Ontology Creation and Population for Natural Language Processing Domain

Niloofar Naderian, Mehrnoush Shamsfard*, Razieh Adelkhah
Faculty of Computer Science and Engineering, Shahid Beheshti University, Tehran, Iran
n.naderian@mail.sbu.ac.ir, m-shams@sbu.ac.ir, r.adelkhah@mail.sbu.ac.ir

Received: 2019/06/08 Revised:2019/06/28 Accepted:2019/07/16

***Abstract*— In this paper, we describe our proposed methodology for constructing an ontology of natural language processing (NLP). We use a semi-automatic method; a combination of rule-based and machine learning techniques; to construct and populate an ontology with bilingual (English-Persian) concept labels (lexicon) and evaluate it manually. This methodology results in a complete ontology in the natural language processing domain with 1333 classes (containing concepts, tools, applications, etc.), 88 object properties, and 2437 annotation assertions for different classes. The built ontology is populated with about 428K NLP related papers and 38K authors, and also about 5M "is Related to" relations between papers and ontology classes and 1M "is Author of" relations between papers and authors. The evaluation results show that the ontology achieved a good result. The instantiation is done to enable applications find experts, publications and institutions (such as universities or research laboratories) related to various topics in NLP field.**

Keywords*— Domain Ontolog, Ontology Construction, NLP Ontology.

1. INTRODUCTION

Ontologies, which are abstract models of a world and specify concepts and relationships between them, can be used to access information appropriately and provide accurate access to information based on meaning [1].

Information access is one of the main requirements for people and organizations. Nowadays, the world faces with the rapid growth in the number and diversity of research activities, scientific resources, publications and experts. Without automatic methods and systems for information access including search engines, expert finders, summarizers, translators, and knowledge extractors accessing and using this huge amount of information is rather impossible. Domain-specific ontologies are one of the essential resources for such systems. They can help us to resolve knowledge-based queries.

In this paper, we focus on the construction of a bilingual ontology for Natural Language Processing (NLP) domain.

To construct and populate the NLP ontology, we employed a semi-automatic method, which will be discussed. The employed method is language and domain independent so can be applied to any domain or language as well. The resulted ontology is revised manually and is going to be used in an expert finding system.

There are some datasets and ontologies in different domains. General Ontology for Linguistic Description (GOLD) ([2]) is the most significant model for the scientific description of human languages. Pisarev. and Kotova [3] have constructed a thematic ontology while representing a method for the automated thesauri development. Amini and colleagues [4] have proposed a method for integrating of multiple domain taxonomies to build a reference ontology to be used in profiling shholrrs' bcckground knowdddge. iii a ontologiss[5] serve as a reference hub for annotation terminology for linguistic phenomena on a great bandwidth of language within the Linguistic Linked Open Data (LLOD) cloud. OnLit [6] is a data model, which can be used to represent linguistic terms and concepts in a semantically interrelated data structure. Despite all of the mentioned works, to the best of our knowledge, there is no NLP ontology with such a wide coverage as ours. The NLP ontology that we provide includes many of the related terms to the domain, that a researcher or author could possibly use or mention in research works. The ontology classifies the domain terms from an academic and technical point of view, which not only can be used in different kinds of applications, but also demonstrate the domain by a complete categorical glossary of the related terms.

For the development of the ontology, we use a revised version of the ontology design and evaluation methodology of [7]. As an extension of our previous work [8], We specify our ontology building methodology in 6 steps in the next sections: determine the scope and provide competency questions in 2, extract concepts in 3, define class hierarchies and relations in 4, completion and integration in 5, final review in 6, and population in 7. The architecture of the system is shown in Figure 1.

2. COMPETENCY QUESTIONS

The first step in ontology construction methodology is determining the domain, scope, application of the ontology and the competency questions it should be able to answer. As it was discussed, although the built ontology can be used in many applications, our aim is to use it in an expert finding application. The application is going to be used for finding reviewers for journals and conferences for a given manuscript, finding supervisor,

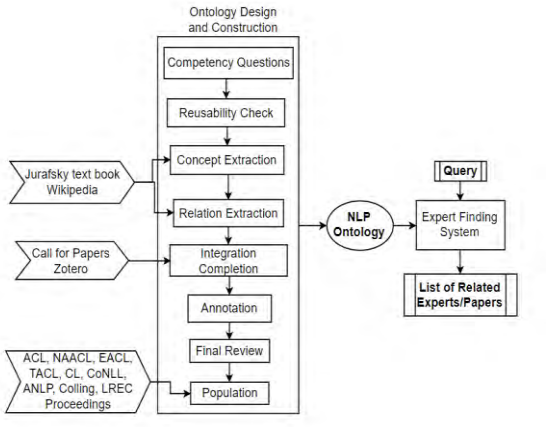

Fig 1 The Architecture of the System

advisor or consultant in a specific sub-domain of NLP and also finding relevant publications in a domain of interest.

Some of the competency questions the ontology should answer are as following:

- What are the main concepts and topics in natural language processing domain?

  What resources could be used in different applications of natural language processing?

- What topics are needed in applications of the domain?

  What approaches are used for solving different problems of the domain?

  What are the expected skills of the expert in each field?

  What is the expertise of an individual?

  What are the expected skills of the author of a publication?

  What is the topic of an application if it uses each tool?

  What are the resources used in a specific approach or a specific piece of work?

  Who is the expert person in each field of the domain?

  What are the most relevant publications in a specific topic?

3. Concept Extraction

3.1. Concept Categories

For building the NLP ontology, we use several resources for different parts of the ontology. We try to create an ontology that contains comprehensive and complete information in a variety of areas that a scientific field can have.

Generally, every scientific domain, independent of the subject, have some core topics of the domain, each of which could have a large number of subtopics and other kinds of resources, that integration of them defines and describes the main scientific domain. Also, each domain could have a number of tools that make the achievements of researches and experiments applicable for other users. In addition, every domain has several usable resources such as data resources (as databases), or research resources (as scientific books or articles) and useful models in experiments (as physical or conceptual representations of a system).

Now consider the scientific domain of Natural Language Processing (NLP).

We know the main steps of processing a language oonssst of “Lxxcdl Anyyysis,, “yyncccccc Anyyysss,, “mmmnn... A yyysss,, “aaagmcccc Anyyysss,, “sss oo. rse Anll ysss” and so on. Each of these steps is one of the main core topics of NLP, which contains a large subset of subtopics describing smaller steps of Natural Language Analysis. These topics populate one of the largest concept categories of the final ontology of NLP domain called “Topsss..

Each topic defines the tools for using the results of experiments and researches in different domain applications. For example, one of the most popular and p ccccdl tooss nn“Lxxllll Anyyysss” oopi,, ss “mmmmmrr”, which has different implementations with common and dsstnnvvvv faauurss. The “Too”” concept category of the ontology should be populated by these kinds of tools.

In addition, as examples for usable resources in the LLP domnnn, we aan rffrr to dfffrr nnt knnds of “Corpus” that could have varyyyyof properiiss (suhh ss “Tggged Corpus” or “uu ttnnngull Corpus”). Therefore, there is another concept category nmmdd “Rooourees” nn hhe ontology to cover the defined resources including corpora, thesauri, ontologies, lexicons, etc.

Also, since the domain uses different natural languages besides formal or computer languages, there should be another surplus category in the ontology for different classifications of languages (as the hypernym of “Natural Language”, “formll lnngugg””, “compuee nnngugg””, … classes).

The next subsection discusses the procedure to extract concepts within these categories.

3.2. . Extraction Process

We use 5 sequential steps to extract NLP related concepts as follows:

1. Initial core creation
2. Addnng “ppeehh nnd Lnngugge Procsssnng” concepts to the core
3. Extracting new concepts from Wikipedia
4. oooeessnng LLP rll tt dd vvnnss’ CPPs
5. Classification of natural languages

In this section, explanations will be provided about the details of each steps.

For building the ontology, first, we construct an initial core from scratch manually. Then, we use NLP textbook [9] to extract key phrases of the domain. After that, we use information about NLP terms available at Wikipedia to complete the list of extracted concepts.

The initial core is constructed by listing 150 most known terms of natural language processing domain, and categorizing them under four categories of concepts, relations, properties, and instances manually. We also create a hierarchy of concepts. The core had 98 concepts, 16 relations, 20 properties and 16 instances.

At the next step, on the extracted text of the NLP textbook [9], first we remove all stop-words, lone characters, numbers, adverbs and verbs. Then, we extract the most frequent words and phrases (1-gram and 2-grams) of the text as candidate phrases, by calculating the tf-idf and consider a threshold for its values. So the most frequent words and phrases of the domain are extracted as candidate concepts.

Finally, we manually review the results and remove superficial ones according to application requirements. As the result, 120 out of 180 concepts, with 67.2% precision are accepted.

Then, to extract concepts from Wikipedia, we use the structured data available at the page entitled "outline of Natural Language Processing"[1], which has a proper category of concepts of NLP. Furthermore, we create a small-scale ontology semi-automatically for each main part of the Wiki page. Then we merge these ontologies with our NLP ontology to extract new concepts. For this purpose, for each concept pair (C_1, C_2) in which C_1 is from the main NLP ontology and C_2 is from the small-scale ontology of Wiki page, we measure the similarity between two phrases by computing syntactic similarity according to (1) presented by [4].

$$SSM\ (c_1.c_2) = \max\left(0, \frac{min\big(Len(c_1).Len(c_2)\big) - EditDistance(c_1.c_2)}{min\big(Len(c_1).Len(c_2)\big)}\right) (1)$$

Where Edit-Distance (C_1, C_2) estimates the number of edits required to change C_1 to C_2. Having similarity more than a predefined threshold of 0.7 tags the pair as similar concepts. Concepts that do not have any similar class in the main ontology are considered as new concepts to be added to the ontology. In the process of comparison, we compare each of the concepts of Wikipedia ontology (from leaf to the root) to NLP ontology classes and find the potential similar ones. Also, we do the same process to find the possible super-class for each new concept to determine the position of them in the ontology.

An exampee of "Appiiaatoons" prrt (one of the sub-scciions of scciion "oooeessss of LL""), from Wikipedia page, is shown in Figure 2. In this example, the "Open Domain Question Answering" didn't have any similar concept in the NLP ontology, but its super-class "Question Answering" is matched with the exact same

New Concept	Matched with	Syntactic Similarity
Open Domain Question Answering	-	-
Question Answering (Super-Class of Open Domain Question Answering)	Question Answering	1.0
Applications (Super-Class of Question Answering)	Application	0.90

Fig 2 Syntactic Similarity between section "Processes of NLP" of the Wikipedia and the NLP ontology

named class in the Ontology and the next level class, their main headline, "Applications" is matched with the "Application" class of the NLP ontology. These comparisons result that the "Open Domain Question Answering" should be added to the NLP ontology, and its suggested super class is "Question Answering".

The "outline of Natural Language Processing" page of Wikipedia, added about 290 new concepts to our ontology.

Furthermore, we used some of the most reputable NLP conferences and journals, to consider their call for papers text for extracting more NLP-related phrases. Some of the publications used for this purpose are as follows:

ACL –Association for Computational Linguistics

EACL -European Chapter of the ACL

EMNLP -Empirical Methods in Natural Language Processing

CMCL -Cognitive Modeling and Computational Linguistics

ANLP -Advances in Natural Language Processing

NLPACC -Natural Language Processing And Cognitive Computing

NAACL HLT -North American Chapter of the ACL: Human Language Technologies

SEM– Lexical and Computational Semantics

LT– Language Technologies

We consddered hhe part "Clll for aaper" of hhe most recent events of these publications and extracted the most frequent and important phrases of them automatically (and revised manually) based on their structure. The results, contain about 700 new phrases, that we pruned them based on the generality of our target ontology.

In addition to the mentioned resources, we also investigated some web pages related to NLP groups and laboratories, preprocessed the data, extracted proper values, removed existing ones, and finally manually checked the remained ones for adding to the main ontology. Some of these web pages are as follows:

http://stanfordnlp.github.io/CoreNLP/

http://nlp.stanford.edu/sentiment/

[1] https://en.wikipedia.org/wiki/Outline_of_natural_language_processing

http://research.microsoft.com/en-us/projects/mt/

http://panacea-lr.eu

The method of finding existing similar concepts is same as what we explained for Wiki pages, but we do not have the hierarchical structure in these resources, so placing new concepts in the main ontology cannot be handled automatically.

To enter the new extracted terms into the ontology, we use lexical similarity and semantic similarity between concepts to compare them and find similar ones to be merged. We also use the similarities to find the right place for concepts that are not under the main hierarchy of the ontology if available.

To calculate lexical similarity, as we did for Wikipedia pages, we use the Levenshtein distance. But for these resources which do not have the hierarchy of terms, we compute the similarity of a term not only with ooncepss' aabess, but ssso whhhhhrrr sub-classes, super-classes, data properties, and object properties. Also, we assign a weight to each of the similarity metrics according to their importance. As an example properties are less important than super-class so it is assigned a lower weight.

To calculate semantic similarity we use WordNet to find all synonyms of a concept. Then we calculate the Levenshtein distance between all pairs of synonyms of two concepts.

Finally, we compare the sum of calculated distances of all concepts and find similar ones. Then we merge the similar concepts and all of their sub-classes.

Given that the NLP domain essentially works on human languages, we need a complete section in the ontology for covering all variety of languages, especially natural languages. For this purpose, numerous searches have been conducted to study different language classes. Some of the resources used in this process are as follows:

Wikipedia different language related web pages

The Language Galper[2]

The Indo-European Family of Languages[3]

At the end of this step, we had a Natural Language class in the ontology, with about 90 languages in different classifications.

4. Relation Extraction

To find hierarchical and non-hierarchical relations between concepts we use non-structured data of NLP textbook [9].

To extract hierarchical relations, we employed a template driven method using Hearst patterns (1992).

NPy including NPx and|or|, NPx

NPy such as NPx and|or|, NPx

NPy like NPx and|or|, NPx

NPx is a/an NPy

NPx and other Npy

NPx or other Npy

In multi-word noun phrases (NPs) we consider the extracted relation for the head of the NP as well as the NP. Also, we consider a hierarchical relation between the heads of noun phrases and the whole noun phrase.

After creating a thousand of candidate relations, 305 proper relations were accepted manually.

To extract non-hierarchical relations, we select four most frequent verbs in NLP domain (Generate/Produce, Use, and Have) which we have defined their corresponding relations in the ontology. We extract all sentences containing these verbs and use Stanford parser [10] to get their dependency tree and extract dependencies between tokens of sentences. Then we select the object and subject argument of verbs as concepts with verb corresponding relation. Some of these extracted relations are as follows:

Grammar ~ have ~ rules

Word ~ have ~ tag

Word ~ have ~ morphemes

Synthesis ~ produce ~ speech

Grammars ~ generate~ language

Algorithms ~ use ~ representation

We extract 500 relations. Then we revised them manually and accepted 148 relations to be added to the ontology.

Also, we checked words of all extracted relations and added them to the ontology if they already didn't exist. In this way, 340 new concepts are added to the ontology.

Subsequently, we define more relations and put some of them in a hierarchical structure. Some defined relations are "is_Expert_in", "is_Related_to", "Evaluated_By", "Evaluates", "Use" and "Related". "Use" and "Related" are two relations that are defined hierarchically.

The hierarchical structure of "Related" relation is shown in Figure 3. Defining relations hierarchically help the user to understand relations between specific source and target separately (for example between a specific pppiiaaiion nnd nnyhlnrg eeee ee ". plaaaiion Rll tt ed Thnrg"), unnccsss ry oo oollcct lll rll iii ons bewwen concepts.

oome of hhe "Rll eeed" rll iii ons rr e . hown nn Figure 4.

[2] http://www.languagesgulper.com/

[3] https://web.cn.edu/kwheeler/IE_Main.html

5. Completion and Integration

In this phase, lots of searches were done to evaluate the completeness of the ontology. By web searching using Google, a number of new concepts were found from different resources and glossaries to be added to the ontology.

As a result, in addition to improving the main categories of the ontology, two new categories are added as "Other_Terms" and "Related_Topic".

We place words or terms that do not have any position in the main hierarchy of the ontology (application, topic, tool, etc.) under "Other_Terms". “Lnngugg__Co suuunnss” ctt ggory, ss a subtree of “hhhrr_Term” ss shown nnFigure 5.

Also, we add other topics rather than NLP topics and their most important sub-classes under "Related_Topic" node. The added topics are closely related to NLP domain and cooperate with it in researches. As an example, a part of the "Linguistic_Topic" is shown in Figure 6.

As another assessment, we evaluate the completeness of the ontology with respect to LREC 2016 topics[4]. The topics have compared to ontology classes to calculate the completeness of NLP ontology. From the 90 topics of LREC 2016, 58 (64%) were found in the ontology. Although most of the not found topics are supplementary titles, not exactly related to NLP, like "web service", "policy issues", "metadata", etc., but some of them need to be appended to the main ontology.

To handle all of these new topics for the ontology, we add some more explanation for each concept as annotations, as well as creating new classes. These annotations are as follows:

"Abbreviation" to show the shortened form of the class label.

"Gloss" to explain the topic or application purpose shortly.

"OtherLabel" to define other labels or phrases with the same meaning (at least in the context of NLP).

"RelatedTerm" to define other terms that almost have same processes or applications. They do not have the exact same meaning, but they can be considered the same in the main application of NLP ontology, expert finding system. For example, "Medicine" and "Medical".

"ExternalLink" to save the website address or other related links to journals, conferences or maybe organizations.

Adding these annotations can improve precision and/or recall of the future applications that uses the ontology. For example, in an expert finding system, if the query contains "Spam Detection" phrase, the system recognizes the relation with "Spam Filtering" too, because "Spam Detection" is the "otherLabel" of "Spam Filtering" class in the ontology. It should be noticed that

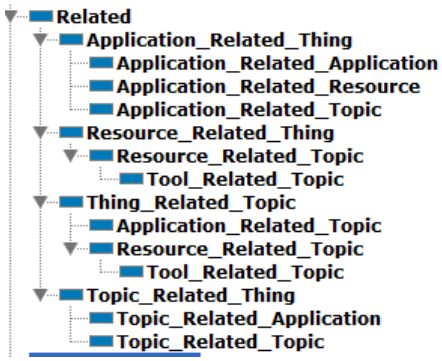

Fig 3 The "Related" Object Property Hierarchy

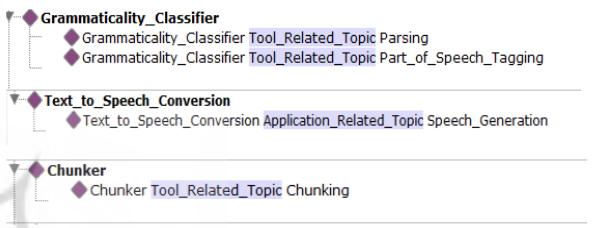

Fig 4 Examples of Relations between the Ontology Classes.

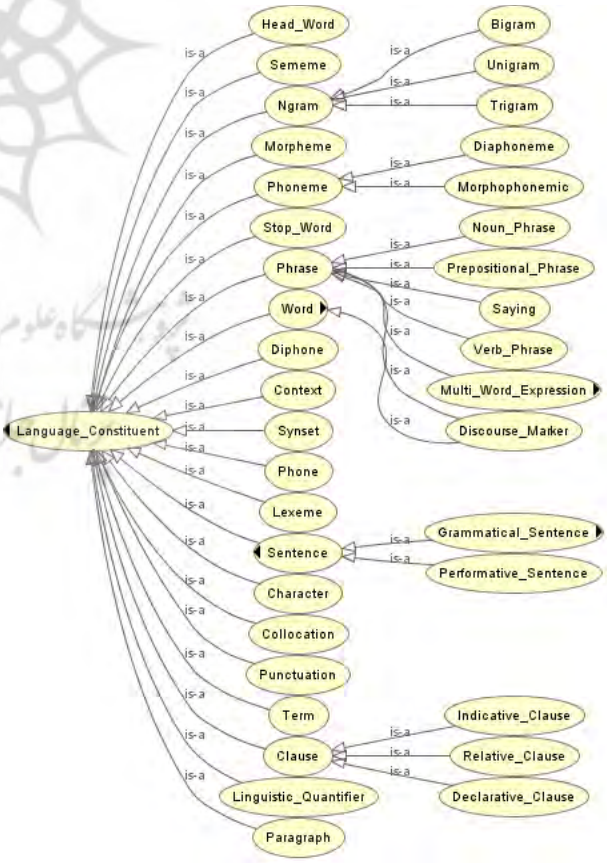

Fig 5 "Language_Constituent" as a subclass of "Other_Terms" in NLP Ontology

[4] www.lrec conf.org/proceedings/lrec2016/topics.html

by adding these annotations, the number of concepts decreases as some of them are merged to others as their "relatedTerm" or "otherLabel".

Some of these annotations are shown in Figure 7.

6. Final Review

At last, we asked NLP experts to review and revise the whole ontology. The Review is done to approve the ontology as a suitable ontology for being used in the expert finding applications. Some needed corrections were applied to the ontology after the revision. Some of the concepts were removed, the location of some were changed in the hierarchy and some were merged.

The Final output is the domain-specific NLP ontology with 1333 concept classes, 88 object properties, and 2437 annotation assertions. The first level of the NLP ontology is shown in Figure 8, and one of the most important parts of the ontology (uubtree of "Txxt_Proeessnrg" Topic) is shown in Figure 9.

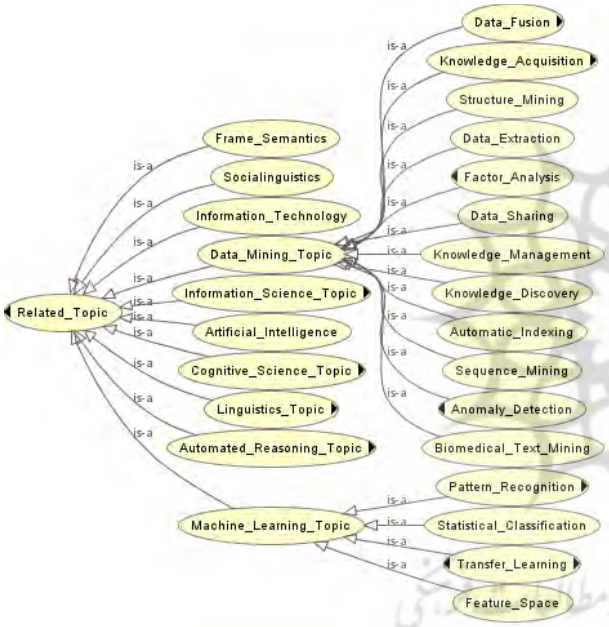

Fig 6 "Linguistic Topic" as a subclass of "Related_Topic" in NLP ontology

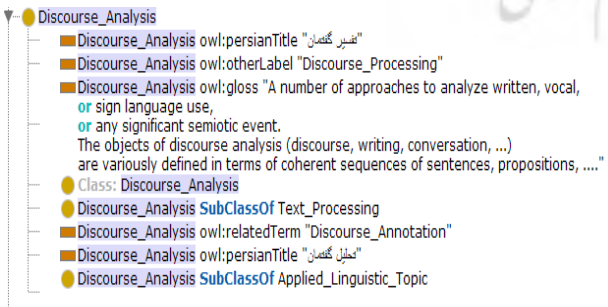

Fig 7 Different Annotations of "Discourse Analysis" Class

7. Ontology Population

After completing the ontology construction, we started populating the ontology.

First, we collect papers of ACL events and Non-ACL events from 2000 to 2017 available at ACL Anthology (TACL, ACL, EACl, SEMEVAL, COLING, HLT, etc.) and their information (title, abstract, venue, publisher, year and authors).

Then we also collect papers of four important journals of Natural Language Engineering (NLE), Language Resources and Evaluation (LRE), IEEE/ACM transactions on Audio Speech and Language processing and Computer Speech and Language from 2000 to 2017.

So, in the first step, we gather all papers from each of above events and all of their authors as primary data of population. Totally about 38K authors and 38K papers gathered from ACL Anthology and journals. Then we design two data structures for papers and authors, and then we fulfill the structures with collected information.

Secondly, we tried to complete each author's profile by collecting more information (citations, h-index, etc.) about authors of the papers from four scientific databases of Google Scholar[5], Research Gate[6], Scopus[7] and ACM digital library[8]. So, all of the authors' names were searched in each of the four resources and potential equivalent profiles were obtained. Also, we gather the

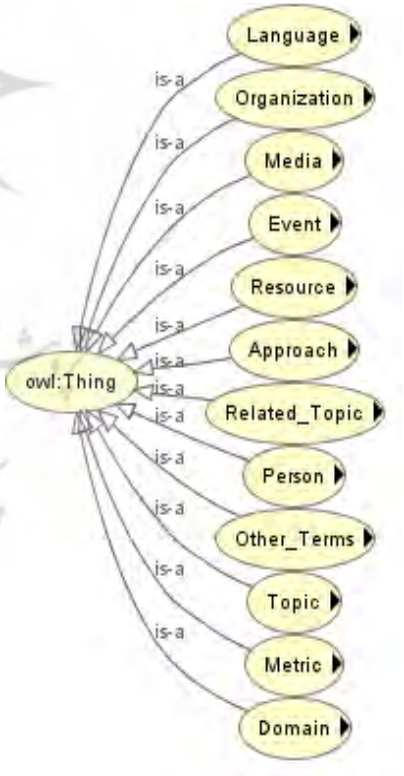

Fig 8 The First Level of NLP Ontology with the Number of Subclasses

[5] https://scholar.google.com
[6] https://www.researchgate.net/
[7] http://scopus.com
[8] https://dl.acm.org/

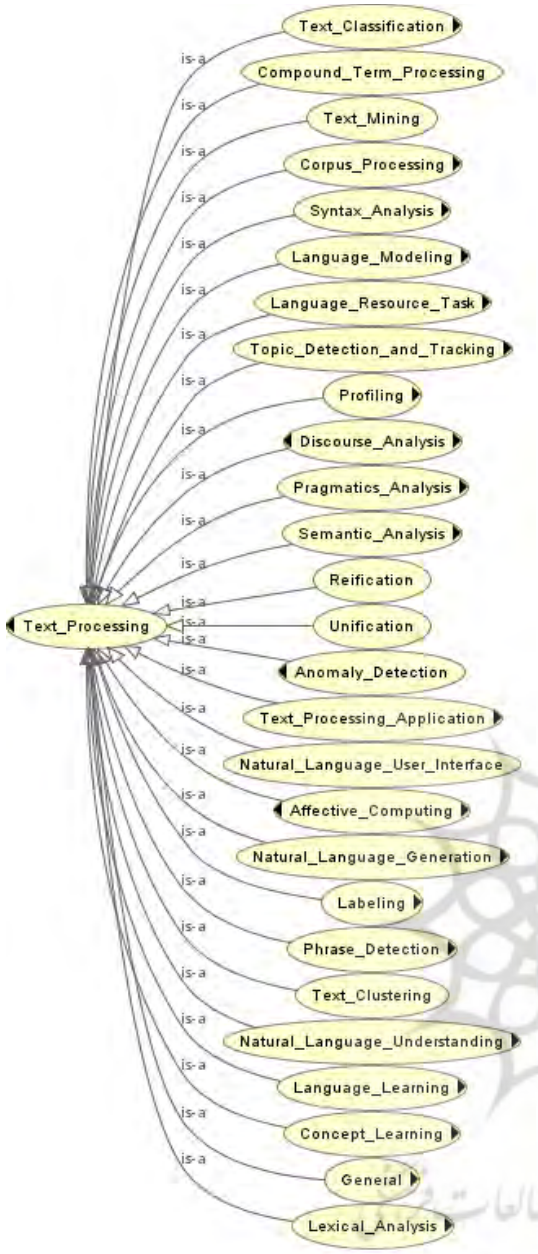

Fig 9 The First Level Subtree of "Text_Processing" Topic

papers assigned to each of the obtained profiles in these resources too.

At the end of this step, a large amount of data is collected that should be pruned to remove inappropriate found profiles. The total number of profiles gathered from ACM digital library, Scopus, Research Gate and Google Scholar was about 28K, 162K, 35K, and 9K respectively.

Also, there were about three million new papers that are related to the collected profiles, which were stored to be used in further operations. The information available at each of the author profiles are shown in Table 1.

In the collection of papers, we assumed that each paper has a unique title, so in case of some unwanted casual differences of a paper title in various resources, we

TABLE 1 INFORMATION AVAILABLE AT AUTHOR PROFILES

ACM	Cit–tion, count, average citation per article, publication count, publication years, affiliation, subject areas, keywords
Scopus	Scopus author id, document size, affiliation, city, country, other names, subject areas
Research Gate	Research Gate score, univer ity, department, location, position, skills
Google Scholar	Citations, Citations since 2013, h-index, h-index since 2013, i00-index, i00-index since 2013emain, affiliation, organization, labels

used an edit distance measurement by assigning a maximum threshold of three differences to consider different titles as unique (if the authors of two different papers are potentially the same).

In the next step, we map the potential profiles collected in the previous step to the primary authors by considering common papers and matching names.

Since authors may declare their names differently in various resources, an algorithm is implemented to check equivalency of two similar names (or Profiles). For example, if there are two authors with names "Richard Wang" and "Richard C. Wang" from two different resources, using name matching algorithm and further procedures, they are assumed as one unique author.

After mapping profiles to the authors, many irrelevant profiles that are not mapped to any author are removed from our database. As the result we have about 18K, 20K, 8K, and 7K profiles remained from ACM digital library, Scopus, Research Gate and Google Scholar respectively, which are mapped to primary authors.

After removing irrelevant profiles, papers assigned to them were also removed and finally, 467422 papers remain which are relevant directly to the authors or indirectly to the profiles mapped to them.

As it is stated, 467K papers are collected from author profiles that may not all of them be related to the NLP domain. So, we classify them into NLP-related and Non-NLP-related ones in order to separate all papers appropriate for the NLP ontology.

As we described in the ontology construction sections, the final ontology contains not only the NLP-related phrases but also other non-NLP-related phrases, which are usable in the domain or somehow related to any procedure or approaches of NLP. Because of that, we extract a second ontology from our main NLP ontology with only NLP-related phrases for papers classification. Each one of the 467422 papers, which contains at least one of the classes of the second ontology in its title, keywords or abstract, marked as NLP-Related paper. Some of the sub-trees of the main NLP ontology that rmmmnndd nn hhe second onoooogy rre “Topi””, “Tool”, “hhhrr_Terms,, “Corpus,, “Lnngugg__Mod”””, “Txx””, etc.

The above procedure results in 428257 papers as NLP-Related and 39165 papers as non-NLP-Related.

After pruning the authors and papers collected from 4 counted resources, we collect other information which can help in Expert Finding System.

We gathered a collection of world's universities with their country rank, world rank, abbreviation, city, country and continent from "nn vvrr stty rrrr hh Engnre" wbbseee[9] to be mapped to profiles according to their value of university or affiliation attribute. Therefore, we have the ranking of universities for most of the authors.

In addition, we gathered another collection of country ranking in Computer Science based on "SJR Counrry Rnnks"[10], so we could use the property of country rank of above collection (Universities) by having each country rank in the world.

Also, we made another collection containing different university positions that an author can have in a university or research program. The collection contains the score for each position, which shows their rank compared to each other, and this information can help in xxpert rnnknrg of fuure work to sssggn a "nn vversyyy ooofssssor" uuhlor more rr ddtt thnn a "Ph.D. uuudnn"".

Moreover, the languages, which the authors work on, extracted from their profiles and papers, and as a result, we know if each author focus on a specific language, except the English language, more than other languages. For example, 1044 authors of 38K focus on the Turkish language and 613 authors focus on the Persian (or Farsi) language.

The relations between countries and different natural languages we obtained using "Lnngugge ppoknn nn Ecch Counrry of hhe World"[11]. This information will help in Expert Finding System, to give priority to authors of the country with native language same as the language in the query of the system.

At the final step, we have a database of 38444 authors, with 18002 ACM profiles, 20519 Scopus profiles, 8974 Research Gate profiles, and 7445 Google Scholar profiles. Also, 428257 NLP-related papers mapped to authors and profiles. In addition, a collection of universities and their ranks, containing 13145 records, and the relationships between the profiles and them gathered. A collection of academic positions and their ranks, containing about 60 different positions, and the relationships between the profiles and them gathered. Moreover, the languages with which the authors work in the NLP domain obtained from keywords, title, and abstracts of papers using the sub-rree of "Lnngugg"" in the NLP ontology.

8. EVALUATION

An ontology can be evaluated through different processes [11, 12, 13 and 14]:

Comparison against a gold standard,

- Data-driven evaluation,

  User-based evaluation,

  Application or task-based evaluation.

Due to the best knowledge of authors, there are no other ontologies for NLP domain, so the first choice that is to compare it with a gold standard cannot be done. Moreover, data-driven evaluation is the process of comparing ontology against existing data about the domain that the ontology models it. This process is exactly the procedure that has been followed to construct the ontology. Also, the user-based evaluation has been done under the supervision of an expert of the domain, during various steps of ontology construction. The last one, application-based evaluation, as it is mentioned before, will be done using the expert finding system we will implement as the further work.

Besides other evaluations, we use the two measures recommended by [15] depth and breadth of the ontology. It is concluded that among different measures of depth and breadth, the most important ones are breadth variance and depth variance, and that the best ontologies are generally those that have higher values of depth and breadth variances in their structure.

The calculated metrics for current ontology are listed in Table 2.

I''s worhh mnnoonnrg hhat these metrics are appropriate to compare more than one ontologies together, and now that there are no other NLP ontologies, hhyy don't hvve any gains unless for future alternatives of ontology, to determine that the changes get the ontology to a better situation or not.

Batet. and Sánchez [16] recommends another measure to evaluate an ontology, which uses the semantic distance between two classes. This metric relies on comparing concepts according to the number of semantic evidence that they have and do not have in common in the ontology.

Based on this principle, a state-of-the-art feature-based measure is proposed [2, 17] that measures the semantic distance as a function of the number of non-common taxonomic ancestors divided (for normalization) by their total number of ancestors, as in (2):

$$d(c_1.c_2) = \log_2\left(1 + \frac{|T(C_1) \cup T(C_2)| - |T(C_1) \cap T(C_2)|}{|T(C_1) \cup T(C_2)|}\right) \quad (2)$$

TABLE 2 DEPTH AND BREADTH MEASURES FOR NLP ONTOLOGY

	Depth	**Breadth**
Minimum	2	3
Maximum	14	473
Average	4.74	4.41
Variance	4.50	2.31

[9] https://www.4icu.org/

[10] https://www.scimagojr.com/countryrank.php?area=1700

[11] https://www.infoplease.com/languages-spoken-each-country-world

Where T (C_1) is the set of taxonomic ancestors, including itself. [17] proposes a semantic dispersion of an ontology relied on above distance (3):

$$Dispersion(O) = \sqrt{\frac{\sum_{c_1 \in C} d(c_1.root(o))^2}{|C|}} \quad (3)$$

As concluded by [15], the higher values of dispersion show the appropriate distribution of concepts in the ontology. The dispersion of NLP ontology has the value of 0.80 that seems to be a reasonable value.

9. Conclusion and Future Work

This paper described our constructed ontology and the methods we used to create it. The evaluation results show that the ontology achieved a good result at the expert's point of view.

Although future work will focus on enhancing the ontology and do more population to cover all resources and experts in NLP. Furthermore we will develop an automatic updating system to make the populated knowledge up to date. Also, we will focus on developing our Expert finding system according to available information using the ontology. It is expected that using the ontology in an expert finding systems will help the results to be more semantically related to the query than other related works.

References

[1] Kara, S., Alan, Ö., Sabuncu, O., Akpınrr, S., Cicekli, N. K., & Alpaslan, F. N. (2012). An ontology-based retrieval system using semantic indexing. *Information Systems*, 37(4), 294-305.

[2] Batet, M., Sánchez, D., & Valls, A. (2011). An ontology-based measure to compute semantic similarity in biomedicine. *Journal of Biomedical Informatics*, 44(1), 118-125.

[3] Pisarev, I., & Kotova,V. (2016). Construction of thematic ontologies using the method of automated thesauri development. In 2016 IEEE NW Russia Young Researchers in Electrical and Electronic Engineering Conference (EIConRusNW) (pp. 305-307). IEEE.

[4] Amini, B., Ibrahim, R., Othman, M. S., & Nematbakhsh, M. A. (2015). A reference ontology for profnnng schorrr's bcckground knowddge nn recommender systems. *Expert Systems with Applications*, 42(2), 913-928.

[5] Chiarcos, C., & Sukhareva, M. (2015). Olia–ontologies of linguistic annotation. *Semantic Web*, 6(4), 379-386.

[6] Klimek, B., McCrae, J. P., Lehmann, C., Chiarcos, C., & Hellmann, S. (2017). OnLiT: An ontology for linguistic terminology. In *International Conference on Language, Data and Knowledge*. pp. 42-57. Springer.

[7] Noy, N. F., & McGuinness, D. L. (2001). Ontology development 101: A guide to creating your first ontology. Stanford knowledge systems laboratory technical report KSL-01-05. Retrieved from http://www.corais.org/sites/default/files/ontology_development_101_aguide_to_creating_your_first_ontology.pdf.

[8] Adelkhah, R., Shamsfard, M., & Naderian, N. The Ontology of Natural Language Processing. 2019 5th International Conference on Web Research (ICWR), Tehran, Iran, 2019, pp. 128-133.

[9] Martin, J. H., & Jurafsky,V. (2009). *Speech and language processing: An Introduction to natural language Processing, Computational Linguistics, and Speech Recognition*. Upper Saddle River, Pearson/Prentice Hall

[10] Chen, D., & Manning, C. (2014). A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 740-750).

[11] Hlomani, H., & Stacey, D. (2014). Approaches, methods, metrics, measures, and subjectivity in ontology evaluation: A survey. *Semantic Web Journal*, 1(5), 1-11.

[12] Hlomani, H., & Stacey, D. A. (2013). Contributing evidence to data-driven ontology evaluation workflow ontologies perspective. In *5th International Conference on Knowledge Engineering and Ontology Development, KEOD* (pp. 207-213).

[13] Ouyang, L., Zou, B., Qu, M., & Zhang, C. (2011). A method of ontology evaluation based on coverage, cohesion and coupling. In 2*011 Eighth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)* (Vol. 4, pp. 2451-2455). IEEE.

[14] Brewster, C., Alani, H., Dasmahapatra, S., & Wilks, Y. (2004). Data driven ontology evaluation. In International Conference on Language Resou*rces and Evaluation (LREC)*, Lisbon, Portugal.

[15] Fernández, M., Overbeeke, C., Sabou, M., & Motta, E. (2009). What makes a good ontology? A case-study in fine-grained knowledge reuse. In *Asian Semantic Web Conference* (pp. 61-75). Berlin, Heidelberg, Springer

[16] Batet, M., & Sánchez, D. (2014). A semantic approach for ontology evaluation. *In 2014 IEEE 26th International Conference on Tools with Artificial Intelligence*(pp. 138-145). IEEE..

[17] Sánchez, D., Batet, M., Isern, D., & Valls, A. (2012). Ontology-based semantic similarity: A new feature-based approach. *Expert Systems with Applications*, 39(9), 7718-7728.

Niloofar Naderian is a graduate student in master of software engineering at the Faculty of Computer Science and Engineering, Shahid Beheshti University, Tehran, Iran. Her research interest is Natural Language Processing.

Dr. Mehrnoush Shamsfard has received her BS and MSc both on computer software engineering from Sharif University of Technology and her PhD in Computer Engineering-Artificial Intelligence from AmirKabir University of Technology, Tehran, Iran. She has been with Shahid Beheshti University from 2004. Dr. Shamsfard is currently the dean of Faculty of computer science and engineering, and also the head of NLP research Laboratory of this faculty. Her main fields of interest are natural language processing, knowledge and ontology engineering, text mining and semantic web.

Razieh Adelkhah is a PhD student Artificial Intelligence at Shahid Beheshti University. Her main research interests are in natural language processing and ontology. She also has received her BS on computer software engineering and MSc on Artificial Intelligence both from Shahid Beheshti University and has been a member of NLP research Laboratory of Faculty of computer science and engineering since 2014.