EUROPEAN COMMISSION
JOINT RESEARCH CENTRE

Institute for the Protection and Security of the Citizen (Ispra)

JRC GRANTHOLDER FINAL REPORT

Grantholder [redacted]

Report Title & Project title GH Final report, "Multi-lingual and multi-functional information extraction methods and tools" project

Date & Status 15/04/2016, Final version

Project Leader:	[redacted]	Contract n.:	2012-IPR-G-30-000-00741
HoU:	[redacted]	Start date:	16/04/2013
		Date due:	end

1. INTRODUCTION/BACKGROUND

The Global Security and Crisis Management Unit of the IPSC supports the Union's policies to strengthen the EU's resilience to crises and disasters as well the EU's aim to promote stability and peace through its research in crisis management technologies and in information mining and analysis. The Unit's EMM (Europe Media Monitor) project (formerly OPTIMA action - Open Source Text Information Mining and Analysis action), develops innovative solutions for retrieving and extracting information from the internet, and especially from online news and social media, serving many Commission Services, EU agencies and some EU Member State authorities.

During these three years, my contribution has been to develop, improve and implement approaches and systems to be integrated in the existing solutions for retrieving and extracting information from the internet. More precisely, I worked in collaboration with different colleagues from the unit, from the JRC and from outside the JRC, on tasks such as the Improvement and development of a Named Entity Guesser, Acronym and multiword entity recognition and extraction, statistical extension of a Named Entity Guesser, convert some existing linguistic resources as Linked Data format. My role was also to promote the work in the team for these tasks by publishing and presenting the developed systems and approaches in scientific journals and conference. Finally, in the last year, I've been involved in a collaborative project between our team, a team from JRC headquarter and the Cern to develop a platform called TIM (Technology Innovation Monitoring). All these contributions will be detailed in the following sections.

2. JRC PROJECTS WORKED ON

The following tasks and projects are linked with the MMA and the OSINT project

a. IMPROVE AND DEVELOP FURTHER THE NAMED ENTITY GUESSER

- Maintain the existing Named Entity guesser and adapt it to some specific needs from external customers (e.g. African Union Commission)
- Improve coverage and precision of the existing Named Entity guesser for languages already processed. e.g. by addressing specific phenomena like inflected languages.
- Extend the Named Entity guesser to new languages based on the needs of external customers.

b. STATISTICAL EXTENSION OF NAMED ENTITY GUESSER

- Developing hybrid method combining statistical and rule based approaches. It includes experiments on automatic cross lingual lexicon extension and experiments on automatic rule creation. Automatic cross lingual lexicon extension consist in harmonizing the lexical resources we have for different languages. For instance, by extending lexicons of the less covered language based on lexicons we have for the well covered languages.
- Automatic rule creation consists in creating new rules for the NE Guesser based on how the person names and organisation names are found in the news texts.

c. ADDRESS THE ACRONYM AND ORGANIZATION EXTRACTION

Carry on the on-going work on Acronym detection. It consists in detecting, from news articles, links between short forms (acronyms) and long form (full name) of the same entity. This includes two sub-tasks:

- Detect links between multiple forms, long forms and short forms, which refer to the same entity from one language.
- Detect links between long form and short form, from multiple languages.

d. CONTRIBUTION TO COLLABORATIVE PROJECT BETWEEN JRC AND CERN

Develop different tools adapted to the CERN needs, related to Named Entities Recognition. It includes Named Entity disambiguation and linking between different types of entities. Linking between persons working together, between person and the company she works for, between persons and research topics.

3. PROGRESS OF THE PROJECT, TIMING AND RESULTS

e. IMPROVE AND DEVELOP FURTHER THE NAMED ENTITY GUESSER

Maintain and develop NE guesser by working on lexical resource optimisation, grammar rule improvement and evaluation method optimisation. This contribution corresponds to address many small subtasks which can be illustrated by the following examples: adding functionalities for a better interaction with the Geomatcher, working on person profession/title to improve person recognition, deal with person name containing specific characters like quotes, contributing to grammar rule adaptation for some specific languages, enrich the reference corpora in order to extend the evaluation dataset.

Based on all these improvements, the new results we obtained are presented in the following table. Ona verage over 9 languages, we obtain a precision of 92.1% and a recall of 50.4%. It corresponds to an improvement of +2.4 points for precision, and +0.6 points for recall.

language code	Reference entities	Entities found by the system	Correct entities	Wrong entities	Added entities	Forgotten entities	Precision	Recall	F-measure	precision improvement	recall improvment	F-measure improvment
de	2770	1616	1488	128	86	1239	92.08%	53.72%	67.85%	2.4%	0.3%	0.9%
nl	4716	2343	2105	238	121	2468	89.84%	44.64%	59.64%	2.8%	0.8%	1.3%
es	4252	2577	2397	180	147	1822	93.02%	56.37%	70.20%	3.7%	1.6%	1.9%
en	6560	3815	3693	122	99	2844	96.80%	56.30%	71.19%	3.1%	-1.0%	-0.5%
ro	336	225	205	20	17	128	91.11%	61.01%	73.08%	2.8%	-0.3%	0.5%
hu	678	596	506	90	75	157	84.90%	74.63%	79.43%	4.8%	2.4%	3.2%
it	4545	1953	1789	164	110	2698	91.60%	39.36%	55.06%	2.0%	-0.4%	0.0%
hu	3746	1988	1790	198	115	1874	90.04%	47.78%	62.43%	2.4%	0.9%	1.5%
pt	740	460	382	78	57	337	83.04%	51.62%	63.67%	1.1%	-0.8%	1.0%
tr	388	139	123	16	13	262	88.49%	31.70%	46.68%	5.8%	-0.3%	0.6%
Ttl	28731	15712	14478	1234	840	13829	92.15%	50.39%	65.15%	2.4%	0.6%	1.1%

f. STATISTICAL EXTENSION OF NAMED ENTITY GUESSER

We pursued our experiments on lexicon expansion for a specific language based on what we have in the other languages. We also pursued other experiments on automatic rule creating for multiword expression recognition, on multiple

languages. A paper describing our approach is currently started. We aim at submitting this paper this year or next year depending on how efficient is the approach.

g. Address the Acronym and organization extraction

Publication of a paper describing our crosslingual linking of multiword expression clusters [redacted]

h. Contribution to collaborative project between JRC and CERN

My contribution to this project is in collaboration with different persons from the three teams, CERN team, JRC-headquarter team and our team.
The different tasks where:

- Automatic Affiliation resource construction
- EntityMatcher adaptation and development
- Geomatcher adaptation
- Entity Organizer development

4. Objectives of the Project, Main Results and Conclusions

4.1. Improve and develop further the Named Entity guesser

In collaboration with [redacted], I contributed to improve and develop the Named Entity (NE) Guesser.

Worked, with [redacted], on and finalized the re-structuration of the NE guesser in order to integrate it better with the other modules of the EMM chain. Namely, I embedded the Geomatcher module as source of information in the NE Guesser. We did some evaluation on the new configuration showing comparable results if not a slight improvement which is what we expected.

Developed, with [redacted], a framework to integrate inflected language information in the process chain.

In collaboration with [redacted], we included a new gold standard corpus for Turkish and strongly improved the grammar, lexicon and rules, for Turkish text processing (Kucuk et al, 2014).

Given the mentioned updates, the global precision of the NE guesser over the 11 evaluated languages increased from 86.2% to 92.13%. F-measure increased from 47.5% to 65.10%.

language code	Reference entities	Entities found by the system	Correct entities	Wrong entities	Forgotten entities	Precision	Recall	F-measure
de	2770	1613	1487	126	1242	92.19%	53.68%	67.85%
nl	4716	2344	2105	239	2468	89.80%	44.64%	59.63%
es	4252	2582	2395	187	1824	92.76%	56.33%	70.09%
en	6560	3815	3694	121	2843	96.83%	56.31%	71.21%
ro	336	226	205	21	128	90.71%	61.01%	72.95%
hu	678	589	501	88	163	85.06%	73.89%	79.08%
it	4545	1954	1789	165	2698	91.56%	39.36%	55.05%
hu	3746	1975	1779	196	1885	90.08%	47.49%	62.19%
pt	740	458	382	76	337	83.41%	51.62%	63.77%
tr	388	140	123	17	262	87.86%	31.70%	46.59%
Ttl	**28731**	**15696**	**14460**	**1236**	**13850**	**92.13%**	**50.33%**	**65.10%**

4.2. Acronym and multiword entity extraction

Multi-word entities, such as organisation names, are frequently written in many different ways. ( ) previously automatically identified over one million acronym pairs in 22 languages, consisting of their short forms (e.g. EC) and their corresponding long forms (e.g. European Commission, European Union Commission). In order to automatically determine which of these multi-word entities are variant spellings of the same conceptual entity, ( ) performed the clustering of those long forms belonging to the same short form.

The clustering results needed to be evaluated and, for that purpose, (presents our effort to automatically evaluate in 22 languages the multiword entity clusters with the use of Wikipedia redirection tables as the gold standard. The evaluation results achieved with this method are convincing and show that the proposed evaluation method is stable enough to measure data quality.

I then developed, implemented and evaluated a multilingual clustering of multi-word entity names (mostly organisation names but not only). Starting from a collection of millions of acronym/expansion pairs for 22 languages where expansion variants were grouped into monolingual clusters, we experiment with several competing methods to link these clusters across languages. Aggregation strategies make use of string similarity distances and translation probabilities, and identify connected clusters according to different similarity measures. The accuracy of the approach is evaluated against Wikipedia's redirection and cross-lingual linking tables. The resulting multi-word entity resource contains 70K multi-word entities with unique identifiers and their 600K multilingual lexical variants.

4.3. Lexical resource expansion

We developed, with , a hybrid method combining statistical and rule based approaches aiming at expanding existing lexical resource based on cross-lingual existing resources. Here is a summary of our contribution:

Named entity recognition (NER) is an important part of the language processing in the EMM chain. In this contribution, we concentrate on the problem of person name recognition, however we believe that the proposed techniques can be also used for other named entities like organizations. Also, we focus on languages with latin scripts, in which are written the majority of the news articles we have to process.

Monolingual person name recognition is a well described and addressed task. In our context, we must address this task in an highly multilingual environment. Therefore we have a by-default configuration of our NER which works for all the languages and we have more specific lexical resources for these languages for which linguistic experts were available. For some of the languages we have large lexical resources and for some others resources are quite scarce. Moreover, we can have heterogeneous resources for each of the covered languages: language l1 can have a large resource of specific person first names but a weak resource of person professions, where it can be the opposite for language l2. Finally, it would be useful for our framework to be able to automatically create a by-default lexical resource for a new language without having a linguist expert for this language. If not perfect, such automatic resource could provide a basis for processing in this language.

Our method aims at addressing these problems by expanding language-specific lexicon based on distributional approach starting from more generic language-independent resources we already have.

We carried out experiments in three languages, namely Spanish, Hungarian and Turkish and we showed that the automatically expanded resources deliver better results than the original ones: between 1.4 points and 6.4 points of F-measure improvement.

	Turkish				Hungarian				Spanish			
	Nb ext. terms	P	R	F	Nb ext. terms	P	R	F	Nb ext. terms	P	R	F
without language specific firstname	0	89.9%	18.3%	30.4%	0	86.9%	32.5%	47.7%	0	89.0%	52.6%	66.1%
with manual language specific firstname	6350	82.7%	32.0%	46.1%	4340	87.6%	43.9%	58.5%	537	89.3%	55.3%	68.3%
with only seeds	20	88.6%	18.0%	30.0%	20	87.0%	33.5%	48.4%	20	89.1%	52.8%	66.3%
with extended language specific firstname filter 1	321	87.7%	18.3%	30.3%	563	87.2%	35.3%	50.3%	220	89.1%	53.4%	66.8%
with extended language specific firstname filter 0.1	1548	75.5%	20.6%	32.4%	2149	87.7%	38.5%	53.5%	1079	88.7%	54.2%	67.3%
with extended language specific firstname filter 0.01	2577	76.1%	21.4%	33.4%	3135	85.8%	39.5%	54.1%	1700	88.0%	54.8%	67.5%

4.4. Contribution to collaborative project TIM (Technology Innovation Monitoring)

My contribution to this project is in collaboration with different persons from the three teams, CERN team, JRC-headquarter team and our team.
The different tasks where:

- Automatic Affiliation resource construction
- EntityMatcher adaptation and development
- Geomather adaptation
- Entity Organizer development

5. Patents and Publications

[redacted]. Cross-lingual Linking of Multi-word Entities and their corresponding Acronyms. 10th edition of the Language Resources and Evaluation Conference, 23-28 May, Portorož, Slovenia.

[redacted] Multilingual Entity Name variants and titles as Linked Data, Semantic Web Journal.

[redacted], "Multilingual Entity Name variants as Linked Data", invited talk during the Talk of Europe Camp, Amsterdam, Netherland. JRC95095

[redacted], "Creation and use of multilingual named entity variant dictionaries", chapter in Traduire aux confins du lexique: les nouveaux terrains de la terminologie. JRC91623

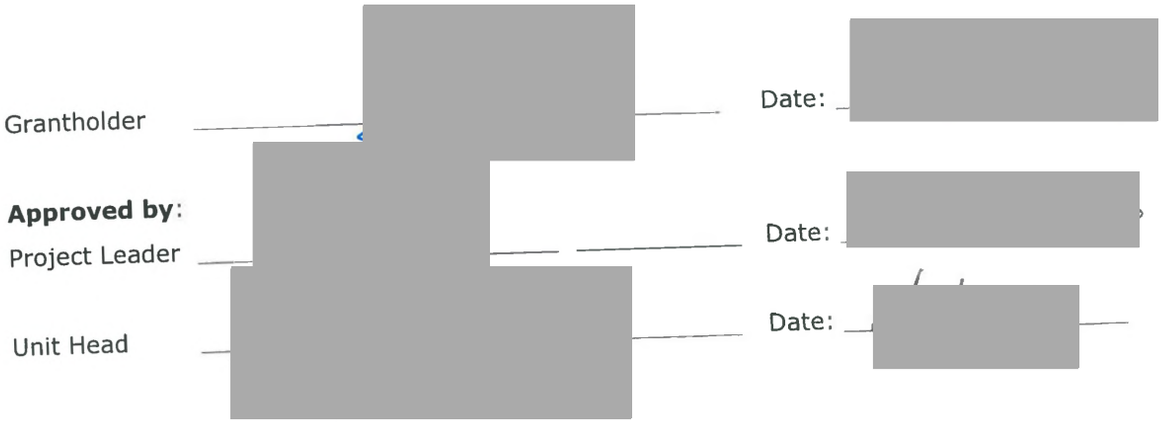
[redacted], "EuroVoc thesaurus and the JEX (JRC Eurovoc Indexer) software", SPLET workshop, LREC conference, Rekjavik, Iceland. JRC89974

[redacted] "Multi-word entities recognition in a multilingual environment", oral presentation during the XRCE scientific seminar, Grenoble, France. JRC89708

[redacted] (2014). "Resource Creation and Evaluation for Multilingual Sentiment Analysis in Social Media Texts", 9th edition of the Language Resources and Evaluation Conference (LREC), Reykjavik, Iceland, 2014.

[redacted] (2014). "Named Entity Recognition on Turkish Tweets". 9th edition of the Language Resources and Evaluation Conference (LREC), Reykjavik, Iceland, 2014.

[redacted] "Clustering of multiword named entity variants : Multilingual evaluation". 9th edition of the Language Resources and Evaluation Conference (LREC), Reykjavik, Iceland, 2014.


Grantholder
Date:
Approved by:
Project Leader
Date:
Unit Head
Date: