JRC GRANThOLDER PROGRESS REPORT

Granholder

Report Title & Project title GH Annual report, “Multi-lingual and multi-functional information extraction methods and tools” project

Date & Status 26/05/2015

<table>
<thead>
<tr>
<th>Project Leader:</th>
<th>Contract n.: 2012-IPR-G-30-000-00741</th>
</tr>
</thead>
<tbody>
<tr>
<td>HoU:</td>
<td>Start date: 16/04/2013</td>
</tr>
<tr>
<td></td>
<td>Date due: Month 24</td>
</tr>
</tbody>
</table>

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.

1. JRC PROJECTS WORKED ON

This annual report concerns the work carried out within the EMM project for the development and maintenance of automatic multilingual text analysis tools. The following objectives are linked with the MMA and the OSINT project.

OBJECTIVE 1: IMPROVE AND DEVELOP FURTHER THE NAMED ENTITY Guesser

Definition

- 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.

OBJECTIVE 2: NAMED ENTITY GUesser: STATISTICAL EXTENSION

Definition

- 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.

European Commission, Via Enrico Fermi 2749, I-21027 Ispra (Varese) - Italy. Telephone: (39)0332-78-
OBJECTIVE 3: CONTRIBUTION TO COLLABORATIVE PROJECT BETWEEN JRC AND CERN

Definition
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.

OBJECTIVE 4: ADDRESS THE ACRONYM AND ORGANIZATION EXTRACTION

Definition
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.

2. PROGRESS OF THE PROJECT, TIMING AND RESULTS

OBJECTIVE 1: IMPROVE AND DEVELOP FURTHER THE NAMED ENTITY GUESSER

• Worked, with [redacted], on and finalized the re-structuring 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 Guessers. 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.

OBJECTIVE 2: NAMED ENTITY GUESSER: STATISTICAL EXTENSION

• We developed, with [redacted], 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.

OBJECTIVE 3: CONTRIBUTION TO COLLABORATIVE PROJECT BETWEEN JRC AND CERN

• Based on the Entity Matcher existing in the EMM chain, I developed and adapted a first pilot of Entity Matcher for the Cern project.
OBJECTIVE 4: ADDRESS THE ACRONYM AND ORGANIZATION EXTRACTION

- I developed, implemented and evaluated a multilingual clustering of multi-word entity names (mostly organisation names but not only). Here is a short description of the contribution: 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.

- The process of integrating the module of acronym/expansion recognition in the EMM chain is on-going.

3. PATENTS AND PUBLICATIONS

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

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

EuroVoc thesaurus and the JEX (JRC EuroVoc Indexer) software", SPLCET workshop, LREC conference, Reykjavík, Iceland. JRC89974

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