems were based on handcrafted rules (Chiticariu et al., 2010), lexicons, orthographic features and ontologies. These systems were followed by NER systems based on feature-engineering and machine learning (Nadeau and Sekine, 2007).

There are not many systems for NER that address inflected languages like the Slavic ones. Among the others, (Piskorski et al., 2007) tackled the task of matching morphological variants of names in Polish text by optimising string similarity calculations for inflections. (Pajzs et al., 2014) experimented with name lemmatisation and inflection variant generation in the highly inflected and agglutinative language Hungarian. (Gareev et al., 2013) describes NER for the highly inflective Russian language. The first edition of the Shared Task on Slavic NER was organised in the context of BSNLP 2017 (Piskorski et al., 2017)

5 Conclusions and Future Work

We presented lightweight method to improve the performance of our in-house NER system NERONE for the recognition and linking of inflected named entities in inflected languages without delving into the morphological rules and proper name declension paradigms of each of the languages. We learnt potential name inflection patterns by searching for suffix variants of known names in large volumes of text. We then changed the known-name lookup part of NERONE by replacing the last letters of each name with wildcards to capture inflectional variants. We used the newly captured potential name inflections to reduce the number of wrong wildcard matches. As expected, we achieved good precision scores, 94.4%, 90.2%, 95.4%, 93.7% respectively for Czech, Russian, Bulgarian and Polish and unbalanced F-measures, from too low (58.0% and 57.4% for Czech and Polish) to reasonably good (73.5% and 74.2% for Russian and Bulgarian). One of the main drive of developing the described extension of NERONE was to contribute to the provision of 'good' baseline systems for the BSNLP-2019 Shared Task.

The proposed systems could be improved in many ways, including, i.a.: (a) expansion of the set of inflection patterns to guess *new* names, (b) integration of a classifier to distinguish the reading of entities that can designate different entity types (e.g. *BBC* as an organisation or as a product, (c) expansion of the lookup of geographical names, (d) integration of a mechanism to distinguish the Czech female gender marker *-ova* from case markers as it behaves differently: Forms such as *Merkelova* are the Czech nominative base form of the German Chancellor *Merkel* and inflections apply to *Merkelova* instead of to our name list's base form *Merkel*, (e) introduction of additional heuristics to narrow down the possible name mention matches, since the automatically generated groups of name inflection variants, from which we

learn the inflection patterns, contain errors because the wildcards match too generously, and (f) updating and completing our list of geographical names as the coverage for different languages currently ranges from over 100,000 geographical names to below 3,000.

References

Laura Chiticariu, Rajasekar Krishnamurthy, Yunyao Li, Frederick Reiss, and Shivakumar Vaithyanathan. 2010. Domain adaptation of rule-based annotators for named-entity recognition tasks. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, pages 1002–1012. Association for Computational Linguistics.

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

Rinat Gareev, Maksim Tkachenko, Valery Solovyev, Andrey Simanovsky, and Vladimir Ivanov. 2013. Introducing baselines for Russian named entity recognition. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 329–342. Springer.

Ralf Grishman and Beth Sundheim. 1996. In *The 16th International Conference on Computational Linguistics*.

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 (SAC 2019)*.

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

Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Júlia Pajzs, Ralf Steinberger, Maud Ehrmann, Mohamed Ebrahim, Leonida Della Rocca, Eszter Simon, and Tamás Váradi. 2014. Media monitoring and information extraction for the highly inflected agglutinative language Hungarian.

Jakub Piskorski. 2007. ExPRESS – Extraction Pattern Recognition Engine and Specification Suite. In *Proceedings of the International Workshop Finite-State Methods and Natural Language Processing 2007 (FSMNLP 2007)*.

Jakub Piskorski, Laska Laskova, Michał Marcińczuk, Lidia Pivovarova, Pavel Přibáň, Josef Steinberger, and Roman Yangarber. 2019. The second cross-lingual challenge on recognition, classification, lemmatization, and linking of named entities across Slavic languages. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, Florence, Italy. Association for Computational Linguistics.

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*, pages 76–85, Valencia, Spain. Association for Computational Linguistics.

Jakub Piskorski, Marcin Sydow, and Karol Wieloch. 2007. Comparison of string distance metrics for lemmatisation of named entities in Polish. In *Language and Technology Conference*, pages 413–427. Springer.

Ralf Steinberger, Martin Atkinson, Teófilo Garcia, Erik Van der Goot, Jens Linge, Charles Macmillan, Hristo Tanev, Marco Verile, and Gerhard Wagner. 2017. EMM: Supporting the analyst by turning multilingual text into structured data. In *Transparenz aus Verantwortung: Neue Herausforderungen für die digitale Datenanalyse*. Erich Schmidt Verlag.

Ralf Steinberger, Guillaume Jacquet, and Leonida Della Rocca. 2015. Creation and use of multilingual named entity variant dictionaries. *Traduire aux confins du lexique: les nouveaux terrains de la terminologie*, 40:113.

Ralf Steinberger, Bruno Pouliquen, Mijail Kabadjov, and Erik van der Goot. 2011. JRC-Names: A Freely Available, Highly Multilingual Named Entity Resource. In *Proceedings of the* 8^{th} *International Conference Recent Advances in Natural Language Processing (RANLP'2011)*, pages 104–110, Hissar, Bulgaria.