A Survey of Natural Language Processing Techniques

Vishal Gupta
Assistant Professor, Computer Science and Engineering
UIET, Panjab University
Chandigarh, India
vishal@pu.ac.in

Abstract— Processing of natural language is branch of linguistics, artificial intelligence & computer science and its purpose is to have interaction among natural language of human beings and computers. We can say it is related to field of computer–human interaction. There are different challenges in this field like understanding of natural language i.e. allowing machines to have understanding from natural language of human beings. Mostly available tasks of natural language processing are: analysis of discourse, morphological separation, machine translation, generation and understanding of natural language, recognition of named entities, part of speech tagging, recognition of optical characters, recognition of speech and analysis of sentiments etc. Current research in NLP is showing more interest on learning algorithms which are either unsupervised or semi-supervised in nature. These techniques of learning can perform this task of learning from data which is not annotated manually with required answers or by applying mixture of non-annotated & annotated data. Normally, this job is very hard as compared to learning which is supervised & usually shows little correct results for particular amount of data as input. But there is large quantity of data is available which is non annotated in nature i.e. whole contents available on world wide web and it normally produces less accurate results. This paper discusses about a survey of different techniques of natural language processing.

Keywords- NLP; natural language processing; natural language understanding; mining of text

I. INTRODUCTION

Processing of natural language [1][6]is branch of linguistics, artificial intelligence & computer science and its purpose is to have interaction among natural language of human beings and computers [9]. We can say it is related to field of computer–human interaction. There are different challenges in this field like understanding of natural language i.e. allowing machines to have understanding from natural language of human beings. Mostly available tasks of natural language processing are: analysis of discourse, morphological separation, machine translation, generation and understanding of natural language, recognition of named entities, part of speech tagging, recognition of optical characters, recognition of speech and analysis of sentiments etc. Current research in NLP is showing more interest on learning algorithms which are either unsupervised or semi-supervised in nature. These techniques of learning can perform this task of learning from data which is not annotated manually with required answers or by applying mixture of non-annotated & annotated data. Normally, this job is very hard as compared to learning which is supervised & usually shows little correct results for particular amount of data as input. But there is large quantity of data is available which is non annotated in nature i.e. whole contents available on world wide web and it normally produces less accurate results.

Before 1980s many systems related to natural language processing were on the basis of complicated rules which were written by hand. But in the era of 1980s revolution has come in field of natural language processing by the invention of learning procedures based on machine learning techniques for processing he natural language. It has become possible only because of introduction of two very popular concepts of theories of Chomskyan related to linguistics (also termed as grammars of transformation) and law by Moore. These two concepts have de-motivated linguistics related to corpus and these concepts relate the techniques of machine learning to processing of language. Very popular techniques of machine learning like decision trees have replaced existing techniques of rules written by hands. Then the research has shown interest in statistics based techniques [4] that can apply decisions which are probabilistic on the basis of values of weights related with different features for formulating input data. These techniques are usually very much robust while we give input which is unfamiliar, particularly the input which involves errors (which are very much usual in case of data belonging to real world) & generates very much dependable results while integrating it with lengthy system containing many subtasks.

Much of earlier success has been attained in case of machine translation, because of good quality of research-work in IBM at which regularly very much complex models of statistics were designed and implemented. These techniques were availing benefits from earlier textual corpus which was multilingual and was developed by European Union of European and Canadian Parliament due to presence of laws for converting all proceedings of government to different official languages used in different systems belonging to government. But, many of other techniques relied on this corpus and it was main drawback for success of those techniques. Due to this, large amount of research has been done for techniques of learning effectively by using small quantity of data. This paper discusses about a survey of different techniques of natural language processing [8].

II. TECHNIQUES OF NATURAL LANGUAGE PROCESSING [1][8]

A. Machine Translation

It is the process of translation [5] of text automatically from any human language to other human language. It is very hard problem & is and associated with problems named as AI-complete. For completely solving the problems of translation, It demands to possess different kinds of knowledge which humans beings have (i.e. Knowledge of semantics, grammar and concepts related to real world).

B. Analysis of Discourse

The task of discourse analysis has a many related jobs to do. One such job is determining structure of discourse of text which is connected, that is kind of relationships of discourse among lines like: contrast and explanation. One more job is identifying & categorizing acts of speech in the particular text. For example content question , yes/no questions, assertion and statement etc.

C. Morphological Splitting

Split terms to separate morphemes & recognize category of corresponding morphemes. Main problem in this job is that it relies largely upon complication of term structures in that language which we are considering. There is very simple morphology of English language, particularly in case of morphology related to inflection & hence it is usually feasible of ignoring the job completely & normally make different feasible forms of any term. For example treating opened, opens and opening as different terms. But in case of Turkish language, this technique is not feasible because every entry in Turkish dictionary can have large number of feasible forms of a word.

D. Generation and Understanding of Natural Language

Generation of natural language involves translate information into easily readable language of beings from computerized databases. Understanding of natural language involves changing text sections to much formal notations like structures related to logic in first order which are very easier to manipulate by programs. Moreover it deals with recognition of semantics using many feasible semantics obtained using expressions of natural language that is normally in form of organized notations present in concepts of natural languages. Generation of ontology & meta-model in language are very suitable solutions and are empirical in nature. Formalization of semantics of natural languages by making assumptions like assumption of closed term vs assumptions of open term, assumptions of objective vs subjective in absence of confusions is required in generation of formalizations of semantics.

E. Identification of Named Entities

With input text, identify terms can be labeled as named entities like places names, people names & also to identify to which types these named entities belong for example organization, location or person. Capitalization is although helpful for identifying the names present in languages like English, but it is not helpful in identifying type of names. Moreover capitalization is not the sufficient criteria for identifying the names because, 1st character in a line is also written in capital case and also these names usually span many terms any few of them are written as capitalized. Moreover many languages which are non western like Hindi, Arabic, Punjabi and Chinese etc. are not possessing feature of capitalization. Also many languages having capitalization feature can not throughout apply it for identifying the names e.g. In German language all noun terms are capitalized irrespective of if they point to names or not, & Spanish and French also not use capitalization for names which treat like adjectives.

F. Marking Part of Speech

Input a line, identify and mark part of speech [3] in case of every term. Many terms, usually common terms may be treated with more than one parts of speech e.g. a term book might be treated as noun or can be treated as verb. Another term set might be treated as verb, noun or adjective. Also many languages can show much of this type of ambiguity. English Language having very less inflectional morphology is very much prone to this ambiguity. Chinese language also show this ambiguity as this language is of type a tonal language while performing verbalization.

G. Optical Character Recognition

Optical character recognition deals with identifying text from images denoting text in printed form. Success of a OCR for any language depends on quality of images denoting the printed text in same language.

H. Recognizing Boundary of Sentences

It deals with finding boundary of lines in given input text. We can normally mark line boundaries by full stop character or other characters like ?, ! etc but problem is that same type of characters may also be used for other motives like forming abbreviations.

I. *Parsing of Text*

Parsing deals grammatical analysis of sentences by forming parse tree of them. We know that usually grammars in natural languages are having ambiguity & particular lines can have more than one feasible analyses. For any particular line there can thousands of possible parses and many of them will be nonsensical for human beings completely.

J. *Recognition of Speech*

Recognition of speech deals with recognizing textual notation of any speech by listening to sound clip of any person. It is very hard problem and is entirely opposite to the task of text to speech conversion. Moreover in case of any natural speech there will be very less number of pauses among consecutive terms & we can say that segmentation of speech is important sub-step of recognition of speech. In many spoken languages, the utterance of sounds denoting consecutive words mix with each other and this process is called as co articulation, that is why changing analog signal of sound into discrete textual characters is very difficult job.

K. *Analysis of Sentiments*

Recognition Sentiment analysis [2] deals with retrieving information subjective in nature normally from collection of text documents like online reviews for finding polarity of particular objects. This process is very much applied for determining It is very much applied in marketing for determining sentiments or reviews of public opinion about social media.

L. *Finding Words Boundary*

It deals with splitting of sections of continuous characters of text to separate terms. In English language, this task is very simple because terms are normally separated using spaces. But in many languages of world like Japanese, Chinese and Thai term boundaries are not marked in this manner & for these languages words segmentation is very difficult job as it demands knowledge and information of terms morphology and vocabulary for these languages.

M. *Word Sense Disambiguation*

In Word sense disambiguation, one word can have many possible senses depending upon its context. It is difficult to identify correct sense of a word used in a particular context in a given sentence. For resolving this problem we can use Word-Net for a particular language which contains a list of words and associated word senses.

III. CONCLUSIONS

Processing of natural language [7] is branch of linguistics, artificial intelligence & computer science and its purpose is to have interaction among natural language of human beings and computers. We can say it is related to field of computer–human interaction. Mostly available tasks of natural language processing are: analysis of discourse, morphological separation, machine translation, generation and understanding of natural language, recognition of named entities, part of speech tagging, recognition of optical characters, recognition of speech and analysis of sentiments etc. Current research in NLP is showing more interest on learning algorithms which are either unsupervised or semi-supervised in nature.

REFERENCES

[1] http://en.wikipedia.org

[2] C. Wu and Y. Chen, "A Survey of Researches on the Application of Natural Language Processing in Internet Public Opinion Monitor," Proceedings of International Confernce on Computer Science and Service System (CSSS), IEEE, 2011, pp. 1035-1038.

[3] B. B. Ali and F. Jarray, "Genetic Approach for Arabic Part of Speech Tagging," International Journal on Natural Language Computing (IJNLC), vol.2, 2013, pp. 1-12.

[4] Z. B. Wu, L. S. Hsu, and C. L. Tan, "A survey on statistical approaches to natural language processing," Technical Report, 1992.

[5] A. Lopez, "Statistical Machine Translation," Iternational Journal of ACM Computing Surveys, vol. 40, 2008.

[6] W. Fan, L. Wallace, S. Rich and Z. Zhang, "Tapping into the Power of Text Mining", International Journal of ACM, Blacksburg, 2005.

[7] R. Mihalcea, H. Liu, and H. Lieberman, "NLP (Natural Language Processing) for NLP (Natural Language Programming), " In Proceedings of CICLING'06, LNCS, Springer, 2006, pp. 319-330.

[8] P. M Nadkarni, L. O. Machado and W. W. Chapman, "Natural Language Processing: An Introduction,"J Am Med Inform Assoc, vol.18, 2011, pp. 544-551.

[9] B. Manaris, "Natural Language Processing: A Human–Computer Interaction Perspective," Appears in Advances in Computers (Marvin V. Zelkowitz, ed.), Academic Press, New York, vol. 47, 1998, pp. 1-66.