Data Extraction from Natural Language Using Universal Networking Language

Conference Paper · September 2017

DOI: 10.1109/CTCEEC.2017.8454920

Dr. Aloke Kumar Saha
University of Asia Pacific
27 PUBLICATIONS **54** CITATIONS

SEE PROFILE

M. Firoz Mridha Ph. D.
Bangladesh University of Business and Technology (BUBT)
60 PUBLICATIONS **108** CITATIONS

SEE PROFILE

J. K. Das
Jahangirnagar University
26 PUBLICATIONS **61** CITATIONS

SEE PROFILE

Data Extraction from Natural Language Using Universal Networking Language

Aloke Kumar Saha[1], M. F. Mridha[1], Jahir Ibna Rafiq[1] and Jugal Krishna Das[2]
[1]Dept. Of Computer Science and Enginerring, University of Asia Pacific, Dhaka, Banglaesh
[1]Dept. Of Computer Science and Enginerring, Jahangirnagar University, Dhaka
mdfirozm@yahoo.com

Abstract— **Data extraction, which falls under the area of Natural Language Processing (UNL), finds specific data from unstructured data. This research paves the way to introduce a unique technique on data extraction – providing the user with exactly what is asked without any mimicry of unsolicited data. The proposal sets logical and symmetrical relation between the search criteria and operational data. Since the data is unstructured and volume can be relatively high, we have emphasized highly on putting the data under categories – defined and used by the researchers for further exploitation of data. Universal Networking Language (UNL) is efficiently used to compare data and merge. A new approach of machine learning is presented herein that essentially augments efficiency of Natural Language Computing (NLC) and Cognitive Computing (CC). This proposed approach uses UNL relationship and successful test data shows much improved results and efficient generalization. Existing machine learning approaches are widely used on numeric data which are producing expected results but one key contention is the limitation of data type that can be handled. Current models fail to properly train on the semantics, logical consistency; many natural language properties are either ignored or prove too much of a task. Consequently, the approach presented herein this paper carries further positive points in producing meaningful and worthwhile result. Moreover, complex data that are consisted of alphanumeric data, sequence and resulting criteria can be executed correctly.**

Keywords—Big Data; Data Extraction; NLP; Universal Networking; Natural Language Computing; Machine Learning.

I. INTRODUCTION

Machine Learning (ML) algorithms such as Probabilistic Learning, Neural Networks, Support Vector, Genetic Algorithms [1] are widely used. They are classified into Evolutionary Learning, Supervised, Reinforcement, and Unsupervised. They are taught through examples, various forms of numeric values, so that they can recognize these trends from their training once unfamiliar but similar data is given. They are provided with numeric values that represent certain features from target objects. For instance, a motion detector placed in front of a supermarket gate only opens when a car comes within ten feet of the gate. Here average car weights are feed in during training so that door remains shut in case a ball rolls near the gate. These numeric values are important property in recognizing data but they are not sufficient for most natural language and their further beneficial applications.

At present more and more scientific and research work is being carried out on the behest of machine functioning as human, giving out proper answer and immaculately executing repetitive tasks. It is more common in the field of numbers. But it leaves out data that are not purely numeric, as of now. Many machine-learning algorithms are limited to one or two set of data types. But dealing with multiple data type is more realistic because real-life data tends to consist of number as well as instructions in plain language; it is quite common to receive signal that are mixed type data on top of plain language in alphabet, they could be alphanumeric data and special characters like mathematical formula. Currently we can hardly focus on data analysis, nitpick the data asked by any customer or sort data on the basis of simple query.

Natural Language (NL) based applications should find the exact linguistic meaning of words in a sentence in order to correctly analysis a query presented in a plain language – for instance, searching for operational amplifiers in a search engine. Indeed, better understanding of words formation and meaning from the context of sentence are important, so are logical steps in the questions. Once these are achieved, data extraction based on query from unstructured data i.e. data that are not homogeneous in nature and formed through many combinations of data types like numbers, alphabet, characters are possible. Properly replying to questions, provide meaningful summary, translation from language to language are only also possible. But many numeric driven ML are simply not on par with these task.

Hence, a new approach of machine learning algorithm needs to be developed. Recent research shows that growing demand of text data has eclipsed that of numeric data type. On the Internet, 80% of the data are text and the remaining 20% data are numeric [2]. Sorting out this huge amount of text and properly find out required text is an important aspect. Hence, this paper suggests ML algorithm to efficiently learn NL semantics. This algorithm can train on logic and semantics, resulting in rewarding improvement in output data and data analysis as part of query in different forms and sizes.

The organization of this paper is as follows: In Section 2, we describe the Literature Review, Section 3 has the short description of Data Extraction Method, Section 4 describes the UNL system, Section 5 depicts the proposed model, Section 6

978-1-5386-3243-7/17/$31.00 ©2017 IEEE.

demonstrates our Results and finally, Section 7 draws the curtains by concluding amid some heeds towards our future work.

II. LITERATURE REVIEW

Data extraction from natural language is the most previous approaches to Natural Language Generation from semantic network which was discussed with verbalizing ontologies [1]. The textual realisation of natural language has paid relatively low attention. Our main interest on finding the data which are generally found in natural language format [2]. Generating UNL expression from natural language like Bangla to UNL [3,4] that is used for finding the target data extraction. Data extraction from trained data (UNL expression), in our proceeding parallel text-knowledge is derived from templates and learning is mostly done through it [5 ,6]. The text data is aligned in a trained database in a 2-step process. Starting with a system process in finding out best matched data that are in string format – the larger the better and as it appears in the unstructured data. Followed by gradual building of a statistical language model which in terms utilizes the entropy of the original query and confirms the length of matched data. Here emphasis is put on generating natural language that closely expresses a matched case in the source data with no prior knowledge of the data. Obtaining this level of accuracy depends how closely we can find an entity that has the property of traceability and matches our query or text input. Where such sentences are not directly derivable from the text, it is possible to modify them to make them transferable[7,8,9]. We adopt a syntactic pruning approach inspired, where sentences are first parsed and then the resulting structures are simplified by applying hand-built rules and filters [10,11,12].

III. DATA EXTRACTION METHOD

Processing natural language is not sufficient based on machine learning algorithms, which are driven by regression, classifications, and matching data when the value is simply presented in numeric form. This is also applicable for understanding and computing of natural language and cognitive computing. Semantics is an inherent and vital property especially when computation of natural language and cognitive learning are involved; it has to be stressed during any analysis by the machine. In other words, the proper meaning of the sentence will largely be dependent on how accurately the semantics are computed by the machine on the context of the sentence. Machines should be taught and trained about semantics of words. The goal should be recognizing new semantics based on training. This is specifically a requirement for cognitive computing.

The task given to machine will have to be completed on the basis of the nature of the query. The appropriate meaning of the sentence lies on the logic used to ask the question. Hence, machine learning for unstructured data should be capable of reasonable logical progression and the actions must reflect the logic. In stark contrast to currently used model of feeding numeric data for training, machine learning for the unstructured data needs to focus on semantic of words and logic. This proposed model is on par with the properties required for successful derivation of semantics and logic.

Instead of feeding big chunk of datasets during the training, the proposed learning model is semantic driven; it is trained on logical connection and explanation of the action to be taken. This is akin to the learning model of human being. Our learning is effective when logics are correctly presented and well understood by the accepting individual rather than committing to memory. By correctly learning single logic, all homogenous logical work can be inferred and executed. On the other hand, conventional machine learning by numeric data (only numbers) follow the path of feeding a large amount of examples but it conveniently leaves out any explanation or logic behind the task. Therefore, it can only execute tasks, which are similar in nature and produce result. The limitation lies in erratic data, data that may not conform to the example and has properties with limited functionalities.

A key factor is how quicker this algorithm is – as far as learning and training is concerned; semantics are instilled while computation is carried on. Both the procedures are conducive to learning and ultimately deriving correct data, bearing in mind that, they are equally important and closely interrelated. In majority of the cases, learning is done through computing except when new semantics are enacted.

However, as newer semantics are proposed and existing semantics research provide us with new knowledge, learning is often differentiated from computing. This paradigm also incorporates the refined meaning of words, mainly from Word Feature (WF) and to some lesser extent from Word Knowledge (WK) tables and NL corpora.

As new paradigm offers to transfer teaching method from examples to semantics and logic, it produces better result in generalization and identified as a key improvement. Because here computing and learning are driven by logical progress and words are divulged in their contextual meaning, there is little appetite for larger data set such NL corpora. While proposed model is independent of such datasets, many other models such as Probability based N-grams are very much dependent on datasets.

Machine Learning in NLP has been tremendously successful in training and computing. It can capitalize on the new paradigms mentioned herein, often learns as we learn as a person and hence, produces result very similar to human judgment and succeeds in giving right answers. The procedures are efficient – ML in NLP takes command as to what is the full query, then divides it according to parts of action or sequences are produced, finally the most accurate result is assumed as correct answer and shown. Let's examine the following examples to clarify the procedures.

[Example -1]
"Please give me celebration picture of last Sunday's FIFA world cup finale from twitter."

The working steps of ML in NLP will be

a. Go to twitter website.
b. Prompts user to login.
c. Determine the exact date of last Sunday considering the date of today.
d. Using hast tag of key words #Celebration #FIFA #Finale, find out the most relevant photos.
e. Produce result where the photos are placed according to the likes that they received. Higher likes receive higher position – thus popular pictures are shown first.

[Example -2]
"Please show me the statistics of people passed away from accidents in M25 motorway in the UK."

The working steps of ML in NLP will be

a. Keywords are identified as UK, M25 highway, accident, death and so on.
b. Here, search engines like Google are preferable to look for appropriate data.
c. Go to Google or similar search engine and search of all combination of key words.
d. Results are cross-checked and verified as they appear from different sources. Most matched data are taken as correct data.
e. Result is produced for user.

IV. UNL SYSTEM

The structure of UNL system consists of three parts namely the **Universal Words**, **UNL Attributes** and **UNL Relations**. Universal word is an English word which is represented by nodes in a hypergraph [13]. Nodes associated with a sentence are connected by a relation known as UNL relation. Each universal word has some attributes that uniquely specifies that word and is placed according to a conceptual hierarchy derives from a UNL knowledge base. However, each of the universal words is comprised of headword along with some constraints list. The headword is considered as the unit form of the English word, known as label, whereas each of the constraints in a constraint list of the universal word corresponds to a concept of that word. The attribute lists associated with the individual universal word are used to represent the subjectivity of word based on their grammatical properties [13]. Fig.1 shows the structure of UNL.

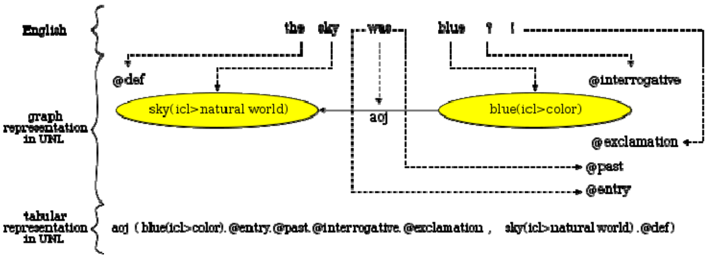
English
the sky was blue ? !
@def
@interrogative
graph representation in UNL
sky(icl>natural world)
aoj
blue(icl>color)
@exclamation
@past
@entry
tabular representation in UNL
aoj (blue(icl>color). @entry. @past. @interrogative. @exclamation , sky(icl>natural world) . @def)

Fig. 1. Structure of UNL

V. PROPOSED MODEL

For finding the target data, we have used the compatibility property of sentence in UNL platform. The relation between the words of a sentence according to the meaning is called compatibility or propriety. As an example: if we work with the sentence: স্টিমার পানিতে চলে (Steamer moves on water) then there are some meaningful relation between the words as the boat can float in water. But if we consider this sentence: স্টিমার আকাশে চলে(Steamer moves on sky), then it doesn't make any sense. As there is no meaningful relation between boat and sky. That's why the last sentence has lack of compatibility. In order to become a perfect sentence, it must have the quality of compatibility.

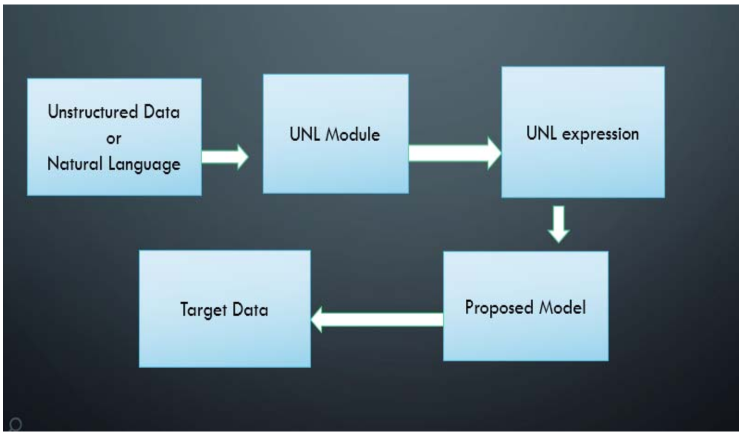

Fig. 2. Proposed UNL based data extraction model

In order to check the compatibility for a sentence, we build tables and populate data there for making relationships if two words have any relation between them. We considered the rules' common property as the elements of the tables. Such as if we are talking about the pronouns of human, then it comes - I, We, He, She, They etc. In UNL these types of pronouns are grouped together in a common property. So, we then just need a very few rows and column comparing the table having each subjects objects and verbs from the real world.

We build a table using UNL relation[4,13] for checking if the subject, object and other words have any perfect relation Fig. 2 shows the Proposed UNL based data extraction model. Now, for checking the compatibility, after a sentence passes the previous requirements, we will find the UNL rule for the sentence. Then from there, we will look up to the tables to find if the relation between subject - verb, object - verb and subject - object all are true. If we find all the relation tables cell of this three are true, then we can just say that this sentence has perfect meaningful relation between words. Therefore, this sentence has the quality of compatibility.

Let us explain an example for more details. We want to find the data "where the cycle move?" the correct answer will be "on road". We consider a sentence গাড়ি রাস্তায় চলে (the car moves on the road). In our proposed model the natural language first convert into UNL expression and then using the UNL relation we will find our target data.
From the UNL module we will find the following output:

```
{unl}
obj(move(icl>occur,equ>displace,plt>thing,plf>thing,obj>thing).@entry.@present,car(icl>wheeled_vehicle>thing))
plc(move(icl>occur,equ>displace,plt>thing,plf>thing,obj>thing).@entry.@present,street(icl>thoroughfare>thing).@def)
{/unl}
```

Now, we will look up to the table. In the "obj relation table" [4,13], we will find the value true in the intersection cell of "icl>wheeled_vehicle>thing" and "icl>occur" or "equ>displace".
In the "plc relation table", we will also find true value in the intersection cell "icl>thoroughfare>thing" and "icl>occur".

Therefore, for the all two combinations, we can find true value, for which we can say that the sentence has meaningful relations among the words and we can find our target data.

Again, if we consider another example: গাড়ি আকাশে চলে(the car moves on the road).
The corresponding UNL expression will be:

```
{unl}
obj(move(icl>occur,equ>displace,plt>thing,plf>thing,obj>thing).@entry.@present,car(icl>wheeled_vehicle>thing).@def)
plc(move(icl>occur,equ>displace,plt>thing,plf>thing,obj>thing).@entry.@present,sky(icl>atmosphere>thing).@def)
{/unl}
```

Now, here, we will find one cell containing true value. For the intersection cell of "icl>wheeled_vehicle>thing" and "icl>occur". But the intersection cell in the "plc relation table" of "icl>wheeled_vehicle>thing" and "icl>atmosphere>thing" doesn't contain true value. So, from here, we will come to the decision that this sentence has no meaningful relations between the words. Then we unable to find the target data.

VI. Result analysis

In this proposed model, natural languages are taken as input and it is converted in UNL expression and then its semantic relations are compared the target data. The accuracy of data extraction is about 93.5% was achieved from our proposed model. During the experiment, it has been scrutinized that the accuracy was getting diminutive when the researchers have tested more than 5000 sentences. The output generated by our system is given in table 1 below:

TABLE 1. ACCURACY CALCULATION OF SENTENCES

	Total
Bangla words and morphemes in word dictionary	20000
Correct words found	18700
Percentage of accuracy	93.5

VII. Conclusion

Unstructured data and mixed data type presents a unique challenge but it remains quite vital to explore these data types in regards to semantics learning, because they are already occupying a major share in the type of data that we have to deal with. In this work, UNL based data extraction has been developed that can be used for searching target data. It is confirmed currently Big Data is four-fifth unstructured data – a big majority are either text or alpha-numeric. Many fundamental operations like text based search can really improve how NLC and CC work towards finding data that are of interest rather than returning huge amount of similar but mostly unrelated data.

References

[1] Liang, S., R. Stevens, D. Scott, and A. Rector (2012). OntoVerbal: a Proteg´e plugin for verbalizing ontology classes. In Proceedings of the Third International Conference on Biomedical Ontology.

[2] Heath, T. and C. Bizer (2011). Linked Data: Evolving the Web into a Global Data Space. Synthesis Lectures on the Semantic Web: Theory and Technology. Morgan & Claypool.

[3] F. Mridha, Molla Rashied Hussein, Md. Musfiqur Rahaman, Jugal Krishna Das "A Proficient Autonomous Bangla Semantic Parser for Natural Language Processing", ARPN Journal of Engineering and Applied Sciences, VOL. 10, NO. 15, AUGUST 2015,ISSN 1819-6608, pp 6398-6403.

[4] M. F. Mridha, Aloke Kumar Saha, Md. Akhtaruzzaman Adnan, Molla Rashied Hussain and Jugal Krishna Das,"Design and Implementation of an Efficient Enconverter for Bangla Language" ARPN Journal of Engineering and Applied Sciences, VOL. 10, NO. 15, AUGUST 2015,ISSN 1819-6608, pp 6543-6548.

[5] Muhammad F. Mridha, Aloke Kumar Saha, Mahadi hasan and Jugal Krishna Das," Solving Semantic Problem of Phrases in NLP Using Universal Networking Language (UNL) "(IJACSA) International Journal of Advanced Computer Science and Applications, Special Issue on Natural Language Processing(NLP) 2014.

[6] Duboue, P. A. and K. R. Mckeown (2003). Statistical acquisition of content selection rules for natural language generation. In Proceedings of the 2003 conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 121–128.

[7] Aloke Kumar Saha, Muhammad F. Mridha, Shammi Akhtar and Jugal Krishna Das,“ Attribute Analysis for Bangla Words for Universal Networking Language(UNL)”, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 4, No.1, 2013.

[8] Cohn, T. and M. Lapata (2009). Sentence compression as tree transduction. Journal of Artificial Intelligence Research 34, 637–674.

[9] Filippova, K. and M. Strube (2008). Dependency tree based sentence compression. In Proceedings of the Fifth International Natural Language Generation Conference, pp. 25–32. Association for Computational Linguistics.

[8] Gagnon, M. and L. Da Sylva (2006). Text compression by syntactic pruning. Advances in Artificial Intelligence 1, 312–323.

[9] Hewlett, D., A. Kalyanpur, V. Kolovski, and C. Halaschek-Wiener (2005). Effective NL paraphrasing of ontologies on the Semantic Web. In Workshop on End-User Semantic Web Interaction, 4th Int. Semantic Web conference, Galway, Ireland.

[10] Klein, D. and C. Manning (2003). Accurate unlexicalized parsing. In Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1, pp. 423–430. Association for Computational Linguistics.

[11] Mendes, P., M. Jakob, and C. Bizer (2012). DBpedia: A multilingual cross-domain knowledge base. In Proceedings of the International Conference on Language Resources and Evaluation (LREC 2012).

[12] Stevens, R., J. Malone, S.Williams, R. Power, and A. Third (2011). Automating generation of textual class definitions from OWL to English. Journal of Biomedical Semantics 2(Suppl 2), S5.

[13] Universal Networking Language (UNL) Specifications Version 2005, http://www.undl.org/unlsys/unl/unl2005/.