ISSN: 2277-9655
(I2OR), Publication Impact Factor: 3.785

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

SIMBIO: A EFFECTIVE APPROACH FOR SIMPLIFYING AGGREGATE MENTIONS IN BIOMEDICAL TEXT

Miss. Sonali V. Gunjal [*], Prof.N.B.Kadu (Guide)[*]
Department of Computer Engineering, Collage, Loni,India

DOI: 10.5281/zenodo.48822

ABSTRACT

One of the major challenge in biomedical named entity recognition (NER) and normalization process is the detection and decision of aggregate(compound) named entities, in which a single entity refers to many concept e.g., SMAD/1/2. Previous research regarding named entity recognition and normalization, some of them have neglected aggregate mentions, apply simply rules for detecting, or perform coordination ellipsis, so that force to require a such method that can easily handle the different types of aggregate mentions. In this paper, we propose a new approach that combines a machine-learning approach with a pattern detection method to recognize the each entity of each aggregate mention. The proposed method effectively handles various types of aggregate mentions. The proposed method provides high performance in detecting and finding aggregate mentions that are: genes, diseases, and chemicals. The proposed system will later increase the performance of sequence as well as unwellness idea recognition, detection and normalization.

KEYWORDS: — Named entity recognition, Simconcept, Composite mention, Gens, Disease.

INTRODUCTION

The proposed system is the only one technique to consistently handle various sorts of aggregate mentions. The proposed method provides high performance in distinguishing and calculating aggregate mentions for various biological things: genes, diseases and chemicals. But the problem is that, these literature researches have centered on just variety of aggregate mention: that is entities with coordination ellipsis. In this paper, it handles six kinds of aggregate entities considered as well as five different varieties and a mixed variety of entities.

Entities with coordination ellipsis: in this type of entities, the entities share part of the entity portion, such as the token “TGF” in “TGF’s 1, 5, and 8.”

Range entity: This is the same like as the Entities with coordination ellipsis, that entities share part of the entity portion, but,in this type , entities provides a range of entities, not a set of entities (e.g., “TGF 1 to 5”).

Independent entity: this is an indivisual single aggregate mention. Total concepts are partitioned into nonoverlapping entity (e.g., “TGF/SMAD/TEC”).

Overlap short pair entity: The long form entity and short form of entity referes to same entity. But the short and long form ponting to the same entity identifier.

Independent short pair entity: this is an independent aggregate entity where the two different enitites pointing to the same entity identifier. (e.g., “ectodermal dysplasia”).

Mixed entity: in this a mixed type of combination two mentioned types, like “TGF 1 and 2”—a mix of type 1 and 4.

The three major contributions in this paper are:

1) A new system is implemented to handle all mentioned types of aggregate entities, that all are not implemented together yet.

2) When system executes, on the various bio medical concepts (i.e., gene, disease, and chemical), the proposed methods provide high performance

3) Due to the system can easily handles more than one mention recognition type, the proposed scheme is robust.

LITERATURE SURVEY

In the field of medical text mining, various researches concentrated on automatically extracting important data from available research. The important information is primarily concentrated on a particular topic, like communication between protein [2], [3], protein transportation and restriction method [4]–[6], medicine-disease relation [7]–[9], or RNA procedure extraction [10]. Among the various methods, the use of text analysis or machine learning approach to detect pattern from the text are the very common approach. One of the complicated step regarding to this motivation is automatically detecting medical mentions like e.g., gene, chemical. Also the named entity recognition (NER) is also crucial method. After detecting biomedical concepts, mapping these to a particular identifier available in database is important task. Various globally medical text mining events particularly focuses on these important tasks [11]–[13].

In the biomedical research work Genes, diseases, and chemicals these are very important things as well as they are most famous things [14], [15]. Various normalization researches have main two things: term variation and unclearness [16]–[22]. Various earlier researches have defined different techniques like machine learning etc. methods to handle these two problems. But, individual type of problem that has not been solved well is aggregate entities, in which a one entity may refer to more than one entity (e.g., "TGF 2, 6, and 8"). These entities particularly refer to many concepts; that they are different from things like protein group and chemical compound in which various entities are added to derive a one physical unit.

According to observation, near about 12%-13 % of gene, disease, and chemical entities are aggregate entities, due to which it is required to work with them properly. The proposed system provides easy way for biomedical concepts in very effective fashion. Most of earlier researches have concentrated on document or paragraph [23]–[27] and sentence [28]–[31].

Buyko, [32] proposed a CRF-based technique having: conjunction, conjuncts, and abbreviation antecedent. For example, in "boy or Horse DNA," where "boy" and "horse" are conjuncts, "or" is a conjunction, and "DNA" is an abbreviation antecedent. Implementation of these are performed using technique GENIA [33] substance, achieving 86% exactness. But this technique not achieve good performance because of complexity, Chae, [34] proposed a template-based system that detect the portion of every component for each entity.

SYSTEM ARCHITECTURE

The architecture of the proposed system is show in Fig.1. The proposed system mainly consist two phase. The main purpose of the first phase is separation. In this conditional random field (CRF) is used. In this stage, the input entity is partitioned into sample. Then to the every token, the label is elected on the basis of the most likely chain of phase through the CRF. The second phase is used gather these tokens as a separate entity using a pattern detection technique.

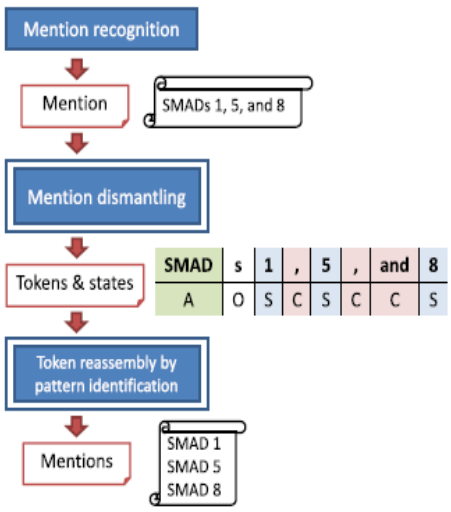

Fig.1. The System Architecture

IMPLEMENTATION DETAILS

As far discussed earlier, our proposed system is considered entity simplification problem as a chain of labelling procedure. To identify the aggregate entities, it detect the aggregation of defined entities and nine phases for implementing a CRF technique [35]:

i) antecedent (A)

ii) strain/suffix(S)

iii) conjunction of entities with coordination ellipsis (C)

iv) conjunction of range entities (CR)

v) left parentheses of abbreviation pair (L)

vi) right parentheses of abbreviation (R)

vii) right parentheses of abbreviation, but the abbreviation and long form cannot be separated (Ro)

viii) conjunction of individual mentions (I)

ix) Redundant (O).

The above mentioned phases are type of conjunction that are use to determine the entity types. If any single entity having more than one state, this entity can be referred as mixed type entity.

A. CRF Features

The proposed system is implemented using tmVar's techniques [37] and by using the some properties of this. The proposed system works same as the tmVar, that it separated characters and digits. Character either uppercase or can be lowercase. For example, "TGFs 1 to 2" can be partitioned as "TGF" "s," "1," "to" and "2.". But the major difference between tmVar and our system is that it tmVar works on document where as our system works on single entity. Therefore, after checking all the possibilities for various types of tokens for entity, the proposed system uses several things like suffixes, prefixes or some types that are used to detecting the entity characteristics. In general, near about all entity suffixes for disease and chemical entities are not number. For e.g."Lung and Mouth cancer" (disease) and "Alcindoromycine" and "Marcellomycin " (chemical), which is very difficult to detect without any related data. Therefore, in proposed system, through the collected the semantic characteristics [37]–[39].

In proposed system three types of features are used token, pattern contextual. Token features provide the total digits, uppercase-lowercase characters, words, and special symbols. Pattern features are implemented by removing uppercase word to "A" and any lower to "a". Any digit is replaced by "0". Also, we combine succeeding character and digits to generate new characteristics, such as "CCC" to "C". Then, we can used in full sentence as characteristics. That is, search entity in all text and search and check whether it is available or not. For e.g., in evaluating that how to separate "A1 and A2 A3,", then it need to check the pair "A1 A3" in all text. If present, then it is reasonable to conclude that "A1A3" is a valid and have some sense entity. Otherwise, it force to that A1 should be separated by itself.

B. Token grouping through pattern detection

By examining the features of aggregate entities in training data, proposed system defined four different sequences to calculate the different types of aggregate biomedical entities. To make task easier of entities, the character in antecedent portion must be available in the all entities, the character in conjuncts portion must be substituted by all possible conjunct entities in this portion and conjuncts portion should consists of at least one conjunction. Entities are defined to one of the provided sequence & then reassembly of the entities is performed.

C. Preprocessing

The proposed system defined various heuristic constraints in post processing. In the first constraint, it focuses on some plural entities, such as "TGFs 2 and 4." If such entity found then the character "s" is neglected when they are fetched from aggregate entities. For e.g., the result of "TGFs 2 and 4" is "TGF 2" and "TGF 4." But, it is not applicable in all cases; sometimes the character "s" is actually part of an entity name in each entity. Due to this, it needs to fetch individual entity that does not contain the character "s" and looking for how many times it's appears in the full text. After that if no match is found in full text, then "s" is added to independent entity The next post processing constraints handles any antecedent and prefix characters that cannot be easily separate by the proposed method, like "tri- and diorganton." In this case, system identifies the prefix and modifies the state of tokens properly.

D. SimConcept Corpus

The SimConcept corpus was compiled using five datasets: three for genes, one for diseases, and one for chemicals. For genes, we tend to integrated the BioCreative II cistron normalization control task coaching (281

abstracts) and take a look at (262 abstracts) corpora and the GIA test collection. (http://ii.nlm.nih.gov/DataSets/index.shtml#GIA).

RESULTS AND DISCUSSION

Our proposed system provides efficient way to solve the challenge of recognizing and determining of aggregate named entities in biomedical name entities recognition and normalization process. The techniques mention in proposed system produce great performance in recognizing and finding aggregate entities for three types of biological mentions: genes, diseases, and chemicals.

CONCLUSION

In this paper we have proposed a SimConcept which is a methodology to handle the task of composite named entity simplification. we have a tendency to integrated a CRF based methodology with a pattern identification strategy to consistently decompose the six sorts of composite mentions.The results show that SimConcept handles composite mention simplification effectively. We more used SimConcept to help the bioconcept standardization task.The results counsel that SimConcept is useful for rising standardization performance. Our approach ought to generalize to alternative entity sorts additionally to the 3 ideas that were the main focus of this study: genes, diseases, and chemicals.

REFERENCES

[1] C.-H. Wei, R. Leaman, and Z. Lu, "SimConcept: A hybrid approach for simplifying composite named entities in biomedicine," in *Proc. ACMConf. Bioinformat. Comput. Biol. Health Informat.*, Newport Beach, CA, USA, 2014, pp. 138–146.

[2] M. Krallinger, M. Vazquez, F. Leitner, D. Salgado, A. Chatr-aryamontri, A.Winter, L. Perfetto, L.Briganti, L. Licata, M. Iannuccelli, L. Castagnoli, G. Cesareni,M. Tyers, G. Schneider, F. Rinaldi, R. Leaman, G. Gonzalez, S. Matos, S. Kim, W. J. Wilbur, L. Rocha, H. Shatkay, A. V. Tendulkar, S. Agarwal, F. Liu, X. Wang, R. Rak, K. Noto, C. Elkan, Z. Lu, R. I. Dogan, J.-F. Fontaine, M. A. Andrade-Navarro, and A. Valencia, "The protein-protein interaction tasks of biocreative iii: Classification/ranking of articles and linking bio-ontology concepts to full text," *BMC Bioinformatics*, Suppl 8:S3, 2011.

[3] W. A. Baumgartner Jr., Z. Lu, H. L. Johnson, J. G. Caporaso, J. Paquette, A. Lindemann, E. K. White, O. Medvedeva, K. B. Cohen, and L. Hunter, "An integrated approach to concept recognition in biomedical text," in *Proc 2nd BioCreative Challenge Eval. Workshop*, 2007, pp. 257–271.

[4] H. Poon and L. Vanderwende, "Joint inference for knowledge extraction from biomedical literature," presented at the Human Language Technologies Annu. Conf. North American Chapter Association for Computational Linguistics, Los Angeles, CA, USA, 2010.

[5] L. Hunter, Z. Lu, J. Firby,W. A. Baumgartner, H. L. Johnson, P. V. Ogren, and K. B. Cohen, "OpenDMAP: An open source, ontology-driven concept analysis engine, with applications to capturing knowledge regarding protein transport, protein interactions and cell-type-specific gene expression," *BMC Bioinformat.*, 9:78, 2008.

[6] S. Bethard, Z. Lu, J. H. Martin, and L. Hunter, "Semantic role labeling for protein transport predicates," *BMC Bioinformat.*, 9:277, 2008.

[7] C. C. Yang, H. Yang, and L. Jiang, "Postmarketing drug safety surveillance using publicly available health-consumer-contributed content in social media," *ACM Trans. Manage. Inf. Syst.*, vol. 5, no. 1, art. no. 2, Apr. 2014.

[8] R. I.Do˘gan, A.N´ev´eol, and Z. Lu, "A context-blocks model for identifying clinical relationships in patient records," *BMC Bioinformat.*, Suppl 3:S3, 2011.

[9] J. Li and Z. Lu, "Systematic identification of pharmacogenomics information from clinical trials," *J. Biomed. Informat.*, vol. 45, pp. 870–878, 2012.

[10] Y. Mao, K. Van Auken, D. Li, C. N. Arighi, P. McQuilton, G. T. Hayman, S. Tweedie, M. L. Schaeffer, S. J. F. Laulederkind, S.-J.Wang, J. Gobeill,P. Ruch, A. T. Luu, J.-j. Kim, J.-H. Chiang, Y.-D. Chen, C.-J. Yang, H. Liu, D. Zhu, Y. Li, H. Yu, E. Emadzadeh, G. Gonzalez, J.-M. Chen, H.-J.Dai, and Z. Lu, "Overview of the gene ontology task at BioCreative IV,"*Database*, vol. 2014, bau086, 2014.

[11] C. N. Arighi, C. H. Wu, K. B. Cohen, L. Hirschman, M. Krallinger, A. Valencia, Z. Lu, J. W. Wilbur, and T. C. Wiegers, "BioCreative-IV virtual issue," *Database*, vol. 2014, bau039, 2014.

[12] Z. Lu, H.-Y. Kao, C.-H. Wei, M. Huang, J. Liu, C.-J. Kuo, C.-N. Hsu, R. T.-H. Tsai, H.-J. Dai, N. Okazaki, H.-C. Cho, M. Gerner, I. Solt, S. Agarwal, F. Liu, D. Vishnyakova, P. Ruch, M. Romacker, F. Rinaldi, S. Bhattacharya, P. Srinivasan, H. Liu, M. Torii, S. Matos, D. Campos, K. Verspoor, K. M. Livingston, and W. J. Wilbur, "The gene normalization task in BioCreative III," *BMC Bioinformat.*, Suppl 8:S2, 2011.

[13] C. N. Arighi, Z. Lu, M. Krallinger, K. B. Cohen,W. J.Wilbur, A. Valencia, L. Hirschman, and C.H.Wu,"Overviewof the BioCreative IIIworkshop," *BMC Bioinformat.*, Suppl 8: S1, 2011.

[14] A. N´ev´eol, R. I. Do˘gan, and Z. Lu, "Semi-automatic semantic annotation of PubMed queries:Astudy on quality, efficiency, satisfaction," *J. Biomed. Informat.*, vol. 44, pp. 310–318, 2011.

[15] R. I. Dogan, G. C. Murray, A.N´ev´eol, and Z. Lu, "Understanding PubMed user search behavior through log analysis," *Database*, vol. 2009, bap018, 2009.

[16] R. Leaman, R. I. Do˘gan, and Z. Lu, "DNorm: Disease name normalization with pairwise learning to rank," *Bioinformatics*, vol. 29, pp. 2909–2917, 2013

[17] C.-H.Wei, H.-Y. Kao, and Z. Lu, "SR4GN: a species recognition software tool for gene normalization," *Plos One*, 7(6): p. e38460, 2012.

[18] T. Rockt¨aschel, M. Weidlich, and U. Leser, "ChemSpot: A hybrid system for chemical named entity recognition," *Bioinformatics*, vol. 28, pp. 1633– 1640, 2012.

[19] C.-H. Wei and H.-Y. Kao, "Cross-species gene normalization by species inference," *BMC Bioinformat.*, vol. 12, S5, 2011.

[20] M. Torii, K. Wagholikar, and H. Liu, "Detecting concept mentions in biomedical text using hidden Markov model: Multiple concept types at once or one at a time?," *J Biomed. Semantics*, 5(1):3, 2014.

[21] R. Leaman, C.-H.Wei, and Z. Lu, "tmChem: a high performance approach for chemical named entity recognition and normalization," *J Cheminformat.*, 7(Suppl 1):S3, 2015.

[22] S. Van Landeghem, J. Bj¨orne, C.-H. Wei, K. Hakala, S. Pyysalo, S. Ananiadou, H.-Y. Kao, Z. Lu, T. Salakoski, Y. Van de Peer, and F. Ginter, "Large-scale event extraction from literature with multi-level gene normalization,"*PloS One*, 8(4): e55814, 2013.

[23] G. Leroy, J. E. Endicott, O. Mouradi, A. Kauchak, Melissa, and L. Just, "Improving perceived and actual text difficulty for health information consumers using semi-automated methods," presented at the American Medical Informatics Association Annu. Symp., Chicago, IL, USA, 2012.

[24] E. Ong, J. Damay, G. Lojico, K. Lu, and D. Tarantan, "Simplifying text in medical literature," *J. Res. Sci., Comput. Eng.*, vol. 4, pp. 37–47, 2007.

[25] A. Siddharthan, "Syntactic simplification and text cohesion," *Res. Lang. Comput.*, vol. 4, pp. 77–109, 2006.

[26] R. Chandrasekar and B. Srinivas, "Automatic induction of rules for text simplification," *Knowledge-Based Syst.*, vol. 10, pp. 183–190, 1997.

[27] D. Kauchak, "Improving text simplification language modeling using unsimplified text data," presented at the 51st Annu. Meet. Association Computational Linguistics, Sofia, Bulgaria, 2013.

[28] S. B. Silveira and A. Branco, "Enhancingmulti-document summaries with sentence simplification," presented at the Int. Conf. Artificial Intelligence, Las Vegas, NV, USA, 2012.

[29] Y. Peng, C. O. Tudor,M. Torii, C. H.Wu, and K. Vijay-Shanker, "Isimp: A sentence simplification system for biomedical text," presented at the IEEE Int. Conf. Bioinformatics Biomedicine, Philadelphia, PA, USA, 2012.

[30] D.Vickrey and D.Koller, "Sentence simplification for semantic role labeling," presented at the 22nd Int. Conf. Computational Linguistics, Stroudsburg, PA, USA, 2008.

[31] L. Vanderwende, H. Suzuki, C. Brockett, and A. Nenkova, "Beyond Sum- Basic: Task-focused summarization with sentence simplification and lexical expansion," *Inf. Proc. Manag.*, vol. 43, pp. 1606–1618, 2007.

[32] E. Buyko, K. Tomanek, and U. Hahn, "Resolution of coordination ellipses in biological named entities using conditional random fields," presented at the 10th Conf. Pacific Association for Computational Linguistics, Melbourne, Australia, 2007.

[33] J.-D. Kim, T. Ohta, Y. Tateisi, and J. I. Tsujii, "GENIA corpus—A semantically annotated corpus for bio-text mining," *Bioinformatics*, vol. 19, pp. i180–i182, 2003.

[34] J. Chae, Y. Jung, T. Lee, S. Jung, C. Huh, G. Kim, and H. Oh, "Identifying non-elliptical entity mentions in a coordinated NP with ellipses," *J. Biomed. Inf.*, pp. 139–152, 2013.

[35] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," presented at the Int. Conf. Machine Learning, Williamstown, MA, USA, 2001.

[36] D. C. Liu and J. Nocedal, "On the limited memory BFGS method for large scale optimization," *Math. Program. B*, vol. 45, pp. 503–528, 1989.

[37] C.-H. Wei, B. R. Harris, H.-Y. Kao, and Z. Lu, "tmVar: A text mining approach for extracting sequence variants in biomedical literature," *Bioinformatics*, vol. 29, pp. 1433–1439, 2013.

[38] D. M. Lowe, P. T. Corbett, P. Murray-Rust, and R. C. Glen, "Chemical name to structure: OPSIN, an open source solution," *J. Chem. Inf. Modeling*, vol. 51, pp. 739–753, 2011.

[39] L. Tanabe andW. J.Wilbur, "Tagging gene and protein names in biomedicaltext," *Bioinformatics*, vol. 18, pp. 1124–1132, 2002.