New Modified Semantic Similarity Measure based on Information Content Approach

Nababteh Mohammed[1] Deri Mohammed[1]
1Computer Center, FESA University, Jordan

Abstract
In recent years, the semantic similarity measure has got a great concern especially in NLP, information retrieval and cognitive science. Several approaches have been introduced for computing the semantic similarity score among concepts. This paper presents a modified semantic similarity measure of LIN measure. The proposed method focuses on solving the low similarity score between synonyms, as well as avoiding zero similarity score when the concept has no occurrence in corpus. The proposed measure computes the similarity score using the parents of the compared concepts and least common subsumer. The experiments show that the proposed measure has achieved high correlation against LIN measure on the miller and Charles benchmark dataset, also the MSE value of the proposed measure was 0.263758, on the other hand; the MSE of LIN measure on the same dataset was 0.344196.

Keyword:
Ontology, WordNet, semantic similarity, similarity measures.

1. Introduction

There are many methods that have been developed to measure the semantic similarity among word pairs. These methods are based on information extracted from structured model of ontology [1].The semantic similarity measures play an important role in natural language processing, question answering [2], information retrieval [3], word sense disambiguation [4], and text segmentation [5]. There are different approaches that traditional measures rely on to compute the similarity; these are: knowledge-based approach which uses well-structured ontology such as WordNet to extract the needed information [6]. In recent years, WordNet has gained great attention in semantic similarity measures since it provides useful information for these measures by organizing nouns and verbs hierarchically. Another important approach that the measures use is corpus-based approach [7]; under this approach, concept relationships are derived from their co-occurrence in a corpus [8]. Information content-based measures use this approach to find how much information each concept contains. Information content-based measures assume that each concept has certain amount of information [9], so the similarity between the two concepts is calculated by quantifing how much they share information between them. The information content value of a concept is computed based on the frequency of the concept in a given corpus [10]. This paper has adapted a well-known information content measure called LIN [11] to compute the semantic similarity between two concepts.
The reminder of the paper is organized as follows: section 2 introduces the related work about the semantic similarity measures. Section 3 describes the proposed semantic similarity measure in detail. The experimental study on the proposed measure and LIN is presented in section 4. The conclusion is presented in section 6.

2. Related work

Measures of semantic similarity have been used to estimate the similarity score between two concepts in a given ontology. Traditional similarity measures can be classified into four categories: path-based measures which rely on the distance between concepts in a taxonomy, information content-based measures which are based on the notion of information content [1], feature-based measures which are based on features of concepts and hybrid measures [12] which combine the approaches from different categories [13]. This paper has focused on the information content approach.

2.1 Information Content-based Measures

These measures take the information content (IC) of concepts in the taxonomy into account. This category assumes that the more shared information between two entities, the more similar they are. The general concepts that are located at the top of taxonomy have low information content. The specific concepts that are located at the bottom of the taxonomy have high information content. Information content-based measures can be divided into two groups: the first group is corpus-dependent measures. The second group is corpus-independent measures (taxonomy-based) [14]. The new proposed measure in this paper is based on corpus dependent approach.

2.1.1 Corpus-dependent

The measures in this group use statistical analysis that is extracted from the corpus to compute the similarity value. Resnik [9] proposed corpus-dependent measure that calculates the similarity between two concepts by finding how much they share information between them. It

Manuscript received March 5, 2017
Manuscript revised March 20, 2017

considers the information content of the least common subsumer of the two concepts. This measure uses the ontology to find the instances of concepts, then corpus is used to obtain the frequencies of these instances. Resnik's measure computes the IC through calculating the probabilities of concepts occurring in the corpus.

$$IC(c) = -log\ p(c) \quad \ldots\ldots (1)$$

Where ***c*** is a given concept and ***p(c)*** is the probability of occurring the instances of the concept ***c.*** Probability of the concept is estimated as:

$$p(c) = \frac{freq(c)}{N} \quad \ldots\ldots\ (2)$$

Where ***N*** is the total number of nouns, and ***freq(c)*** is the frequency of instance of concept c occurring in the taxonomy. Resnik's measure estimates the similarity between two concepts by calculating the IC value of the concepts that subsume both of them as follows:

$$sim_{Res}\ (c1, c2) = IC(\ LCS(c1, c2)\ ) \quad \ldots. (3)$$

Where LCS(c1,c2) is the least common subsumer of two compared concepts, the drawback of this measure is that all pairs of concepts with the same LCS will have the same similarity score.

Jiang & Conrath [15] extended Resnik's measure by considering the IC of the individual concepts. This measure computes semantic distance to obtain semantic similarity as follows:

$$Dist_{jcn}(c1, c2) = IC(c1) + IC(c2) - 2 * IC(LCS(c1, c2)) \quad \ldots\ldots\ldots\ldots (4)$$

Semantic similarity is the opposite of the distance:

$$sim(c1, c2)_{jcn}\ \frac{1}{Dist_{jcn}(c1,c2)} \quad \ldots\ldots (5)$$

LIN proposed measure that takes into account the information content of two compared concepts. This method assumes that the information content weight of compared concepts should be considered to measure the similarity score[11]. The similarity between c1 and c2 is calculated by the ratio between the amount of information needed to state the commonality of c1 and c2 and the information needed to fully describe what c1 and c2 are.

$$sim_{Lin}(c1, c2\ ) = \frac{2\ logP(LCS(c1,c2\ ))}{logP(c1)+logP(c2\ )} \quad \ldots. (6)$$

The IC value of LCS is less than or equal to the IC of both concepts c1 and c2. Therefore, the values of this measure vary between 1 and 0. As noted from formula (5), if the IC of LCS, c1 or c2 is zero, then the similarity score will be zero.

The problem with probability calculations for concepts is that a huge number of annotations are needed in order to provide fair coverage of the main taxonomy to get acceptable estimates. Another problem is that the annotations vary from one corpus to another. Furthermore, missing annotations in some concepts causes less accuracy in probability calculations. Therefore, corpus-dependent approach suffers from sparse data and ambiguity problem.

2.1.2 Corpus-independent

Corpus-independent measures have been proposed to avoid sparse data and an ambiguity problem. Unlike the first group, this group doesn't rely on the corpus, and uses information sources that are extracted from structured ontology.

Seco [14] used WordNet as a statistical knowledge base instead of using a corpus to obtain IC value of concepts. This measure assumes that the more the concept has hyponyms, the more abstract it is. Therefore, the concepts with many children hold less information than concepts that are leaves. Since the root node has the largest number of hyponyms, it is the least informative. Thus, leaf concepts located at the bottom of the tree have the maximum information content value [14]. The IC of root node is zero, and the IC of a leaf is one. The IC value of a given concept can be calculated as follows:

3. The proposed semantic similarity measure

As discussed above, the LIN measure computes the similarity between concepts by finding the fraction between the IC of the least common subsumer and the IC of the two compared concepts. As known, the IC of the concepts is computed by finding the occurrence of the concept and their descendant in the corpus; thus, the LIN measure computes the percentage of the occurrence of the two compared concepts to the occurrence of the LCS.

During the experiments on LIN measure, if one of the two compared concepts has IC equal to zero, the LIN measure assigns zero as semantic similarity score between concepts despite of the high similarity between concepts according to the human rating.

If the two concepts are tested by using LIN measure, and they are synonyms for the same concept and one of them has large IC value because it is rarely used in corpus; in contrast, the other concept has a small IC value because it is used in the corpus frequently, the LIN measure in this case will give a small similarity score against Human rate.

If the length between the two concepts C12 and C7 in the taxonomy is large as shown in the Figure 1 and the concepts have big IC score because they are rarely used in the corpus but the synonyms of them are used frequently, then they have small IC value. Thus, the LIN measure returns low semantic similarity score between concepts.

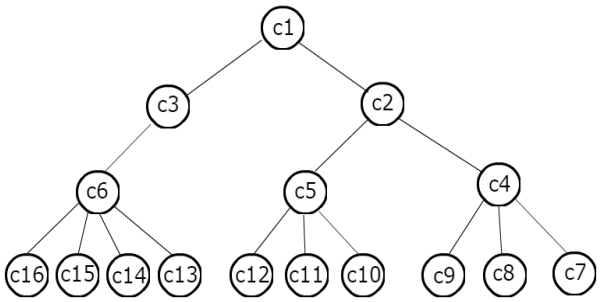

Figure 1. simplified represenatation of WordNet

The proposed measure tries to solve the problem of returning zero as a semantic score when one concept has zero IC score. The method computes the IC of the first parent of the concept that has non zero IC value as well as the proposed method tries to solve the problem of the low similarity score between synonyms in LIN against human rate which occurs from low frequency of concepts in corpus. The proposed measure considers the synonyms as one word and gives them the highest similarity. In this research, the proposed updated semantic similarity measure, which shown in **equation 7,** computes the semantic similarity measure by referring to the first parent which has IC greater than zero to compute the similarity.

The proposed method is extracted from LIN measure but with some modification to make the results near to the human rate.

Equation 7 computes the semantic similarity by dividing the IC of the least common subsumer multiplied by 2 on the sum of IC(P(C1)) and IC(P(C2)), where p(C) denotes to the parent of concept.

$$SIM(C1, C2) = \frac{2*IC(LCS(C1,C2))}{IC(P(C1))+IC(P(C2))} \quad \ldots.(7)$$

4. Experimental Results

The experiments that we have conducted in this research evaluate the proposed semantic similarity measure by comparing the similarity scores produced by the proposed measure against Miller-Charles benchmark dataset [16]; in addition, we compared the similarity scores of LIN measure on the same dataset benchmark against the proposed measure.

Table 1: The result of applying the new modified measure and LIN measure

Word pairs		**Human Rate**	LIN	Err	Sqr. Err	proposed measure	Err	Sqr. Err
Magician	Glass	**0.0275**	0.0663	0.0388	0.001505	0.09335	0.06585	0.004336
Lad	Wizard	**0.105**	0.2241	0.1191	0.014185	0.34155	0.23655	0.055956
Crane	Bird	**0.7425**	0	-0.7425	0.551306	0.850943	0.108443	0.01176
Cock	bird	**0.7625**	0.7881	0.0256	0.000655	0.930014	0.167514	0.028061
Automobile	Car	**0.98**	1	0.02	0.0004	1	0.02	0.0004
Boy	Lad	**0.94**	0.6433	-0.2967	0.088031	1	0.06	0.0036
Monk	oracle	**0.275**	0.1828	-0.0922	0.008501	0.191595	-0.08341	0.006956
Noon	string	**0.02**	0	-0.02	0.0004	0	-0.02	0.0004
Voyage	rooster	0.02	0	-0.02	0.0004	0	-0.02	0.0004
Cord	smile	**0.0325**	0	-0.0325	0.001056	0	-0.0325	0.001056
coast	hill	**0.2175**	0.127	-0.0905	0.00819	0.19414	-0.02336	0.000546
Brother	Monk	0.705	0	-0.705	0.497025	1	0.295	0.087025
Journey	Voyage	**0.96**	0.8277	-0.1323	0.017503	0.857335	-0.10267	0.01054
Forest	graveyard	**0.21**	0.1119	-0.0981	0.009624	0.1706	-0.0394	0.001552
Food	Fruit	**0.77**	0.0956	-0.6744	0.454815	0.103839	-0.66616	0.44377
Jem	Jewel	**0.96**	0.2434	-0.7166	0.513516	0.31453	-0.64547	0.416632
Coast	Shore	0.92	0.96	0.04	0.0016	1	0.08	0.0064
Vegetable	countryside	0.0775	0.0642	-0.0133	0.000177	0.076639	-0.00086	7.41E-07
Monk	Slave	**0.1375**	0.2011	0.0636	0.004045	0.34281	0.20531	0.042152
Food	Rooster	**0.2225**	0.0762	-0.1463	0.021404	0.095302	-0.1272	0.016179
Car	Journey	0.29	0	-0.29	0.0841	0	-0.29	0.0841
Brother	Lad	0.415	**0.24**	-0.175	0.030625	0.29735	-0.11765	0.013842
Furnace	Stove	0.7775	**0.2294**	-0.5481	0.300414	0.26674	-0.51076	0.260876
Implement	Crane	0.42	0	-0.42	0.1764	0.513459	0.093459	0.008735
Asylum	Madhouse	0.9025	0.769	-0.1335	0.017822	0.879	-0.0235	0.000552
Magician	Wizard	0.875	0.1958	-0.6792	0.461313	0.28158	-0.59342	0.352147
implement	Tool	0.7375	0.914	0.1765	0.031152	1	0.2625	0.068906
Midday	Noon	0.855	1	0.145	0.021025	1	0.145	0.021025
		MSE	0.344196			0.263758		

The results of the experiments on the LIN and proposed measure are shown in table 1. Table 1 includes in first column the 28 word pairs which represent Miller-Charles benchmark dataset[16]; column two shows **Human**

Rating which represents the human judgment similarity score of the word pairs. The results of LIN and the proposed measures have been shown in column 3 and 6 respectively. The two columns (**Err**, **Sqr_Err**) in the table contain the **error** which is the difference between the computed similarity scores and human rating as well as the **square error** to compute the mean square error.

The evaluation process in this paper was carried out by finding two factors, namely **correlation** between the score of similarity measure and human rating, in addition to the **mean square error** (MSE). The table 1 shows the MSE for LIN and the proposed measure, as shown the MSE of the proposed method is .263758. In contrast, the MSE of LIN is 0.344196, and that indicates that the error of the proposed method in computing the similarity is less than the error of LIN.
The proposed measure has achieved a high correlation coefficient with the value of 0.76 against LIN that achieved 0.68. This value of correlation coefficient indicates good performance for the proposed measure**.**

5. Conclusion

The paper presented modified semantic similarity measure based on popular measure that is called LIN. The modified measure computes the similarity score between two concepts based on the first parent of the concepts that has IC value greater than 0 in order to solve some problems that exists in LIN measure. The experimental results of applying the proposed measure on Miller-Charles benchmark dataset found out that this measure has achieved a high correlation coefficient (0.76) value with human rating. Furthermore, the proposed measure has got low MSE= 0.263758 against LIN which got MSE= 0.344196. These results indicate good performance for the proposed measure.

References

[1] McInnes, B., & Pedersen, T. (2013). Evaluating measures of semantic similarity and relatedness to disambiguate terms in biomedical text. Journal of Biomedical Informatics, 46(6), 1116-1124.

[2] Tapeh, A.G., Rahgozar, M. (2008). A knowledge-based question answering system for B2C eCommerce, Knowledge-Based Systems 21(8), 946-950.

[3] Srihari, Rohini K., Zhongfei Zhang, and Aibing Rao. "Intelligent indexing and semantic retrieval of multimodal documents." Information Retrieval 2.2-3 (2000): 245-275.

[4] Patwardhan, Siddharth, Satanjeev Banerjee, and Ted Pedersen. "Using measures of semantic relatedness for word sense disambiguation." In International Conference on Intelligent Text Processing and Computational LINguistics, pp. 241-257. Springer BerLIN Heidelberg, 2003.

[5] Li, Y., Bandar, Z. A., & McLean, D. (2003). An approach for measuring semantic similarity between words using multiple information sources. Knowledge and Data Engineering, IEEE Transactions on, 15(4), 871-882.

[6] Atoum, I., Otoom, A., & Kulathuramaiyer, N. (2016). A Comprehensive Comparative Study of Word and Sentence Similarity Measures. International Journal of Computer Applications, 135(1), 10-17.

[7] Mihalcea, R., Corley, C., & Strapparava, C. (2006, July). Corpus-based and knowledge-based measures of text semantic similarity. In AAAI (Vol. 6, pp. 775-780).

[8] Michelizzi, J. (2005). Semantic relatedness applied to all words sense disambiguation (Doctoral dissertation, University of Minnesota).

[9] Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. In Proceedings of the 14th International Joint Conference on Artificial Intelligence, 448–453.

[10] Slimani, T. (2013). Description and Evaluation of Semantic Similarity Measures Approaches. International Journal of Computer Applications, 80(10), 25-33. http://dx.doi.org/10.5120/13897-1851

[11] LIN, D. (1998, July). An information-theoretic definition of similarity. In ICML (Vol. 98, pp. 296-304).

[12] Zhou, Z., Wang, Y., & Gu, J. (2008, November). New model of semantic similarity measuring in wordnet. In Intelligent System and Knowledge Engineering, 2008. ISKE 2008. 3rd International Conference on (Vol. 1, pp. 256-261). IEEE.

[13] Meng, L., Huang, R., & Gu, J. (2013). A review of semantic similarity measures in wordnet. International Journal of Hybrid Information Technology, 6(1), 1-12.

[14] Seco, N., Veale, T., & Hayes, J. (2004, August). An intrinsic information content metric for semantic similarity in WordNet. ECAI (Vol. 16, p. 1089).

[15] Jiang, R., Gan, M., & Dou, X.,. (2013). From ontology to semantic similarity: calculation of ontology-based semantic similarity. The Scientific World Journal,2013.

[16] G.A. Miller. G.A, & Charles W.G. (1991). Contextual correlates of semantic similarity, Language and Cognitive Processes, vol. 6, pp.1–28.