Usage of IT Terminology in Corpus Linguistics

Mavlonova Mavluda Davurovna

Abstract: Corpus linguistic will be the one of latest and fast method for teaching and learning of different languages. Proposed article can provide the corpus linguistic introduction and give the general overview about the linguistic team of corpus. Article described about the corpus, novelties and corpus are associated in the following field. Proposed paper can focus on the using the corpus linguistic in IT field. As from research it is obvious that corpora will be used as internet and primary data in the field of IT. The method of text corpus is the digestive approach which can drive the different set of rules to govern the natural language from the text in the particular language, it can explore about languages that how it can inter-related with other. Hence the corpus linguistic will be automatically drive from the text sources.

Keywords: Corpus, linguistics, Corpus language, Novelties, IT (Information Technology)

I. INTRODUCTION

Corpus linguistic is suggested as sample of real-world language along with their study. Method named text-corpus is a good approach to derive the rule can be used to govern the IT terminologies or natural language from the text in the particular language. Hence corpora are derived automatically from text source and that is compare with origin time while corpus is manually obtained.

Also, it suggests about analysis of language is very reliable in order to achieve with the sample already collect along with experiment. The field of corpus scientific study has tried to differentiate the views of annotation values ranging from Jon McHardy Sinclair who advocated the minimal annotation where text would speak for their own-self while writing the usage survey of IT terminology for the team who support annotation and allow the understanding of linguistic via recording rigorously. Propose article can discussed about the terminologies of IT in Corpus Linguistic.

II. CORPUS LINGUISTICS ORIGIN

Text was collecting from middle age to study languages. In middle age word are listed in a book along with their context which know as concordance. Researcher collect different words and text then counted their frequency and note down which words come frequently and determine most frequently used words.

At this period corpora were used in semantics and related field. At this period corpus linguistics was not commonly used, no computers were used. Corpus linguistic history were opposed by Noam Chomsky. Now a day's modern corpus linguistic was formed from English work context and now it is used in many other languages. Alongside was corpus linguistic history and this is considered as methodology [Abdumanapovna, 2018]. A landmark is modern corpus linguistic which was publish by Henry Kucera.

Corpus linguistic introduce new methods and techniques for more complete description of modern languages and these also provide opportunities to obtain new result with the help of these methods and research techniques [Andersen, 2012]. Corpus of text helps to identify the range of semantic in broad context. Corpus of text is act like an array, it not only building contextual dictionary but also for distinguishing between different language variants.

III. TYPES OF CORPUS

Computer linguistic and corpus linguistic are directly correlated with each other. Computer linguistic build languages model and corpus linguistic is also same that's why they help each other. Application of computer linguistic help in automatic translation, extract information form natural text, help to design an interface between human and machine and it is quantities description of communication in natural languages [Bennett, 2010]. Computer linguistic creates tool for corpus linguistic. So, they also complement each other. Corpora are the collection text and word form written work of an author. Text corpora is dividing in to different categories depending on how it relates to other corpora. Corpora fall in to different categories that are mentioned below:

Monolingual Corpus

This corpus type is most frequent, and it contain only one language in the entire text. It is mostly used for studding the syntax function in the language [Abdumanapovna, 2018]. It is used by different user in different task for styling.

Parallel Corpus

This type of corpus contains paired monolingual corpus in which each corpus is a reciprocal of other. For example, the book that has its translation, both correspond to each other and aligned in a segment [Breyer, 2008]. The person is observed how these word and text are translated form one language to other language, one is searching the word from one language and result show same word translation in to other language.

Revised Manuscript Received on May 05, 2019.

Mavlonova Mavluda Davurovna, Tashkent University of Information Technologies named after, Mukhammad al-Khwarizmi, Foreign Languages Department, Phd candidate, Tashkent city, Yunusabad district, Amir Temur, Tashkent city, Yakkasaray district, massive Bashlik

Multilingual Corpus

This type is same as parallel corpus because the two are used interchangeably. This type of corpus contains languages in different context, and these are the copies of book and in line just like a parallel corpus [Bowman, 2015]. The engine tool enables the researchers to choose one or more than one aligned corpora and other devise name search and it display what it has been translated into other languages at once. If pair of expression is selected, then this corpus act like a parallel corpus.

Comparable Corpus

This type of corpus is made up of many monolingual corpora which is same topic related to their text, they are not translated to each other and not aligned just like previous type. Corpora have corresponding metadata when the searcher searching for these comparable corpora [Abdumanapovna, 2018]. The most common example of this corpora is childes corpora in sketch engine, the other example is different corpora found in Wikipedia.

Learners Corpus

This type of corpora is the corpus of word that are language produce by the learner to study and identify different problems that leaner face while they are studding foreign languages. This corpus annotated with multiple type of errors [Andersen, 2012]. It provides interface to learner in which learner search for failure, done correction of mistakes and to explore kind of errors and combination of errors.

Diachronic Corpus

This type of corpus contains different words form different time period and these words are used by the researcher to investigate about changes and development of speech with the passage of time. The sketch engine allows this corpus to search as whole or only be include in an identified time at interim into search [Bennett, 2010]. There is another toll in this corpus that is used to identify the word whose usage change over the selected period.

Specialized

This type of corpus is one which contain one or more than one topic, subject area and domain and this corpus is used to study how the specialized language used. In sketch engine searcher creates a sub-corpus which are special from general corpora. The most common example of specialized corpora is notes, document and articles [Breyer, 2008]. For the practice, verity of curriculum contains the text form specific academic subjects just like sociology, religion and economics etc.

Multimedia

This type of corpus has words that is improved with multimedia content just like audiovisual materials. In this corpus transcribed files must be organized in such a way that information about the source is available. Just like information about age, education, place of birth, place of record, people, their gender etc. The entry transcription coding keys includes:

<S1>, <S2>, etc. - designation of individual speakers;
+ - cases of interruption ("latched turns");
<?> - an unintelligible statement (unintelligible, utterance);
= - truncated utterances;
<SE> laugh <\ SE> - extra linguistic information laughter, cough, etc.

Oral speech can be coded to the level of reflection of replica to speaker turns, speaker overlaps, cases of interruptions, vocalization, truncated utterances, hesitation, the inclusion of extra linguistic information, laughter and extraneous sounds [Abdumanapovna, 2018].

IV. LINGUISTICS OPERATING IN THIS FIELD

A corpus language is such language which study by using corpus scientific methods. The corpus linguistic examine the recorded of the word that are living. According to researcher not all the languages are corpus languages. Most of the style are vanished with no recorded production of their speaker and adequate [Abdumanapovna, 2018].Examples of these corpus languages are Latin, Egyptian Language, Medieval English, and Ancient Greek. British national corpus is an example of text corpus that are mostly written in English language and spoken by the people around the world. British English was expressed during the late 20th century and the corpus include the British English language with the intension to be a sample representative of spoken and written English at that time period [Allen, 1995]. The used of this technology especially high computer system will help in the retrieval of any needed data.

V. LANGUAGE TECHNOLOGY

In hybrid linguistic and IT field, language technologist work for service and tool development which are used in the imaginable product and service along with linguistics dealing with language. Business and industry make technology language extensive. Spellings and Grammar correction program are modern development in language technology as well as it can underline program for automatic recognition of speech, speech synthesis, technology language including intelligent search engines. Technologist of language can work with theoretical problem in research field i.e. development of algorithm, lexicographical question and analysis of syntactic [Bowman, 2015], Typically, language technology and application research are large collection of text and speech in organization. These digital sections can be called as corpora. Corpus is assembled by the collection of text from different source i.e. fiction, newspaper and internet. It can consist of sound recordings with transcription associated. A corpus makes it possible for analyzing the enormous data quantities while uncovering the pattern of speech or writing.

VI. IT POSSIBILITIES AND DATA WORK

Data work or corpus is very natural of the Information Technology. All of the computers can rapidly and painlessly sort, count as well as so forth the volumes of vast material as they are the increasing text which can be readable by the machine already, so there is no effort for the data entry of linguist [Alshawi, 1992].

IT appear now will be to have to much offer. Point given below refers to the primarily or naturally independent produced text instead of data elicited, although they would be applied on latter; manipulation of automatic data is very useful for the material market by the linguist. Corpus work is value at three different levels which are: Observational level, derivative level and validation level.

In observation level corpora is process very simple by the routines of cordance and display usefully as the language phenomena of both recording and drawing the attention. This will be the one of recent use of IT for study of linguistic, can remain important as the corpora will get larger so that become tough in order to digest the concordance of information [Allen, 1995]. At this level the important issue is about the representativeness versus corpus coverage. While the obvious corpora use is basics of grammar because they become important in lexicographers' field. There will be the presumption that they are large enough to mass of the miscellaneous stuff which is taken from newspaper and it can act as the representative of mainstream, regular as well as common phenomena. Information Technology can make this possible by applying the low-level process of linguistic. At same time when the simple concordance is very useful then Information Technology can make that easily possible in order to apply the low-level process of linguistic of uncontroversial but it will be kind of helpful for example tagging of the categories of syntactic, lemmatization and local syntactic constituents labelling i.e. group of very/noun and some of the work which can make sense (referring to the different set of the sense of dictionary.

Using corpus derivative level is very useful and interesting but it is more challenging. As it is foreshadowed by the collation at statistics first level simple frequency but its main aim to perform the analysis much more about data in order to derive patterns automatically. Lexical collocations, subcategorization, terminology structure, grammar induction and behavior [Biber, 1994]. Following analysis type presuppose some notion intuitively by name of game as basis of choosing the formal and attribute model having the specification of what is automatically sought and instance of model is discovered by actual algorithm. The example however can illustrate about range of outputs which are useful for the process of the principle to deliver linguist will not be indicative merely the set of discourses actually, but the genre which are high in order can get the definition based on the membership of class (by using the centroid vectors in analogy) also the word for genre is label in lexicon.

Validation level is that in which there are two area of IT utility for linguistic overlap. IT Principle can offer big opportunity here by making the possible theory evaluation for linguistic phenomenon in systematic way i.e. objective and comprehensive way which are against natural corpus [Biber, 1994]. Additionally, using IT for theory validation against corpus required theory application automatic procedure. That will be obvious that corpus analysis will be for lexicon ad text which are bring together, for example, in order to select the sub-lexicon domain, which would be linked with semantic as well as syntactic preference of grammar domain will be ground in the corpus of text.

Comparison

The law in other provision which can gives the people of the British the right for obtaining data is the sweeping right for the US to accessing the data. The experiment would be conducted by using the 20years-old law of British which can entitles about the individual to see that the data about them by the country companies. Law can provide the similar access of data right from the rule of European which is called GDPR (Growth Data Protection) this can offer the sense of how will be new law would be implemented or play out. However, [Bowman, 2015] in this field it will be more usual that British National Corpus for developing different set for the selection criteria which can draw the both intuitively and conventional notion of acceptable genre along with their samples. But this will be very far from the basics of rigorous or scientific to claim for the proper status which result in the facts of linguistic.

VII. IT ACTUALITIES

Now having the rehearsed the potential utilities of IT in general for the linguistic we can ask that how far has IT has actually developed the impact on utility? Additionally, has any direct impact through the data computationally-derived or through the validation of model? Or was it will be in-directed through the computation paradigm recognition? In relation with the data this can influence as the clearest respect of statistics and allow that used language behavior would be influenced by the frequency [Black, 1995]. This may seem as an obvious language property, but acknowledge the computational paradigm which bring it into the open. At theory level, the computation paradigm focuses not too much on the rules: a familiar desideratum of linguistic as a rule application. When the computation work adopts the declarative instead of procedural approach concern will be always with what happens when the declarations are executed [Boguraec, 1989]. Hence overall though this will be the informal judgement in which the impact of IT on the linguistics as a whole have been light and more peripheral than the substantive.

VIII. CONCLUSION

However, it may be hard to discern the significant impact of IT on linguistic outside the area which labelled as computational linguistics, because IT is now pervasive generally, whether the Chomsky's Programme of minimalist which is appeared to be invoke with some notion to encountered in the computational linguistic, and it can also demonstrate there will be any material which is influence from IT.

However, NLP will be force by tackling some of the task i.e. interactive inquiry, to address the topic i.e. dialogue structure, and language processing and automation speech continue to make a progress often with the surprising success by the alien means as in use of Hidden Markov Model for the recognition of speech, there will be much for the linguistic to gain from looking at both what if finds and how the computation can do things. It is something belong to caricature to see who engaged with the computation as crass technoracts for the non-computational theory for whom the express will be an oxymoron as well as linguists as toffee-nosed unwillingly snobs to inspect the rude levers and cranks of mechanical and huge chase between them. But there will be gap which need to be bridged because of linguists and especially the theorists instead of other those whose metaphysics are anti-computational resolutely in any sense whether there will be everything learnt from the appreciating the distinction between real, assumed and ideal computation.

REFERENCES

1. Andersen, G. (Ed.). (2012). Exploring Newspaper Language: Using the web to create and investigate a large corpus of modern Norwegian (Vol. 49). John Benjamins Publishing.
2. Bennett, G. R. (2010). Using corpora in the language learning classroom: Corpus linguistics for teachers. University of Michigan Press.
3. Bowman, S. R., Angeli, G., Potts, C., & Manning, C. D. (2015). A large annotated corpus for learning natural language inference. ArXiv preprint arXiv: 1508.05326.
4. Breyer, Y. (2008). Learning and teaching with corpora: Reflections by student teachers. Computer Assisted Language Learning, 22(2), 153-172. Brown, L. D., Hua, H., and Gao, C. 2003. A widget framework for augmented interaction in SCAPE. In *Proceedings of the 16th Annual ACM Symposium on User Interface Software and Technology* (Vancouver, Canada, November 02 - 05, 2003).
5. Allen, J. (1995), Natural Language Understanding, Second Edition, (Palo Alto, CA: Benjamin/Cummings).
6. Alshawi, H. (1992) (ed.), The Core Language Engine, (Cambridge MA: MIT Press).
7. Ballim, A., Wilks, Y. and Barnden, J. (1991), ‘Belief Ascription, Metaphor, and Intensional Identification’, Cognitive Science, 15 (1), 133-171.
8. Biber, D. (1994), ‘Representativeness in Corpus Design’, in Zampolli, Calzolari and Palmer (1994).
9. Black, E. et al. (1996), ‘Beyond Skeleton Parsing: Producing a Comprehensive Large-Scale General-English Treebank with Full Grammatical Analysis’, COLING96, Proceedings of the 16th International Conference on Computational Linguistics, Copenhagen, 107-112.
10. Boguraev, B. K. and Briscoe, E. J. (1989) (eds.), Computational Lexicography for Natural Language Processing, (London: Longman).
11. Abdumanapovna, S. A. (2018). The Contemporary Language Studies with Corpus Linguistics. *Proceedings of the 2nd International Conference on Digital Technology in Education - ICDTE 2018*.