Analysis of word co-occurrence in human literature for supporting semantic correspondence discovery

Jorge Martinez-Gil
Software Competence Center Hagenberg
Softwarepark 21, 4232
Hagenberg, Austria
jorge.martinez-gil@scch.at

Mario Pichler
Software Competence Center Hagenberg
Softwarepark 21, 4232
Hagenberg, Austria
mario.pichler@scch.at

ABSTRACT

Semantic similarity measurement aims to determine the likeness between two text expressions that use different lexicographies for representing the same real object or idea. In this work, we describe the way to exploit broad cultural trends for identifying semantic similarity. This is possible through the quantitative analysis of a vast digital book collection representing the digested history of humanity. Our research work has revealed that appropriately analyzing the co-occurrence of words in some periods of human literature can help us to determine the semantic similarity between these words by means of computers with a high degree of accuracy.

1. INTRODUCTION

Semantic similarity measurement is a well established field of research whereby two terms or text expressions are assigned a quantitative score based on the likeness of their meaning [25]. Automatic measurement of semantic similarity is considered to be one of the pillars for many computer related fields since a wide variety of techniques rely on determining the meaning of data they work with. In fact, for the research community working in the field of Linked Data, semantic similarity measurement is of vital importance in order to support the process of connecting and sharing related data on the Web.

In the past, there have been great efforts in finding new semantic similarity measures mainly due it is of fundamental importance in many application-oriented fields of the modern computer science. The reason is that computational techniques for semantic similarity measurement can be used for going beyond the literal lexical match of words and text expressions by operating at a conceptual level. Past works in this field include the automatic processing of text and email messages [15], healthcare dialogue systems [5], natural language querying of databases [13], question answering [21], and sentence fusion [2].

In the literature, this problem has been addressed from two different perspectives: similarity and relatedness; but nowadays there is a common agreement about the scope of each of them [3]. Firstly, semantic similarity states the taxonomic proximity between terms or text expressions. For example, automobile and car are similar because both are means of transport. Secondly, the more general concept of semantic relatedness considers taxonomic and relational proximity. For example, blood and hospital are related because both belong to the world of health, but they are far from being similar. Due to the impact of measuring similarity in modern computer science we are going to focus on semantic similarity for the rest of this paper, but it should be noted that many of the presented ideas are also applicable to the computation of relatedness.

The usual approach for solving the semantic similarity problem has consisted of using manually compiled dictionaries such as WordNet [23] to assist researchers when determining the semantic similarity between terms, but an important problem remains open. There is a gap between dictionaries and the language used by people, the reason is a balance that every dictionary must strike for: to be comprehensive enough for being a useful reference but concise enough to be practically used. For this reason, many infrequent words are usually omitted. Therefore, how can we measure semantic similarity in situations where terms are not covered by a dictionary? We investigate Culturomics as an answer.

Culturomics is a field of study which consists of collecting and analyzing large amounts of data for the study of human culture. Michel et al. [19] established this discipline by means of their seminal work where they presented a corpus of digitized texts representing the digested history of human literature. The rationale behind this idea was that an analysis of this corpus could enable people to investigate cultural trends quantitatively.

The study of human culture through digitized books has had a strong positive impact on our core research since its inception. We know that it is difficult to measure semantic similarity for terms usually omitted in traditional dictionaries, but it is highly improbable for these terms not having ever appeared in any book from the human literature. For this reason, we decided to open a new research line for finding quantitative methods to assist us in the process of measuring semantic similarity automatically using world literature. We have tested many methods, but this work is

intended to describe one of the most promising approaches we have found. This approach consists of studying the co-occurrences of words in a significant book sample from the human literature. Therefore, the main contributions presented in this work are the following:

1. We propose to study the co-occurrence of words in the human literature for trying to determine the semantic similarity between words.

2. We evaluate our proposal according to the word pairs included in the Miller & Charles benchmark data set [20] which is one of the most widely used on this context.

The rest of this paper is organized as follows: Section 2 describes related approaches that are proposed in the literature currently available. Section 3 describes the key ideas to understand our contribution. Section 4 presents a qualitative evaluation of our method, and finally, we draw conclusions and put forward future lines of research in Section 5.

2. RELATED WORK

The notion of semantic similarity is a widely intuitive concept. Miller and Charles wrote: *...subjects accept instructions to judge similarity of meaning as if they understood immediately what is being requested, then make their judgments rapidly with no apparent difficulty* [20]. This view has been reinforced by other researchers who observed that similarity is treated as a property characterized by human perception and intuition [26]. In general, it is assumed that not only are the participants comfortable in their understanding of the concept, but also when they perform a judgment task they do it using the same procedure or at least have a common understanding of the attribute they are measuring [22].

In the past, there have been great efforts in finding new semantic similarity measures mainly due to its fundamental importance in many computer related fields. The detection of different formulations of the same concept is a key method for solving a lot of problems. To name only a few, we can refer to a) data clustering where semantic similarity measures are necessary to detect and group the most similar subjects [4], b) data matching which consists of finding some data representing the same concept across different data sources [16], c) data mining where using appropriate semantic similarity measures can facilitate the processes of text classification and pattern discovery in large texts [11], or d) machine translation where the detection of term pairs expressed in different languages but referring to a same idea is of vital importance [9]. Semantic similarity is also of vital importance for the community working on Linked Data paradigms since software tools for automatically discovering relationships between data items within different Linked Data sources can be very useful.

According to Sanchez el al. [28], most of existing semantic similarity measures can be classified into one of these four main categories:

1. *Edge-counting measures* which are based on the computation of the number of taxonomical links separating two concepts represented in a given dictionary [16].

2. *Feature-based measures* which try to estimate the amount of common and non-common taxonomical information retrieved from dictionaries [24].

3. *Information theoretic measures* which try to determine similarity between concepts as a function of what both concepts have in common in a given ontology. These measures are typically computed from concept distribution in text corpora [14].

4. *Distributional measures* which use text corpora as source. They look for word co-occurrences in the Web or large document collections using search engines [6].

There are also several related works that try to aggregate or combine semantic similarity measures [8]. These methods come from the field of semantic similarity aggregation. For instance COMA, where a library of semantic similarity measures and friendly user interface to aggregate them are provided [12], or MaF, a matching framework that allow users to combine simple similarity measures to create more complex ones [17].

These approaches can be even improved by using weighted means where the weights are automatically computed by means of heuristic and meta-heuristic algorithms [18]. In that case, most promising measures receive better weights. This means that all the efforts are focused on getting more complex weighted means that after some training are able to recognize the most important atomic measures for solving a given problem [17]. There are two major problems that make these approaches not very appropriate in real environments: First problem is that these techniques require a lot of training efforts. Secondly, these weights are obtained for a specific problem and it is not easy to find a way to transfer them to other problems.

Our proposal is a distributional semantic similarity measure since, as it will be explained in more depth, we try to look for co-occurrences of words in the same text corpus. In fact, we are going to get benefit from a corpus of digitized texts containing 5.2 million books which represent about four percent of all books ever printed [19]. Achieving good results could represent an improvement over traditional approaches since our approach does not incur in the drawbacks from the heuristic and meta-heuristic methods, and does not require any kind of training or knowledge transfer. Please note that we are not analyzing the entire corpus for determining the co-occurrence of words, but a digested representation (in the form of a time series) of the word occurrences along the time. This means that the processing time by means of a computer is very short.

3. CONTRIBUTION

Semantic similarity measurement is a well established field of research whereby two text entities are assigned a score based on the likeness of their meaning. More formally, we can define a semantic similarity measure as a function μ_1 x $\mu_2 \rightarrow$ R that associates the degree of similarity for the text

entities μ_1 and μ_2 to a score $s \in \mathrm{R}$ in the range $[0, 1]$ where a score of 0 stands for no similarity at all, and 1 for total similarity of the meanings associated to μ_1 and μ_2.

Our key contribution is based on the idea of exploring culturomics for designing such a function, thus the application of quantitative analysis to the study of human culture, for trying to determine the semantic similarity between terms or text expressions. The main reason for preferring this paradigm rather than a traditional approach based on dictionaries is obvious; according to the book library digitized by Google, the number of words in the English lexicon is currently above a million. The lexicon is in a period of enormous growth with the addition of thousands of words per year. Therefore, there are more words from the data sets we are using than appear in any dictionary. For instance, the Webster's Third New International Dictionary[1], which keeps track of the contemporary American lexicon, lists much less than 400,000 single-word word forms currently [19]. This means that one of the advantages of this technique in relation to the traditional ones is that it can be applied on more than 600,000 single-word word forms on which dictionary-based techniques cannot work.

One of the problems we have to address is that all information from the book library is stored in data sets which are currently 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 world literature. Therefore, each word which has appeared at least once will have a number sequence (time series) associated. These number sequences represent the records for the total number of word occurrences per year in the books digitized. This allows us to compute the frequencies of words along human history, but it is necessary to have quantitative algorithms for helping us to get benefit from this information.

The method that we propose consists of measuring how often two terms appear in the same text sentence. Studying the co-occurrence of terms in a text corpus has been usually used as an evidence of semantic similarity in the scientific literature [6, 28]. In this work, we propose adapting this paradigm for our purposes. To do this, we are going to calculate the joint probability so that a text expression may contain the two terms together over time. Equation 1 shows the mathematical formula we propose (being a and b are the two words that are going to be compared):

$$sim(a, b) = \frac{time\ units\ a\ and\ b\ co-occur}{time\ units\ considered} \qquad (1)$$

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). Moreover, this output value for stating the degree of semantic similarity can be fuzzificated in case a great level of detail may not be needed. Now, let us see some examples of application of this technique:

Example 1. Compute the similarity for the terms lift and elevator in the time range [1850, 1950] taking five-year periods as a time unit.
We query the database using the expression "lift elevator" OR "elevator lift". We got that there is, at least, a co-occurrence on 14 different time units. Moreover, we know that 100 years have 20 periods of 5 years, so we have that $sim(lift, elevator)^{5}_{1850-1950} = \frac{14}{20} = 0.7$ what means that these terms are quite similar. This result has sense in the real life since both terms are used as synonyms very frequently.

Example 2. Compute the similarity for the terms beach and drink in the time range [1920, 2000] taking ten-year periods as a time unit.
We query the database using the expression "beach drink" OR "drink beach". We got that there is not any co-occurrence on the different time units that have been specified. Moreover, we know that 80 years have 8 periods of 10 years, so we have that $sim(beach, drink)^{10}_{1920-2000} = \frac{0}{8} = 0.0$ what means that these terms are not similar at all. This result has sense again since a beach is a kind of expanse of sand along a shore, meanwhile a drink is a liquid than can be drunk.

The great advantage of using culturomics instead of classic techniques is that it can be used for measuring the semantic similarity for more than 600,000 single-word forms on which dictionary-based techniques cannot work. Some examples of these words are: actionscript, bluetooth, dreamweaver, ejb, ipod, itunes, mysql, sharepoint, voip, wsdl, xhtml or xslt. It is possible to find lack of accuracy with very new words. The reason is that we still need some years in order these words can appear in the literature more frequently. However, the major part of the vocabulary ever in use will be covered. However, the mere fact of being able to work with this vast amount of single words cannot be considered as a great advantage if the quality achieved is not, at least, reasonable. For this reason, we think that it is necessary to asses the quality of our method by using classical evaluation techniques. If our proposal succeeds when solving traditional benchmark data sets, we can suppose that it will also perform well when dealing with other less popular terms since our technique does not make any kind of distinction between them.

4. EVALUATION

We report our results using the data set offered by Google[2]. It is important to remark that only words that appear over 40 times across the corpus can be considered. The data used has been extracted from the English between 1900 and 2000. The reason is that there are not enough books before 1900 to reliably quantify many of the modern terms from the data sets we are using. On the other hand, after year 2000, quality of the corpus is lower since the book collection is subject to many changes.

[1] http://www.merriam-webster.com

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

WordA	WordB	Human
rooster	voyage	0.08
noon	string	0.08
glass	magician	0.11
chord	smile	0.13
coast	forest	0.42
lad	wizard	0.42
monk	slave	0.55
shore	woodland	0.63
forest	graveyard	0.84
coast	hill	0.87
food	rooster	0.89
cementery	woodland	0.95
monk	oracle	1.10
car	journey	1.16
brother	lad	1.66
crane	implement	1.68
brother	monk	2.82
implement	tool	2.95
bird	crane	2.97
bird	cock	3.05
food	fruit	3.08
furnace	stove	3.11
midday	noon	3.42
magician	wizard	3.50
asylum	madhouse	3.61
coast	shore	3.70
boy	lad	3.76
journey	voyage	3.84
gem	jewel	3.84
automobile	car	3.92

Table 1: Miller-Charles data set. Semantic similarity 30 for word pairs is assessed by humans. Human ratings are between 0 (not similar at all) and 4 (totally similar)

Results are obtained according Miller-Charles benchmark data set [20] which is a widely used reference data set 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 goal is to replicate human behavior when solving tasks related to semantic similarity without any kind of supervision. Table 1 lists the complete collection of word pairs from this benchmark data set. This collection of word pairs ranges from words which are not similar at all (rooster-voyage or noon-string, for instance) to word pairs that are synonyms according to human judgment (automobile-car or gem-jewel, for instance). Columns called Human represent the opinion provided by the people who rated the term pairs. This opinion was originally given in numeric score in the range [0, 4] where 0 stands for no similarity between the two words from the word pair and 4 stands for complete similarity. There is no problem when artificial measures assess semantic similarity using values belonging to the interval [0, 1] since the Pearson Correlation Coefficient is invariant against a linear transformation. This means that results are independent from the scales used for assessing semantic similarity.

Table 2 shows the results that have been obtained by using our method for the range 1900-2000 using 5 years as a time unit. The overall fitness we have obtained by measuring the correlation between human judgment and our approach is 0.458. For this benchmark dataset, a result greater than 0.3 is statistically significant [25]. This result is achieved since our technique is not able to detect very clear synonym pairs like gem-jewel that despite to be considered very similar from the human judgment, do not appear together in the human literature very frequently. Errors of this kind penalize the final score and should be study objects in the future.

If we focus on the results publicly available in the literature, and despite this is only the first study performed using this paradigm, we have that these results are significantly better than most of techniques reported by Bollegala et al. [7]. In this way, our technique beats Jaccard, Dice, and Overlap Coefficient. However, the results are still far from those reported by Sahami [27], CODC [10], and SemSim [7] which is a complex method involving great efforts in previous optimization and training.

One of the reasons for these results 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.

In this context, the conversion into linguistic variables comprises the process of transforming the numeric values we have obtained in the previous experiment into grades of membership for linguistic terms. As we mention before, this process is useful in cases where an approximate reasoning is preferred to an exact one. In order to proceed, 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 (the closest to the word we are comparing with). 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 blurred in the original process of conversion where we have only a limited number of linguistic variables to describe all degrees of semantic similarity.

Therefore, we repeated our experiment with some modifications through some kind of fuzzification for the numerical values. This means we have transformed the numeri-

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

Table 2: Results for the Miller-Charles benchmark dataset. Columns called Wordpair represent the words being evaluated. Columns called Human represent the opinion provided by people. Columns called Machine represent the result achieved by our approach. The fitness is 0.458

Wordpair	Human	Machine	Wordpair	Human	Machine
rooster-voyage	not similar	right	crane-implement	not similar	right
noon-string	not similar	right	brother-monk	similar	right
glass-magician	not similar	right	implement-tool	similar	right
cord-smile	not similar	right	bird-crane	similar	right
coast-forest	not similar	wrong	bird-cock	similar	right
lad-wizard	not similar	right	food-fruit	similar	right
monk-slave	not similar	right	furnace-stove	similar	right
shore-woodland	not similar	wrong	midday-noon	similar	right
forest-graveyard	not similar	wrong	magician-wizard	similar	right
coast-hill	not similar	wrong	asylum-madhouse	similar	wrong
food-rooster	not similar	right	coast-shore	similar	right
cementery-woodland	not similar	right	boy-lad	similar	right
monk-oracle	not similar	right	journey-voyage	similar	right
car-journey	not similar	wrong	gem-jewel	similar	wrong
brother-lad	not similar	right	automobile-car	similar	right

Table 3: Results for the Miller-Charles benchmark data set. Columns called Wordpair represent the words being evaluated. Columns called Human represent the opinion provided by people. Columns called Machine represent the result achieved by our approach. There are 23/30 hits, this means we have been able to achieve 76.67% of accuracy

cal values into linguistic variables. In fact, these numerical values have been fuzzificated into two linguistic variables (not similar and similar) since a great level of granularity is not often needed, but it would be possible to define additional categories if necessary. Therefore, the columns called Wordpair in Table 3 represent the words being evaluated, columns called Human represent the opinion provided by people, columns called Machine indicate if our approach has been able to guess the semantic similarity of the word pair or not. We have found that there are 23/30 hits, this means we have been able to achieve 76.67% of accuracy. Now, it is possible to perceive much better results than in the previous experiment.

It is necessary to take into account that results from Table 2 and Table 3 are not comparable since they are not expressed in the same units. The result presented in Table 2 is a correlation coefficient that tell us the degree of linear correlation between the opinion expressed by people and the opinion expressed by our algorithm. Results presented in Table 3 represent the number of times that our algorithm is able to correctly guess if a term pair is semantically similar or not. This means that we are working with binary values, and therefore, it has less sense to use a correlation coefficient to assess the quality of the given results.

4.1 Discussion

Our results tell us that this technique can be very useful when supporting a number of tasks that have to be manually done currently. One of the clearest examples belongs to the field of Human Resources Management Systems (HRMS). One of the major problems in this domain consists of automatically matching job offers and applicant profiles. This problem can be addressed by means of an automatic matching process that use semantic similarity measurement for determine the degree of correspondence between those applicant profiles and job offers. Solutions of this kind are good for employers which can make the recruitment process may become cheaper, faster and more successful, but also for job applicants who can receive feedback about the recruitment

decisions concerning their applications.

But there are still some major problems that have to be faced. One of these major problems for these systems is that, mainly due to the high dynamism of the job market, both job offers and applicant profiles contain terms that are not usually covered by dictionaries (emergence of new programming languages, software tools, artifacts for automation of tasks, and so on). This means that it is very difficult to identify any kind of semantic correspondence between them, and therefore to compute a fitness score for the correspondence between the job offer and the applicant profile. However, the great amount of technical literature publicly available can be used to support this process. If we are able to find the most appropriate algorithms for discovering semantic similarity on basis of large book libraries, then we do not need to use dictionaries for supporting the process.

Another interesting field of application could be to support the process of adding new terms to dictionaries. This case is becoming more and more usual since new technologies and social networks are adopting new terms in a very quick way. Currently, this task has to be manually performed. The process is quite tedious since requires big efforts to be made by linguistic experts. We think that our technique can partially support processes of this kind since it can be possible to look for similar terms covered by the dictionary to be extended. This information can help a lot in the process of categorizing and defining this new term. Since the results of this technique are not perfect yes, techniques of this kind cannot be used to develop perfect solutions in this field, but semi-automatic tools providing suggestions to people responsible for taking the final decisions.

5. CONCLUSIONS & FUTURE WORK

In this work, we have described how we have got benefit from a new paradigm called culturomics for automatically determining the degree of semantic similarity between words. We aim to go through the quantitative analysis of a vast digital book collection representing a significant sample of the history of literature to solve problems related to the semantic similarity measurement of words. In fact, semantic similarity measurement is of vital importance for the Linked Data and other related communities since it is in order to support the process of connecting and sharing related data on the Web.

We have shown that appropriately studying the co-occurrence of words along human literature can provide very accurate results when measuring semantic similarity between these words. Moreover, the major advantage of this technique in relation to the traditional ones is that it can be applied on more than 600,000 single-word forms on which dictionary-based techniques cannot work. While we carried out just a first study, results have outperformed a number of traditional techniques. However, there is still much research work to do. In fact, it is necessary to further research what are the best time ranges and time units to compute the semantic similarity using the huge book library.

It should also be noted that this work focuses on the study of single words, but our plans include researching about the similarity of short text expressions. We assume that using new algorithms for word occurrence or for statistical transformation of data could be beneficial since positive results in this context could lead to the ability of computers to recognize and predict the semantic similarity between words ever appeared in the human literature without requiring any kind of human intervention. We also think that could be very interesting to research towards the creation of benchmark data sets reflecting time issues and constraints. The reason is that semantic similarity is not a fixed notion and can vary along the years. For example, nowadays we consider that the term pair car-automobile is quite similar, but maybe this fact has not been always true along the history of humanity or maybe it is not going to be true in the future. For this reason, we think that it is important not only to asses the semantic similarity of terms, but also the temporal validity of this semantic similarity. In this sense, computing optimal historic windows for determining semantic similarity could have a great impact.

Acknowledgments

This work has been funded by Vertical Model Integration within Regionale Wettbewerbsfaehigkeit OOE 2007-2013 from the European Fund for Regional Development and the State of Upper Austria, and by the project ACEPROM (Proj.Nr. 841284) in the frame of the 17th BRIDGE Call managed by the Austrian Research Promotion Agency (FFG).

6. REFERENCES

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

[2] R. Barzilay, K. McKeown. Sentence Fusion for Multidocument News Summarization. Computational Linguistics 31(3). 297-328 (2005).

[3] M. Batet, D. Sanchez, A. Valls. An ontology-based measure to compute semantic similarity in biomedicine. J. Biomed. Inform. (44): 118-125 (2010).

[4] M. Batet. Ontology-based semantic clustering. AI Commun. 24(3): 291-292 (2011).

[5] T.W. Bickmore, T. Giorgino. Health dialog systems for patients and consumers. Journal of Biomedical Informatics: 556-571 (2006).

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

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

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

[9] B. Chen, G.F. Foster, R. Kuhn. Bilingual Sense Similarity for Statistical Machine Translation. ACL 2010:834-843.

[10] H.H. Chen, M.S. Lin, Y.C. Wei. Novel Association Measures Using Web Search with Double Checking.

ACL 2006.
[11] F.M. Couto, M.J. Silva, P. Coutinho. Semantic similarity over the gene ontology: family correlation and selecting disjunctive ancestors. CIKM 2005:343-344.
[12] Do H.H., Rahm, E. COMA - A System for Flexible Combination of Schema Matching Approaches. VLDB 2002: 610-621.
[13] G. Erozel, N.K. Cicekli, I. Cicekli. Natural language querying for video databases. Inf. Sci. 178(12): 2534-2552 (2008).
[14] J.J. Jiang, D.W. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. ROCLING 1997: 19-33.
[15] L. Lamontagne, G. Lapalme. Textual Reuse for Email Response. ECCBR 2004: 242-256.
[16] J. Martinez-Gil, J.F. Aldana-Montes. Evaluation of two heuristic approaches to solve the ontology meta-matching problem. Knowl. Inf. Syst. 26(2): 225-247 (2011).
[17] J. Martinez-Gil. An overview of textual semantic similarity measures based on web intelligence. Artif. Intell. Rev. 42(4): 935-943 (2014).
[18] J. Martinez-Gil, J.F. Aldana-Montes. Semantic similarity measurement using historical google search patterns. Inf. Syst. Frontiers 15(3): 399-410 (2013).
[19] J.B. Michel, Y. Shen, A. Aiden, A. Veres, M. Gray, J. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, S. Pinker, M. Nowak, E. Aiden. Quantitative analysis of culture using millions of digitized books, Science 331(6014): 176-182 (2011).
[20] G.A. Miller, W.G. Charles. Contextual correlates of semantic similarity. Language and Cognitive Processes, 6(1): 1-8 (1991).
[21] A. Moschitti, S. Quarteroni. Kernels on Linguistic Structures for Answer Extraction. ACL (Short Papers) 2008: 113-116.
[22] J. O'Shea, Z. Bandar, K.A. Crockett, D. McLean. Benchmarking short text semantic similarity. IJIIDS 4(2): 103-120 (2010).
[23] T. Pedersen, S. Patwardhan, J. Michelizzi. WordNet::Similarity - Measuring the Relatedness of Concepts. AAAI 2004: 1024-1025.
[24] E.G.M. Petrakis, G. Varelas, A. Hliaoutakis, P. Raftopoulou. X-similarity: computing semantic similarity between concepts from different ontologies. J. Digit. Inf. Manage. 233-237 (2003).
[25] G. Pirro. A semantic similarity metric combining features and intrinsic information content. Data Knowl. Eng. 68(11): 1289-1308 (2009).
[26] P. Resnik. Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. J. Artif. Intell. Res. (JAIR) 11: 95-130 (1999).
[27] M. Sahami, T.D. Heilman. A web-based kernel function for measuring the similarity of short text snippets. WWW 2006: 377-386.
[28] D. Sanchez, M. Batet, D. Isern. Ontology-based information content computation. Knowl.-Based Syst. 24(2): 297-303 (2011).