A Mixed Approach in Recognising Geographical Entities in Texts

Dan Cristea[1,2], Daniela Gîfu[1], Ionuţ Pistol[1], Daniel Sfirnaciuc[1], and Mihai Niculiţă[3]

[1]Faculty of Computer Science, "Alexandru Ioan Cuza" University of Iaşi
[2]Institute for Computer Science Romanian Academy - the Iaşi branch
[3]Faculty of Geography, "Alexandru Ioan Cuza" University of Iaşi
{dcristea,daniela.gifu,ipistol,daniel.sfirnaciuc}@info.uaic.ro,
mihai.niculita@uaic.ro

Abstract. The paper describes an approach for automatic identification in Romanian texts of name entities belonging to the geographical domain. The research is part of a project (*MappingBooks*) aimed to link mentions of entities in an e-book with external information, as found in social media, Wikipedia, or web pages containing cultural or touristic information, in order to enhance the reader's experience. The described name entity recognizer mixes ontological information, as found in public resources, with handwritten symbolic rules. The outputs of the two component modules are compared and heuristics are used to take decisions in cases of conflict.

Keywords: name entity recognition (NER), annotation conventions, geonames, geo-ontology, gazetteer, symbolic rules, pattern matching techniques, mixed approaches in NER

1 Introduction

MappingBooks is an on-going project[1] aiming to develop a new type of electronic product with a high impact in education and tourism. The main envisioned users are school pupils and students. The technology mixes methods from natural language processing, web cartography, web mapping, mixed reality techniques and ambient intelligence/ubiquitous computing to link mentions of geographical entities existing in school manuals onto data existing on the web, to localise these entities on 2D and 3D hypermaps [11] and to put them in correlation with the reader's location and related data. The toponyms can be supplemented with different type of information, diagrams or any related graphic materials. For example, if a reader is focusing a mention of the Mount Ceahlău in a school book, not only that a localisation of the Ceahlău Mountain will be signalled to her/him on an electronic map, but also information about this mountain, as a function of the context in which the toponym appears will be fetched

[1] Financed by the Romanian Ministry of Education and Research (UEFISCDI) under the Partnerships Programme (PN II Parteneriate, competition PCCA 2013), project code: PN-II-PT-PCCA-2013-4-1878.

and displayed on the user's mobile screen. If the school book covers the topic of physical geography, the localisation on a general map will be supplemented with thematic maps (geologic, landforms, climate maps), and associated thematic information.

What is most important in *MappingBooks* is that the connections from the book onto the virtual space would have to be realised by a technology, therefore automatically, and not by a human annotator. It is clear that this ambitious goal ought to be sustained by a powerful tool that manipulates with accuracy mentions of geographical entities in free texts. In this paper we describe that part of this project that deals with the recognition of name entities. Our approach mixes brute force methods (as provided by the use of a large collection of proper nouns) with symbolic methods (a collection of regular expressions, or rules, intended to discover the significance of proper names using the local context).

The paper has the following structure. In section 2 we briefly present the state of the art in named entity recognition. Section 3 gives definitions for entities, as a semantic concept, and their realisation in texts. Section 4 shows the overall architecture of the system. Then, sections 5, 6 and 7 briefly describe the three component modules of the system, including some remarks about the evaluation. Finally, section 8 states some conclusions, while the Appendix gives a number of details about the other sources of free geographical data used in the project.

2 Background

The problem of extracting text information represents a constant concern in Natural Language Processing (NLP), now already for more than two decades. It has a wide range of applications in different domains (geography, biomedical sciences, business intelligence, etc.). The Message Understanding series of Conferences (MUC) has been launched at the beginning of 1990s to face the larger and larger interest of companies to extract, from unstructured text (such as newspaper articles), structured information about their activities or products. A few years later, at the 6th edition of MUC, the term *named entity* was introduced [8] and their recognition was considered a problem of classification. Named Entity Recognition (NER) and Classification became a task of Information Extraction. The important issue was the classification of words or word groups that signify proper names [18].

NER has become a very important topic for many other sub-fields of NLP [20], [2] and [15]. Initially, the most important results were obtained using rule-based systems created manually. To overcome the tedious work of writing rules manually and to improve the rate of recognition, the researchers started to use statistical models, based on machine learning techniques, which have proved to be very effective. Here the manual effort was transferred in the direction of manual annotation, in order to build large corpora of positive examples. The most efficient techniques of our days, combine rule-based grammars with statistical (maximum entropy) models. An example of this type is the LTG system [17], presented at MUC-7. The FIJZ system [6], presented at the CONLL-2003 uses four different classifiers (robust linear classifier, maximum entropy, transformation-based learning, and hidden Markov model), which, combined under special conditions, produce very good results. Another notorious information extraction system was ANNIE (A Nearly-New Information Extraction System) [16],

included in the GATE (General Architecture for Text Engineering) framework [5]. ANNIE recognizes person, location, organization, money, percent, data, address, identifier and unknown. ANNIE, was used with success for many languages, including Romanian, being a perfect example of combination of a lexical resource (gazetteer) and a rule-based approach in information extraction (a set of pattern/action rules written as JAPE grammars).

We mention, also, the MUSE system, incorporating ANNIE resources, processing also Romanian texts, as an example of a fast and cheap adaptation of an existing system to deal with new applications.

Another important project is TTL (Tokenizing, Tagging and Lemmatizing), a text processing platform developed at RACAI[2], trained to deal mainly with Romanian and English, which recognizes entities, does sentence splitting, tokenizing, chunking, etc. This platform works with techniques based on the use of regular expressions. In TTL, the NER function precedes the sentence splitter, avoiding thus the dangers of considering the dot in an abbreviation as signalling the limit of a sentence. Another NER system for Romanian that combines a collection of linguistic grammar rules and a set of resources is described in [10]. Other tasks are focused on: personal name disambiguation [14], named entity translation [7], [9] and acronym identification [19].

In the early 2000s, a priority in the research based on analysis of geographical references, focused on the named entity, was the geographic instances classification in text. For instance, the geographical references classification by assuming consecutive proper nouns as named entity candidates, using a co-training algorithm [3]. Also, a classification could be based on the fine-grained sub-types of geographical entities, knowing they refer generalized names as well as locations [23]. Some researchers suggested unsupervised learning methods in the NER area, related to bootstrapping learning algorithms [13], [22], [12]. Note that most bootstrapping approaches start with incomplete annotations and patterns derived from selected seeds, which imply possible annotation errors that can be included in the learning process. These errors could be avoided by designing statistical measures of control.

3 Entities in text

In [4], we addressed the issue of annotating relations linking entities in texts. In the mentioned paper we say that any mention of an entity (restricted only to persons, gods, group of persons and of gods) is a mapping from a text expression to a corresponding 'container'[3]. In all corpus-based approaches, mentions, not containers, are annotated, but if semantic reasoning is tried on these annotations, then containers and their contents are recreated as semantic representations.

Following the usual tendency in the literature, we consider entities as being semantic categories expressed at the textual level by noun phrases (NPs). In certain contexts, proper names could be parts of NPs. It is therefore important to make the distinction

---

[2] The Romanian Academy Center for Artificial Intelligence, in Bucharest.

[3] A box or a container is associated with each character (entity), which in a text is, at the first mention, partially filled in with pieces of information and, subsequently, complemented with details (name, sex, kinship connections, composition, beliefs, religion, etc.).

between proper names and name entities. Proper names represent part-of-speech categories, therefore are manifested at the text level, while entities, among which also name entities, are semantic categories, therefore presented at a representational level. For instance, in the sequence *muntele Ceahlău* (*mount Ceahlău*), *Ceahlău* is a proper name, part of the NP *muntele Ceahlău*, but there is only one entity at the representational level, and this could be noted either `[muntele Ceahlău]` or `[mount Ceahlău]` (semantic representations are usually language independent). But NPs could have also a recursive structure, such that one NP may include one or more other NPs. In situations of the type $_{NP2}[...{}_{NP1}[...]_{NP1}...]_{NP2}$ (in our examples an entity is noted as a span of text in-between square brackets and marked with a double label to ease reading: $_{NP1}$`[span]`$_{NP1}$ and heads are underlined words), `NP1` and `NP2` will both be marked if and only if head(`NP1`) ≠ head(`NP2`), as here: $_{NP3}$`[clădirea` $_{NP2}$`[Universității din` $_{NP1}$`[Iași]`$_{NP1}$`]`$_{NP2}$`]`$_{NP3}$ ($_{NP3}$`[the building of` $_{NP2}$`[the University of` $_{NP1}$`[Iași]`$_{NP1}$`]`$_{NP2}$`]`$_{NP3}$). Also, when talking about "imbricated entities" we will mean entities realised in text by imbricated (or nested) NPs. In the above example, two of the three entities are of a geographical semantic nature (GE)s: `NP1` (a city) and `NP2` (an organisation).

A discussion may arise in the case of complex expressions such as *Western and Central Europe*, which could be seen as a group entity `[Western and Central Europe]` or the juxtaposition of two simple entities, `[Western Europe]` and `[Central Europe]`. In order to avoid the proliferation of group entities, by combining in all possible ways the elements of similar geographical sets, we adopted the second solution, by disambiguating *Western* in the context of *Europe*, even if the component parts are separated by other tokens.

4 Approach and architecture

A geographical entity is defined in our approach as a concept which can be associated with geographical characteristics, usually coordinates (point or bounding box) that are able to place it on the map, but potentially also: height, surface, population and others (see the Appendix for a comprehensive list of sources of additional information). In the context of our work, we look for geographical entities as referenced in texts, each textual reference being associated with specific geographical characteristics. Thus, from the perspective of our system, a text reference is equivalent to a geographical entity (which can have multiple equivalent text references). For all geographical entities annotated, we specify a type (general classification) and a subtype, specifying variations within a general type. In order to identify a geographical entity and its type and subtype, a three-step approach is used. The three steps are performed sequentially, as follows:

- a pre-processing phase performed over the original target document;
- a parallel application of a gazetteer module and a pattern-matching module;
- a merging and validation phase.

In *MappingBooks*, the pre-processing phase (PRE) involves several steps. First, the initial text is extracted from the original document (usually a PDF file including im-

ages and other non-textual content). This step involves the application of the iText[4] package, which leaves behind the text without formatting. Further on, the text is prepared by correcting diacritics and special characters, and eliminating end-of-line separators and other remains from the original format. Then, the corrected text is used as input for a chain of linguistic processes, adding the following markings: borders of lexical tokens, noun phrases and sentences, and part-of-speech categories and lemmas attached to tokens and compounds. For linguistic markings, the NLP-Group@UAIC-FII web-services[5] are used. The resulted annotated document (in stand-off XML format) serves as input for the next step, which passes through three other modules.

The gazetteer-applier (GAZ-APP) uses lists of toponyms and other geographic names, grouped by categories (usually called gazetteers – GAZ), to identify potential entity candidates. The result of this process is a document containing annotations for those surface names which are mentioned in GAZ, and where the type, subtype, coordinates, and other related geographical data, as found in the external resource, are added. Where ambiguous, a name will contain multiple tags, one for each category/subcategory and the disambiguation process is postponed.

In parallel, the patterns-applier module (PAT-APP) uses a set of patterns, described in terms of the markings left in the document by the PRE module, to discover potential geographical entities. The difference between PAT-APP and GAZ-APP is that the gazetteer makes use of strictly proper names, while the patterns include also contextual words that appear in their vicinity and which are used to reduce the ambiguities.

Finally, the merging and validation module (MER) compares the two annotated files to take final decisions of all markings.

5 The gazetteer

We have looked for a gazetteer that includes as many Romanian names as possible. In a first step, we identified a set of types and their subtypes, and then we attached to them lists of relevant names. Our list includes 15 major types:

There are nine types of relationships between two tags:

1. LOCATION (with 23 subtypes, covering all locations that are usually referenced on the map of a region: cities, ports, streets, etc.);
2. GEO_POSITION (with 6 subtypes, corresponding to map references: parallel, meridian, cardinal point);
3. GEOLOGY (with 6 subtypes, indicating geological formations visible on a specific map);
4. LANDFORM (with 16 subtypes, covering types of physiographic formations usually indicated on maps: mountain, valley, cave, etc.);
5. CLIME (with 5 subtypes, covering meteorological data shown on some types of maps);

[4] http://itextpdf.com/
[5] http://nlptools.info.uaic.ro/

6. WATER (with 11 subtypes for each variation of surface aquatic formation: river, lake, strait, etc.);
7. DIMENSION (with 9 subtypes, corresponding to the various ways in which geographical entities can be accompanied by (exact or approximated) values in text: height, depth, surface, etc.);
8. PERSON (names of people, accompanied by professions, where specified);
9. ORGANISATION (with 5 subtypes: military, education, etc., indicating also possible locations associated with a particular organisation type);
10. URL (web references);
11. TIMEX (dates, moments of time, intervals, etc.);
12. RESOURCE (with 4 subtypes, for natural resources associated with locations);
13. INDUSTRY (with 4 subtypes, for industrial areas: factories, electrical plants, etc.);
14. CULTURAL (with 6 subtypes, for cultural areas: museums, parks, etc.);
15. UNKNOWN (for other geographical entities not covered by the above types).

In total, we identified 103 types+subtypes, out of which 67 are of a geographical nature. To populate our gazetteer organized around the above types, we consulted a number of freely available resources. Among them, Geonames[6] is commonly used by many developers who need accurate geographical reference data. Developed on the base of various governmental and educational data sources and completed with user contributed and verified data, this open resource provides now gazetteer data for over 2.8 million entities, with 5.5 million alternative names. For Romania, the focus of our developments, Geonames includes 25.951 names, with over 45.000 alternative names, with a density of ~ 0.108 toponyms/km^2, and 1 toponym to ~ 1,000 inhabitants. The names are grouped in 9 types, with 654 subtypes. The 9 types are identified by letters:

- A: country, state, region, …
- H: stream, lake, …
- L: parks, area, …
- P: city, village, …
- R: road, railroad, …
- S: spot, building, farm, …
- T: mountain, hill, rock, …
- U: undersea, …
- V: forest, heath, …

For each of these types, besides the geographical coordinates, Geonames offers values for specific attributes, such as population (for P), surface (for A, H, L), height (for T), depth (for U), etc. In order to use these data to populate our gazetteer, we mapped our types/subtypes to those used by Geonames. As such, we found a many-to-one mapping between the 652 subtypes in Geonames and the 67 types/subtypes referring to the geography domain in our categorisation, to which are added the ones outside the domain of geography (DIMENSION, PEOPLE, ORGANISATION, URL,

[6] http://www.geonames.org/

etc.). For example, any of the following type.subtype in Geonames is categorised as our type.subtype LANDFORM.HILL:

- T.BUTE - butte(s): a small, isolated, usually flat-topped hill with steep sides;
- T.HLL - hill: a rounded elevation of limited extent rising above the surrounding land with local relief of less than 300m;
- T.HMCK - hammock(s): a patch of ground, distinct from and slightly above the surrounding plain or wetland;
- T.MND - mound(s): a low, isolated, rounded hill;
- T.PROM - promontory(-ies): a bluff or prominent hill overlooking or projecting into a lowland;
- T.MRN - moraine: a mound, ridge, or other accumulation of glacial till;
- U.HLLU - under-see hill: an elevation rising generally less than 500 meters.

The reduced number of types and subtypes in our classification theoretically should improve the precision of the GAZ-APP module, because of a lower classification ambiguity for each potential entity.

6 The pattern-matching module

The set of patterns (PAT) of the PAT-APP module were manually written using the Graphical Grammar Studio (GGS)[7] tool [21]. GGS is a framework for the development and processing of grammars, which has incorporated a constraint description language allowing the implementation of composite features, of look-ahead and look-behind assertions, and placing priority scores on arcs, forcing thus a preference order in processing paths. GGS has been designed with the main purpose to perform syntactical and sub-syntactical analysis. Its networks consume and annotate sequences of tokens or other XML elements. The input tokens can include any number of associated attributes (usually denoting part of speeches, lemmas, articles in cases of nouns and adjectives, tokens IDs, etc.), which are mentioned in the GGS networks to specify acceptance conditions over the sequences they receive in input.

GGS networks are structured as directed graphs. The nodes of these graphs express token consuming conditions and are linked by directed edges. Some nodes can make jumps to other sub-graphs. The networks are meant to be integrated into NLP chains, since they usually require some sort of pre-processed input (tokens annotated in some form). A GGS network is basically a finite state machine whose nodes can be associated with states. The PAT-APP module is a matching process that takes as input a sequence of XML elements and a GGS network and tries to find a path in the network from its starting node to its ending node.

An example of how such a pattern can be viewed in GGS is shown in Fig. 1. The sequence *Bărăganul este cea mai mică câmpie* (EN: *Bărăganul is the smallest plain*) is parsed by the above pattern following path 5, resulting in the first word *Bărăganul* as being annotated as `ENTITY` with `TYPE="LANDFORM"` and `SUBTYPE="PLAIN"`. Path 3 would match expressions like *câmpia cea mai mică*

[7] http://sourceforge.net/projects/ggs/

este Bărăganul (EN: *The smallest field is Bărăganul*), with the same annotations being added for *Bărăganul*.

The graphs are organized according to predetermined hierarchy of types, representing the 15 major categories, each of these presupposing the existence of other derivatives, with a total of 93 subcategories. A rule acts simultaneously for the identification and classification, combining contextual features found in tokens (like lemma, flexed word, etc.).

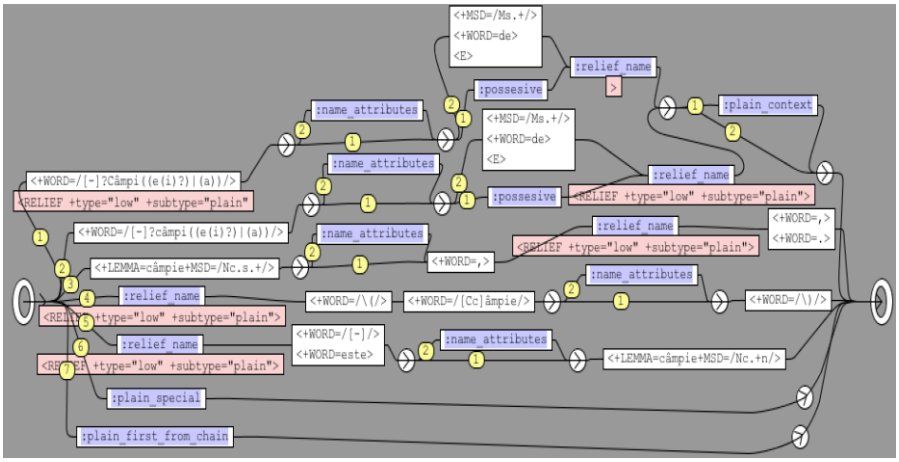


Fig. 1. A GGS network intended to recognize text references for geographical entities.

A concrete example when the GGS priority rules are applied is in the case of the sequence: *Universitatea Alexandru Ioan Cuza*. When processing this input, the pattern for the personal name recognition (*Alexandru Ioan Cuza* is a historic personage) has the lowest priority, and the preceding word *Univesitatea* forces the grammar to prefer a solution in which the whole expression is considered an educational institution: `TYPE="ORGANISATION", SUBTYPE="EDUCATION"`.

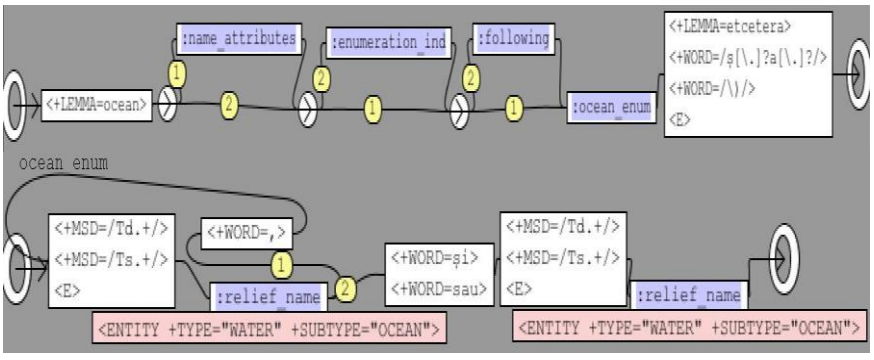


Fig. 2. A graph used to recognize an ocean name or an enumeration of such names, with the corresponding sub-graph for oceans that match elements of a list.

Enumerations are treated for each of the types/subtypes considered. An example matching this rule is the sequence (see Fig. 2.): *oceanele, de la mare la mic, sunt după cum urmează: Pacific, Atlantic, Indian, Antarctic și Arctic* (*the oceans, from biggest to smallest, are as follows:…*). In this case, every ocean is annotated as an `ENTITY` with `TYPE="WATER"` and `SUBTYPE="OCEAN"`.

7 Merging, validation and evaluation

Table 1 shows a comparison between the GAZ-APP and the GGS-APP with respect to the number of occurrences they are able to recognise in the same text (only the main types are counted).

Table 1. Comparison of recalls for the GSS-APP and the GAZ-APP modules

Types	GGS	GAZ
câmpie (field)	47	56
chei (canyon)	2	-
continent (continent)	31	-
deal (hill)	31	-
deltă (delta)	29	58
depresiune (depression)	58	-
județ (county)	51	-
lac (lake)	34	28
luncă (meadow)	1	-
mare (see)	57	-
munte (mountain)	294	74
ocean (ocean)	2	-
oraș (city)	344	756
persoană_f_x_m (person)	52	-
persoană_feminin (person_fem)	8	-
persoană_masculin (person_masc)	35	-
podiș (plateau)	41	4
râu (river)	360	128
regiune (region)	114	220
sat (village)	240	934
țară (country)	129	-
vârf_montan (peak)	21	118
Total	1981	2376

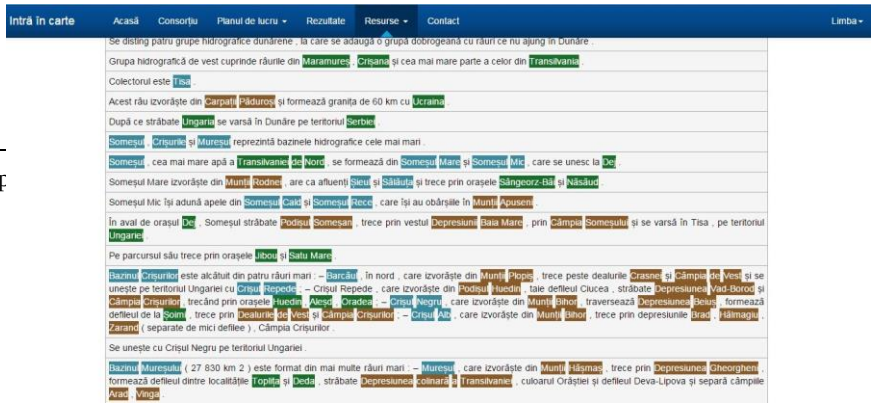
As can be seen, in general GGS-APP covers better certain categories than GAZ-APP, although, globally, GAZ-APP supersedes GGS-APP. This means that a proper treatment would be to combine the two processes. This observation actually let to the decision to include a merging and validation module, which follows both in the processing chain. The outputs from the two parallel processes, GAZ-APP and PAT-APP, are compared and validated by the MER module. The following cases are examined by MER:

- both GAZ-APP and PAT-APP annotate the same text span and the tag left by PAT-APP is among those left by GAZ-APP ⇒ the common tag is copied in the output file;
- both GAZ-APP and PAT-APP annotate the same text span and the tag left by PAT-APP is not among those left by GAZ-APP ⇒ the PAT-APP tag is copied in the output file;
- the text span annotated by GAZ-APP is included in the one annotated by PAT-APP and the tag left by PAT-APP is among those left by GAZ-APP ⇒ the common tag is copied on the largest text span in the output file;
- the text span annotated by GAZ-APP is included in the one annotated by PAT-APP and the tag left by PAT-APP is not among those left by GAZ-APP ⇒ the PAT-APP tag is copied on the largest text span in the output file;
- these is an intersection between the text spans annotated by the two modules and the tag left by PAT-APP is among those left by GAZ-APP ⇒ the common tag is copied on the union of the text spans in the output file;
- these is an intersection between the text spans annotated by the two modules and the tag left by PAT-APP is not among those left by GAZ-APP ⇒ the PAT-APP tag is copied on the union of the text spans in the output file;
- only one of the two modules annotate a certain text span with one or more tags ⇒ one tag out of those annotated is chosen randomly for that text span in the output file.

The criteria above show that, generally, more credibility is given to the PAT-APP module than to the GAZ-APP module, on the base that it uses the context to disambiguate names.

8 Web applications and services

In order to make an easy and better analysis of the system, a Name Entity Viewer was developed as a web application. The viewer is hosted on the *MappingBooks* project page[8]. It presents entities by highlighting them in different colors, as seen in Fig. 3.



[8] http

Fig. 3: The interface of the Entity Viewer

Usually an annotator uses the entity viewer in tandem with a Name Entity Editor, also hosted by the *MappingBooks* project pages[9]. The user has the possibility to upload a new text file, let the system identify the entities and then correct them, by changing entity boundaries (continuous strings of tokens), their types and subtypes.

The default downloading format is *stand-off*. For example, for the text *Relieful **României** este definit de mai multe caracteristici* (*Romania's relief is defined by several characteristics*), the resulted XML is as follows:

```
<DOCUMENT>
<P ID="p1" offsetStart="0" offsetStop="58"/>
<S ID="s1" offsetStart="0" offsetStop="58"/>
<W Case="direct" Definiteness="yes" Gender="masculine" ID="w1.1"
 LEMMA="relief" MSD="Ncmsry" Number="singular" POS="NOUN" Type="
common" offsetStart="0" offsetStop="8" text="Relieful"/>
<W Case="oblique" Definiteness="yes" Gender="feminine" ID="w1.2"
 LEMMA="românie" MSD="Ncfsoy" Number="singular" POS="NOUN" Type=
"common" offsetStart="9" offsetStop="17" text="României"/>
<W EXTRA="intranzitiv" ID="w1.3" LEMMA="fi" MSD="Vmip3s" Mood="i
ndica-
tive" Number="singular" POS="VERB" Person="third" Tense="present
" Type="predicative" offsetStart="18" offsetStop="22"text="este"
/>
<W Case="direct" Definiteness="no" EXTRA="ParticipleLemma:defini
(tranzitiv)" Gender="masculine" ID="w1.4" LEMMA="definit" MSD="A
fpmsrn" Number="singular" POS="ADJECTIVE" offsetStart="23"offset
Stop="30" text="definit"/>
<W ID="w1.5" LEMMA="de" MSD="Sp" POS="ADPOSITION" offsetStart="3
1" offsetStop="33" text="de"/>
<W ID="w1.6" LEMMA="mai" MSD="Rg" POS="ADVERB" offsetStart="34"
offsetStop="37" text="mai"/>
```

[9] http://85.122.23.18:8181/MappingBooks/resources/editor

```
<W Case="direct" Gender="feminine" ID="w1.7" LEMMA="mult" MSD="D
i3fpr" Number="plural" POS="DETERMINER" Person="third" Type="ind
efinite" offsetStart="38" offsetStop="43" text="multe"/>
<W Case="direct" Definiteness="no" Gender="feminine" ID="w1.8" L
EMMA="caracteristică" MSD="Ncfprn" Number="plural" POS="NOUN" Ty
pe="common" offsetStart="44" offsetStop="58"text="caracteristici
"/>
<ENTITY ID="e0" SUBTYPE="VILLAGE" TYPE="LOCATION" WORDSID="w1.2"
 offsetStart="9" offsetStop="17"/>
</DOCUMENT>
```

Developed as a web application, the Entity Editor is also platform-independent (available for a variety of operating systems including Windows, Mac OS and Linux).

An API allows the user to process books or large pieces of text and upload them on the site for subsequent queries. This implementation allows "live" entity type classifications, initiated by queries directly addressed by users[10]. The presented approach was adopted to make it suitable for online querying of huge texts.

The web services are integrated in the MultiDPS platform [1], which is a Service-Oriented-Architecture that provides also tools for visualization of annotations, in a user-friendly manner.

9 Conclusions

We have presented in this paper an approach to build a sophisticated NER module for geographical entities that appear in Romanian texts. Its design is based on a combination between a brute-force approach (the use of an extended list of proper names) and a regular expressions approach (the use of a collection of manually written rules). The final decision to accept or reject an annotation over a span of words as being a geographical entity depends on the acceptance of more constraints, which are verified by a merge and validation module. For evaluation, the output of the merge module is compared against a test corpus, manually annotated. The results of the comparison are used to raise the quality of both resources (the gazetteer and the collection of patterns), in a bootstrapping enhancement loop, which is still on-going.

The work reported at this point is still preliminary and we don't want to risk conclusions regarding the accuracy of our system. However, the whole architecture is built on the presumption that the NER module could be made perfectible within the constraints imposed by the *MappingBooks* application.

We have a number of ideas that could guide an enhancement process: first, the more credibility that we give now to the PAT-APP module in its competition with the GAZ-APP module should be better contextualised and parameterised, by using more examples and training. Then, the random decision that we take now when we are left with more solutions should also be replaced by a biased decision, based on a thorough statistical study. Furthermore, the borders established for each entity should corre-

[10] The API allows also identification of relations between entities, a facility not described in this paper.

spond to one of the NPs borders, i.e. the span of a name entity should always be equal to one existent noun phrase (conclusion left after examining the manually annotated corpus). But the NP-chunker is itself prone to errors and we believe that the result of the NER could correct the decisions previously taken by this module. Apparently, this kind of corrections should also be left at the handle of the MER module. When this study will be finished we hope to provide a thorough individual evaluation of the three component modules in relation with the manually annotated corpus.

Appendix

It is worth mentioning that, in *MappingBooks*, the identified geographical entities are intended to be used as location points on the document, linking them with actual maps or external web links, or participating in relevant semantic relations. As a repository of spatial data GeoNetwork[11] was used, an open source platform that allows creating catalogues of spatial data, searching and storing their spatial metadata. The application is based on the principles of FOSS (Free and Open Source Software) and implements international standards (ISO / TC211 and OGC). The GeoNetwork application, running as a service server, stores the data in a database and provides a web interface through which the user can access catalogues of view spatial data and publishing spatial data, or can enter, visualise and edit metadata associated with the geospatial data.

Our intention is to attach to the recognised geographical entities different types of information, found on public sources. For this we are spotting a number of possible sources of free geospatial data: *Natural Earth*[12] – a set of cultural, physical and raster layers data, generalized for three spatial scales: 1:10 millions, 1:50 millions and 1:110 millions; *Romanian geomorphological regionalization*[13], digitised after a number of analogic versions; *Open Street Map*[14] – a dataset created by the community, open to anyone for contribution and editing, containing points of interests (POI), lines and polygons representing different types of spatial entities complemented with more information; *Bing Maps*®[15] – a product of Microsoft®, providing a WMS service with maps and aerial images and a Geocoding service, with a suite of data licenses, which, to some extent, can be used for personal and educational purposes; *Wikipedia*[16] containing in addition to the related text for each word, a location, as geographical coordinates, for toponyms; the Romanian SDI and the Romanian INSPIRE geoportal[17] crated by the *National Agency for Cadastre and Registration*[18], through the National Geodetic Fund and several collaborations; *Data.gov.ro* – a portal of partially

[11] http://geonetwork-opensource.org/
[12] http://www.naturalearthdata.com/
[13] http://earth.unibuc.ro/download/harta-unitati-relief-romania
[14] http://www.openstreetmap.org
[15] www.bing.com/maps
[16] http://ro.wikipedia.org
[17] http://geoportal.ancpi.ro
[18] ANCPI – http://www.ancpi.ro

geospatial data produced by Romanian government agencies (the SIRUTA national codes for administrative units); statistical data provided by the *National Statistics Institute*[19] – the Romanian national statistics service, linkable to geospatial boundaries: the TEMPO database[20], the eDemos database[21], the IDDT database[22] of sustainable development indexes[23], etc.

Also, *compiled datasets* can be produced by linking statistical databases with geospatial data, or through generalisation or other kinds of spatial analysis. For most datasets global processing is needed for cutting the region of interest, or to possibly change the format and projection.

Acknowledgement

The work was published with the support of the PN-II-PT-PCCA-2013-4-1878 Partnership PCCA 2013 grant *MappingBooks – Jump in the Book!*, having as partners the "Alexandru Ioan Cuza" University of Iași, SIVECO S.R.L. Bucharest and "Ștefan cel Mare" University of Suceava.

References

1. Anechitei A. Daniel (2014). *MultiDPS - A multilingual Discourse Processing System*. In Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations (COLING-2014), Dublin, Ireland, August 2014, pages 44-47
2. Borthwick, A., Sterling, J., Agichtein, E., Grishman, R.: Exploiting Diverse Knowledge Sources via Maximum Entropy in Named Entity Recognition. In Proceedings of the 6th Workshop on Very Large Corpora (1998)
3. Collins, M., Singer, Y.: Unsupervised Models for Named Entity Classification. In: Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC), College Park, MD, Association for Computational Linguistics, pp. 100–110 (1999)
4. Cristea, D, Gîfu, D., Colhon, M., Diac, P., Bibiri, A.-D., Mărănduc, C. and Scutelnicu A.-L.: Quo Vadis: A Corpus of Entities and Relations. In N.Gala, R.Rapp and G.B.Enguix (eds.): *Language Production, Cognition, and the Lexicon*, Springer International Publishing Switzerland (2015)
5. Cunningham, H., Maynard, D., Bontcheva, K., and Tablan, V.: GATE: A framework and graphical development environment for robust NLP tools and applications. In Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (2002)
6. Florian, R., Ittycheriah, A., Jing, H., Zhang, T.: Named entity recognition through classifier combination, in Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4, pp. 168-171 (2003)

[19] http://www.insse.ro
[20] https://statistici.insse.ro/shop/?lang=ro
[21] http://edemos.insse.ro/portal
[22] http://www.insse.ro/cms/files/IDDT%202012/index_IDDT.htm
[23] http://www.insse.ro/cms/files/Web_IDD_BD_ro/index.htm

7. Fung, P.: A Pattern Matching Method for Finding Noun and Proper Noun Translations from Noisy Parallel Corpora. Proc. Association for Computational Linguistics (1995)
8. Grishman, R. and Sundheim, B.: Message understanding conference - 6: A brief history. In Proceedings of COLING (1996)
9. Huang, F.: Multilingual Named Entity Extraction and Translation from Text and Speech. Ph.D. Thesis. Carnegie Mellon University (2005)
10. Iftene, A., Trandabăţ, D., Toader, M., Corîci, M.: Named Entity Recognition for Romanian. In Proceedings of the 3th Conference on Knowledge Engineering: Principles and Techniques Conference (KEPT2011). In Studia Universitatis, Babeş-Bolyai, Vol. 2, Cluj-Napoca, Romania, pp.19-24 (2011)
11. Kraak, M.-J., Rico, V.D.: Principles of hypermaps, Computers & Geosciences, 23(4), pp. 457-464 (1997)
12. Lee, S. and Lee, G.G.: A Bootstrapping Approach for Geographic Named Entity Annotation in Information Retrieval Technology Lecture Notes in Computer Science Volume 3411, pp 178-189 (2005)
13. Li, H., Srihari, R.K., Niu, C., Li, W.: InfoXtract location normalization: a hybrid approach to geographic references in information extraction. In: Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References, Alberta, Canada, pp. 39-44 (2003)
14. Mann, Gideon S. and Yarowsky, D.: Unsupervised Personal Name Disambiguation. Proceedings of the 9th Conference on Computational Natural Language Learning (2003)
15. Masayuki, A., Matsumoto, Y.: Japanese: Named Entity Extraction with Redundant Morphological Analysis. In Proceedings of the Human Language Technology conference – North American chapter of the Association for Computational Linguistic (2003)
16. Maynard, D., Tablan, V., Ursu, C., Cunningham, H., and Wilks, Y.: Named Entity Recognition from Diverse Text Types. In Recent Advances in Natural Language Processing 2001 Conference, pp. 257–274, Tzigov Chark, Bulgaria (2001)
17. Mikheev, M., Grover, C. and Moens, M.: Description of the LTG System Used for MUC-7. In Proceedings of 7th Message Understanding Conference (MUC-7) (1998)
18. Nadeau, D., Sekine, S. A.: Survey of Named Entity Recognition and Classification (2007)
19. Nadeau, David and Turney, P. A.: Supervised Learning Approach to Acronym Identification. Proceedings of the 18th Canadian Conference on Artificial Intelligence (2005)
20. Sekine, S., Grishman, R., Shinnou, H.: A Decision Tree Method for Finding and Classifying Names in Japanese Texts. In: Proceedings of the Sixth Workshop on Very Large Corpora (1998)
21. Simionescu, R.: Graphical grammar studio as a constraint grammar solution for part of speech tagger", in Proceedings of the International Conference *Resources and Tools for Romanian Language* – ConsILR-2011, Bucharest, "Alexandru Ioan Cuza" University of Iaşi Publishing House, (2011)
22. Smith, D.A., Mann, G.S.: Bootstrapping toponym classifiers. In: Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References, Alberta, Canada, pp. 45–49 (2003)
23. Yangarber, R., Lin, W., Grishman, R.: Unsupervised Learning of Generalized Names. In: Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002), Taipei, Taiwan, pp. 1135–1141 (2002)