Word Sense Disambiguation for Biomedical Text Mining Using Definition-Based Semantic Relatedness and Similarity Measures

Ahmad Pesaranghader, Ali Pesaranghader, and Norwati Mustapha

***Abstract*—Automatically identifying the intended sense of ambiguous words improves the performance of clinical and biomedical applications. This paper by proposing Optimized Gloss Vector relatedness and Adapted Gloss Vector similarity measures, two enhanced semantic measures based on Gloss Vector relatedness measure (GV), evaluates their effectiveness over the task of word sense disambiguation (WSD) in the biomedical domain. Generally, GV measure, after computation of the concepts' gloss vectors using their definitions and an external corpus, quantifies the degree of relatedness as the cosine of the angle between two input concepts' computed gloss vectors. We use Pointwise Mutual Information (PMI) and Medical Subject Heading (MeSH) Structure for GV optimization and similarity adaptation respectively. The experimental result on the WSD dataset shows the proposed definition-based measures outperform other semantic measures in terms of accuracy.**

***Index Terms*—Word sense disambiguation, PMI, semantic relatedness, semantic similarity, MeSH, biomedical text mining, bioinformatics.**

I. Introduction

Word Sense Disambiguation (WSD) task attempts automatically identify the true sense of an ambiguous word based on its context. In our work, the set of possible meanings for a term (word) is the Concept Unique Identifiers (CUIs) associated with that term in the Unified Medical Language System (UMLS). Thus, when performing WSD of biomedical terms, our goal is to assign a term one of its possible CUIs considering its context. This identification of the intended sense of ambiguous terms improves the performance of clinical and biomedical applications such as medical coding and indexing for quality assessment, cohort discovery and other secondary uses of data. In this paper, by introducing two definition-based measures (optimized and adapted versions of Gloss Vector relatedness measure) we try to enhance WSD in the biomedical textual data.

Manuscript received February 8, 2014; revised April 17, 2014.

Ahmad Pesaranghader is with the Faculty of Creative Multimedia, Multimedia University, Jalan Multimedia 63100, Cyberjaya, Selangor, Malaysia (e-mail: ahmad.pgh@sfmd.ir).

Ali Pesaranghader and Norwati Mustapha are with the Faculty of Computer Science and Infromation Technology, Universiti Putra Malaysia, 43400 UPM, Serdang, Malaysia (e-mail: ali.pgh@sfmd.ir, norwati@upm.edu.my).

II. Related Works

A. Related Works in Biomedical WSD

Existing methods for automatically disambiguating words in biomedical text are divided into four groups: supervised, semi-supervised, unsupervised, and knowledge-based. Considering knowledge-based method, UMLS::SenseRelate is an open source Perl package[1] designed in [1] to assign UMLS concepts to ambiguous terms in biomedical text, as each possible sense of a term gets a score by summing up the similarity between that sense and terms surrounding the ambiguous term in a given window of context. The sense with the highest score gets assigned to the target term. Once the terms surrounding the target term are identified, the algorithm computes the similarity between the different senses of the target term and each of the surrounding terms using other open source Perl package UMLS::Similarity [2] developed to calculate the similarity between biomedical terms. In our experiments we employ UMLS::SenseRelate for the WSD task. We evaluate our proposed semantic measures against existing ones after adding our developed measures into the UMLS::Similarity package.

B. Semantic Similarity Measures

Existing semantic similarity measures get divided into three groups: 1) path-based, 2) path and depth-based and 3) path and information content (IC) based. Path relies on the shortest path in a taxonomy while depth considers depth of concepts, and IC incorporates the probability of the concept occurring in a corpus of text.

1) Path-based measure

Rada *et al.* [2],

$$sim_{path}(c_1, c_2) = \left(shortest\ is-a\ path(c_1, c_2)\right)^{-1} \quad (1)$$

2) Path-based and depth-based measures

Wu and Palmer [3], where LCS is the least common subsumer concept of the two concepts c_1 and c_2.

$$sim_{wup}(c_1, c_2) = \frac{2 \times depth(LCS(c_1, c_2))}{depth(c_1) + depth(c_2)} \quad (2)$$

Leacock and Chodorow [4], where minpath is shortest path,

[1] http://search.cpan.org/dist/UMLS-SenseRelate/
[2] http://search.cpan.org/dist/UMLS-Similarity/

and D is the taxonomy's total depth

$$sim_{lch}(c_1, c_2) = -\log\left(\frac{\text{minpath}(c_1, c_2)}{2 \times D}\right) \quad (3)$$

Nguyen and Al-Mubaid [5], where *D* is the total depth of the taxonomy, and $d = depth(LCS(c_1,c_2))$.

$$sim_{nam}(c_1, c_2) = \log(2 + (\text{minpath}(c_1, c_2) - 1) \times (D - d)) \quad (4)$$

3) Path-based and IC-based measures

Resnik [6] uses the LCS's information content (IC) as the estimated similarity while $IC(c) = -\log(P(c))$.

$$sim_{res}(c_1, c_2) = IC(LCS(c_1, c_2)) \quad (5)$$

Jiang and Conrath [7] is actually a distance measure which is convertible to a similarity measure.

$$dis_{jcn}(c_1, c_2) = IC(c_1) + IC(c_2) - 2 \times IC(LCS(c_1, c_2)) \quad (6)$$

Lin [8],

$$sim_{lin}(c_1, c_2) = \frac{2 \times IC(LCS(c_1, c_2))}{IC(c_1) + IC(c_2)} \quad (7)$$

The IC notion for similarity measures are so intuitive and straightforward that it has been used in many studies whether directly [9] or indirectly [10].

C. Gloss Vector Semantic Relatedness Measure

Patwardhan and Pedersen [11] introduced the Gloss Vector relatedness measure (GV) by combining concepts' definitions from a thesaurus and co-occurrence data from a corpus. In their approach, every word in the definition of a concept gets replaced by its context vector from the co-occurrence data to compute that concept's gloss vector; then the relatedness is the cosine of the angle between two input concepts associated with gloss vectors. In brief, Gloss Vector measure gets completed by five steps: 1) constructing co-occurrence matrix from a corpus, 2) removing insignificant words by low/high frequency cut-off, 3) exploiting concepts definitions from a thesaurus, 4) calculating concepts' gloss vectors by having step 2 and step 3's results, and finally 5) estimating semantic relatedness for concept pairs.

Due to the vast application of Gloss Vector measure many studies like [12], [13] have attempted to improve this measure's accuracy and performance even further. In [14] the authors have combined this measure idea with IC idea to take advantage of both measures' benefits.

Nevertheless, the Gloss Vector based measures suffer from two important drawbacks: 1) in cut-off step only bigrams frequencies are considered, without noting frequencies of individual terms. We tackle this problem in Gloss Vector measure using Pointwise Mutual Information (PMI), 2) these measures are developed for relatedness measurement rather than similarity. By considering the UMLS hierarchy we adapt Gloss Vector measure for similarity estimation.

III. Data

A. UMLS

In our experiments, the Unified Medical Language System (UMLS) gets used for concepts' definition extraction and also, by taking its hierarchy of concepts into account, we adapt calculated gloss vectors for similarity measurement. The UMLS generally is a knowledge representation framework designed to support biomedical and clinical research. Its fundamental usage is provision of a database of biomedical terminologies for encoding information contained in electronic medical records and medical decision support. It comprises over 160 terminologies and classification systems, and Medical Subject Headings (MeSH) used in this study is one of them. We have limited the scope to 2012AB release of the UMLS.

B. MEDLINE

For the current study we used MEDLINE abstracts as the corpus to calculate IC and to build a first-order term-term co-occurrence matrix for later computation of optimized and adapted gloss vectors used in the proposed semantic relatedness and similarity measures. We employ the 2013 MEDLINE; it contains over 20 million biomedical articles from 1966 to the present. The MEDLINE database covers journal articles from almost every field of biomedicine including medicine, nursing, pharmacy, dentistry, and healthcare.

C. WSD Evaluation Data

We evaluate existing measures on NLM's MSH-WSD dataset [15]. This dataset includes 203 ambiguous terms and acronyms from the 2010 Medline baseline known as target words. Out of these 203 target words, 106 are terms, 88 are acronyms, and 9 have possible senses that are both acronyms and terms. For example, the target word "*cold*" has the acronym "*Chronic Obstructive Pulmonary Disease*" (CUI: C0024117) as a possible sense, and the term "*Cold Temperature*" (CUI: C0009264). The total number of instances is 37,888.

IV. Methods

A. Stage 1: PMI and Optimized Gloss Vector Semantic Relatedness Measure

In computational linguistics, Pointwise Mutual Information (PMI) for two given terms indicates the likelihood of finding one term in a text document that includes the other term. PMI is formulated as:

$$PMI(t_1, t_2) = \log \frac{P(t_1, t_2)}{P(t_1) \times P(t_2)} \quad (8)$$

TABLE I: MEAN ACCURACY OF THE SEMANTIC RELATEDNESS AND SIMILARITY MEASURES ON MSH-WSD

Window Size	SIMILARITY								RELATEDNESS	
	path	lch	wup	nam	res	jcn	lin	AdpGV	GV	OptGV
1	0.538	0.533	0.540	0.543	0.548	0.552	0.556	0.573	0.533	0.573
5	0.653	0.649	0.663	0.671	0.683	0.685	0.689	0.736	0.629	0.744
10	0.675	0.668	0.678	0.682	0.691	0.698	0.698	0.775	0.646	0.787
25	0.693	0.682	0.695	0.698	0.715	0.715	0.721	0.808	0.668	0.812
50	0.704	0.701	0.714	0.716	0.721	0.724	0.726	**0.816**	0.685	**0.828**
75	0.695	0.693	0.709	0.711	0.712	0.715	0.718	**0.809**	0.673	**0.822**

We employ PMI for insignificant features (words) removal in the proposed measures. In order to integrate this statistical association measure into the Gloss Vector measure procedure, in our approach we: 1) ignore the low/high frequency cut-off step in Gloss Vector measure, 2) construct normalized second order co-occurrence matrix using concepts' definitions and first order co-occurrence matrix directly, 3) build PMI-on-SOC matrix by enforcing PMI on the normalized second order co-occurrence matrix to find relative association between concepts (rows of matrix) and words (columns of matrix), and 4) apply low/high level of association cut-off on PMI-on-SOC matrix. As PMI has a limitation for being biased towards low frequency words, the add-one technique is considered. In this technique, before applying PMI on a matrix, all the elements of the matrix are incremented by 1 unit. After this stage, the Optimized Gloss Vector relatedness measure (OptGV) is available.

B. Stage 2: MeSH Hierarchy and Adapted Gloss Vector Semantic Similarity Measure

Here, by considering the concepts' taxonomy, we would construct an enriched second order co-occurrence matrix used for measurement of the similarity between concepts. For this purpose, by using the optimized PMI-on-SOC matrix, we can calculate enriched gloss vectors of the concepts in a taxonomy (MeSH in our case). Our proposed formula to compute the enriched gloss vector for a concept is:

$$Vector(c_j) = \frac{ca_i(c_j) + ia_i(c_j)}{\sum_{k=1}^{m}\left(ca_i(c_k) + ia_i(c_k)\right)} \qquad \forall i <= n \quad (9)$$

$Vector \in R^n$, *n: the quantity of features*
$1 \le j \le m$, *m: the quantity of concepts*

where $ca_i(c_j)$ (concept association) is the level of association between concept c_j and feature i (already computed by PMI in the previous stage); $ia_i(c_j)$ (inherited association) is the level of association between concept c_j and feature i inherited from c_j's descendants in the taxonomy calculable by summation of all c_j's descendants' levels of association with feature i; and finally, the denominator is the summation of all augmented concepts (concepts plus their descendants) levels of association with feature i. All of these levels of association' values are retrievable from optimized PMI-on-SOC matrix. The enriched gloss vectors get used for final similarity estimation of concept pairs in the Adapted Gloss Vector similarity measure (AdpGV).

V. EXPERIMENTS AND RESULTS

We evaluate the relatedness and similarity measures for the WSD task by using UMLS::SenseRelate. We use terms (compound words) instead of single words surrounding the ambiguous term. Different window sizes (without stop-words) for these surrounding terms are tested. The experiments employ MeSH taxonomy in the UMLS Metathesaurus because the possible senses of each of the target terms in the MSH-WSD dataset are from this source. Table I shows the mean of the disambiguation accuracy produced by various measures.

The achieved results indicate while IC measures yield higher accuracy comparing to the path and depth measures, the Adapted Gloss Vector similarity (AdpGV) and Optimized Gloss Vector relatedness (OptGV) measures produce the highest disambiguation accuracy. Moreover, the accuracy obtained for the Gloss Vector measure (GV) shows the weakness of this measure in the WSD task. It should be mentioned that for the GV, OptGV and AdpGV the optimum cut-off points for the elimination of insignificant features (words) are considered. A comparison between OptGV and AdpGV denotes OptGV superiority on the task of WSD.

VI. CONCLUSION

This paper proposed and evaluated Adapted Gloss Vector similarity and Optimized Gloss Vector relatedness measures over the task of word sense disambiguation in the biomedical domain. The experiment results showed these definitions based on semantic measures, which outperform other existing similarity measures in terms of accuracy. For the future works, since the proposed measures belong to the knowledge-based method of WSD, they can be evaluated against supervised, unsupervised and semi-supervised WSD approaches.

REFERENCES

[1] B. McInnes, T. Pedersen, Y. Liu, S. Pakhomov, and G. Melton, "Knowledge-based method for determining the meaning of ambiguous biomedical terms using information content measures of similarity," in *Proc. the American Medical Informatics Association Symposium*, 2011.

[2] R. Rada, H. Mili, E. Bicknell, and M. Blettner, "Development and application of a metric on semantic nets," *IEEE Trans on Systems, Man and Cybernetics*, vol. 19, pp. 17-30, 1989.

[3] Z. Wu and M. Palmer, "Verb semantics and lexical selections," in *Proc. the 32nd Annual Meeting of the Association for Computational Linguistics*, 1994.

[4] C. Leacock and M. Chodorow, "Combining local context and WordNet similarity for word sense identification in WordNet: An electronic lexical database," pp. 265-283, 1998
[5] H. A. Nguyen and H. Al-Mubaid, "New ontology-based semantic similarity measure for the biomedical domain," in *Proc. IEEE Eng Med Biol.*, pp. 623–628, 2006.
[6] P. Resnik, "Using information content to evaluate semantic similarity in a taxonomy," in *Proc. 14th International Joint Conference on Artificial Intelligence*, pp. 448-453, 1995.
[7] J. J. Jiang and D. W. Conrath, "Semantic similarity based on corpus statistics and lexical taxonomy," presented at the International Conference on Research in Computational Linguistics, 1997.
[8] D. Lin, "An information-theoretic definition of similarity," presented at 15th International Conference on Machine Learning, 1998.
[9] A. Pesaranghader, A. Pesaranghader, A. Rezaei, and D. Davoodi, "Gene functional similarity analysis by definition-based semantic similarity measurement of GO terms," in *Proc. 27th Canadian Conference on Artificial Intelligence*, 2014.
[10] A. Pesaranghader, A. Pesaranghader, N. Mustapha, and N. M. Sharef, "Improving multi-term topics focused crawling by introducing term frequency-information content (TF-IC) measure," presented at the 3rd International Conference on Research and Innovation in Information Systems, 2013.
[11] S. Patwardhan and T. Pedersen, "Using WordNet-based context vectors to estimate the semantic relatedness of concepts," in *Proc. EACL 2006 Workshop*, Trento, Italy, 2006.
[12] A. Pesaranghader, A. Pesaranghader, and A. Rezaei, "Applying latent semantic analysis to optimize second-order co-occurrence vectors for semantic relatedness measurement," in *Proc. 1st International Conference on Mining Intelligence and Knowledge Exploration*, 2013.
[13] A. Pesaranghader, A. Pesaranghader, and A. Rezaei, "Augmenting concept definition in gloss vector semantic relatedness measure using Wikipedia articles," in *Proc. 1st International Conference on Data Engineering*, 2013.
[14] A. Pesaranghader and S. Muthaiyah, "Definition-based information content vectors for semantic similarity measurement," in *Proc. 2nd International Multi-Conference on Artificial Intelligence Technology*, 2013.
[15] A. Jimeno-Yepes, B. T. McInnes, and A. R. Aronson, "Exploiting MeSH indexing in MEDLINE to generate a dataset for word sense disambiguation," *BMC Bioinformatics*, 2011.

Ahmad Pesaranghader is a researcher in the fields of computational intelligence, health informatics and computational linguistics. He held a bachelor's degree in computer software engineering. He also obtained a master's degree in knowledge management with multimedia from Multimedia University. His primary research interests are social and semantic web. Next to his working experience, he is the main author of many research papers all related to the foregoing areas of research.

Ali Pesaranghader is a researcher in the fields of intelligent computing, information retrieval, machine learning and computational linguistics. Beside industrial experiences in these disciplines, he held a master's degree in computer science from Universiti Putra Malaysia. He also obtained a bachelor's degree in computer engineering from University of Kashan, Iran. He has been involved in a variety of research works related to his fields of interest.

Norwati Mustapha received her B.Sc. degree in computer science from Universiti Putra Malaysia and the M.Sc. degree in information systems from University of Leeds, England. She also obtained her Ph.D. in artificial intelligence from Universiti Putra Malaysia. She is an active researcher in the areas of data mining, web mining, social networks and intelligent computing. Moreover, she has been working as a senior lecturer and an associate professor at Universiti Putra Malaysia.
Luminous international awards and great experiences in intelligent computing are seen in her resume.