NER-FL: A Novel Named Entity Recognizer of Farsi Language using the Web-Based Natural Language Processors and Semantic Annotations

Babak Farhadi
Department of Computer engineering, University of Tehran
Tehran, Iran

ABSTRACT

Named Entity Recognition is a main task in the NLP area that has yielded multiple web-based natural language processors gaining popularity in the Semantic Web community for extracting knowledge from web data. These processors are generally located as pipelines, using dedicated APIs and various taxonomy for extracting, classifying and disambiguating named entities. In this paper, we address the problem of NER on Farsi language by proposing NER-FL, a novel semantic framework which unifies three popular named entity extractors available on the web, and the NER-FL ontology which provides a rich set of axioms aligning the taxonomies of these web natural language processors automatically on the LOD-cloud.

KEYWORDS

Web-based natural language processor, named entity, semantic web, ontology, Farsi language

1. INTRODUCTION

The Web of Data is often illustrated as a fast growing cloud of interconnected dataset representing information about barely everything [4]. The web hosts millions of unstructured data such as news articles, scientific papers, as well as forum and archived mailing list threads that have written to Farsi language. This information has usually a semantically structure which is obvious for the human being but that remains mostly hidden to computing machinery. On the other hand, web-based Natural Language Processing (NLP) tools aim to extract such a structure from those free texts (on a few languages, except the Farsi language). They provide algorithms for analyzing nuclear information elements which occur in a sentence and identify Named Entity (NE). They also classify these entities according to predefined schema increasing discoverability and reusability of information. In this regard, Named Entity Recognition (NER) labels sequences of words in a text which are the names of things, such as person and company names, or Place and cities names.

Newly, research and commercial communities have spent efforts to publish web-based natural language processors on the web of data. Beside the common task of POS tagging, they provide more disambiguation facility with URIs that describe web resources, leveraging on the web of real world objects. Further, these processors classify such information using common ontologies (e.g. DBpedia ontology) exploiting the large amount of knowledge available from the web of

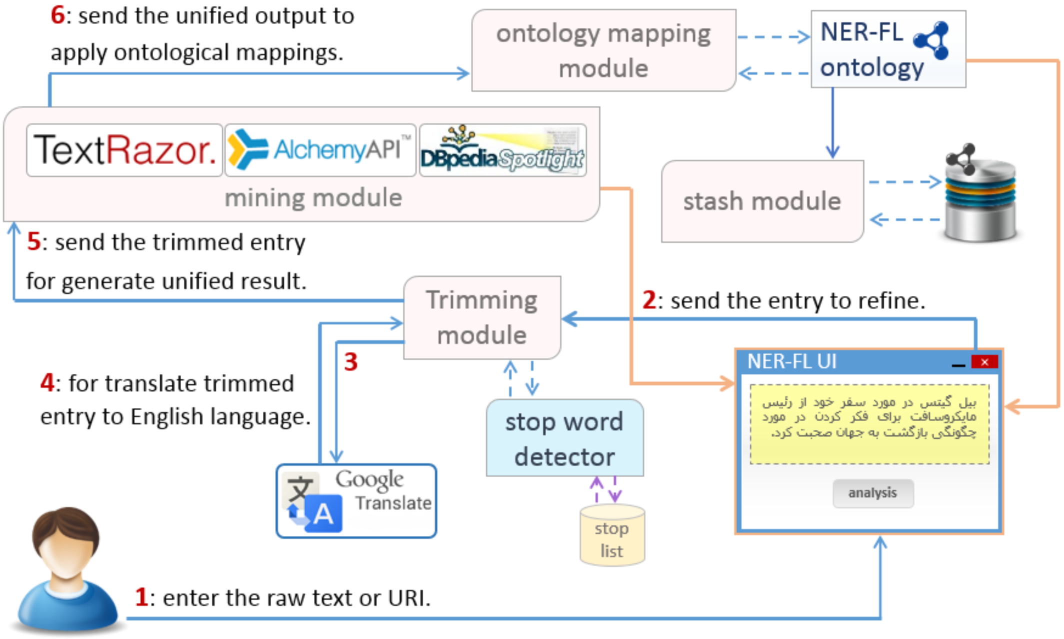

6: send the unified output to
apply ontological mappings.
ontology mapping
module
NER-FL
ontology
TextRazor.
AlchemyAPI
DBpedia Spotlight
mining module
stash module
5: send the trimmed entry
for generate unified result.
Trimming
module
2: send the entry to refine.
3
4: for translate trimmed
entry to English language.
stop word
detector
Google
Translate
stop
list
NER-FL UI
بیل گیتس در مورد سفر خود از رئیس
مایکروسافت برای فکر کردن در مورد
چگونگی بازگشت به جهان صحبت کرد.
analysis
1: enter the raw text or URI.

Fig 1: The basic data flow process in proposed NER-FL.

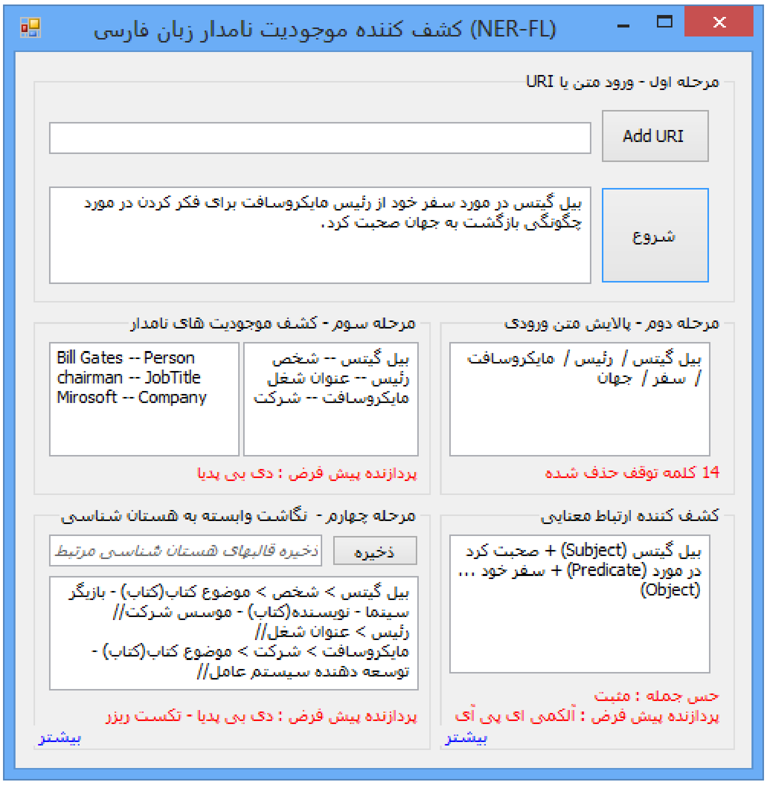
data. The famous web-based natural language processors such as AlchemyAPI, DBpedia Spotlight and TextRazor represent a clear opportunity for the web community to increase the volume of interconnected semantic data. Although these extractors share the same goals such as extract NE and relation from text, classify and disambiguate this information, but they make use of different algorithms and provide different outputs.

In here, role of semantic web technologies is to make implicit meaning of content explicit by providing suitable metadata annotations based on formal knowledge representations. In this area, Linked Data (LD) means to expose, share and connect pieces of data, information and knowledge on the semantic web using Uniform Resource Identifiers (URI) for identification of resources and RDF as a structured data format. It is creating the relationships from the data to other sources on the Web. These data sets are not only accessible by human beings, but also readable for machines. LOD can aims to publish and connect open but heterogeneous databases by applying the LD principles. The aggregation of all LOD data set is denoted as LOD-cloud [1].

This paper presents NER-FL (Named Entity Recognizer of Farsi Language), a novel framework that unifies the output of three diverse web-based natural language processors publicly available on the web. Our input text is in Farsi language and the framework relies on the development of the NER-FL ontology that provides an effective interface for annotating Farsi elements, and a web REST API which is used to access the unified output of these processors.

2. RELATED WORK

One of the first research papers in the NLP field, aiming at automatically identifying named entities in texts, was proposed by [5]. A natural evolution of this approach, mainly driven by the Semantic Web community, consists in disambiguating named entities with data from the LOD cloud. In [6], the authors proposed an approach to avoid named entity ambiguity using the DBpedia dataset. Many approaches have already presented NER as LD, which offers experience for us on textual resource representing. However, all the proposed approaches are presented ontological features full-manually. On the other hand, all the papers that proposed on Farsi language NER, don't use the utilization of the well-known web-based natural language processors and related semantic annotations. In [3] proposed an approach using the 10 web-based natural language processors on English language. They only used the portion of NE extraction of these processors and in very restricted types. By having an overview on their demo, end client realizes that he/she can’t use advantages RDF syntaxes outputs, SPARQL GUI and triple store module, it’s unlike the proposed approach in [2]. About the NE extraction, in [3] haven’t grabbed any NE subtypes and other derivatives of entity extraction portion. Hence, their NER module has answers of chopped and with restricted types of dependent on NEs. Using all the main portions of the web-based natural language processors and efficient ontological mappings, boosting resulted NEs and also utilizing outcomes by RDF-GM and attached modules of related to it, can be eventuated in an enriched entity set.


کشف کننده موجودیت نامدار زبان فارسی (NER-FL)
مرحله اول - ورود متن یا URI
Add URI
شروع
مرحله دوم - پالایش متن ورودی
مرحله سوم - کشف موجودیت های نامدار
کشف کننده ارتباط معنایی
مرحله چهارم - نگاشت وابسته به هستان شناسی
ذخیره
بیشتر
بیشتر

بیل گیتس در مورد سفر خود از رئیس مایکروسافت برای فکر کردن در مورد
چگونگی بازگشت به جهان صحبت کرد.

بیل گیتس / رئیس / مایکروسافت
/ سفر / جهان

14 کلمه توقف حذف شده

بیل گیتس -- شخص
رئیس -- عنوان شغل
مایکروسافت -- شرکت

Bill Gates -- Person
chairman -- JobTitle
Mirosoft -- Company

پردازنده پیش فرض : دی بی پدیا

بیل گیتس (Subject) + صحبت کرد
در مورد (Predicate) + سفر خود ...
(Object)

حس جمله : مثبت
پردازنده پیش فرض : آلکمی ای پی آی

ذخیره قالبهای هستان شناسی مرتبط

بیل گیتس > شخص > موضوع کتاب(کتاب) - بازیگر
سینما - نویسنده(کتاب) - موسس شرکت//
رئیس > عنوان شغل//
مایکروسافت > شرکت > موضوع کتاب(کتاب) -
توسعه دهنده سیستم عامل//

پردازنده پیش فرض : دی بی پدیا - تکست ریزر

Fig 2: Some overviews on NER-FL user interface and sample results in a raw text

3. FRAMEWORK OVERVIEW

The NER-FL web application is in visual C# 2013 and it requires to input a raw text or URI of a web document which is analyzed in order to extract its main textual content. NER-FL (proposed) architecture follows the REST principles and provides a web HTML access for humans and a semantic integrator module for machines to exchange semantic content in all of the structured formats such as XML, JSON, RDF and RDFS. Both interfaces are powered by the NER-FL integrator module. The Figure 1 shows the workflow of an interaction among end clients (humans or machines), the NER-FL integrator module and different web-based natural language processors which are used by NER-FL for extracting NEs and providing a type and disambiguation URIs pointing to real world objects as they could be defined in the web of data.

3.1 NER-FL in interface level

Starting from the raw text, the NER-FL interface drives every three processors to extract and integrate the list of Named Entity, classification and ontological mappings that disambiguate these entities. The main purpose of this interface is to enable a human user to distinguish the quality of the extraction results collected by those processors. We've used Google translate API (with auto spell checking feature) for converting and translating of the inputted Farsi text to the appropriate English text. In addition, we've collected 3600 Farsi stop words which would appear to be of little value in helping select texts matching a user need are excluded from the vocabulary entirely. The general strategy of NER-FL (proposed) for determining a stop list is to sort the text terms by text collection frequency (the total number of times each text term appears in the text collection), and then to take the most frequent text terms, often hand-filtered for their semantic content relative to the domain of the texts being processed, as a stop list, the human users of which are then discarded during processing.

The API interface is developed following the REST principles and aims to enable programmatic access to the NER-FL framework. GET, POST and PUT methods manage the requests coming from end clients to retrieve the list of NEs, classification types and URIs for a specific processor or for the combination of them.

3.2 NER-FL integrator module

The NER-FL integrator module is composed of four main modules, namely: trimming, mining, ontology mapping and stash. The trimming module takes as input the raw text or URI and refine it to the main textual content. Mining is the module designed to invoke the web-based natural language processors and collect the unified results. Each processor provides its own taxonomy of named entity types it can recognize. Therefore we designed the NER-FL ontology that provides a set of mappings between these various classifications. The ontology mapping is the module in charge to mapping the classification type retrieved to the NER-FL ontology. The stash module saves all of the structured outputs and semantic user annotations toward the user disk storage, according to the approach in [2].

3.3 NER-FL Ontology

As mentioned, the web-based natural language processors use different algorithms and their own classification taxonomies which makes complex their comparison. To solving this problem, we've developed the NER-FL ontology which is a set of mappings established automatically between the ontological features of the NE categories. Concepts included in the NER-FL ontology are collected from various ontological features of DBpedia Spotlight, AlchemyAPI, and TextRazor. The NER-FL ontology becomes a reference ontology on Farsi language for comparing the classification and ontological tasks of the well-known NE processors. For instance, The DBpedia Ontology is a shallow, cross-domain ontology, which has been manually created based on the most commonly used info boxes within Wikipedia. Its ontology currently covers 529 classes which form a subsumption hierarchy and are described by 2,333 various properties. It currently contains about 3,220,000 instances and they are in classes of Place, Person, Work, Species and Organization. In this regard, NER-FL ontology automatically place entities into an ontology of thousands of categories derived from LD resources.

4. EVALUATION

We handled an experiment to evaluation of the alignment of the NER-FL framework according to the ontology we developed. For this purpose, we collected 1200 news articles (in Farsi version) of the Fars News Agency from 01/03/2014 to 30/04/2014 and we performed the extraction of NEs with the processors supported by NER-FL. The goal is to explore the NE extraction patterns with this dataset and to evaluation of the commonalities and differences of the classification ontological features used. We propose the alignment of the four common main types recognized by all the processors using the NER-FL ontology.

Table 1: Number of axioms aligned for all the processors involved in the comparison according to the NER-FL ontology

	DBpedia Spotlight	TextRazor	AlchemyAPI
Place	1,340	593	2,840
Person	8,422	6,434	1,760
Organization	5,200	1,365	871
Country	568	237	182

To handle this experiment, we used the default configuration for all processors used. We define the following variables: the number n_d of evaluated documents, the number n_w of words, the total number n_e of entities, the total number n_c of categories and n_u URIs. in addition, we compute the following metrics: word detection rate $r(w, d)$, i.e. the number of words per document, entity detection rate $r(e, d)$, i.e. the number of entities per document, entity detection rate per word, i.e. the ratio between entities and words $r(e, w)$, category detection rate, i.e. the number of categories per document $r(c, d)$ and URI detection rate, i.e. the number of URIs per document $r(u, d)$. The evaluation we performed concerned n_d = 1200 documents that amount to n_w = 834,487 words. The word detection rate per document $r(w, d)$ is equal to 695.4 and the total number of recognized entities n_e is 375,268 with the $r(e, d)$ equal to 312.7. Finally $r(e, w)$ is 0.4496 and $r(u, d)$ is 62.354.

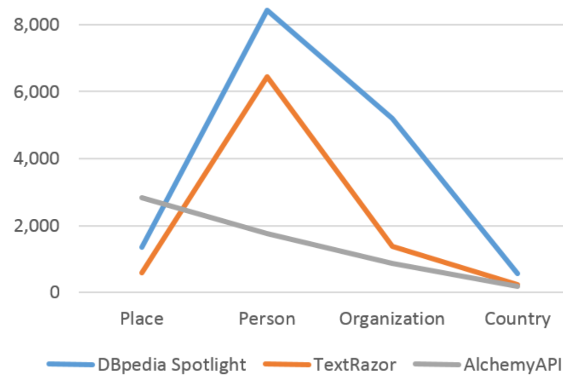


Fig 3: The experimental results on NER-FL ontology per four type (obtained from table1).

6. CONCLUSION

In this paper, we implemented NER-FL (proposed), a novel framework for apply NER features of the web-based natural language processors on Farsi language developed following REST principles. Further, we've presented NER-FL ontology, a reference ontology to mapping several NER processors publicly accessible on the web of data. We propose a preliminary comparison results where we investigate the importance of a reference ontology in order to assess the strengths and weaknesses of the NER processors. We'll investigate whether the combination of extractors may overcome the performance and efficiency of a single processor or not. For future works, we plan to apply our proposed approach to video classification field.

7. ACKNOWLEDGMENTS

I'd like to acknowledge the contributors to this project such as service supporters of the famous web-based natural language processors such as AlchemyAPI, DBpedia Spotlight, TextRazor and many others that helped me in implementing this project.

8. REFERENCES

[1] B. Farhadi and M. B. Ghaznavi Ghoushchi, "Creating a Novel Semantic Video Search Engine through Enrichment Textual and Temporal Features of Subtitled YouTube Media Fragments, " in 3rd International conference on Computer and Knowledge Engineering, 2013.

[2] B. Farhadi, "Enriching Subtitled YouTube Media Fragments via Utilization of the Web-Based Natural Language Processors and Efficient Semantic Video Annotations," Global Journal of Science, Engineering and Technology, pp. 41-54, 2013.

[3] Rizzo G. and Troncy R, " NERD: A Framework for Evaluating Named Entity Recognition Tools in the Web of Data, " 10th International Semantic Web Conference (ISWC'11), 2011.

[4] R. Cyganiak and A. Jentzsch, "Linking open data cloud diagram," LOD Community (http://lod-cloud. net/), 2011.

[5] L. F. Rau, "Extracting company names from text," in Artificial Intelligence Applications, 1991. Proceedings., Seventh IEEE Conference on, 1991, pp. 29-32.

[6] P. N. Mendes, M. Jakob, A. García-Silva, and C. Bizer, "DBpedia spotlight: shedding light on the web of documents," in Proceedings of the 7th International Conference on Semantic Systems, 2011, pp. 1-8.

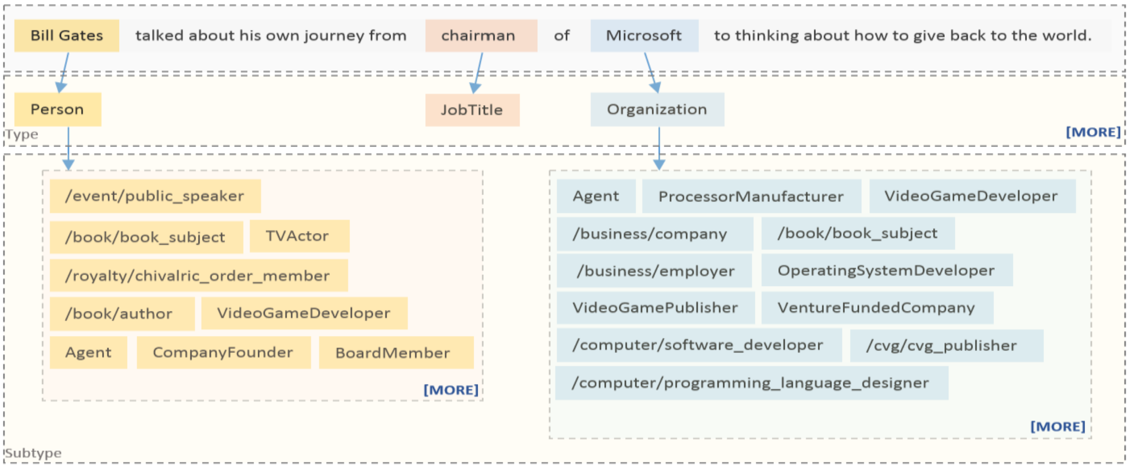


Fig 4: Ontological types and sub types of the entry text in figure 2