Constituency Parser for Clinical Narratives using NLP

Anjali Kedawat, A. Senthil,

Abstract: Clinical parsing is useful in medical domain .Clinical narratives are difficult to understand as it is in unstructured format .Medical Natural language processing systems are used to make these clinical narratives in readable format. Clinical Parser is the combination of natural language processing and medical lexicon .For making clinical narrative understandable parsing technique is used .In this paper we are discussing about constituency parser for clinical narratives, which is based on phrase structured grammar. This parser convert unstructured clinical narratives into structured report. This paper focus on clinical sentences which is in unstructured format after parsing convert into structured format. For each sentence recall ,precision and bracketing f- measure are calculated .

Keywords :clinical narratives constituency parser, phrase structured grammar, probabilistic context free grammar

I. INTRODUCTION

Clinical Parser is basically used for parsing clinical text. Natural Language Processing system is used for unlock information from the EHR (electronic heath records). Clinical narratives are differ according to domain i.e. narratives for radiology reports, cardiology reports etc. are different from each other. Clinical information can be extracted from narrative reports through Natural Language Processing which support automated decision-support and clinical research. Clinical medical records are in free text form. Bikel[4], Charniak[5] are some clinical parser which is used for clinical text parsing. All of these based on lexicalized parsing. Stanford Parser is first parser which is based on lexicalized as well as unlexicalized parsing. In lexicalized parsing parent node is labelled by its main child nodes. It is more accurate but require lot of manual efforts while in unlexicalized parsing data are not labelled.

Stanford parser [1] used both constituency parsing and dependency parsing. Stanford parser used for general English domain, but not suitable for medical domain. In this paper we use constituency parser instead of dependency parser because in constituency parser phrase structured grammar is used in which probability context free grammar is used to include the probability of each grammatical rule which is useful of precision and recall.

Revised Manuscript Received on January 06, 2020.

Anjali Kedawat*, Department of CSE, Gitam University, Visakhapatnam, India .Email:akedawat15@gmail.com

Dr. A. Senthil, Department of CSE, Mody University, Laxmangarh, India. Email: asenthil.cet@modyuniversity.ac.in

It is suitable for medical domain as we use UMLS(Unified Medical Language systems) .UMLS contains software and files which is useful for biomedical vocabularies

Stanford parser is not suitable for biomedical text. Furthermore it mix noun phrases with normal sentence in laboratory test and treatments. Symbols and prepositional phrases are error prone.Finkel et.al[2] use Stanford parser for parsing and GENIA corpus. The disadvantage of it that it have confusion in cell type/cell line and DNA/protein. Huang et al.[3] not define NP in parse tree format and mapping it to corresponding UMLS concept.Bikel use lexicalized parsing only.Mc.Closky et al.[6] give parser which can't handle unknown words as well as prepositional phrases. Gabriel J. Ferrer [7] take regular expression as terminal symbols due to this named abstractions and recursive structures not possible. Alessandro Warth et. Al.[8] give Ohm parser which not use incremental parsing which causes small change in input reparsed whole input. Foster R. Goss.et. al. [9] developed an allergy module on MTERMS NLP system which is used to encode and identify food, environmental and drug allergies and allergic reactions. They use standard terminologies, and novel disambiguation algorithms. Ricky K. Taira [10] develops a natural language processor for radiology reports. They define deficiencies of symbolic methods over statistical natural language processor. The goal of parser module is to create a dependency diagram between words in input sentences. They describe methodology in three ways: first collect a large number of documents from domain of interest, second create the training data by manually indicating the dependency graph, third for each word in sentence define resonance condition occurred or not. It achieves recall of 90% and precision of 89%. Ernestina Menasalvas et al. [11] analyze several tools and frameworks to extract medical entities by applying NLP with NER process ,P.O. El Guedj et al. [12] build Chart parser which is used to analyze large medical corpora.Nuala A. Bennett et al. [13] extract noun phrase from Medline using general parser.

II. PROPOSED METHODOLOGY

In this paper we include USG reports of same format and generate grammar and syntax tree for sentences .This grammar belongs to constituency parser which is easy to generate and include phrase structured grammar .In this we use PCFG (probability context free grammar) where probability is assigned for each production of grammar according to number of occurrences in grammar.

Approximity of each tree can be calculated according to probability of phrase . Clinical narratives are in unstructured format. This unstructured format is converted into structured format by using parser. Clinical narratives are differ report to report and hospital to hospital also. So in this paper we only considering USG report of neck and cheeks of an hospital .For e.g. " There is evidence of small hypo echoic lesion of size 6x2 mm seen in left side of cheek superficial facia", this is one of the sentence in cheek USG after parsing ,before parsing verb ,noun, preposition not included in clinical narratives .By using below PCFG the sentence will come in meaningful form. When for above sentence parse tree will be generated by using following PCFG approximation of tree is $3.6e^{-30}$.First unstructured clinical narratives will give as an input to text maker after it by using rules it will convert into meaningful sentence. Steps for conversion clinical narratives into structured format is shown in Fig.1.

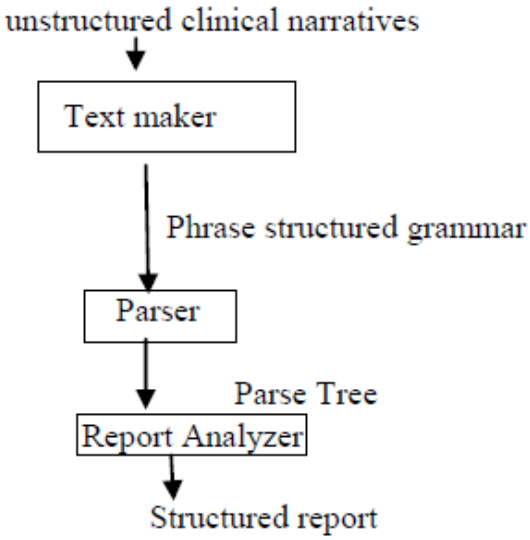

Fig.1. Block Diagram for Clinical Parser Process

S->NPVP 1
NP->EX .1/NPPP .2/NN .1/NPVP .1//NNCDNN .1/DTNP .1/NNS .1/JJNP .1/ NNNP .1
VP->VBZNP .5/VBNPP .3/VBPADJP .2
IN->OF .5/IN .5
PP->INNP 1
EX->There 1
CD->6X2 1
DT->NO .2/THE .5/BOTH .3
ADJP->JJ 1
VBZ->is 1
VBP-> appear .5/are .5
VBN->seen 1
NN->evidence .5/size .05/mm .05/side .05/cheek .05/facia .05 /lymphadenopathy .05/neck .05/region .05/thyroid .05/lesion .05
NNS->lobes .5/glands .5
JJ->normal .5/small .05/hypo .05/echoic .05/superficial .05/left .05 /Submandibular.25

Stanford parser is not suitable for biomedical text. Furthermore, it mixes noun phrases with normal sentence in laboratory test and treatments. Symbols and prepositional phrases are error prone.Finkel et.al use Stanford parser for parsing and GENIA corpus. The disadvantage of it that it has confusion in cell type/cell line and DNA/protein.Huang et al. not define NP in parse tree format and mapping it to corresponding UMLS concept.Bikel use lexicalized parsing only.Mc. Closky et al. give parser which can't handle unknown words as well as prepositional phrases. Gabriel J. Ferrer take regular expression as terminal symbols due to this named abstractions and recursive structures not possible.Alessandro Warth et. al give Ohm parser which not use incremental parsing which causes small change in input reparsed whole input.

For sentence "Both lobes thyroid gland normal" convert it by given grammar in readable form .First this unstructured format will go in text maker and after this by using PCFG which is phrase structured grammar with probability it will converted in to meaningful sentence "Both lobes of thyroid glands appears normal".

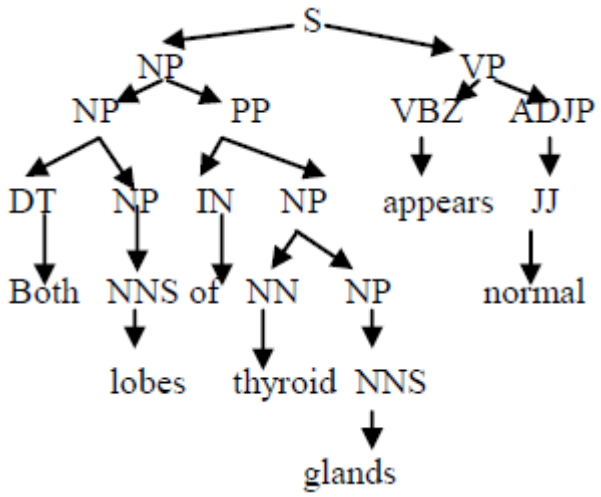

Fig.2. Flow Chart for unstructured to structured clinical narratives

In PCFG we mention probability for each sentence , according to that probabilities approximity of tree can be calculated. The same can be implemented by dependency parser. Constituency parser can also be implemented by shift reduce parser and if conflict arises in any production then it can be removed by selecting production which comes first in sequence .By taking more than one possibility recall, precision and bracketing f-measure can be calculated .If this is compared by gold standard then recall, precision and f-measure can be calculated.

In other example we consider unstructured clinical sentence which will take as an input to text maker and by using rule based techniques which is used on constituency parser converted in to structured format. These rules are generated so that clinical sentence will be converted into meaningful sentences

Kidney normal episodes shapes normal dimensions

↓

Text Maker

Now we will generate rules so that this sentence will be converted into meaningful sentence

Rule 1:noun phrase should followed by verb phrase

Rule 2:Noun singular, noun phrase followed by preposition phrase or noun phrase followed by coordinating conjunction followed by noun singular

Rule 3:verb third person followed by noun phrase

Rule 4:Preposition phrase or subordinating conjunction followed by noun phrase

After applying above rules sentence is converted in meaningful sentence "Kidney has normal episodes and shapes with normal dimensions "For each sentence rules are different. This sentence format depends on clinical reports which also differ from hospital to hospital.

III. RESUT ANALYSIS

Clinical sentence recall and precision calculated by comparison it to Stanford parser parse tree and System generated parse tree(based on our approach) .For sentence "Both lobes of thyroid glands appears normal" .

System Generated parse Tree

(S
(NP(NP(DT Both) (NP(NNS lobes)))
(PP (IN of)(NP(NN thyroid)(NP (NNS glands))))

(VP (VBZ appears)(ADJP (JJ normal)))
)

Stanford Parser Parse Tree

(S
(NP(NP(DT both)(NNS lobes))
(PP(IN of)(NP(NN thyroid)(NNS glands)
)))
(VP(VBZ appears)(ADJP(JJ normal)))
)

System Generated parse Tree has 9 brackets while Stanford has 7 brackets.7of 9 brackets of system generated parse tree correct in comparison to Stanford parser.BR=7/7(100%),BP=7/9(77.7%),BF=2x100x77.7/(100+77.7)(87.45%)

Table1:Bracketing f-measure

Recall	Precision	Bracketing F-Measure
100%	77.7%	87.45%

IV. CONCLUSION

This paper contain rule based technique to convert clinical narrative into structured format. This rule based technique is in form of phrase structured grammar. By using this unstructured clinical narrative will convert in to structured clinical narrative. PCFG is used for approximation of tree. Further we will do it by dependency parser. In dependency parser many techniques available to convert these clinical narratives in structured format.

REFRENCES

1. Hua Xu, Samir Abdel Rahman, Min Jiang, Jung-wei Fan and Yang Huang, An Initial Study of Full Parsing of Clinical Text using the Stanford Parser. In IEEE International Conference on Bioinformatics and Biomedicine Workshops,2011, pp. 607-614.
2. Finkel, J., S. Dingare, and H. Nguyen, Exploiting context for biomedical entity recognition: From syntax to the web. In Proceedings of the Joint Workshop on Natural Language Processing in Biomedicine and its applications, 2004.
3. Huang Y. et al., Improved identification of noun phrases in clinical radiology reports using a high-performance statistical natural language parser augmented with the UMLS specialist lexicon. J Am Med Inform Assoc., 2005. 12(3): pp. 275-85.
4. Bikel, D.M., On the parameter space of generative lexicalized statistical parsing models. 2004.
5. Charniak, E. and M. Johnson, Coarse-to-fine n-best parsing and Maxent discriminative reranking, in Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics: Ann Arbor., Michigan,2005, pp. 173-180.
6. McClosky, D., E. Charniak, and M. Johnson, Effective self-training for parsing, in Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics. Association for Computational Linguistics: New York,2006, pp. 152-159.
7. Gabriel J. Ferrer, An Interactive Parser Generator for context-free grammars. In ACM Digital Library Journal of Computing Sciences in Colleges, Volume 28 Issue 2,2012, pp. 103-110
8. Alessandro Warth and Tony Garnock, Modular Semantic Actions. In ACM digital Library ,2016, pp. 108-119.
9. Foster R. Goss, Joseph M. Plasek, Jason J. Lau, Diane L. Seger, Frank Y. Chang, and Li Zhou, An Evaluation of a Natural Language Processing Tool for Identifying and Encoding Allergy Information in Emergency Department Clinical Notes, AMIA Annu Symp Proc.2014, pp. 580–588.
10. Ricky K. Taira and Stephen G. Soderland, A Statistical Natural Language Processor for Medical Report, AMIA Proc.1999, pp. 970-974..
11. Ernestina Menaselvas et al., Clinical Narrative Analytical Challenges, Springer International Publishing,2016, pp. 23-32
12. P.O. El Guedj and P. Nugues, A Chart Parser to analyze large medical corpora, IEEEXplore Digital Library,2002.
13. Nuala A. Bennett, Qin He, Kevin Powell, Bruce R. Schatz, Extracting Noun Phrases for all of MEDLINE,1999, AMIAInc., pp. 671-675.

AUTHORS PROFILE

Ms. Anjali Kedawat, Assistant professor, Department of CSE,Gitam University, Visakhapatnam

Dr. A. Senthil, Head and Associate Professor, Department of CSE, School of Engineering and Technology,Mody University,Laxmangarh,Rajasthan