On the effects of using linguistic variables when assessing textual semantic similarity

Jorge Martinez-Gil
Group of Knowledge Representation & Semantics
Software Competence Center Hagenberg
Softwarepark 21, A-4232 Hagenberg, Austria
jorge.martinez-gil@scch.at

Abstract

Evaluating the quality of semantic similarity measures is often performed by computing the degree of correlation between the results obtained by means of an implementation of the given measure and ground truth datasets extracted from human judgments. In this paper we show that, by using linguistic variables for comparing semantic similarity measures and ground truth datasets, the human perception about the accuracy of a given semantic similarity measure can be notably improved. This is particularly relevant in scenarios where a high degree of granularity is not required.

Keywords: Knowledge engineering, Knowledge integration, Semantic similarity measurement

1 Introduction

Textual semantic similarity measurement is a field of research whereby two (sets of) terms are assigned a score based on the likeness of their meaning [13]. Being able to accurately measure semantic similarity is considered of great relevance in many computer related fields since this notion fits well in a number of particular scenarios. The reason is that textual semantic similarity measures can be used for understanding beyond the literal lexical representation of words and phrases. For example, it is possible to automatically identify that specific terms (e.g., Finance) yields matches on similar terms (e.g., Economics, Economic Affairs, Financial Affairs, etc.) or an expert on the treatment of cancer could also be considered as an expert on oncology or tumor treatment.

Traditionally, this problem has been addressed from two different points of view: semantic similarity and relational similarity. However, there is a common agreement about the scope of each of them [2]. Semantic similarity states the taxonomic proximity between (sets of)

terms [8]. For example, automobile and car are similar because they represent the same notion concerning means of transport. On the other hand, the more general notion of relational similarity considers relations between concepts [14]. For example, nurse and hospital are related (since they belong to the healthcare domain) but they are far from represent the same real idea or concept. Due to its importance in many computer-related fields, we are going to focus on semantic similarity for the rest of this paper.

In the field of semantic similarity measurement, results are often validated according the Miller-Charles benchmark dataset [11] which is a widely used dataset for evaluating the quality of new semantic similarity measures for word pairs. The rationale behind this way to evaluate quality is that each result obtained by means of artificial techniques may be compared to human judgments. Therefore, the ultimate goal is to replicate human behavior when solving tasks related to semantic similarity without any kind of supervision. The problem is that evaluation is often performed using the Pearson Correlation Coefficient [1] which involves providing very precise real numbers for qualifying each degree of similarity. However, there are many real cases (fuzzy based systems, question/answering systems, etc.) where semantic similarity is assessed using vague qualifications such as “similar”, “moderately similar”, “not similar at all”, etc. This is possible because in these cases a high degree of granularity is not required since an approximate reasoning is preferred to an exact one.

Our contribution is the analysis and description of a phenomenon which can change the human perception on the accuracy of a given semantic similarity measure. This phenomenon occurs when using linguistic variables for assessing the values from semantic similarity measures and benchmark datasets at the same time, and implies a notably improvement on the overall quality achieved in some specific cases that will be detailed in the core sections of our manuscript.

The rest of this paper is organized as follows: Section 2 describes briefly the traditional process of creating and evaluating a new textual semantic similarity measure. Section 3 describes the effects of using linguistic variables when assessing the accuracy of a new textual semantic similarity. Finally, we draw conclusions and put forward future lines of research.

2 Designing new semantic similarity measures

We are going to describe how to design a semantic similarity measure based on the idea of using quantitative analysis to the study of human literature. Then, we are going to evaluate it according the traditional way, and finally we are going to convert these results into linguistic variables in order to observe the effects on the quality perceived.

One of the main reasons for designing a new semantic similarity measure is trying to solve a problem belonging to a specific domain [5, 6, 7, 9, 10, 15, 16]. However, we are going to

focus on a novel general purpose paradigm that tries to use the human literature rather than an approach based on traditional dictionaries or thesaurus; using human literature is promising since according the book library digitalized by Google[1], the number of words in the English lexicon is currently above a million. Therefore, there are more words from the datasets we are using than appear in any existing dictionary [10].

All information from the book library is stored in datasets represented by means of time series. These time series are sequences of points ordered along the temporal dimension. Each point represents the number of occurrences of a word in a year of the human literature. Therefore, these time series represent the records for the total number of word occurrences per year in the books digitized.

The method that we propose consists of measuring how often two terms appear in the same text statement. Studying the co-occurrence of terms in a text corpus is not a novel idea, since it has been usually used as an evidence of semantic similarity in the scientific literature [4]. We propose adapting this paradigm for designing a new semantic similarity measure by computing the joint probability so that a text expression may contain the two terms together over time. Equation 1 shows the mathematical formula we propose:

$$similarity\,(a,b) = \frac{\text{time units a and b co-ocurr}}{\text{time units considered}} \quad (1)$$

This formula is appropriate because it computes a similarity score so that it is possible to take into account if two terms never appear together or appear together in the same text expressions each time unit. Due to the way data are stored, the minimum time unit that can be considered is a year. Moreover, the result from this similarity measure can be easily interpreted since the range of possible values is bounded by 0 (no similarity at all) and 1 (totally similar).

Table 1 shows us the results using the 1-gram dataset offered by Google, that it is to say the whole set of single text strings. We have repeated the experiment twice. In the first experiment, we have used data from literature written in English between 1800 and 2000. The reason is that there are not enough books before 1800 to reliably quantify many of the queries from the dataset we are using. In the second experiment, we have used books written in English between 1900 and 2000, since there are some modern words (such as car or automobile) that were not invented before. In both cases, we have chosen year 2000 as an upper bound since after this year, the book collection is subject to many changes

[1] http://books.google.com/ngrams/

Table 1. Results for the Miller & Charles benchmark datasets by using the traditional way

		human	1800	1900			human	1800	1900
rooster	voyage	0.08	0.00	0.00	crane	implement	1.68	0.00	0.00
noon	string	0.08	0.00	0.00	brother	monk	2.82	1.00	1.00
glass	magician	0.11	0.00	0.00	implement	tool	2.95	0.45	0.50
chord	smile	0.13	0.00	0.00	bird	crane	2.97	0.40	0.75
coast	forest	0.42	0.80	1.00	bird	cock	3.05	1.00	1.00
lad	wizard	0.42	0.00	0.00	food	fruit	3.08	0.85	1.00
monk	slave	0.55	0.00	0.00	furnace	stove	3.11	0.80	1.00
shore	woodland	0.63	0.70	0.70	midday	noon	3.42	0.55	0.55
forest	graveyard	0.84	0.45	0.85	magician	wizard	3.50	0.50	0.35
coast	hill	0.87	0.60	0.75	asylum	madhouse	3.61	0.00	0.00
food	rooster	0.89	0.00	0.00	coast	shore	3.70	0.80	1.00
cementery	woodland	0.95	0.00	0.00	boy	lad	3.76	0.60	0.65
monk	oracle	1.10	0.00	0.00	journey	voyage	3.84	0.60	0.90
car	journey	1.16	0.70	1.00	gem	jewel	3.84	0.00	0.00
brother	lad	1.66	0.00	0.00	automobile	car	3.92	0.55	1.00
						Fitness		**0.458**	**0.443**

3 On the effects of using linguistic variables

In this context, the conversion into linguistic variables comprises the process of transforming numeric values into grades of membership for linguistic terms. This process is useful in cases where an approximate reasoning is preferred to an exact one. In order to proceed with this process, the numeric values observed in the previous section have to been transformed into a linguistic variable. In many applications it is also possible to assign a value to two or more linguistic variables. This is the case for words with two or more meanings (also known as polysemy), but in this case this kind of assignation has not sense since we assume that each word represents only one object from the real world. Therefore, this transformation is made by assigning to each linguistic variable a balanced interval from the range of possible real values. After converting all the numeric values, it is necessary to represent the values with real values in order to get a numeric value for the fitness. Despite of this process seems to be just the opposite process to the original one, thus, transforming grades of membership for linguistic terms into numeric values before to apply the Pearson Correlation Coefficient, this process does not restore the original values since some information was missed in the original process of conversion where we have only a limited number of linguistic variables to describe all degrees of semantic similarity.

Table 2 shows us the effects of the conversion process when using two linguistic variables. This means that each given word pair can only be similar or not. Results obtained in the two experiments are better that the results achieved according the traditional way. It is also important to remark that using data from 1900 to 2000 is more convenient.

Table 2. Results for the Miller & Charles benchmark datasets by using two linguistic variables

		human	1800	1900
rooster	voyage	not similar	not similar	not similar
noon	string	not similar	not similar	not similar
glass	magician	not similar	not similar	not similar
chord	smile	not similar	not similar	not similar
coast	forest	not similar	similar	similar
lad	wizard	not similar	not similar	not similar
monk	slave	not similar	not similar	not similar
shore	woodland	not similar	similar	similar
forest	graveyard	not similar	not similar	similar
coast	hill	not similar	similar	similar
food	rooster	not similar	not similar	not similar
cementery	woodland	not similar	not similar	not similar
monk	oracle	not similar	not similar	not similar
car	journey	not similar	similar	similar
brother	lad	not similar	not similar	not similar
crane	implement	not similar	not similar	not similar
brother	monk	similar	similar	similar
implement	tool	similar	not similar	similar
bird	crane	similar	not similar	similar
bird	cock	similar	similar	similar
food	fruit	similar	similar	similar
furnace	stove	similar	similar	similar
midday	noon	similar	similar	similar
magician	wizard	similar	similar	similar
asylum	madhouse	similar	not similar	not similar
coast	shore	similar	similar	similar
boy	lad	similar	similar	similar
journey	voyage	similar	similar	similar
gem	jewel	similar	not similar	not similar
automobile	car	similar	similar	similar
	Fitness		**0.464**	**0.548**

Table 3 shows us the effects of the conversion process when using three linguistic variables, this means that each given word pair can be similar, moderately similar or not similar at all. Our hypothesis concerning the fact that information that is missed in the first process can have a positive impact in the final quality perceived seems to be valid. The reason is that once again, results are able to improve the quality obtained by means of traditional evaluation methods. In this case, we have that using data from 1800 to 2000 is more convenient, so at this moment, it is not possible to determine which of the two time series is better in this scope.

Table 3. Results for the Miller & Charles benchmark datasets by using three linguistic variables (not similar, moderately similar and similar)

		human	1800	1900
rooster	voyage	not similar	not similar	not similar
noon	string	not similar	not similar	not similar
glass	magician	not similar	not similar	not similar
chord	smile	not similar	not similar	not similar
coast	forest	not similar	similar	similar
lad	wizard	not similar	not similar	not similar
monk	slave	not similar	not similar	not similar
shore	woodland	not similar	moderately similar	moderately similar
forest	graveyard	not similar	moderately similar	similar
coast	hill	not similar	moderately similar	moderately similar
food	rooster	not similar	not similar	not similar
cementery	woodland	not similar	not similar	not similar
monk	oracle	not similar	not similar	not similar
car	journey	not similar	moderately similar	similar
brother	lad	moderately similar	not similar	not similar
crane	implement	moderately similar	not similar	not similar
brother	monk	similar	similar	similar
implement	tool	similar	moderately similar	moderately similar
bird	crane	similar	moderately similar	moderately similar
bird	cock	similar	similar	similar
food	fruit	similar	similar	similar
furnace	stove	similar	similar	similar
midday	noon	similar	moderately similar	moderately similar
magician	wizard	similar	moderately similar	moderately similar
asylum	madhouse	similar	not similar	not similar
coast	shore	similar	similar	similar
boy	lad	similar	moderately similar	moderately similar
journey	voyage	similar	moderately similar	similar
gem	jewel	similar	not similar	not similar
automobile	car	similar	moderately similar	similar
	Fitness		**0.499**	**0.436**

Table 4 shows us the effects of conversion process when using four linguistic variables. In this case, although the quality achieved is still higher than the traditional one, the values are following a descending trend. This fact makes us start thinking that perhaps a large number of variables can be counterproductive. This has to be confirmed by performing more experiments.

Table 4. Results for the Miller & Charles benchmark datasets by using four linguistic variables (not similar, little similar, quite similar or similar)

		human	1800	1900
rooster	voyage	not similar	not similar	not similar
noon	string	not similar	not similar	not similar
glass	magician	not similar	not similar	not similar
chord	smile	not similar	not similar	not similar
coast	forest	not similar	similar	similar
lad	wizard	not similar	not similar	not similar
monk	slave	not similar	not similar	not similar
shore	woodland	not similar	quite similar	quite similar
forest	graveyard	not similar	little similar	similar
coast	hill	not similar	quite similar	quite similar
food	rooster	not similar	not similar	not similar
cementery	woodland	not similar	not similar	not similar
monk	oracle	little similar	not similar	not similar
car	journey	little similar	quite similar	similar
brother	lad	little similar	not similar	not similar
crane	implement	little similar	not similar	not similar
brother	monk	quite similar	similar	similar
implement	tool	quite similar	little similar	quite similar
bird	crane	quite similar	little similar	quite similar
bird	cock	similar	similar	similar
food	fruit	similar	quite similar	similar
furnace	stove	similar	quite similar	similar
midday	noon	similar	quite similar	quite similar
magician	wizard	similar	quite similar	little similar
asylum	madhouse	similar	not similar	not similar
coast	shore	similar	similar	similar
boy	lad	similar	quite similar	quite similar
journey	voyage	similar	quite similar	similar
gem	jewel	similar	not similar	not similar
automobile	car	similar	quite similar	similar
	Fitness		**0.483**	**0.456**

Table 5 shows us the effects of the conversion process when using five linguistic variables (not similar, little similar, moderately similar, quite similar or similar). Once again, the values obtained follow a descending trend. Moreover, the current results are even worse than those obtained by means of traditional evaluation techniques.

Table 5. Results for the Miller & Charles benchmark datasets by using five linguistic variables (not similar, little similar, moderately similar, quite similar or similar)

		human	1800	1900
rooster	voyage	not similar	not similar	not similar
noon	string	not similar	not similar	not similar
glass	magician	not similar	not similar	not similar
chord	smile	not similar	not similar	not similar
coast	forest	not similar	similar	similar
lad	wizard	not similar	not similar	not similar
monk	slave	not similar	not similar	not similar
shore	woodland	not similar	quite similar	quite similar
forest	graveyard	little similar	moderately similar	quite similar
coast	hill	little similar	quite similar	quite similar
food	rooster	little similar	not similar	not similar
cementery	woodland	little similar	not similar	not similar
monk	oracle	little similar	not similar	not similar
car	journey	little similar	quite similar	similar
brother	lad	moderately similar	not similar	not similar
crane	implement	moderately similar	not similar	not similar
brother	monk	quite similar	similar	similar
implement	tool	quite similar	moderately similar	moderately similar
bird	crane	quite similar	moderately similar	quite similar
bird	cock	quite similar	similar	similar
food	fruit	quite similar	similar	similar
furnace	stove	quite similar	similar	similar
midday	noon	similar	quite similar	moderately similar
magician	wizard	similar	quite similar	little similar
asylum	madhouse	similar	not similar	not similar
coast	shore	similar	similar	similar
boy	lad	similar	quite similar	quite similar
journey	voyage	similar	quite similar	similar
gem	jewel	similar	not similar	not similar
automobile	car	similar	moderately similar	similar
	Fitness		**0.432**	**0.391**

Table 6 shows us the effects of the conversion process when using six linguistic variables, this means that each given word pair can be: not similar, not very similar, slightly similar, fairly similar, very similar, and completely similar. The results obtained after converting the benchmark dataset are the worst from our experiments. This fact confirms the decreasing trend, and therefore, our hypothesis that a large number of linguistic variables could be negative for the overall quality perceived. Additionally we have to remark, that this fact is valid for the two time series considered along all the experiments.

Table 6. Results for the Miller & Charles benchmark datasets by using six linguistic variables (not similar, not very similar, slightly similar, fairly similar, very similar, and completely similar)

		human	1800	1900
rooster	voyage	not similar	not similar	not similar
noon	string	not similar	not similar	not similar
glass	magician	not similar	not similar	not similar
chord	smile	not similar	not similar	not similar
coast	forest	not similar	very similar	completely similar
lad	wizard	not similar	not similar	not similar
monk	slave	not similar	not similar	not similar
shore	woodland	not similar	very similar	very similar
forest	graveyard	not very similar	slighty similar	completely similar
coast	hill	not very similar	fairly similar	very similar
food	rooster	not very similar	not similar	not similar
cementery	woodland	not very similar	not similar	not similar
monk	oracle	not very similar	not similar	not similar
car	journey	not very similar	very similar	completely similar
brother	lad	slighty similar	not similar	not similar
crane	implement	slighty similar	not similar	not similar
brother	monk	very similar	completely similar	completely similar
implement	tool	very similar	slighty similar	slighty similar
bird	crane	very similar	slighty similar	very similar
bird	cock	very similar	completely similar	completely similar
food	fruit	very similar	very similar	completely similar
furnace	stove	very similar	very similar	completely similar
midday	noon	very similar	fairly similar	fairly similar
magician	wizard	completely similar	fairly similar	slighty similar
asylum	madhouse	completely similar	not similar	not similar
coast	shore	completely similar	very similar	completely similar
boy	lad	completely similar	fairly similar	fairly similar
journey	voyage	completely similar	fairly similar	completely similar
gem	jewel	completely similar	not similar	not similar
automobile	car	completely similar	fairly similar	completely similar
	Fitness		**0.430**	**0.400**

From the results obtained in this experiment, it is not possible to determine the optimum number of linguistic variables for a general case. This means in case we wish to use this way to determine semantic similarity in a real system, we should perform a preliminary study in advance for knowing the most appropriate number of degrees of freedom. Nevertheless, these results come from only one scenario. For this reason we report the results from additional experiments in the following subsection.

3.1 Additional experiments

Table 7 shows us the results for additional experiments we have performed. These experiments have been carried out by using well-known semantic similarity measures for solving the Miller-Charles benchmark dataset. These semantic similarity measures are Jaccard, Dice, Overlap and Pointwise Mutual Information (PMI). These measures try to determine the semantic similarity for word pairs by means of their co-occurrences in the same websites from the Web using some search engines. A detailed explanation for these measures can be found in [3]. On the other hand, the first column indicates the results obtained using the traditional evaluation techniques. The rest of columns show us the values after converting the ground truth datasets and the raw values from the measures by using a range between two and six different linguistic variables.

Table 7. Results from additional experiments. Column Traditional shows the results after applying traditional evaluation techniques. Rest of columns shows the results after converting the numeric results. Values in bold represent an improvement over the traditional evaluation technique

	Traditional	2-variables	3-variables	4-variables	5-variables	6-variables
Jaccard	0.259	0.175	**0.279**	0.179	**0.354**	0.174
Dice	0.267	0.175	**0.279**	0.222	0.252	0.226
Overlap	0.382	**0.466**	0.174	0.375	0.373	**0.411**
PMI	0.548	0.433	0.522	0.488	**0.594**	0.522

After these experiments we have to reject the hypothesis stating that a large number of linguistic variables are counterproductive for the final fitness achieved. Unlike the previous experiments, it has been necessary up to five and six linguistic variables to improve the traditional fitness score for some semantic similarity measures.

However, our hypothesis stating that information missed in the process of conversion can have a positive impact in the final quality perceived seems to be correct. This is due to the fact that for every experiment performed using linguistic variables, we have been able to beat the quality obtained by means of the traditional techniques.

4 Conclusions

In this work, we have analyzed and described a phenomenon which can change the human perception on the accuracy of a given semantic similarity measure. This phenomenon happens when transforming the values from semantic similarity measures and benchmark datasets into linguistic variables at the same time.

Our hypothesis concerning the fact that information that is missed in the process of converting real values into linguistic variables can have a positive impact in the final quality perceived seems to be valid for all experiments we have performed.

On the other hand, this work focuses on the study of single words, but we would like to explore the foundations for dealing with more complex expressions [12] as a part of our future work. In fact, our short-term plans include researching about the semantic similarity of short text expressions.

Acknowledgements

We would like to thank in advance the reviewers for their time and consideration. This work has been funded by Vertical Model Integration within Regionale Wettbewerbsfahigkeit OO 2007-2013 by the European Fund for Regional Development and the State of Upper Austria.

References

[1] Ahlgren, P., Jarneving, B., Rousseau, R.: Requirements for a cocitation similarity measure, with special reference to Pearson's correlation coefficient. JASIST (JASIS) 54(6):550-560 (2003).

[2] Batet, M., Sánchez, D., Valls, A., Gibert, K.: Semantic similarity esmation from multiple ontologies. Appl. Intell. 38(1): 29-44 (2013).

[3] Bollegala, D., Matsuo, Y., Ishizuka, M.: Measuring semantic similarity between words using web search engines. WWW 2007: 757-766.

[4] Bollegala, D., Matsuo, Y., Ishizuka, M.: A Web Search Engine-Based Approach to Measure Semantic Similarity between Words. IEEE Trans. Knowl. Data Eng. (TKDE) 23(7):977-990 (2011).

[5] Couto, F.M., Silva, M.J., Coutinho, P.: Measuring semantic similarity between Gene Ontology terms. Data Knowl. Eng. (DKE) 61(1):137-152 (2007).

[6] Chaves-González, J.M., Martinez-Gil, J.: Evolutionary algorithm based on different semantic similarity functions for synonym recognition in the biomedical domain. Knowl.-Based Syst. 37: 62-69 (2013).

[7] Martinez-Gil, J., Aldana-Montes, J.F.: An overview of current ontology meta-matching solutions. Knowledge Eng. Review 27(4): 393-412 (2012).

[8] Martinez-Gil, J.: An overview of textual semantic similarity measures based on web intelligence. Artif. Intell. Rev. 42(4): 935-943 (2014).

[9] Martinez-Gil, J., Aldana-Montes, J.F.: Reverse ontology matching. SIGMOD Record 39(4): 5-11 (2010).

[10] Michel, J.B., Shen, Y., Aiden, A., Veres, A., Gray, M., Pickett, J., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M., and Aiden, E: Quantitative analysis of culture using millions of digitized books. Science 331(6014): 176-182 (2011).

[11] Miller, G., Charles, W.: Contextual correlates of semantic similarity. Language and Cognitive Processes 6(1): 1–28 (1998).

[12] O'Shea, K.: An approach to conversational agent design using semantic sentence similarity. Appl. Intell. 37(4): 558-568 (2012).

[13] Pirro, G.: A semantic similarity metric combining features and intrinsic information content. Data Knowl. Eng. 68(11): 1289-1308 (2009).

[14] Punuru, J., Chen, J.: Learning non-taxonomical semantic relations from domain texts. J. Intell. Inf. Syst. 38(1): 191-207 (2012).

[15] Sánchez, D., Batet, M., Isern, D., Valls, A.: Ontology-based semantic similarity: A new feature-based approach. Expert Syst. Appl. 39(9): 7718-7728 (2012).

[16] Wang, Z., Li, J., Zhao, Y., Setchi, R., Tang, J.: A unified approach to matching semantic data on the Web. Knowl.-Based Syst. 39: 173-184 (2013).

Biography

Jorge Martinez-Gil is a Spanish-born computer scientist working in the Knowledge Engineering field. He got his PhD in Computer Science from University of Malaga in 2010. He has held a number of research positions across some European countries (Austria, Germany, Spain). He currently holds a Team Leader position within the group of Knowledge Representation and Semantics from the Software Competence Center Hagenberg (Austria) where he is involved in several applied and fundamental research projects related to knowledge-based technologies. Dr. Martinez-Gil has authored many scientific papers, including those published in prestigious journals like SIGMOD Record, Knowledge and Information Systems, Information Systems Frontiers, Knowledge-Based Systems, Artificial Intelligence Review, Knowledge Engineering Review, Online Information Review, Journal of Universal Computer Science, Journal of Computer Science and Technology, and so on.